Biostatistical Algorithms for OMICS Data in oncology. Application to DNA copy number microarray Experiments

Philippe Hupé

▶ **To cite this version:**

Philippe Hupé. Biostatistical Algorithms for OMICS Data in oncology. Application to DNA copy number microarray Experiments. Mathematics [math]. AgroParisTech; Ecole Doctorale Agriculture Alimentation Biologie Environnement Santé, 2008. English. NNT : . tel-02819845

HAL Id: tel-02819845
https://hal.inrae.fr/tel-02819845

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture Agriculture Alimentation Food Biologie Biology Environnement Environment Santé Health

N° /_/_/_/_/_/_/_/_/_/_/

THÈSE

pour obtenir le grade de

Docteur

de

l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité: Mathématiques

présentée et soutenue publiquement par

Philippe Hupé
14/11/08

Biostatistical algorithms for omics data in oncology

Application to DNA copy number microarray experiments

Directeur de thèse: **Jean-Jacques Daudin**
Codirecteur de thèse: **Emmanuel Barillot**
Institut Curie, Service de Bioinformatique, F-75005 Paris
U900 INSERM - Mines Paris Tech - Institut Curie, F-75005 Paris

Devant le jury

M. Emmanuel Barillot, Docteur	Institut Curie	Codirecteur de thèse
M. Philippe Besse, Professeur	Institut National des Sciences Appliquées	Examinateur
M. Max Chaffanet, Docteur	Institut Paoli Calmettes	Rapporteur
M. Jean-Jacques Daudin, Professeur	Agro Paris Tech	Directeur de thèse
M. Xavier Gidrol, Docteur	Commissariat à l'Energie Atomique	Examinateur
M. Yves Moreau, Professeur	Katholieke Universiteit Leuven	Rapporteur

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech) est un Grand Etablissement dépendant du Ministère de l'Agriculture et de la Pêche, composé de l'INA PG, de l'ENGREF et de l'ENSIA
(décret n° 2006-1592 du 13 décembre 2006)

Abstract

Cancer is a major cause of death and lots of effort must be made to defeat the disease. Microarray technology is a powerful tool very helpful in oncology in order to better understand the molecular mechanisms involved in tumoral progression. We know that cancer is due to a modification of the gene regulation. Then, the study of gene expression in tumours is a valuable information in order to understand the biology of the disease and to identify new prognostic and predictive factors so that the clinician can tailor the therapy for each patient. Besides the modification of gene expression, tumours have chromosome alterations and especially a change of their DNA copy number. There are microarrays which allow the quantification of DNA copy number. The raw data obtained from the microarray technology need appropriate statistical processing so that they can be biologically and clinically meaningful. This is precisely the goal of the present work. Thus, statistical methods have been developed in order to normalise and extract the biological information from microarrays devoted to the study of DNA copy number in tumours. The methods have been applied in uveal melanoma in order to identify high-risk tumours. The integrative analysis of different types of molecular profiles is a challenge in biostatistics. Therefore, a statistical method able to combine both gene expression and DNA copy number data has been developed in the framework of supervised classification. The statistical properties of the method have been studied and its performance has been evaluated on both simulated and real data.

Title: Biostatistical algorithms for omics data in oncology - Application to DNA copy number microarray experiments.

Keywords: DNA copy number, microarray, supervised classification, biostatistics, oncology.

Field: mathematics.

Laboratory: Institut Curie, Service de Bioinformatique, 26, rue d'Ulm, F-75005 Paris. U900 INSERM - Mines Paris Tech - Institut Curie (Cancer et génome : bioinformatique, biostatistiques et épidémiologie d'un système complexe), 26, rue d'Ulm, F-75005 Paris.

Résumé

Le cancer est une cause principale de décès et d'importants efforts doivent être réalisés pour vaincre la maladie. La technologie des microarrays est un puissant outil de recherche en oncologie pour comprendre les mécanismes de la progression tumorale qui est due à une perturbation de la régulation des gènes. Par conséquent, l'étude de leur niveau d'expression dans les tumeurs offre une perspective pour comprendre les mécanismes biologiques de la maladie et identifier de nouveaux facteurs pronostiques et prédictifs qui aideront le clinicien à choisir la thérapie de chaque patients. Par ailleurs, les tumeurs présentent un changement du nombre de copies d'ADN dont la quantification est aussi possible par microarray. L'utilisation des données de microarray nécessite un traitement statistique approprié permettant de transformer les données brutes en données interprétables biologiquement et cliniquement. Ainsi, nous avons développé des méthodes statistiques qui visent à normaliser et extraire l'information biologique issue des microarrays dédiés à l'étude du nombre de copies d'ADN des tumeurs. Nos méthodes ont permis la caractérisation des tumeurs de haut-risque métastatique dans le mélanome uvéal. Par ailleurs, un des enjeux de l'analyse biostatistique des données de microarrays consiste en l'analyse intégrée de différents types de profils moléculaires. Ainsi, une méthode statistique qui combine les données d'expression de gènes et du nombre de copie d'ADN obtenues par microarrays a été développée dans un contexte de classification supervisée. Les propriétés statistiques de la méthode ont été étudiées et ses performances estimées sur des données simulées et réelles.

Titre: Algorithmes biostatistiques pour les données omiques en oncologie - Application à l'étude du nombre de copies d'ADN à partir des expériences de microarray.

Mots-clé: nombre de copies d'ADN, microarray, classification supervisée, biostatistiques, oncologie.

Discipline: mathématiques.

Laboratoire: Institut Curie, Service de Bioinformatique, 26, rue d'Ulm, F-75005 Paris. U900 INSERM - Mines Paris Tech - Institut Curie (Cancer et génome : bioinformatique, biostatistiques et épidémiologie d'un système complexe), 26, rue d'Ulm, F-75005 Paris.

Résumé substantiel

Le cancer est une cause principale de décès et d'importants efforts doivent être réalisés pour vaincre la maladie. La technologie des microarrays (appelés également puces) est un puissant outil de recherche en oncologie pour comprendre les mécanismes de la progression tumorale. Cette technologie permet notamment de quantifier les aberrations chromosomiques qui sont une caractéristique commune à tous les cancers. Le travail de thèse présenté ici a contribué aux développements d'algorithmes biostatistiques et d'outils bioinformatiques dédiés à l'analyse des données de puces qui permettent la quantification des aberrations chromosomiques dues à un changement du nombre de copies d'ADN. Le manuscrit de thèse se divise en trois chapitres.

Le **Chapitre 1** introduit le contexte clinique, biologique et biotechnologique. Tout d'abord, les enjeux de la recherche en oncologie sont présentés. Ceux-ci consistent en trois principaux axes. Le premier est l'identification de nouveaux facteurs pronostiques et prédictifs pour aider le clinicien à personnaliser la prise en charge thérapeutique de chaque patient. Le deuxième est la recherche de nouvelles cibles thérapeutiques pour traiter les cancers incurables. Le troisième est l'amélioration de la connaissance des mécanismes de la progression tumorale. Le travail présenté ici s'intéresse au premier axe dont l'objectif est de définir des règles de prédiction qui guideront les choix du clinicien en utilisant les données moléculaires obtenues par la technique des microarrays. Après avoir décrit les mécanismes biologiques qui contrôlent le fonctionnement d'une cellule normale, nous expliquons comment ceux-ci sont altérés dans la cellule cancéreuse. Le cancer est une maladie des gènes dont la régulation a été perturbée suite à une séquence d'accumulation de mutations notamment au niveau de gènes critiques appelés oncogènes et gènes suppresseur de tumeur. L'altération de la régulation des gènes due à cette accumulation de mutations se traduit par une modification de la fonction des facteurs de transcription, des changements des propriétés épigénétiques, des aberrations au niveau de l'épissage alternatif et une perturbation du rôle régulateur des ARNs non-codants. Le réseau de régulation des gènes qui définit le fonctionnement de la cellule normal en réponse à son environnement est alors totalement incapable d'assurer un comportement socialement responsable de la cellule cancéreuse à l'égard des autres cellules de l'organisme. Au final, la cellule cancéreuse acquiert de nouvelles caractéristiques qui lui confèrent entre autre la capacité de proliférer de manière anarchique et d'envahir les tissus adjacents. Une des caractéristiques liée à ce chaos général qui règne au sein de chaque cellule cancéreuse est la présence d'aberrations chromosomiques. En effet, le génome des tumeurs présente un caryotype anormal dans lequel des chromosomes entiers ou seulement certaines portions ont été perdus, gagnés ou amplifiés (*i.e.* gain de plus de quatre copies). Ainsi, ces modifications du nombre de copies d'ADN représentent une information précieuse pour permettre la caractérisation moléculaire des tumeurs et définir les règles de prédiction indispensables pour répondre au premier axe des enjeux de la recherche en oncologie. La quantification du nombre de copies d'ADN dans le génome des tumeurs est possible grace à la technique des microarrays. Initialement, cette technologie a été développée pour étudier le niveau d'expression des gènes mais celle-ci a été adaptée pour assurer l'étude des aberrations chromosomiques à grande échelle. Ainsi, il est désormais possible de mesurer, en un très grand nombre de loci du génome, le nombre de copies d'ADN en utilisant des sondes qui sont spécifiques de chaque position génomique. Cette technique est dite à haut-débit car elle permet en un temps rapide de générer un volume important de données à l'échelle d'un génome complet. La technique des puces CGH (Comparative Genomic Hybridisation) avec des sondes de type BAC (Bacterial Artificial Chromosome) a été la première à permettre une approche à haut-débit de quantification du nombre de copies d'ADN mais celle-ci tend

désormais à être remplacée par les puces CGH avec des sondes à oligonucléotides qui assurent une couverture bien plus grande du génome permettant ainsi l'identification d'aberrations d'une taille de plus en plus petite. Dans ce chapitre, nous introduisons la nécessité des traitements biostatistiques afin de rendre exploitables, tant d'un point de vue biologique que clinique, les données issues de microarray en reprenant les travaux pionniers de l'analyse de l'expression des gènes puis nous justifions le besoin de nouvelles méthodes indispensables au traitement des puces CGH.

Le **Chapitre 2** décrit les algorithmes biostatistiques et les outils bioinformatiques qui ont été développés afin d'analyser les données obtenues par la technique des puces CGH. Des étapes systématiques sont nécessaires pour rendre les données brutes interprétables biologiquement et cliniquement. Les méthodes que nous avons développées traitent de ces différentes étapes. Tout d'abord, il faut normaliser les données brutes, c'est-à-dire corriger les sources de variabilité systématiques afin d'améliorer la qualité du signal. Sur les données de puces CGH avec des sondes de type BAC, des artefacts spatiaux ont été observés. En effet, on observe des zones sur ces puces dans lesquelles le signal mesuré est anormalement fort ou faible. Par conséquent, nous avons développé une méthode statistique (package MANOR) qui corrige et/ou élimine ces zones spatiales aberrantes. Pour cela, une étape de lissage des données suivie d'une segmentation spatiale du signal permet une normalisation efficace. Une fois la puce normalisée, il faut identifier de manière automatique les régions chromosomiques qui présentent un changement du nombre de copies d'ADN. Sur le génome, ce changement du nombre de copies d'ADN se produit en des loci particuliers appelés points de cassure. Nous avons développé une méthode statistique (package GLAD) qui permet leur identification automatique. Pour cela, l'algorithme considère la proximité géographique le long du génome des différents loci et estime une fonction en créneaux dont les sauts correspondent aux points de cassure. Chaque créneau représente une région chromosomique au sein de laquelle le nombre de copies d'ADN est constant. Ensuite, à chaque région est attribué un statut perdu, normal, gagné ou amplifié. Chaque technologie de puce CGH a des spécificités qui lui sont propres. Dans, le cas des puces CGH avec des sondes à oligonucleotides, nous avons développé une méthode (package ITALICS) qui normalise et détecte le nombre de copies d'ADN de manière alternative et itérative en considérant des sources de variabilité à corriger spécifiques de ce type de puce. L'efficacité de cette approche a été montrée sur des données de cancer du sein. Les trois méthodes précédentes qui ont été développées analysent un seul profil moléculaire à la fois. Cependant, il est indispensable d'utiliser des analyses transversales afin d'étudier les aberrations communes au sein de chaque type de cancer. Des méthodes permettent d'extraire de manière transversale les aberrations du nombre de copies d'ADN qui sont informatives. Sur des données de puces CGH du mélanome uvéal, nous avons appliqué nos algorithmes de normalisation spatiale et d'identification des points de cassure. Les aberrations du nombre de copies d'ADN informatives ont ensuite été extraites transversalement et ont permis de prédire efficacement les tumeurs de haut-risque métastatique. Les algorithmes que nous avons développés sont implémentés sous forme de packages R et disponibles dans le projet Bioconductor. De plus, une plate-forme d'analyse web (CAPweb) a été mise en place qui permet aux biologistes d'utiliser simplement nos méthodes sans compétence informatique et de visualiser et analyser avec un logiciel (VAMP) les profils moléculaires du nombre de copies d'ADN. Par ailleurs, nous avons mis en place une base de données (ACTuDB) qui intègre des données d'études des aberrations chromosomiques de différents types de cancer qui ont été mises à disposition publiquement par les chercheurs. Un traitement biostatistique et bioinformatique a été appliqué à chaque jeux de données de sorte qu'ils soient tous comparables entre eux. Cette base de données permet aux nouvelles études en cours de comparer leurs résultats par rapport à ceux qui ont déjà été obtenus par d'autres chercheurs.

Le **Chapitre 3** décrit une méthode d'analyse intégrée de données d'expression de gènes et du nombre de copies d'ADN dans un contexte de classification supervisée. L'analyse intégrée de profils moléculaires de nature différente est un des enjeux de l'analyse biostatistique des données à haut-débit. Désormais, de nombreuses études analysent conjointement les données d'expression des gènes avec le nombre de copies d'ADN. La confrontation de ces deux niveaux d'information permet d'avoir une meilleure compréhension des mécanismes impliqués dans la progression tumorale. Par ailleurs, l'utilisation simultanée de ces deux niveaux d'information représente un atout pour définir de nouvelles règles de prédiction qui aideront le clinicien dans le choix de la thérapie de chaque patient. La méthode que nous avons développée permet d'utiliser des données mixtes, c'est-à-dire qu'elle combine à la fois des variables continues (les données d'expression de gènes) et des données discrètes (les données des aberrations du nombre de copies d'ADN informatives identifiées transversalement) issues des deux types de profils moléculaires. Tout d'abord, les principes généraux de la classification supervisée ont été introduits. Ensuite, les différentes méthodes de classification supervisée capables de prendre en compte des données mixtes ont été décrites. Nous avons justifié le choix de la méthode de Location Model pour répondre à notre problématique de classification supervisée. Cette méthode est une généralisation de l'analyse discriminante linéaire qui permet de prendre en compte une hétérogénéité des variables continues liée aux variables discrètes. De plus, le Location Model permet de modéliser la probabilité d'appartenir à une classe donnée compte-tenu des variables discrètes observées. Les données d'expression de gènes présentent la particularité d'avoir un grand nombre de variables ce qui rend inapplicable la technique choisie sans traitement préalable des données. Ainsi, nous avons proposé une méthode de classification supervisée capable d'utiliser des variables mixtes tout en prenant en compte la haute dimensionnalité des données d'expression de gènes. Pour cela, notre méthode intègre une réduction de la dimension des données d'expression de gènes par PLS (Partial Least Squares) puis utilise les premières composantes PLS comme nouvelles variables continues dans le Location Model. Les performances de prédiction ont été évaluées à la fois sur des données simulées et réelles. Les résultats ont montré l'efficacité de notre méthode sur les données simulées. Toutefois, les résultats ne sont pas convaincants sur des données réelles. En effet, les performances de prédiction ne sont pas meilleures en combinant les données d'expression des gènes et d'aberrations du nombre de copies d'ADN informatives qu'en considérant seulement les données d'expression de gènes. Les raisons peuvent être soit statistique soit biologique. Le choix du nombre de composantes PLS à considérer pour la classification supervisée est une problématique importante. Des techniques de choix de modèles existent et les principaux critères utilisés classiquement ont été rappelés. Parmi ces critères, les critères théoriques présentent l'avantage de limiter les temps de calcul par rapport aux critères empiriques s'appuyant sur une validation croisée. Toutefois, l'application directe des critères théoriques aux composantes PLS s'avère inefficace. Un critère théorique dédié aux choix du nombre de composantes PLS serait dans l'idéal nécessaire mais sa recherche dépasse le cadre du présent travail. Nous avons néanmoins proposé un critère statistique qui teste la significativité du signal capturé par la première composante. Les résultats ont montré l'efficacité du critère statistique sur des données simulées avec toutefois une diminution de ses performances lorsque l'on introduit un forte corrélation entre les variables continues. Ceci explique les faibles performances obtenues sur des données réelles à cause de la présence de corrélation entre les différents gènes. Des perspectives ont été proposées pour améliorer la méthode de classification supervisée et le critère statistique développés afin qu'ils soient applicables à des jeux de données réelles.

Remerciements

Tout d'abord, je tiens à remercier Emmanuel Barillot et François Radvanyi sans qui mon aventure passionnante au sein de l'Institut Curie n'aurait jamais commencé. Merci de la confiance que vous avez pu me témoigner et de votre expérience précieuse que vous avez su partager.

Je remercie Jean-Jacques Daudin et Emmanuel Barillot d'avoir accepté de m'encadrer pendant tout mon travail de thèse. Merci, de votre disponibilité, de votre soutien et de tout ce que vous avez pu m'apprendre.

Je remercie Max Chaffanet et Yves Moreau d'avoir accepté d'être les rapporteurs de ma thèse. Je remercie également Philippe Besse et Xavier Gidrol d'avoir accepté d'évaluer mon travail.

Je remercie tout particulièrement François Radvanyi, Emmanuel Barillot et Jean-Jacques Daudin pour la relecture attentive de mon manuscrit et leurs remarques constructives. Un grand merci également à toutes les personnes qui m'ont fourni une aide précieuse dans la réalisation de ce manuscrit: Bernard Asselain, Marc Bollet, Laurence Calzone, Jérome Couturier, Pierre Neuvial, Priscilla Signorini, Joan Sobota, Vassili Soulemis, Marc-Henri Stern, Anne Vincent-Salomon et Yann de Ricke.

Je tiens à exprimer mes sincères remerciements à l'ensemble des mes collègues présents et passés de l'Institut Curie pour leur compétence, leur disponibilité, leur dynamisme, leur soutien et leur bonne humeur. Merci tout d'abord aux personnes du Service de Bioinformatique qui ont contribué de près à la réalisation de ce travail: Anna Biedak, Isabel Brito, Philippe La Rosa, Séverine Lair, Stéphane Liva, Pierre Neuvial, Guillem Rigaill, Nicolas Servant, Julien Trolet et Eric Viara. Merci ensuite aux personnes membres des unités du Centre de Recherche, de l'Hôpital et du Département de Transfert avec lesquelles ce travail a été réalisé en étroite collaboration: Anna Almeida, Alain Aurias, Marc Bollet, Caroline Brennetot, Jérome Couturier, Charles Decraene, Olivier Delattre, Isabelle Janoueix-Lerosey, Elodie Manié, Jean-Philippe Meyniel, Gaëlle Pierron, Sophie Piperno-Neumann, Céline Rouveirol, Simon Saule, Nicolas Stransky, Xavier Sastre, Vassili Soumelis, Marc-Henri Stern, Jean-Paul Thiery et Elisabetta Volpe. Sans vous, rien de ce qui est présenté ici n'aurait été réalisé. Merci également à tous ceux avec qui de nombreux projets ont pu et seront réalisés: Sabrina Carpentier, David Gentien, Pierre Gestraud, Eléonore Gravier, Georges Lucotte, Eugène Novikok, Patrick Poullet et Andrei Zynoviev.

Un grand merci à tous les membres du Service de Bioinformatique pour l'ambiance chaleureuse qui y règne. Merci également à tous les amateurs de blagues et de bilboquet (je pense qu'ils se reconnaîtront eux-mêmes car mieux vaut ne pas citer de noms): sachez d'ailleurs que mon record de 61 tient toujours, à bon entendeur ...

Je remercie enfin mes parents, ma famille et mes amis de leur soutien.

To my Grandfather

Contents

List of Figures

List of Tables

List of Acronyms

aCGH	array Comparative Genomic Hybridisation
ACTuDB	Array CGH Tumour DataBase
AIC	Akaike Information Criterion
BAC	Bacterial Artificial Chromosome
BIC	Bayesian Information Criterion
CAPweb	CGH Array Pipeline on the web
CART	Classification and Regression Tree
CGH	Comparative Genomic Hybridisation
CNV	Copy Number Variant
CSC	Cancer Stem Cell
DAIC	Discriminant AIC
DISQUAL	DIScriminante analysis with QUALitative variables
DNA	Deoxyribonucleic acid
FISH	Fluorescence In Situ Hybridisation
GA	Genetic Algorithm
GLAD	Gain and Loss Analysis of DNA
GLM	Generalised Linear Model
GSEA	Gene Set Enrichment Analysis
HMM	Hidden Markov Model
ICA	Independent Component Analysis
ITALICS	ITerative and Alternative normaLIsation and Copy number calling for affymetrix Snp arrays
LDA	Linear Discriminant Analysis
LM	Location Model
LOH	Loss of Heterozygosity
LOO	Leave-One-Out
MANOR	MicroArray NORmalisation
MANOVA	Multivariate Analysis of Variance
MCA	Multiple Correspondence Analysis
miRNA	microRNA
mRNA	messenger RNA
ncRNA	non-coding RNA

PCA	Principal Component Analysis
PLS	Partial Least Squares
PLS-LM	Partial Least Squares Location Model
PRESS	Prediction Residuals Sum of Squares
QDA	Quadratic Discriminant Analysis
QDA	Regularised Discriminant Analysis
RNA	Ribonucleic acid
siRNA	small interfering RNA
SVD	Singular Value Decomposition
VAMP	Visualisation and Analysis of Molecular Profiles

Notations

Typographical conventions:

$\mathbf{X}$ $\mathbf{\Sigma}$		bold letters denote matrix or vector

Mathematical notations:

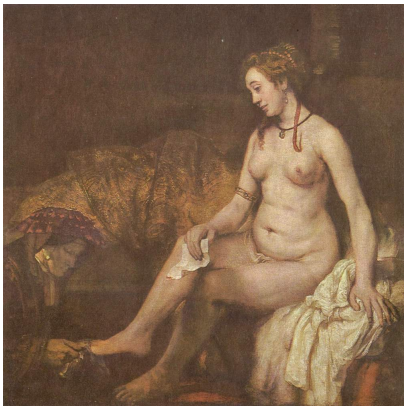
$\mathbf{X}'$		transpose of $\mathbf{X}$
$\mathbf{I}$		identity matrix
$\langle \mathbf{X}, \mathbf{Y} \rangle$		dot product
$\|\mathbf{X}\|^2$	$=$	$\langle \mathbf{X}, \mathbf{X} \rangle$
$rk(\mathbf{X})$		rank of the matrix $\mathbf{X}$
$\log$	$=$	Neperian logarithm
$card(\mathcal{G})$	$=$	cardinality of a set $\mathcal{G}$
$diag\{a_1, \cdots, a_K\}$	$=$	square matrix of size K with $a_1, \cdots, a_K$ on the diagonal and 0 elsewhere

Random distributions:

$X \sim \mathcal{L}$	X follows the distribution $\mathcal{L}$
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathcal{U}(a, b)$	uniform distribution between a and b
$\mathcal{LN}(\mu, \sigma^2)$	log-normal distribution with $\log(\mathcal{LN}(\mu, \sigma^2)) \sim \mathcal{N}(\mu, \sigma^2)$
$\mathcal{T}(k)$	centered student distribution with k degrees of freedom
$E(X)$, $V(X)$	expectation and variance of X
$COV(X, Y)$	covariance between X and Y
$COV_n(X, Y)$	empirical covariance between X and Y for a sample size n
$X_n \xrightarrow{d} X$	X_n converges in distribution to X
$X_n \xrightarrow{p} X$	X_n converges in probability to X
$P(X\|Y)$	conditional probability of X given Y

Notations used in chapter 3 for the supervised classification:

n	$=$	number of individuals
p	$=$	number of continuous predictor variables (*i.e.* number of genes)
q	$=$	number of discrete predictor variables
$\tilde{\mathbf{X}}$	$=$	$(\tilde{\mathbf{X}}_1, \cdots, \tilde{\mathbf{X}}_p) = (\tilde{X}_{ij})$, (n, p) matrix of continuous predictor variables
$\mathbf{X}$	$=$	$(\mathbf{X}_1, \cdots, \mathbf{X}_p) = (X_{ij})$ centered and scaled matrix $\tilde{\mathbf{X}}$
$\mathbf{Z}$	$=$	$(\mathbf{Z}_1, \cdots, \mathbf{Z}_q)$, (n, q) matrix of discrete predictor variables
$\tilde{\mathbf{Y}}$	$=$	vector of class labels
$\mathbf{Y}$	$=$	centered and scaled vector $\tilde{\mathbf{Y}}$
$\mathcal{G}_1, \cdots, \mathcal{G}_K$		K classes



Bathsheba at Her Bath
Rembrandt, *1654*

A qui donc sommes-nous? Qui nous a? qui nous mène?
Vautour fatalité, tiens-tu la race humaine?
Oh ! parlez, cieux vermeils,
L'âme sans fond tient-elle aux étoiles sans nombre?
Chaque rayon d'en haut est-il un fil de l'ombre
Liant l'homme aux soleils?

Victor Hugo, *Les Contemplations*

Preamble

The oldest description of cancer in humans was found in an Egyptian papyrus written between 3000-1500 BC. Hippocrates (460-370 BC), *the father of medicine*, described cancer in detail and used the Greek terms *carcinos* and *carcinoma* to refer to chronic ulcers or growths which seemed to be malignant tumours and *squirr(h)e* to refer to a type of cancer with a hard consistency. In Greek, *carcinos* means *crayfish, canker, cancer, tumour* and *skirros* means *solid tumour* as a noun and *hard, hardened* as an adjective. Celsus (28 BC - 50 AD), a Roman doctor, translated the Greek word *carcinos* into the word *cancer*, a Latin word meaning *crab, crayfish, dunce* and *cancer, canker*. There was a theory which claimed that the form of some cancerous lesions recalled the form of a crab. This explains the origin of the words *carcinos* and *cancer* to refer to these diseases. Galien (131-201), used the Greek term *oncos*, meaning *mass*, to refer to a growth or a tumour which looked malignant. In art history, testimonies of this disease can also be found. Rubens and Rembrandt were main baroque painters who practised realism, which means that they painted whatever their eyes captured. This has allowed physicians to discover alterations which suggest tumour in the breast of the models they painted. One of the most famous paintings which depicts breast tumour is the oil-on-canvas piece by Rembrandt, *Bathsheba at Her Bath*: an Italian surgeon first suggested that Rembrandt might have depicted breast tumour in his painting, accurately showing the clinical signs (the dark shadow on her left breast) of the fatal disease from which his model and mistress, Hendrickje Stoffels, was suffering (Vaidya, 2007). Therefore, cancer is not a modern disease but has passed through the ages for a long time and very likely from the origin of life. Although an old disease, cancer still remains complex and undeciphered. Moreover, cancer is a major cause of death in developed countries. In metropolitan France, a study has shown that, for the first time, cancer is the leading cause of death (30%) preceding cardiovascular diseases (28.9%) (Aouba et al., 2007). Therefore, lots of efforts must be made to defeat the disease.

In the field of life sciences, the twentieth century has seen the discipline of genetics deliver solutions to the most profound problem of biology: hows does the genetic constitution of a cell and organism determines its appearance and function? In 1944, Erwin Schrödinger (Nobel prize in physics, 1933) asked *"What is life?"* (Schrödinger, 1944) and wrote *"We believe a gene - or perhaps the whole chromosome fiber - to be an aperiodic solid"*. The nature of this *aperiodic crystal* was discovered in 1953 by Watson and Crick (Nobel prize in medicine, 1962): the DNA double helix. The advent of biotechnologies and molecular biology has

allowed researchers to sequence the human DNA. This project, called the Human Genome Project, was initiated in 1990 and completed in 2003. Man can now faces his biological destiny and the new knowledge arising from the field of molecular biology offers promising insights into medical sciences and especially in *oncology.

For centuries, the vocation of medicine was first *curative*. The improvement in both technologies and scientific knowledge has made possible to go towards a *predictive* medicine: *an ounce of prevention is worth a pound of cure*. Medicine is an art. Medicine is a science. The term *predictive medicine* was first introduced by Jean Dausset (Nobel prize in medicine, 1980). It is a medical approach which assesses the risk for any individual to develop a given disease based on factors of genetic predisposition. Predictive medicine is therefore statistical medicine. This new era of medicine is evolving, but before predicting what will happen to healthy individuals, this kind of medicine must help to predict the evolution of different pathologies. Let us take the example of breast cancer: to prevent the occurrence of distant *metastasis, the tumoral ablation is followed by adjuvant treatments (such as chemotherapy) whose secondary effects can damage the quality of life of the patient or even patient's health. In many cases, the adjuvant treatment could have been avoided but we do not yet have efficient factors for reliable prediction. This is the reason why we need to identify new factors in order to have an individually tailored therapy. In statistics, this issue of prediction is named supervised classification. Therefore, new biological factors which can help the clinician to choose the most suitable treatment are needed. The field of biotechnologies offers new insights to provide these new biological factors.

High throughput technologies make it possible to quantify molecular profiles at different levels inside the cancer cells. The most widely used and famous technique is the *microarray* which allows the simultaneous monitoring of the expression of all the genes in a sample. This technique has been adapted to investigate other molecular profiles and more particularly the DNA copy number of cancer cells. Indeed, a common characteristic among the different types of cancer is the modification in the number of their chromosomes. Therefore, the characterisation of the modification of DNA copy number in cancer gives valuable new biological information to better understand the disease and to help the clinician in the choice of therapy. The aim of the thesis was to develop statistical methods to analyse DNA copy number microarray experiments for oncology purposes. The outline is the following:

- **Chapter 1** introduces the clinical, biological and biotechnical contexts of the thesis.
- **Chapter 2** presents the statistical methods which have been developed during the thesis to analyse DNA copy number microarray experiments. A clinical application using the statistical methods we had developed is also described.
- **Chapter 3** describes the methodology we have developed to combine both gene expression and DNA copy number microarray experiments in order to perform supervised classification. As already mentioned, the goal of supervised classification is to help the clinician to choose an individually tailored therapy.

At the end of the manuscript the following information is provided:

- The **list of publications** is available.
- The definition of words flagged by * is available in the **glossary**.
- In the **Annexes** are provided the publications which have not been included in **Chapter 2**. Statistical and mathematical theorems which have been used to prove our mathematical propositions are also supplied in this part. Supplementary materials of **Chapter 3** are also supplied.

Galacidalacidesoxyribonucleicacid
Salvador Dalí, *1963*

Toutes les idées des arts ont leurs modèles dans la production de la nature: Dieu a créé et l'homme imite.

George-Louis Leclerc, Comte de Buffon, *Histoire naturelle, Premier discours*

1

Cancer and high-throughput technologies in oncology

Contents

This chapter is an introduction to the clinical, biological and biotechnical contexts of the thesis. Firstly the main challenges in *oncology research are presented. Secondly, the basics of biology for both the normal and cancer cell are described in order to show how molecular biology can help to solve the challenges. Then, the high-throughput technologies currently used in cancer studies are presented with a particular focus on the characterisation of the alterations in DNA copy number and the quantification of gene expression from microarray experiments. Finally, the need for bioinformatics analysis for DNA copy number microarray experiments is justified.

1.1 Challenges in oncology research

The current challenges in oncology research are dominated by three main aspects:

Identifying new prognostic and predictive factors When a cancer is detected, the clinician has to choose the most suitable therapy for each patient. Let us take for example breast cancer as described by Lønning (2007) and Cianfrocca and Goldstein (2004). This is not restrictive as the principles can be extended to other types of cancer (Sawyers, 2008). The questions the clinician must answer to tailor the therapy are *Who needs treatment (or how aggressively to treat)?* and *Which treatment is best?* (see **Figure 1.1**). The factors used by the clinician to answer these questions are termed *prognostic* and *predictive* respectively. Prognostic factors assess patient's risk of relapse and are traditionally used to identify patients who can be spared unnecessary therapy. Predictive factors, in contrast, determine the responsiveness of a particular tumour to a specific treatment. The factors, which can be used for both prognostication and prediction, are collected as follows. Typically, a *pathologist looks at histological sections[1] of the tumour (see **Figure 1.2**) and assesses parameters such as the appearance of the cells, the size and the shape of the cancer cell nuclei, the number of mitoses, the invasiveness of adjacent tissues and then determines the histological type. Tumour size and axillary lymph-node status are used to define the stage (which quantifies the extension and the size of the tumour) and the grade (which reflects the size of the nuclei, the proliferative activity within the tumour evaluated on ten high power field images and the differentiation of the tumour - Elston and Ellis, 1991). Besides these histological parameters, the presence of specific markers is detected by immunohistochemical methods: for example, in breast cancer, the presence of oestrogen (ER), progesterone (PR) and *HER2/neu*[2] receptors are evaluated. Other clinical parameters such as the age of the patient can also be used. All these parameters are referred to as clinico-histopathological criteria and currently determine the choice of the therapy made by the clinician. Yet, taking into account all these criteria does not allow efficient and individually tailored therapies due to the following limitations:

- Some tumours which have similar clinico-histopathological criteria may not have the same clinical outcome. This can be explained by both the existence of cancer subtypes which need to be discovered and the lack of efficient clinico-histopathological prognostic factors. As a consequence, some patients might be given a treatment while they should not or others might not be given a treatment while they should; the trend being rather the first case. Thus, the most important benefit of new prognostic factors may be to help clinicians identify patients in whom therapy could be avoided, sparing these patients treatment-related side effects. Indeed, Van't Veer and Bernards (2008) report that recurrence is likely in 20-30% of young women with early-stage (lymph-node-negative) breast cancer who only undergo surgery and localised radiation treatment. Yet, in the United States, 85-95% of women with this type of cancer receive adjuvant chemotherapy, mostly because conventional clinicopathological criteria fail to identify reliably those patients who are likely to relapse. Therefore, 55-75% of women with early-stage breast cancer in the United States undergo a toxic therapy from which they will not benefit but of which they will experience the side effects.

- While the identification of reliable predictive factors has the potential to spare patients ineffective treatment and unnecessary side effects, the reverse (that a factor may guarantee therapeutic success) may be more difficult to achieve. Thus, while ER negativity

[1] Tumoral tissue from the patient can be obtained using biopsy techniques or after surgical removal of the tumour.

[2] see **Subsection 1.3.4 - Page 26**

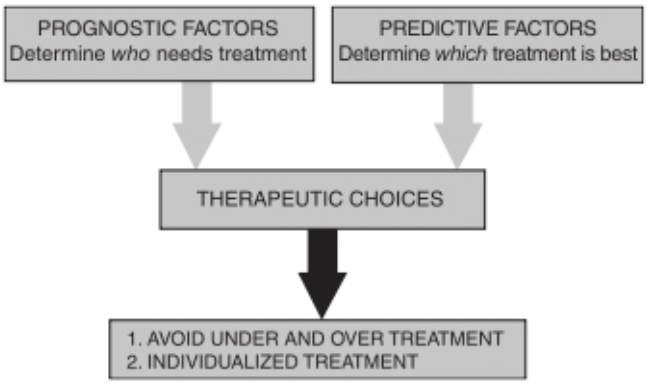


Figure 1.1: Need for prognostic and predictive factors in breast cancer (image from Lønning, 2007).

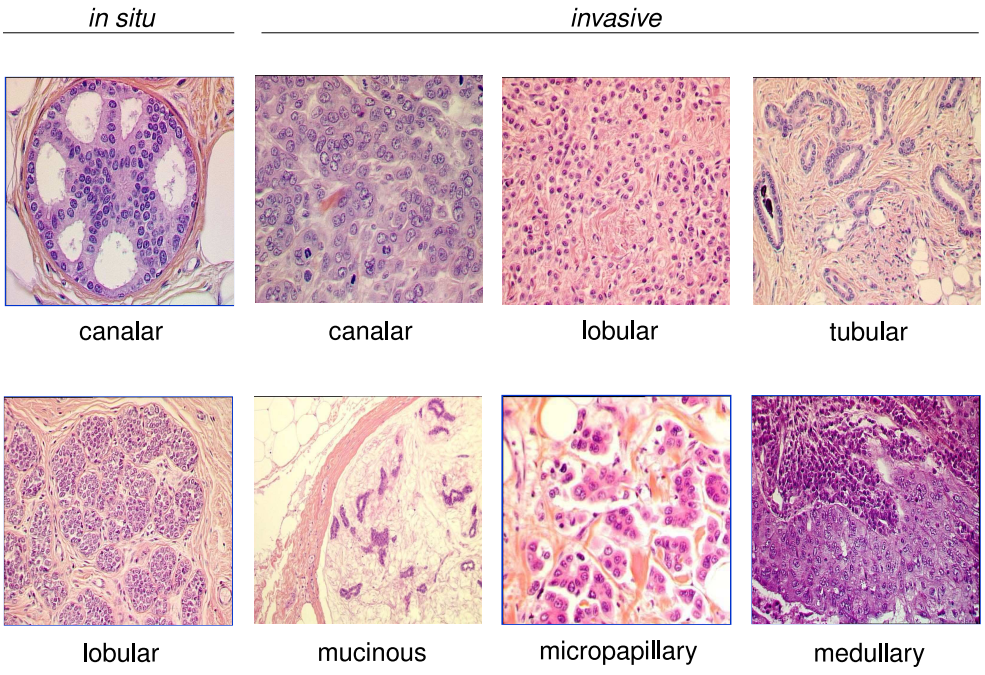


Figure 1.2: Histological sections of breast cancer - The histological sections are used by the pathologist to classify the tumours into histological types (Ellis et al., 1992) and to determine the stage and grade of the tumour. *In situ* tumours do not spread to the surrounding tissues while *invasive* tumours have started to break through normal breast tissue barriers and invade surrounding areas (images provided by Anne Vincent-Salomon, Institut Curie).

is associated with lack of response to endocrine treatment (hormonotherapy), not all patients with ER positive tumours may benefit from such therapy. Similarly, while the absence of *HER2/neu* overexpression has been established as a predictive factor for non-responsiveness to herceptin therapy, not all *HER2/neu*-overexpressing tumours are herceptin sensitive, reflecting the complexity of breast cancer genetics.

- The expertise varies from one pathologist to another: Billerey and Boccon-Gibod (1996) have shown on bladder cancer (and this is very likely to be true in any cancer) that the concordance for the grade and stage assessment between different pathologists was about 67% and 69% respectively. Therefore, depending on the pathologist, the patient might not be given the same therapy and thus the reproductibility for the grade and stage assessment needs to be improved.

Finding new therapeutic targets There are some subtypes of cancer for which no efficient treatment is available and therefore it is important to identify new therapeutic targets. This is the case for basal-like tumour in breast cancer.

Better understanding *tumoral progression Cancer is a disease characterised by different key properties[3] which have been well described in the literature. Nevertheless, the underlying molecular mechanisms involved during the tumoral progression are far from being understood and cancer still remains a complex and undeciphered disease.

This thesis is devoted to the first point of the three challenges: we are interested in the identification of new prognostic and predictive factors to help the clinician to provide patients with better tailored therapies. Besides current clinico-histopathological criteria, these new factors must allow researchers *(i)* to discover new subtypes of tumours, *(ii)* to accurately assess the risk of relapse *(iii)* to predict the response to treatment, and *(iv)* to improve the assessment of the grade and stage. We will particularly focus on points *(ii)*, *(iii)* and *(iv)* which are supervised classification problems. They will be raised in **Section 2.5** and **Chapter 3**. The essential issue to answer is what are the new factors which can help the clinician to take a decision? Clearly, molecular profiling provides promising insights into this issue (Van't Veer and Bernards, 2008; Thiery et al., 2006, this latter article is supplied in the **Annexes**). The problem is clearly to define the most relevant molecular levels we need to investigate in order to efficiently identify the new factors also called biomarkers. Therefore, an overview of the molecular biology of both the normal and cancer cell is necessary to understand what are the key molecular levels to study. This is the scope of the following two sections.

1.2 *Biology of the cell* for beginners

To understand the mechanism of tumoral progression, it is necessary to understand how normal cells work and how they are integrated at the level of the whole organism. A cell is the smallest unit of an organism which is classified as living, and it is sometimes called the building block of life (Alberts et al., 2002) (see **Figure 1.3**). A living organism can be seen as an ecosystem whose members are cells, reproducing by cell division and organised into collaborative assemblies or tissues. This ecosystem is very particular since in a healthy organism there is no competition between the different cell populations: each cellular type completes its specialised function which ensures that the organism can live and reproduce.

[3] See **Subsection 1.3.5**.

To coordinate their behaviour, the cells send, receive and interpret an elaborate set of signals which serve as social controls, telling each of them how to act. As a result, each cell behaves in a socially responsible manner, resting, dividing, differentiating[4] or dying as needed for the good of the organism and the maintenance of its integrity. In cancer cells, we will see that this harmony is broken: the collaboration between cells disappears and a competition and selection between cancer cells appear which can lead to the death of the organism. To complete its specialised function, the cell follows a specific program which is described in the next subsection.

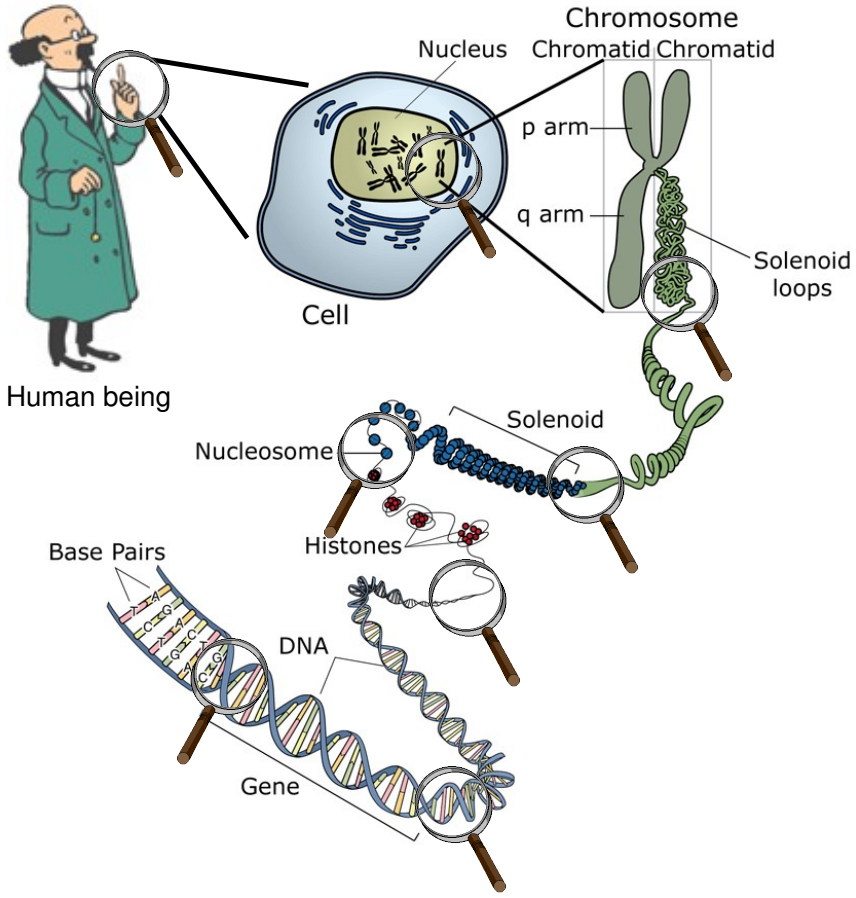


Figure 1.3: Hierarchical representation of a multi-cellular living organism - A living organism consists of building blocks of life called cells. In a cell, there are chromosomes packing the DNA into a solenoid conformation called *chromatin (p and q define the short and long chromosome arms respectively). In the chromatin, DNA is wrapped around *nucleosomes. The DNA is the molecule which carries the genetic information. A gene is a DNA segment which encodes for a specific cellular function. A gene is a sequence of bases A, T, C and G.

1.2.1 *Central dogma of molecular biology*

Each cell follows a specific program which involves different molecular partners. The information flow which allows the program to be completed inside the cell has been formalised by Crick (1970): *"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid"*. The flow of biological information is presented in **Figure 1.4** and the general principles can be summarised as follows:

[4]Cellular differentiation is the process by which a less specialised cell becomes a more specialised cell type.

the molecular partners: **DNA** stores the information in a linear fashion[5] and can be split into segments or genes which encode for a specific function of the cell. **RNA** can be viewed as the template which allows the synthesis of the **protein**. This last partner is the effector within the cell of the function encoded by the gene. In cells, the DNA is packed into entities called chromosomes (see **Figure 1.3**).

the flows: the step which converts DNA into RNA is called **transcription** and the step which converts RNA into protein is called **translation**. DNA can also be duplicated during the **replication**. This process occurs during the cell cycle in which a parent cell reproduces into two daughter cells. This process allows the conservation of the program information in daughter cells.

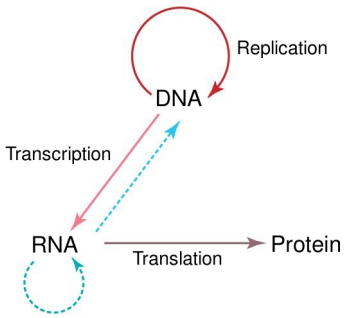


Figure 1.4: Central dogma of molecular biology - The central dogma of molecular biology holds that information flows from DNA to RNA to protein. *Solid arrows* indicate information flows which occur in all cells, through DNA replication, transcription of DNA into RNA, and translation of RNA into protein. *Dotted arrows* indicate flows which are seen occasionally, through reverse transcription and replication of RNA. Crucially, information cannot flow from protein back into nucleic acid sequence (adapted from Crick, 1970).

Towards a new paradigm: expanding the *central dogma* The central dogma of biology holds that genetic information normally flows from DNA to RNA to protein. As a consequence, it has usually been assumed that genes generally encode for proteins, and that proteins are those which fulfil the functions, in all cells, from microbes to mammals. However, the fact that genes encode for proteins may not be the case in complex organisms. Indeed, recent evidence suggests that the majority of the genome of mammals and other complex organisms is in fact transcribed into RNA which does not encode a protein: such RNA is termed non-coding RNA (ncRNA) but this does not mean that it does not contain information nor has a function (Mattick, 2003; Mattick and Makunin, 2006). To distinguish RNA which does not encode protein from RNA which encodes protein, the latter is termed messenger RNA (mRNA). ncRNAs can be divided into two classes: the infrastructural and the small regulatory ncRNAs. Among the infrastructural ncRNAs, there are transfer RNAs, ribosomal RNAs and small nuclear RNAs. They can be involved in regulatory processes. Small regulatory ncRNAs interact with mRNA via RNA interference mechanisms[6] (Mello and Conte Jr, 2004) and inhibit gene expression at the stage of translation (see **Figure 1.5**). Among the different types of small regulatory ncRNAs, microRNAs (miRNA) are naturally produced in human cells, and small interfering RNAs (siRNA) have also be recently identified to be produced endogeneously in mouse oocytes (Watanabe et al., 2008). The world of ncRNAs gives new insights into the understanding of gene regulation and is a very active field of research. In **Subsection 1.2.2**, we will give more details about the role of miRNAs in gene expression regulation. ncRNAs can be viewed as non-coding genes.

[5] The sequence of bases A, T C and G.

[6] In 2006, Andrew Z. Fire and Craig C. Mello received the Nobel Prize in medicine for the discovery of the process RNA interference.

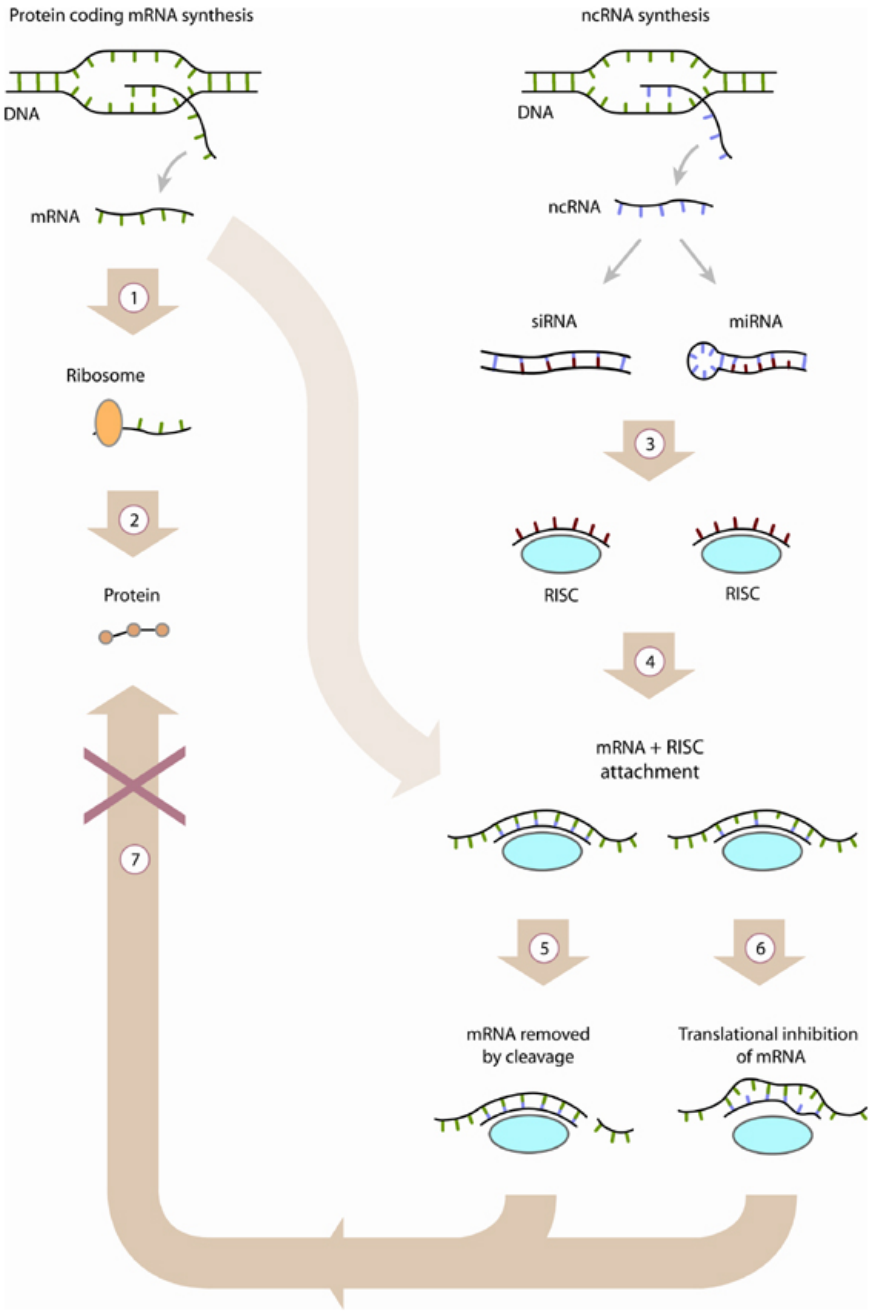


Figure 1.5: Expanding the central dogma: ncRNA and RNA interactions - RNA transcribed from a protein coding gene (1) may be translated into a protein (2). siRNA and miRNA interact with RISC (3). RISC ncRNA complex may then interact with the transcribed RNA (4). Perfectly complementary ncRNA and RNA binding will then subject the RNA to cleavage and degradation (5). Imperfectly complementary ncRNA:RNA may allow the RNA to bind with ribosomes, but interferes with translation (6). Either mechanism precludes protein expression (7). RISC is the RNA-induced silencing complex: it is a multi-protein complex which binds to ncRNAs. A ribosome is the translation machinery. (image and legend from Perkins et al., 2005)

The genetic information carried by a gene is called the **genotype** and when the function encoded by the gene is effective within the cell it is called the **phenotype**. The genotype corresponds to the genetic description of the cell while the phenotype corresponds to the expression of the function encoded by the gene. It is possible that this function is never expressed if the cell does not need it. Then, how does the cell decide to express or not the function encoded by the gene? This is determined by the interaction between the conditions in which the cell lives and the genetic properties of the cell: the cell has many sensors which are sensitive to **environmental stimuli** which are either external (temperature, pH, nutrients, light, pathogen molecules, signals sent from other cells, *etc.*) or internal (DNA damage, length of the telomere[7], osmotic pressure, *etc.*). Thus, the expression of the phenotype is determined by the simple equation **genotype** + **environmental stimuli** → **phenotype**. Environmental stimuli depend not only on the stimuli at a given time but also on the stimuli the cell has been submitted to during its life. As a consequence, the cell has specific characteristics due to its life history[8]. Therefore, the environmental stimuli received by the cell will trigger or not the expression of the phenotype: these stimuli play a key-role in the regulation of the information flow. Indeed, although the dogma appears to be a simple sequential information flow, the reality is much more complex as many interactions between the cell and its environment impact the control of replication, transcription and translation. We will see in what follows that in a cancer cell, the mechanisms which control the flow of the central dogma are altered in such a way that the cell can no longer complete its original program. As a result, the cell cannot complete its specialised function. We will describe in the next subsection the mechanisms involved in the regulation of the information flow in a normal cell.

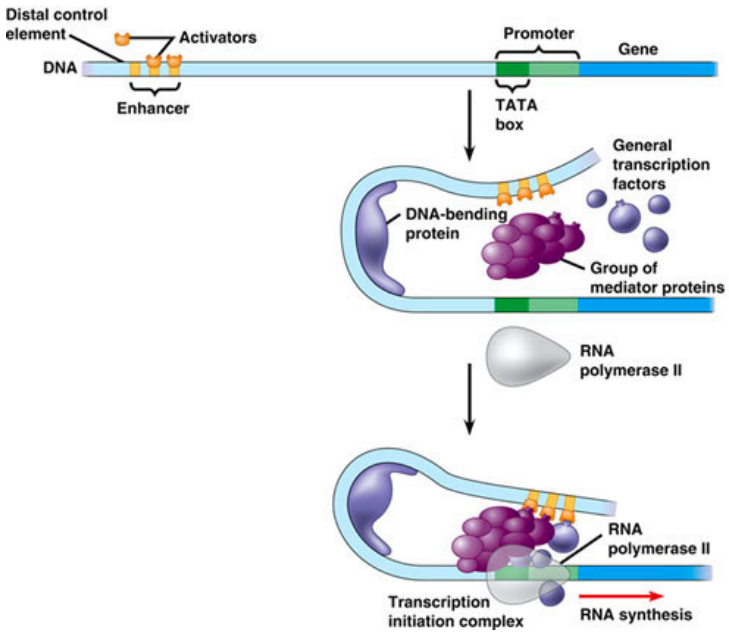


Figure 1.6: Role of transcription factor in gene expression regulation - The transcription factor binds to specific DNA sequences of the promoter located in the upstream region of the gene. This allows the formation of a transcription initiation complex including the RNA polymerase which starts the transcription of the gene into RNA (from http://fig.cox.miami.edu/ cmallery/150/gene/c7.19.6.activators.jpg).

[7] The telomere is the extremity of the chromosome. Its length is an indicator of the number of divisions a cell has undergone.

[8] For example, the differentiation signals have conferred the cell with a specialised function.

1.2.2 Regulation mechanisms of the information flow

The regulation of the flow is mediated by a large set of molecular entities and mechanisms which are detailed below:

Transcription factors Groups of genes must be coordinately expressed while other genes must be repressed so that the cells display complex and tissue-specific phenotypes. Such coordination of expression is the role of **transcription factors**. They are proteins which regulate transcription: they bind to specific sequences of DNA using DNA binding domains and contribute to the regulation of gene expression (see **Figure 1.6**). Transcription factors perform this function alone, or by using other proteins in a complex, by increasing (as an activator), or preventing (as a repressor) the presence of RNA polymerase, a protein which transcribes genetic information into RNA. One transcription factor might have several target genes. Active research is currently going on based on the sequence analysis of promoters in order to discover new target genes for each transcription factor (Tompa et al., 2005). Among transcription factors, let us mention *p53*, also known as *the guardian of the genome*: it plays a key-role in preserving the integrity of the genome during the cell cycle in order to ensure that the specific program of the cell is correctly transmitted into daughter cells.

Epigenetic regulations Classical genetics alone cannot explain the diversity of phenotypes within a population. Nor does classical genetics explain how, despite their identical DNA sequences, monozygotic twins or cloned animals can have different phenotypes and different susceptibilities to a disease. The concept of epigenetics offers a partial explanation of these phenomena. First introduced by Conrad Hal Waddington in the forties to name *"the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being"* (Jablonka and Lamb, 2002; Speybroeck, 2002), epigenetics was later defined as heritable changes in gene expression which are not due to any alteration in the DNA sequence (Esteller, 2008). Epigenetics refers to features such as *chromatin and DNA modifications which are stable over series of cell cycle but do not involve changes in the underlying DNA sequence of the organism. These modifications play an important role in gene silencing at the level of transcription. The main modifications are the following (see **Figure 1.7**):

- **DNA methylation** is a common epigenetic mechanism of gene silencing. Methylation is a chemical modification of the DNA which can be either inherited, created or modified in response to environmental stimuli without changing the DNA sequence. DNA methylation occurs in cytosines which precede guanines in dinucleotide called CpGs. CpG sites are not randomly distributed in the genome but are located in CpG-rich regions known as *CpG islands which span the $5'$ end[9] of the regulatory region of many genes. These islands are usually not methylated in normal cells. DNA hypermethylation is required in particular cases such as *genomic imprinting and the X-chromosome inactivation in females[10]. DNA hypermethylation inside repeated sequences could also have a role in the protection of chromosomal integrity, by preventing chromosomal instability and translocations[11].

- ***histone modifications** is another common epigenetic mechanism. Transcription of DNA is dictated by the structure of the chromatin. In general, the density of its packing is indicative of the frequency of transcription. Octameric protein complexes called

[9] A DNA sequence is oriented from $5'$ end to $3'$ end.

[10] In mammalian females, one X chromosome is inactivated.

[11] Translocation is defined in **Subsection 1.3.6**.

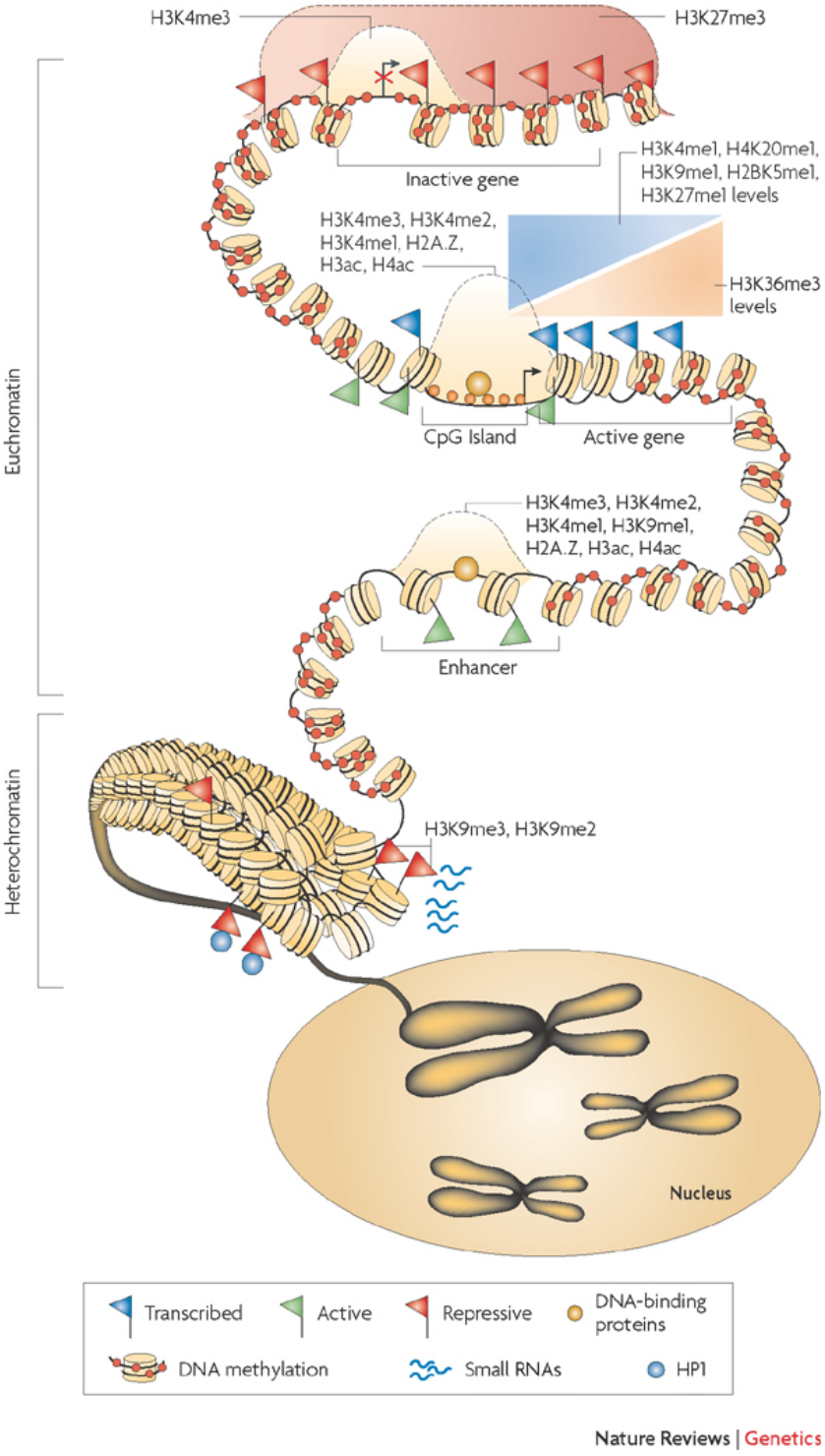


Figure 1.7: Characteristics of epigenome - The interaction of DNA methylation, histone modification, nucleosome positioning and other factors such as small RNAs contribute to an overall epigenome that regulates gene expression and allows cells to remember their identity. Chromosomes are divided into accessible regions of euchromatin and poorly accessible regions of heterochromatin. Heterochromatic regions are marked with histone H3 lysine 9 di- and trimethylation (H3K9me2 and H3K9me3), which serve as a platform for HP1 (heterochromatic protein 1) binding. Small RNAs have been implicated in the maintenance of heterochromatin. DNA methylation is persistent throughout genomes, and is missing only in regions such as CpG islands, promoters and possibly enhancers. The H3K27me3 modification is present in broad domains that encompass inactive genes. Histone modifications including H3K4me3, H3K4me2, H3K4me1 as well as histone acetylation and histone variant H2A.Z mark the transcription start site regions of active genes. The monomethylations of H3K4, H3K9, H3K27, H4K20 and H2BK5 mark actively transcribed regions, peaking near the 5' end of genes. The trimethylation of H3K36 also marks actively transcribed regions, but peaks near the 3' end of genes. (image and legend from Schones and Zhao, 2008).

histones are responsible for chromatin packing, and these complexes can be temporarily or more permanently modified by processes such as methylation and acetylation. These modifications lead to a high degree of packing which prevent genes from being accessible by the transcriptional machinery. Therefore, the modifications act as a gene silencing process. The emerging model is that specific combinations of histone modifications confer the overall expression status of a region of chromatin, a theory known as the *histone code* hypothesis (Turner, 2002).

Post-transcriptional regulations These regulations occur at the RNA level once the DNA has been transcribed. They are the following:

- **alternative splicing** is the mechanism in which the exons of the primary gene transcript, the pre-RNA, are separated and reconnected to produce alternative RNA rearrangements. These linear combinations then undergo the process of translation, resulting in isoform proteins. Alternative splicing increases mRNA and protein diversity by allowing generation of multiple RNA products from a single gene (see **Figure 1.8**). For a given gene, only some splicing variants exist and not all exon combinations are possible. This is another plausible mechanism for the paradoxical inconsistency between the number of genes transcribed and the diversity of phenotypes.

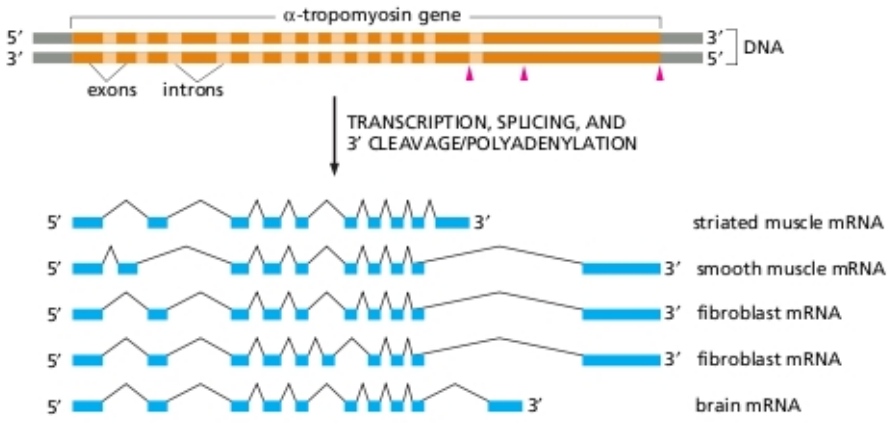


Figure 1.8: Alternative splicing of the α-tropomyosin gene from rat - α-Tropomyosin is a protein which regulates contraction in muscle cells. The primary transcript can be spliced in different ways, as indicated in the figure, to produce distinct mRNAs, which then give rise to variant proteins. Some of the splicing patterns are specific for certain types of cells. For example, the α-tropomyosin made in striated muscle is different from that made from the same gene in smooth muscle. The arrowheads in the top part of the figure mark the sites where cleavage and poly-A addition form the 3′ ends of the mature mRNAs (image and legend from Alberts et al., 2002).

- **miRNAs** are about 21-nucleotide-long single-stranded RNA molecules which regulate gene expression. They are encoded by genes which are transcribed from DNA but not translated into protein. miRNAs are processed from precursor molecules which fold into hairpin structures containing imperfectly base-paired stems. The precursor is processed by enzymes into a mature miRNA which is a single-strand RNA molecule (see **Figure 1.9**). Functional studies indicate that miRNAs participate in the regulation of almost every cellular process: in human a thousand miRNAs are predicted which would regulate about 30% of all protein-coding genes (Filipowicz et al., 2008). miRNAs control gene expression post-transcriptionally by regulating RNA translation or stability in the cytoplasm: they act similarly to siRNAs operating in RNA interference binding imperfectly to its RNA sequence target. The most stringent requirement is

a contiguous and perfect base pairing of the miRNA nucleotides 2-8, representing the *seed* region, which nucleates the interaction with the RNA. The mechanistic details of the function of miRNAs in repressing protein synthesis are still poorly understood. The paramount open question is whether miRNAs inhibit protein synthesis by a primary single mechanism or by different mechanisms. In other words, is it possible that miRNAs trigger an initial event which is then amplified by different mechanisms? A simple, alternative mechanistic model posits that the earliest event in protein-synthesis repression is the inhibition of the translation and that secondary effects of this inhibition could then be manifested at other steps, such as RNA degradation or proteolysis of the nascent polypeptide chains (Filipowicz et al., 2008).

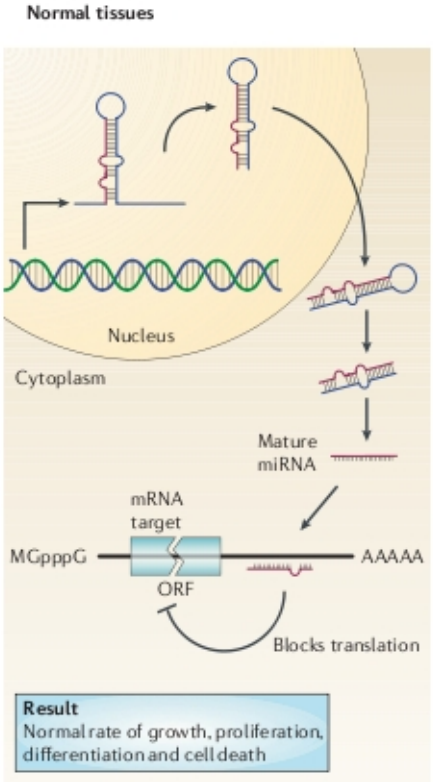


Figure 1.9: Role of miRNA in a normal cell - In normal tissues, proper miRNA transcription, processing and binding to complementary sequences on the target RNA results in the repression of target-gene expression through a block in protein translation or altered RNA stability (not shown). The overall result is normal rates of cellular growth, proliferation, differentiation and cell death. ORF, open reading frame (image and legend from Esquela-Kerscher and Slack, 2006).

Signal transduction We have seen in **Subsection 1.2.1** that once the DNA is transcribed into RNA, and the RNA is translated into protein, then the protein can play its biological function in the cell. This is in fact not completely true. Indeed, some proteins which are present within the cell are present in an inactive state. Modifications, called **post-translational**, are needed so that the protein can express its biological function. Why is this mechanism necessary? In the cell, some proteins are present just in case they are needed to ensure a quick and efficient response to environmental stimuli[12]. Indeed, the transcription and translation machineries take time to complete (on the order of hours or much longer)

[12]We remind the reader that environmental stimuli are either external (temperature, pH, nutrients, light, pathogen molecules, signals send from other cells, *etc.*) or internal (DNA damage, length of the telomere, osmotic pressure, *etc.*). These stimuli are defined by the conditions in which the cell lives.

and the cell cannot wait for the protein to be present if needed immediately. Therefore, post-transcriptional events (mainly phosphorylations[13]) allow the protein to go from an inactive state to an active state in a process lasting a few minutes. The goal of signal transduction is to perform a cascade of phosphorylation in response to an environmental stimulus which implies the protein is active within the cell. The cascade also allows an amplification of the signal so that a relatively small stimulus elicits a large response: once activated, a protein can activate many other proteins involved in the next step so that the signal grows exponentially (see **Figure 1.10**). In this process, **kinase proteins** play a major role since they are the proteins which catalyse the phosphorylation reaction by opposition to **phosphatases** which reverse the reaction.

To illustrate the signal transduction mechanism, let us take for example *p53* which is a transcription factor. *p53* only plays a role when the cell has been exposed to a stress such as a DNA damage. In unstressed cells, once *p53* has been produced, it binds to another protein called *MDM2* which inactivates *p53*: when associated with *MDM2*, *p53* cannot bind onto DNA and therefore cannot bind to its target genes. The protein complexes *p53*-*MDM2* are exported into the cytoplasm where they are degraded in proteasomes. In some circumstances (especially when cells are suffering certain types of stress or damage), *p53* protein molecules must be protected from *MDM2* so that they can accumulate to functionally significant levels in the cell. This protection is achieved by phosphorylation of *p53* by protein-kinases *ATM* or *ATR*: they give the signal to *p53* to play its transcription factor activity. The *ATM* protein is also activated by a phosphorylation reaction as a result of a phosphorylation cascade initiated by stress or DNA damage (see **Figure 1.10**).

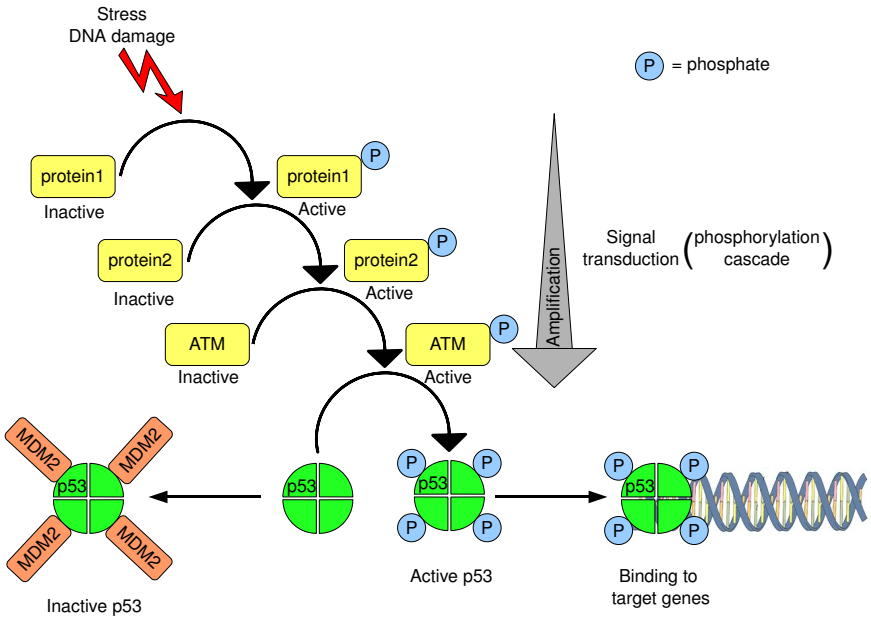


Figure 1.10: Signal transduction cascade - In response to stress (such as DNA damage) a first phosphorylation reaction activates the function of a first protein which can catalyse the phosphorylation of a second protein and so on. Finally, the phosphorylated *ATM* protein activates *p53* by a phosphorylation reaction too. Once phosphorylated, *p53* binds to its target genes and initiates the transcription. If not phosphorylated, *p53* binds to *MDM2* which prevents its transcription factor activity (adapted from Weinberg, 2007, chap. 9 and Nakamura, 1998).

Different signal transduction modules are involved in response to specific stimuli and are related to specific functions of the cell. These different modules are named **signaling pathways** and complex interactions between them exist as illustrated in **Figure 1.11** which is already a simplified view of the reality.

[13] Phosphorylation is the addition of a phosphate group to a protein molecule or a small molecule.

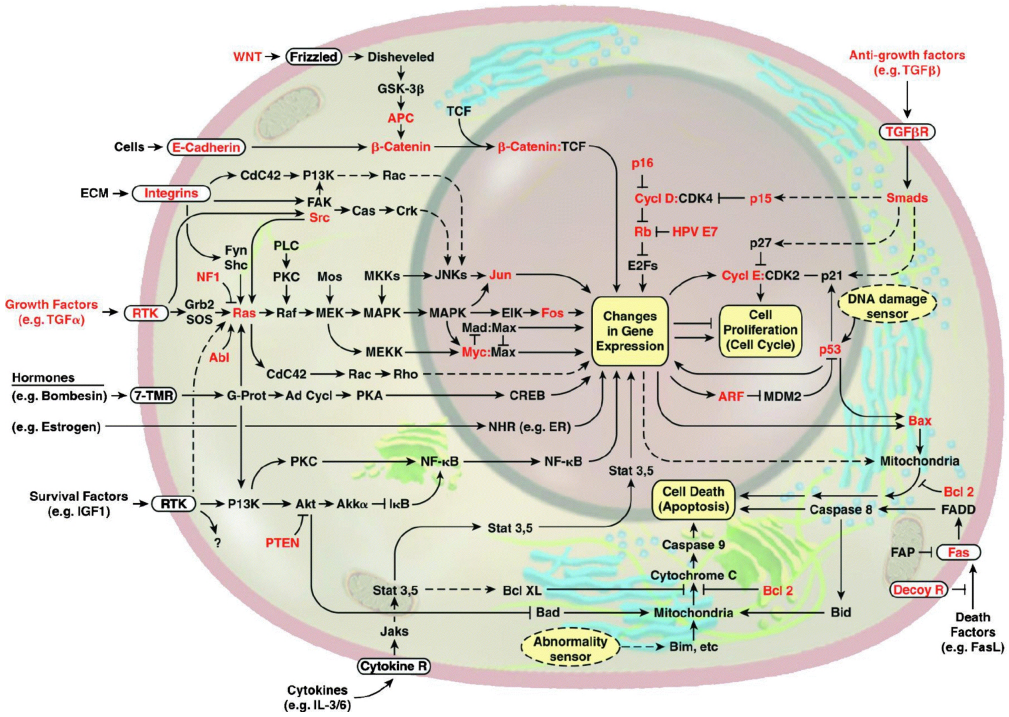


Figure 1.11: Emergent integrated circuit of the cell - Progress in dissecting signaling pathways has begun to lay out a circuitry that will likely mimic electronic integrated circuits in complexity and finesse, where transistors are replaced by proteins (e.g., kinases and phosphatases) and the electrons by phosphates and lipids, among others. In addition to the prototypical growth signaling circuit centered around *Ras* and coupled to a spectrum of extracellular cues, other component circuits transmit antigrowth and differentiation signals or mediate commands to live or die by apoptosis. As for the genetic reprogramming of this integrated circuit in cancer cells, some of the genes known to be functionally altered are highlighted in red (image and legend from Hanahan and Weinberg, 2000).

1.2.3 Life of a normal cell

To conclude this section, let us sum up the different stages of the life of a normal:

1. The cell performs its specialised function.
2. If needed, it reproduces during the cell cycle.
3. It dies after a limited number of cell cycles, a phenomenon called *senescence*.

These different steps are defined inside a program in which the sequential information flow has been formalised in the central dogma of molecular biology. As we have seen, the execution of the specific program of the cell involves complex regulation mechanisms in response to environmental stimuli: the control of gene expression by the transcription factors and alternative splicing, the epigenetic mechanisms, the regulatory function of ncRNAs and the signal transduction are key processes in the normal behaviour of the cell. In a cell, its specific program also includes permanent monitoring systems to check its ability to always behave in a responsible manner. If this is not the case, the cell must disappear and die in a process called *apoptosis*. Among the monitoring systems we can mention the *cell cycle checkpoints*. Indeed, it is important that after cell division, the daughter cells will be the exact copy of the parent cell in order to complete the same function otherwise the cell must enter the apoptosis process. We will see in the next section that the cancer cells which derive from normal cells are not able to perform the original specific program due to a sequential accumulation of events which have disturbed the monitoring system and regulatory mechanisms.

1.3 *Biology of cancer* for beginners

This section describes the events occurring during tumoral progression which transforms a normal cell into a cancer cell (see Weinberg (2007) and Alberts et al. (2002) for details).

1.3.1 Progressive accumulation of mutations

During a lifetime, cells die and have to be replaced to maintain the integrity of the organism. We have seen that cells reproduce during the cell cycle in which DNA is duplicated. In a normal human body, there are an estimated 10^{16} cell divisions which take place in the course of a lifetime. During cell division, there are fundamental limitations on the accuracy of DNA replication and repair so that mutations will occur spontaneously at an estimated rate of about 10^{-6} mutations per gene per cell division. This rate can be increased due to exposure to mutagenic agents such as chemical agents (*e.g.* tobacco smoke), physical agents (*e.g.* UV light) or biological agents (*e.g.* viruses). Thus, in a lifetime, every single gene is likely to have undergone mutation on about 10^{10} separate occasions in any individual human being. Among the resulting mutant cells, one might expect that there would be many which have disturbances in gene regulation. As a consequence, the harmonious behaviour of the cell with respect to its neighbours will be affected. Here, mutation is referred to as genetic change and corresponds to a modification of the DNA sequence. Non-genetic changes which are transmitted in the cell progeny can also occur during the cell cycle. They correspond to epigenetic characteristic modifications (see **Subsection 1.2.2**) and are called epimutations. Thus, the number of mutations (including genetic mutations and epimutations) is likely to be greater than 10^{10}. From this figure of 10^{10}, the problem of cancer does not seem to be why it occurs but why it occurs so infrequently. Clearly, if a single mutation was enough to convert a typical healthy cell into a cancer cell which proliferates without restraint, we would not be viable organisms. Therefore, many mutations are needed to cause cancer. Why are so many mutations needed? One reason is that cellular processes are controlled in complex and interconnected ways: cells employ redundant regulatory mechanisms to help them maintain tight and precise control over their behaviour. Thus, many different regulatory systems have to be disrupted before a cell can throw off its normal restraints and behave defiantly as a malignant cancer cell. In addition, tumour cells may meet new barriers to further expansion at each stage of the evolutionary process and therefore need to acquire additional mutations. Typically, the progressive accumulation of mutations allows the cell to acquire the six capabilities of a tumour (see **Subsection 1.3.5**) partly due to deregulations of apoptosis and senescence mechanisms. The accumulation of mutations which causes cancer explains why its prevalence increases with the age of individuals. Nevertheless, there exists paediatric cancers which involve less complex mechanisms.

Familial form of cancer In some cases, mutations which have been inherited from parents can be already present in the cell and transmitted from one generation to another. In breast cancer, we consider that about 15% of cancer cases can be attributed to inherited predisposition due to the presence of gene mutations. The well-known examples in breast cancer are the mutations of *BRCA1* and *BRCA2* genes which are involved in DNA repair during the cell cycle. These two genes account for about 16% of the familial risk of breast cancer. Since a mutation is constitutively present in the cells, the normal function of *BRCA* genes relies only on the remaining wild-type allele. Therefore, a patient who carries *BRCA1* or *BRCA2* mutations has a 10- to 20-fold higher risk of developing breast cancer (Stratton and Rahman, 2008). Whatever the cancer, the identification of new susceptibility alleles has direct application in the implementation of cancer prevention strategies.

1.3.2 Clonal origin of the tumour and stemness of cancer cells

Clonal origin The clonal evolution of tumour cell populations was proposed by Nowell in 1976 (cited in Vermeulen et al., 2008) and claims that tumour cells arise from a single cell of origin which has acquired an accumulation of mutations as described in the previous subsection. The model argues that tumour development proceeds via a process formally analogous to Darwinian evolution, in which a succession of genetic changes, each conferring one or another type of growth advantage, leads to the progressive conversion of normal human cells into cancer cells (Hanahan and Weinberg, 2000). From an initial population of slightly abnormal cells, descendants of a single mutant ancestor evolve from bad to worse through successive cycles of mutation and natural selection. At each stage, one cell acquires an additional mutation which gives it a selective advantage over its neighbours, making it better able to thrive in its environment (an environment which, inside a tumour, may be harsh, with low levels of oxygen, scarce nutrients, and the natural barriers to growth presented by the surrounding normal tissues). The offspring of this well-adapted cell will continue to divide, eventually taking over the tumour and becoming the dominant clone in the developing lesion. Thus, tumours grow in fits and starts, as additional advantageous mutations arise and the cells bearing them flourish. Their evolution involves a large element of chance and usually takes many years (Alberts et al., 2002).

Stemness of cancer cells In the Cancer Stem Cell (CSC) model, malignancies are viewed as abnormal organs with a *stem cell compartment which drives the growth. CSCs have been defined in analogy to normal stem cells, as cells which have the capacity to self-renew, meaning undergo divisions which allow the number of CSCs to remain constant and give rise to the variety of differentiated cells found in the malignancy (see **Figure 1.12**). The CSC model implies that, in a malignancy with a defined set of genetic alterations, cells with a different malignant potential are present. In a tumour both differentiated cells which have lost the capacity to propagate a tumour, and cells which retain a clonogenic capacity exist. The proposed hierarchical organisation of a malignancy could be easily integrated into the classical clonal selection theory of Nowell. As explained before, this theory views a malignancy as a clonally-derived cell population, which acquires new potentially advantageous mutations and gives rise to new more rapidly proliferating clones (see **Figure 1.13**). When one integrates the CSC theory into this model, the selection pressure is predicted to act at the level of the CSC compartment, implying that certain new traits in CSCs result in an increase in expansion of the CSCs due to self-renewal by symmetrical divisions. This does not mean, however, that certain features present only in more differentiated cells in the tumour could not be subject to selection, especially if this increases the expansion rate of the CSCs from which they are derived (Vermeulen et al., 2008). This theory is still under debate and the question is whether the cell of origin of the CSC has to be a stem cell or whether the accumulation of mutation convert differentiated cells back into stem cells. In this theory, the only cells capable of initiating and driving tumour growth are CSCs and it is logical to assume that a *metastasis arise from CSCs. This genetic framework for metastasis conflicts somewhat with new insights gained by molecular profile studies which point towards the fact that gene expression profiles of the complete tumour can predict metastatic behaviour of the malignancy.

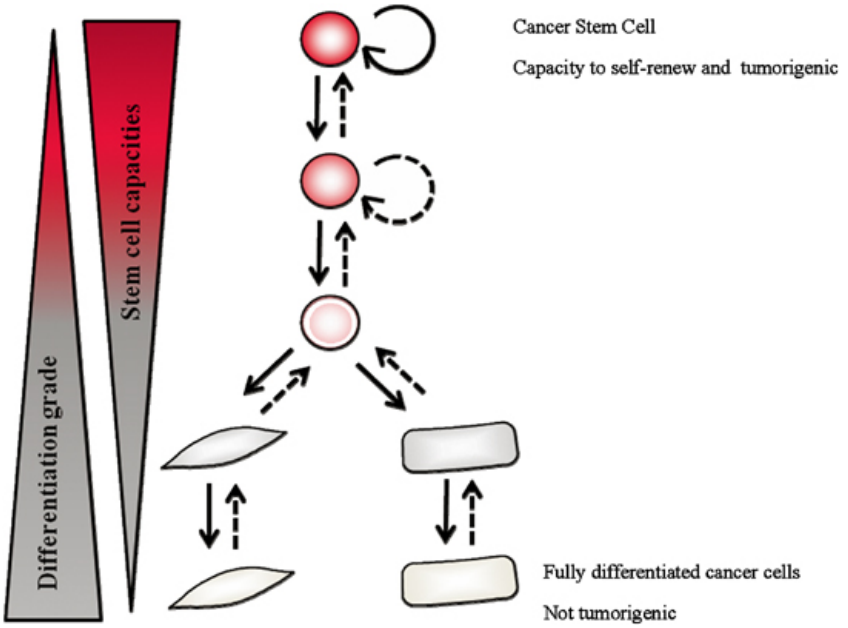


Figure 1.12: Hierarchical organisation of a malignant clone - Depicted is the proposed organisation of a malignant clone as predicted by the CSC model. The CSC on the top of the hierarchy (red) has the ability to self-renew, meaning generating more CSCs, and to spin off more differentiated cells (grey). It is to date not clear if the more differentiated cells can revert back and regain a more stem cell phenotype (image and legend from Vermeulen et al., 2008).

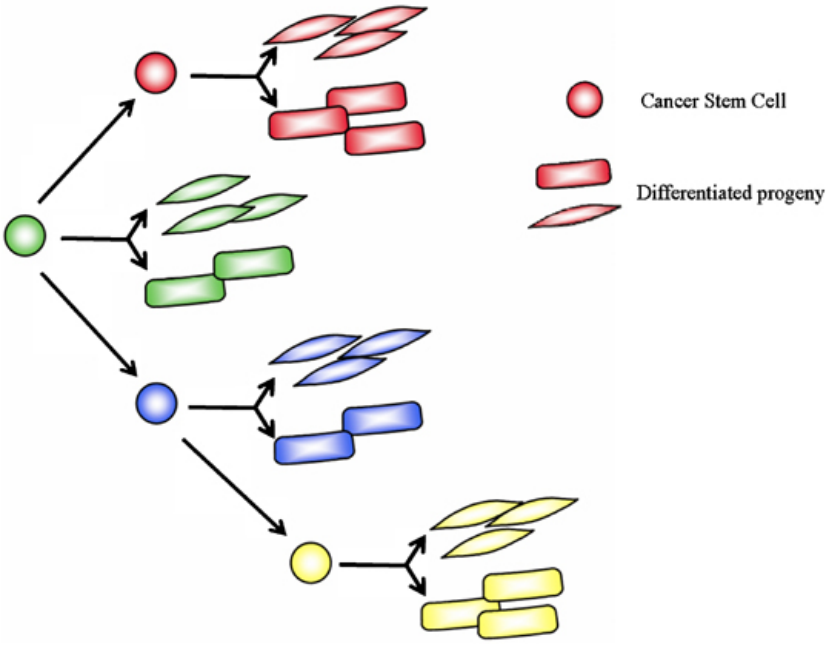


Figure 1.13: Clonal selection of hierarchical organised clones - Cancer stem cells with tumour initiating and tumour growth driving capacity give rise to more differentiated non-tumorigenic offspring. In this model selection pressure is predicted to act on the CSC level. CSCs acquire additional genetic alterations (here depicted by different colours) that can be beneficial for the clone *blue* and *yellow* or dreaded *red* (image and legend from Vermeulen et al., 2008).

1.3.3 Oncogenes and tumour-suppressor genes

As we have seen in **Subsection 1.3.1**, cancer is a disease of genes caused by the accumulation of mutations. The most important genes whose alterations are causal in tumoral progression are named cancer-critical genes (see Vogelstein and Kinzler, 2004). They are grouped into two broad classes, according to whether the cancer risk arises from too much activity of the gene product, or too little.

Oncogenes Genes of the first class, for which a gain-of-function mutation drives a cell toward cancer, are called **proto-oncogenes**. Their mutant and overactive forms are called **oncogenes** (see **Figure 1.14a**). Oncogenes encode proteins which control cell proliferation, apoptosis, or both. They can be activated by structural alterations resulting from mutation or gene fusion, by juxtaposition to enhancer elements and by amplification or translocations[14] (Croce, 2008; Alberts et al., 2002) (see **Figure 1.15**). Translocations and mutations can occur as initiating events or during tumour progression, whereas amplification usually occurs during tumour progression. The products of oncogenes can be classified into six broad groups: transcription factors (*e.g.* amplification of *MYCN*[15] in *neuroblastoma), chromatin remodelers (*e.g.* gene fusion of *MLL* in *leukemia), growth factors (*e.g.* gene fusion of *ABL1* in leukemia), growth factor receptors (*e.g.* mutation of *FGFR3*[16] in bladder cancer), signal transducers (*e.g.* mutation of *HRAS* in colon cancer), and apoptosis regulators (*e.g.* amplification of *MDM2*[17] in *sarcoma).

Tumour-suppressor genes Genes of the second class, for which a loss-of-function mutation creates the danger, are called **tumour-suppressor genes**. They have cancer-preventive effects which usually require the presence of only a single functional gene. To give rise to cancer these genes have to undergo biallellic inactivation in tumours: this is known as the Knudson two-hit model (Knudson, 1971) (see **Figure 1.14b**). Inheritance of a single mutant allele accelerates tumour susceptibility, because only one additional mutation is required for complete loss of gene function. This is why some tumour-suppressor genes have been identified in familial forms of cancer such as *RB1* in retinoblastoma (Knudson, 1971). As for oncogenes, tumour-suppressor genes are involved in many functions (Sherr, 2004). Among tumour-suppressor genes let us mention *TP53* (the gene which encodes the *p53* transcription factor involved in genome integrity maintenance - see **Page 11** and **Figure 1.10**), *ATM* (the gene which encodes a protein kinase involved in DNA damage signal transduction - see **Figure 1.10**), *BRCA1* and *BRCA2* (the genes which encode proteins involved in DNA repair during the cell cycle - see **Page 17**) and *RB1* (the gene which encodes the *pRb* protein involved in the control of cell cycle).

Non-protein-coding critical cancer genes Besides protein-coding genes, ncRNA and especially miRNAs can act as either an oncogene or a tumour-suppressor gene (Esquela-Kerscher and Slack, 2006; Fabbri et al., 2008). A defect in miRNA gene regulation leading to a loss or an amplification of miRNAs has been reported in a variety of cancers (Calin and Croce, 2006a). Typically, an under-expression of a miRNA which targets an oncogene (see **Figure 1.16a**) or an over-expression of a miRNA which targets a tumour-suppressor gene (see **Figure 1.16b**) will have an impact on cancer development.

[14] See **Subsection 1.3.6** for details about chromosome aberrations.

[15] Special dedication to Isabelle Janoueix-Lerosey.

[16] Special dedication to François Radvanyi.

[17] Special dedication to Alain Aurias.

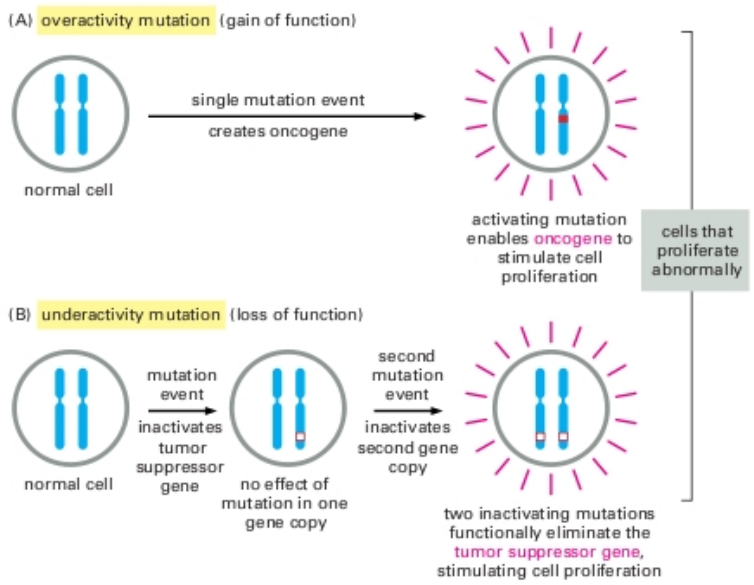


Figure 1.14: Oncogene and tumour-suppressor gene - (A) Oncogenes act in a dominant manner: a gain-of-function in a single copy of gene can drive a cell toward cancer. (B) Tumour-suppressor genes, on the other hand, generally act in a recessive manner: the function of both the alleles of the gene must be lost to drive a cell toward cancer. In this diagram, activating mutations are represented by solid red boxes, inactivating mutations by hollow red boxes (image and legend from Alberts et al., 2002).

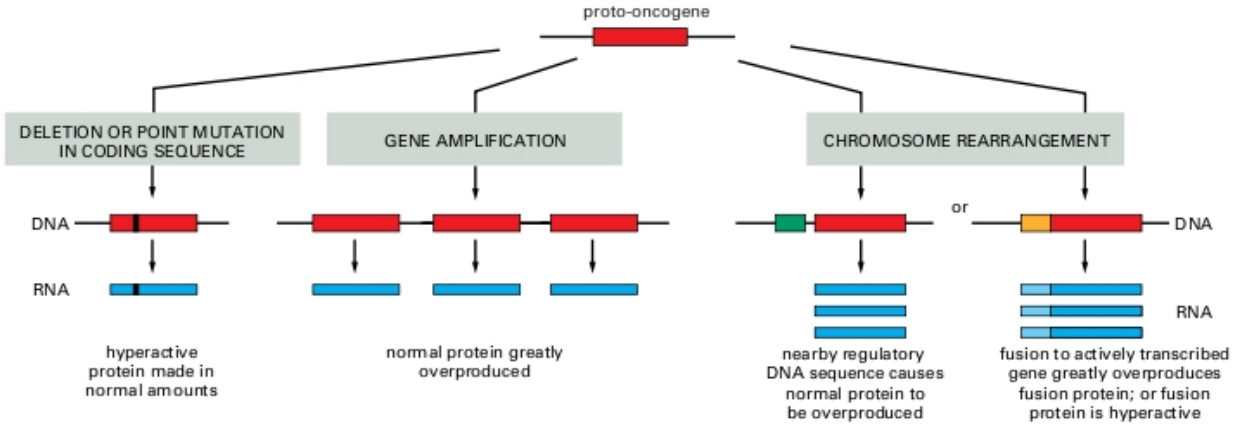


Figure 1.15: From proto-oncogene to oncogene - Three ways in which a proto-oncogene can be made overactive to convert it into an oncogene (image and legend from Alberts et al., 2002).

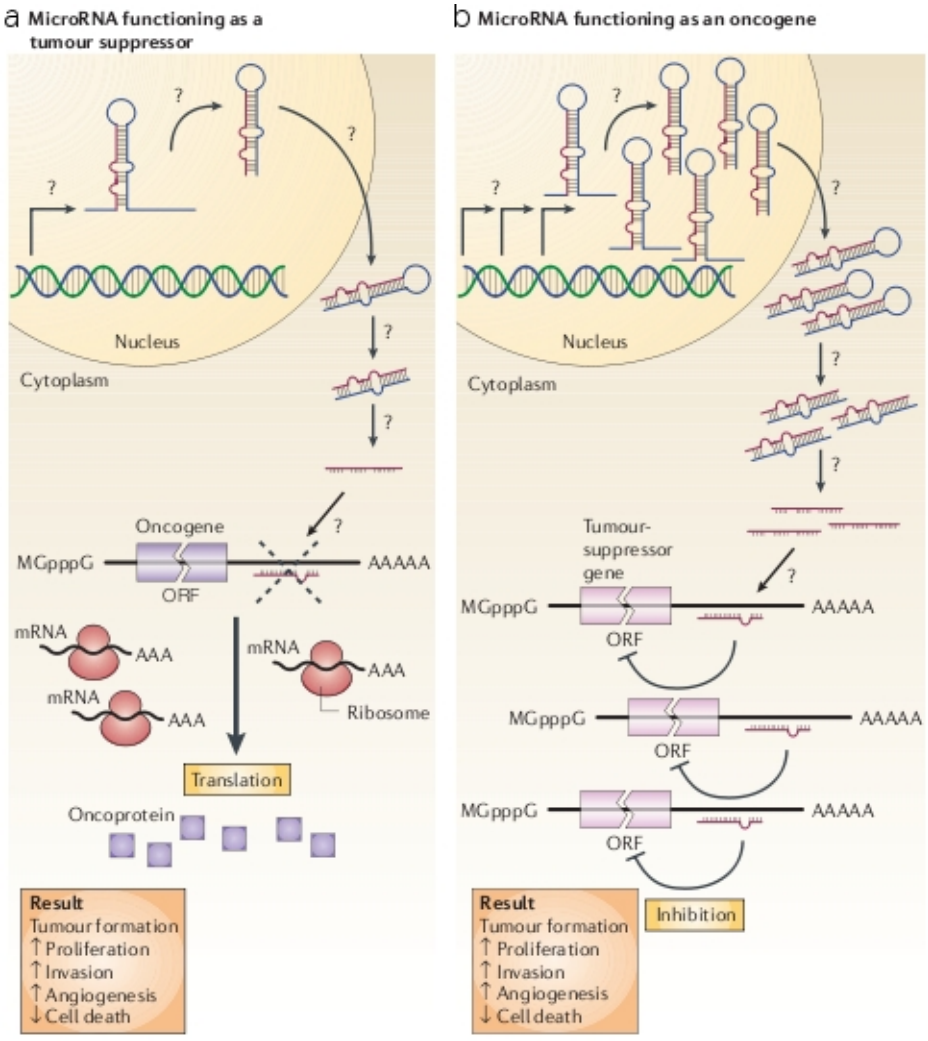


Figure 1.16: Role of miRNA in a cancer cell - (a) The reduction or deletion of a miRNA that functions as a tumour-suppressor leads to tumour formation. A reduction in or elimination of mature miRNA levels can occur because of defects at any stage of miRNA biogenesis (indicated by question marks) and ultimately leads to the inappropriate expression of the miRNA-target oncoprotein (purple squares). The overall outcome might involve increased proliferation, invasiveness or angiogenesis, decreased levels of apoptosis, or undifferentiated or de-differentiated tissue, ultimately leading to tumour formation. (b) The amplification or overexpression of a miRNA that has an oncogenic role would also result in tumour formation. In this situation, increased amounts of a miRNA, which might be produced at inappropriate times or in the wrong tissues, would eliminate the expression of a miRNA-target tumour-suppressor gene (pink) and lead to cancer progression. Increased levels of mature miRNA might occur because of amplification of the miRNA gene, a constitutively active promoter, increased efficiency in miRNA processing or increased stability of the miRNA (indicated by question marks). ORF: open reading frame (image and legend from Esquela-Kerscher and Slack, 2006).

1.3.4 Alterations of the regulation mechanisms of the information flow

In **Subsection 1.2.2**, the different mechanisms involved in the regulation of the information flow in normal cells have been described. Here we will give some examples of alterations of the regulation mechanisms which have occurred in cancer cells due to the progressive accumulation of mutations described in **Subsection 1.3.1**. The outline of this subsection will be the same as in **Subsection 1.2.2**.

Modification of transcription factor activity In cancer, many transcription factors are involved in the tumoral progression mechanism. The most famous one is incontestably the tumour-suppressor gene *TP53*, also known as *the guardian of the genome*: it plays a key-role in preserving the integrity of the genome (see **Subsection 1.3.5**). *p53* is directly involved in many cancers due to the presence of mutations (30 to 50% of common human cancers have a *p53* mutation - figures from Weinberg, 2007) in the gene which encodes this protein. As a consequence, mutated *p53* loses a part of its transcription factor functions since it can no longer bind to all its target genes which can also no longer be transcribed into mRNA (Vogelstein and Kinzler, 2004). Therefore, when mutated, *p53* can no longer play its guardian's role efficiently.

Modification of the epigenetic properties The low level of DNA methylation in tumours as compared to the level of DNA methylation in their normal-tissue counterparts was one of the first epigenetic alteration found in human cancers. The loss of methylation is mainly due to hypomethylation of repetitive DNA sequences and extensive hypomethylated genomic regions in gene-poor areas. During the development of the disease, the degree of hypomethylation of genomic DNA increases as the lesion progresses from a benign proliferation of cells to an invasive cancer (see **Figure 1.17**). This hypomethylation increases chromosome instability leading to deletion, translocations and chromosome rearrangements (see **Subsection 1.3.6**). This was observed by Shann et al. (2008) in breast cancer cell lines. They have also shown that genes with intragenic hypomethylation had low level of expression. The loss of methyl group from DNA can also cause loss of *genomic imprinting and leads to gene activation in some types of cancers (see **Figure 1.18**): this is the case for *IGF2* which increases the risk factor for colorectal cancer (Esteller, 2008). Besides hypomethylation, CpG-island-promoter hypermethylation is a key process in tumoral progression and leads to the transcriptional silencing of tumour-suppressor genes (see **Figure 1.17** and **Figure 1.18**). The profiles of hypermethylation of the CpG islands in tumour-suppressor genes are specific to the type of cancer. Hypermethylation can be the second lesion in Knudson's two hit model. Global alterations of histone modification patterns (see **Figure 1.17** and **Figure 1.19**) have the potential to affect the structure and the integrity of the genome, and to disrupt normal patterns of gene expression, which, like alterations in DNA methylation, may be causal factors in cancer (see **Figure 1.19**). In bladder cancer, Stransky et al. (2006) have shown that histone methylation can occur in a large scale genomic region and lead to the loss of expression of neighbouring genes inside this region: this phenomenon is known as long-range epigenetic silencing. In colorectal cancer, Frigola et al. (2006) have found that this long-range epigenetic silencing at the level of histone methylation could also be associated with DNA methylation.

Modification of the post-transcriptional regulations We have seen that after transcription, the RNA can be processed into different mRNAs to increase the diversity of proteins in a process called alternative splicing (see **Page 13** and **Figure 1.8**). In cancer cells,

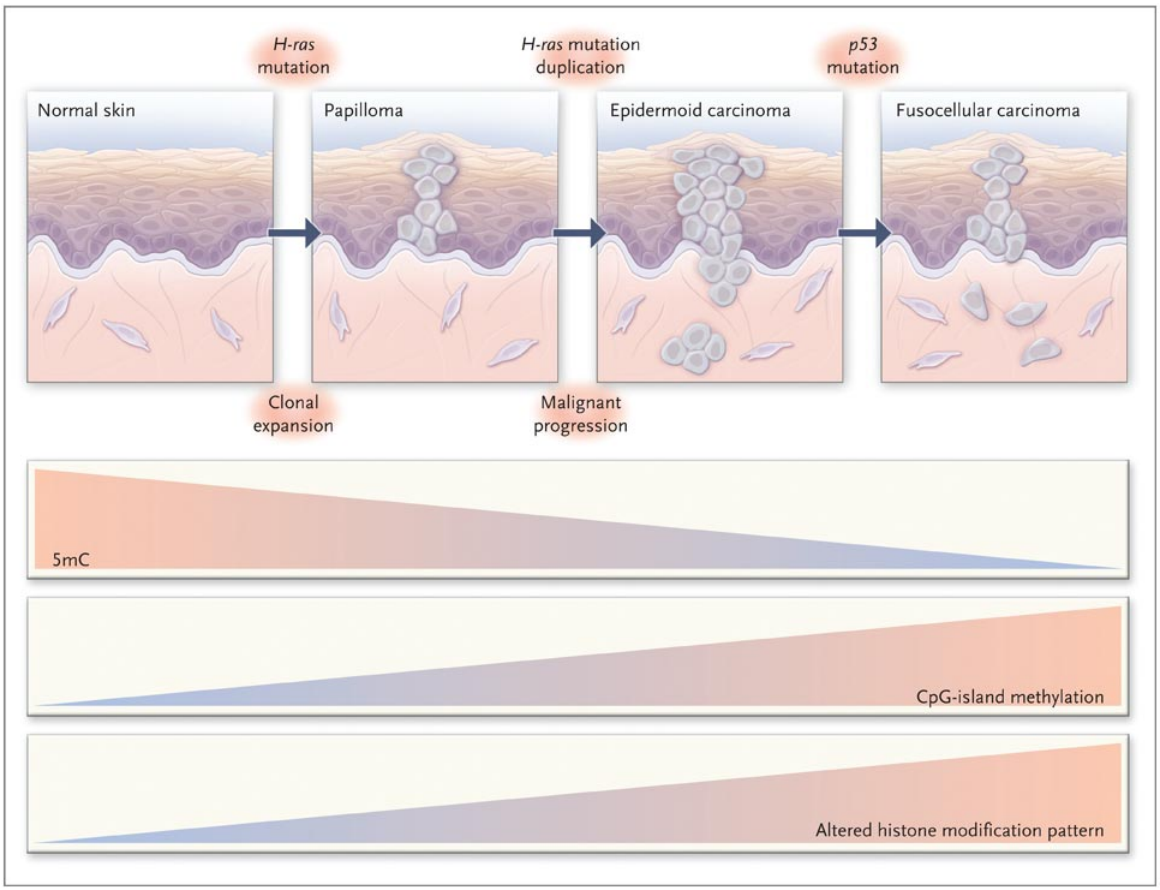


Figure 1.17: Epigenetic alterations in tumoral progression - A multistage model of *carcinogenesis in skin is shown. In conjunction with phenotypic cellular changes and the accumulation of genetic defects, there is a progressive loss of total DNA methylation content, an increased frequency of hypermethylated CpG islands, and an increased histone-modification imbalance in the development of the disease. *H-ras* denotes Harvey-ras oncogene, and 5mC 5-methylcytosine which is the result of the DNA methylation. (image and legend from Esteller, 2008)

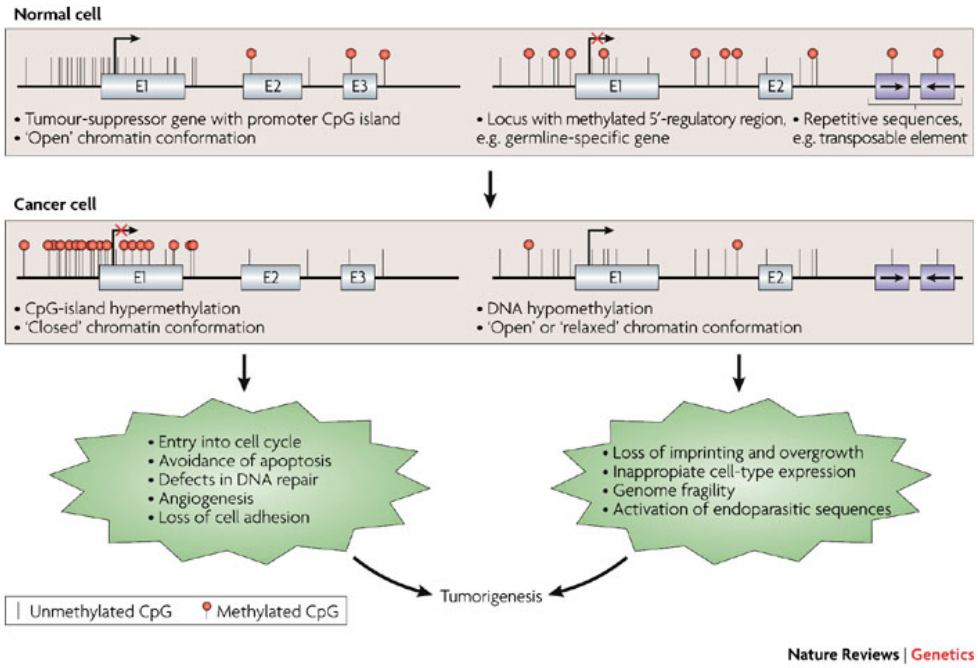


Figure 1.18: DNA methylation in cancer - The hypermethylation of CpG islands of tumour-suppressor genes is a common alteration in cancer cells, and leads to the transcriptional inactivation of these genes and the loss of their normal cellular functions. This contributes to many of the hallmarks of cancer cells. At the same time, the genome of the cancer cell undergoes global hypomethylation at repetitive sequences, and tissue-specific and imprinted genes can also show loss of DNA methylation. In some cases, this hypomethylation is known to contribute to cancer cell phenotypes, causing changes such as loss of imprinting, and might also contribute to the genomic instability that characterises tumours (image and legend from Esteller, 2007).

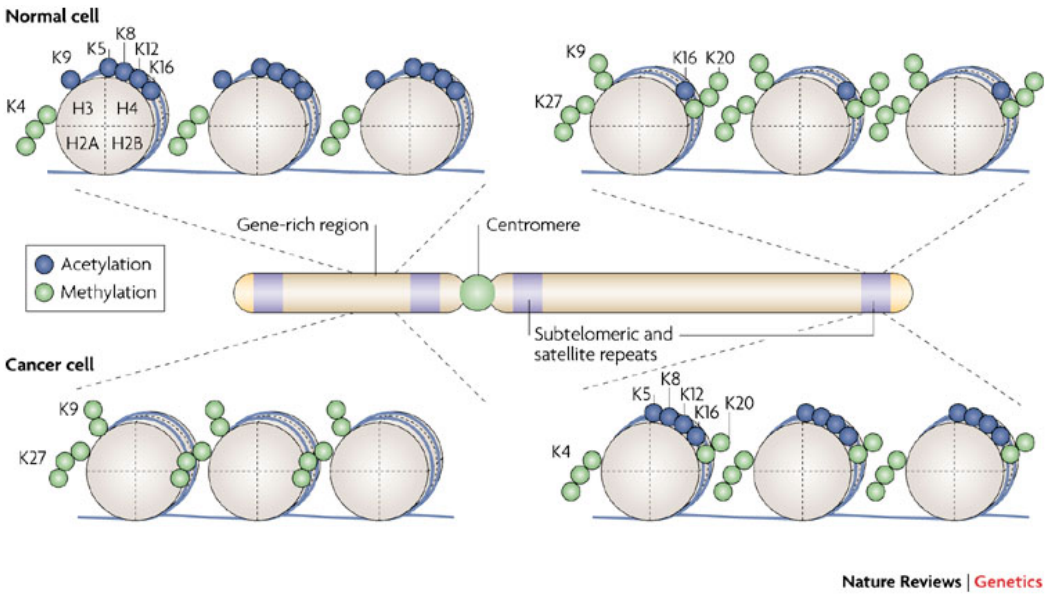


Figure 1.19: Histone modification in cancer - Nucleosomal arrays are shown in the context of chromosomal location and transcriptional activity. Octamers consisting of histones H2A, H2B, H3 and H4 are represented as grey cylinders. Histone acetylation and methylation (di- and tri-) are shown. In *normal* cells, genomic regions that include the promoters of tumour-suppressor genes are enriched in histone-modification marks associated with active transcription, such as acetylation of H3 and H4 lysine residues (for instance K5, K8, K9, K12 and K16) and trimethylation of K4 of H3. In the same cells, DNA repeats and other heterochromatic regions are characterised by trimethylation of K27 and dimethylation of K9 of H3, and trimethylation of K20 of H4, which function as repressive marks. In transformed cells, this scenario is disrupted by the loss of the *active* histone-marks on tumour-suppressor gene promoters, and by the loss of repressive marks such as the trimethylation of K20 of H4 or trimethylation of K27 of histone H3 at subtelomeric DNA and other DNA repeats. This leads to a more *relaxed* chromatin conformation in these regions (image and legend from Esteller, 2007).

there are aberrant splicing variants which are not found in normal cells (Venables, 2004; Srebrow and Kornblihtt, 2006; Kim et al., 2008). They provide the cell with new functions. Moreover, as we mentioned in **Subsection 1.3.3**, miRNAs can act either as oncogenes or tumour-suppressor genes losing the role they had in protein-coding gene regulation (Calin and Croce, 2006a).

Disruption of signal transduction Many signaling pathways are involved in cancer development. Here we will give two illustrations of such disruptions. In **Figure 1.10** we have seen that the activation of *p53* relies on the efficiency of the signal transduction cascade. Any alteration which damages this signal transmission will impact the integrity of the cell in case of DNA damage. For example, an inactivation of the tumour-suppressor kinase gene *ATM* will prevent *p53* from being activated. The second example is the *HER2/neu* oncogene (also known as *ErbB-2*, *ERBB2*) which is a cell membrane surface-bound receptor tyrosine kinase and is normally involved in the signal transduction pathways leading to cell growth and differentiation. In some breast cancer, the *HER2/neu* kinase is overexpressed[18] due to an amplification and induces a huge activation of the signal transduction cascade which makes the cancer very aggressive with high metastatic risk. In the general case, the alteration of the signal transduction can be either an absence of the signal amplification while the signal should be amplified or an over-amplification of the signal while it should not. Protein kinases control the signal transduction within the cell and therefore represent therapeutic targets for drugs: for example, drugs such as Gleevec, Iressa and Herceptin[19] are kinase inhibitors.

1.3.5 *Hallmarks of cancer*

In the previous subsections, we have seen that a progressive accumulation of mutations in cancer-critical genes modifies the regulation mechanisms of the information flow described in the central dogma. Although a large variety of mutations can occur, we will see in this subsection that all cancers have traits in common. Indeed, to lead successfully to a tumour, a cell must acquire a whole range of aberrant properties (a collection of subversive new skills) as it evolves. Different cancers require different combinations of properties. Nevertheless, we can draw up a short list of the key behaviours of cancer cells in general. Hanahan and Weinberg (2000) suggest that the vast catalog of cancer cell genotypes is a manifestation of six essential alterations in cell physiology which collectively dictate malignant growth. These six capabilities are shared by most and perhaps all types of human tumours (see **Figure 1.20**). To simplify, these six capabilities can be merged into the three main following properties (the number in brackets corresponds to the capability detailed in **Figure 1.20**):

Defective control of the cell cycle (1, 2) The cell cycle normally ensures that the number of cells within the organism remains constant so that when a cell dies a new one is born. Therefore, the cell cycle must be precisely controlled. In cancer, this control is inefficient and the cells keep on reproducing. Many mechanisms are involved in the cell cycle. We will only mention here the key-role of *pRb* which controls the initiation of the cell cycle. When in a hypophosphorylated state, *pRb* blocks proliferation by sequestering and altering the function of *E2F* transcription factor which controls the expression of banks of genes essential for progression from G1 into S phase[20] (for a detailed description see Calzone

[18] A normal cell has 20000 *HER2* receptors while there are about 1.5 million in a *HER2/neu* positive cancer cell.

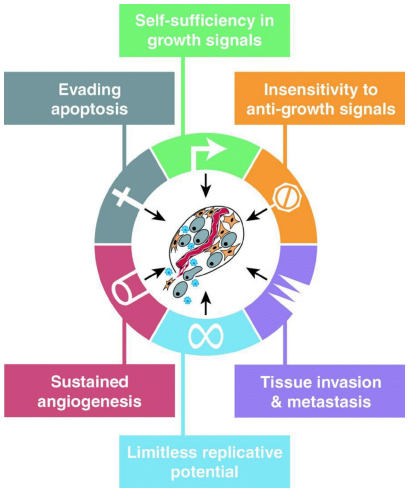
[19] Herceptin is an antibody which interferes with the kinase activity of *HER2/neu* and recruits immune effector cells which are responsible for antibody-dependent cytotoxicity. Herceptin can also induce complement-dependent cytolysis and enhance phagocytosis by Fc-receptors bearing antigen-presenting cells (Hudis, 2007).

[20] G1 phase corresponds to the deployment of a cell cycle program and S phase corresponds to the step of DNA replication.

et al., 2008). It has been shown that mutation, promoter hypermethylation or LOH[21] of *pRb* occur in cancer.

Defective control of cell death (4, 6) The cell death process is the process of cell destruction. This mechanism occurs when cells have reached a limited number of cell cycles (this process is called *senescence*) or when the genome integrity is compromised during the cell cycle (this process is called *apoptosis*). In the apoptosis process, the machinery can be broadly divided into two classes of components: sensors and effectors. The sensors are responsible for monitoring the extracellular and intracellular environment for conditions of normality or abnormality leading to life or death of a cell. These signals regulate the second class of components, which function as effectors of apoptotic death. One of the major players among effectors is *p53*, the guardian of the genome as already mentioned: in case of DNA damage during replication, *p53* triggers apoptosis. Defective apoptosis can be due to a mutation of both *p53* and the sensors (the proteins which are involved in the signal transduction like *ATM*) (see **Subsection 1.3.4**).

Invasiveness and metastatic potentials (3, 5) The uncontrolled proliferation of cells leads the cell population to damage the function of an organ in which the cancer cell originates, to colonise adjacent tissues, and may also damage the function of the neighbouring organs. Moreover, the tumour has the ability to form new blood vessels (this process is called *angiogenesis*) and therefore the tumour has the possibility to spawn cancer cells in the blood. These cells are called metastases and are able to colonise distant tissues. By spreading throughout the body, a cancer becomes almost impossible to eradicate surgically or by localised irradiation, and thus can become deadly. These distant settlements of tumour cells, called metastases, are the cause of 90% of human cancer deaths (Hanahan and Weinberg, 2000).



1. cancer cells can free themselves from dependence on exogenous growth signals
2. cancer cells are insensitive to signals that block cell proliferation
3. the tumour can invade adjacent tissues (they are invasive) and spawn pioneer cells that move out (they metastasise)
4. cancer cells become immortal
5. the tumour can produce new blood vessels
6. the cell death machinery is inefficient

Figure 1.20: Acquired capabilities of cancer - Most if not all cancers have acquired the same set of functional capabilities during their development, albeit through various mechanistic strategies (image and legend from Hanahan and Weinberg, 2000).

Besides the six capabilities described in **Figure 1.20**, cancers have another common characteristic which is the presence of chromosome aberrations. This is described in the next subsection.

[21] See **Subsection 1.3.6 - Page 29**.

1.3.6 Chromosome aberrations in cancer

The normal configuration of chromosomes is often termed the **euploid** karyotype state. Euploidy implies that each of the chromosomes is present in normally structured pairs (see **Figure 1.21b**). Deviation from the euploid karyotype (**aneuploidy**) is observed in many cancers (see **Figure 1.21a**). Often, this aneuploidy is merely a consequence of the general chaos which reigns within a cancer cell (Weinberg, 2007) due to the progressive accumulation of mutations. Indeed, once a sufficient number of mutations is reached, the cell cannot correctly process the duplication and the segregation of the chromosomes because of defects in DNA repair and cell cycle checkpoints leading to a genome instability (Aguilera and Gómez-González, 2008). This instability occurs at both nucleotidic (imperfect copy of the DNA sequence) and chromosomal levels (improper number of chromosomes). As a result, the daughter cells will not be able to perform the specific function they were supposed to do. In 1914, Theodor Boveri proposed the hypothesis that cancer cells derive from cells with an irreparable defect within the chromosomes. This hypothesis of a chromosomal or genetic cause of cancer was only reconsidered in recent decades in the light of new findings on genomic rearrangements and cancer genetics (Satzinger, 2008). Rearrangements occur due to mechanisms such as DNA breaks and fusions and lead the daughter cells to have an abnormal karyotype. In neuroblastoma, such breaks have been shown to occur preferentially within early replicating regions during S phase (Janoueix-Lerosey et al., 2005, this article is supplied in the **Annexes**). Typical chromosomal aberrations which produce an abnormal karyotype are illustrated in **Figure 1.22a** and explained below:

polyploidy: instead of being present in two copies, each chromosome is present in p copies where $p > 2$ represents the ploidy.

aneuploidy: there are extra abnormal copy numbers of some chromosomes.

translocation: a chromosome translocation is a chromosome abnormality caused by rearrangement of parts between homologous or nonhomologous chromosomes. The translocation can be reciprocal (or balanced, meaning that there is an exchange of two extremities of chromosomes) or non-reciprocal (or imbalanced, meaning that one extremity is gained and/or lost during the rearrangement).

amplification: the same small part of a chromosome (such as a gene or a group of a few contiguous genes) is present in a high number of copies (from 4 to more than 50 copies) either as acentric fragments (double minute), incorporated into chromosomes in nearly contiguous homogeneously stained regions (HSRs) or interspersed in the genome.

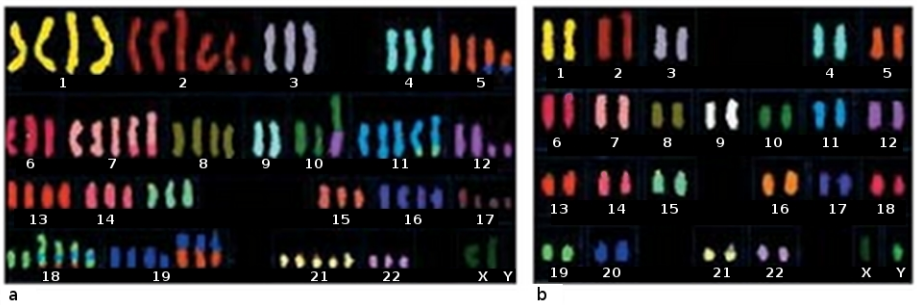


Figure 1.21: Karyotype of a colon cancer cell and normal cell - (a) The karyotype of a typical cancer shows many abnormalities in chromosome number and structure. (b) Karyotype of a normal cell where each autosomal chromosome is present in two copies (adapted from Alberts et al., 2002).

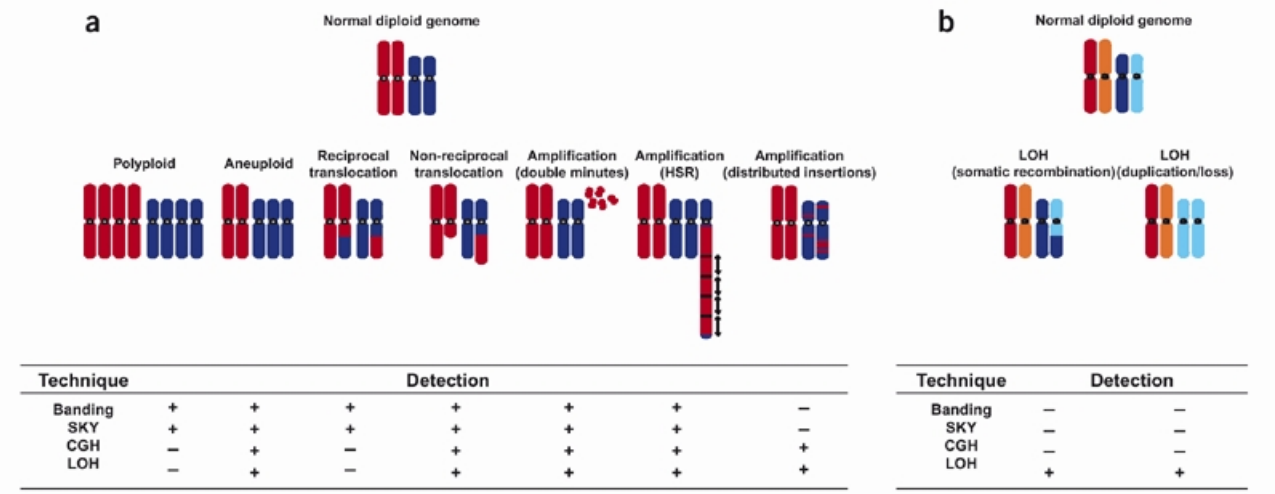


Figure 1.22: Chromosomal aberrations - Schematic illustration of mechanisms by which chromosomal aberrations arise plus a summary of the ability of commonly applied technologies to detect the aberrations. (a) Aberrations that lead to aneuploidy. (b) Aberrations which leave the chromosome apparently intact (image and legend from Albertson et al., 2003).

The following terms can also define the copy number for a given region of the chromosome: deletion (no copy is present anymore), monosomy (one copy is present), trisomy (three copies are present), tetrasomy (four copies are present), *etc.*. These rearrangements can be either complete or partial.

Interestingly, some chromosome aberrations do not produce an abnormal karyotype. They are summarised in **Figure 1.22b**. In this case, the chromosomes appear to be present with the expected number of two copies. In a normal genome, there is one copy from the father (dark blue chromosome) and one copy from the mother (light blue chromosome). Therefore, in a normal cell, chromosomes are heterozygote: they come from two distinct parental origins. In cancer, it has been noticed that some chromosomes come from the same parental origin as illustrated in **Figure 1.22b**. In this case, the two copies of the chromosome are the same and therefore the chromosome is homozygote. This phenomenon is termed Loss of Heterozygosity (LOH) without DNA copy number change since one parental chromosome (or only a portion) has been lost in a first event and the missing chromosome (or only the missing portion) has been duplicated from the remaining parental chromosome. When the LOH rearrangement concerns the whole chromosome we speak about duplication also call isodisomy and when it concerns only portion we speak about somatic recombination or partial isodisomy.

Why are these chromosomal aberrations so important in cancer? In fact, in many cases they cause tumoral progression. To illustrate this purpose, we can take as examples the critical genes involved in cancer:

- **example of a tumour-suppressor gene:** let us take for example *pRb*. We have seen that this protein plays an important role in the control of cell cycle. Let us imagine the following situation: a mutation occurs in *pRb* which inactivates the function of the protein (the mutation can have been inherited from parents or arise spontaneously during replication). Then, a second event occurs (this is the Knudson two-hit model presented in **Subsection 1.3.3**) so that the duplication of the chromosome which carries the mutation combined with the loss of the chromosome which carries the func-

tional gene (*i.e.* the LOH event depicted on the bottom-right part of **Figure 1.22b**). The result is that *pRb* is now present in two copies of inactive forms and will no longer be able to control the cell cycle. From the different aberration configurations depicted in **Figure 1.22** we can imagine other combinations of alterations which also lead to the same effect.

- **examples of oncogenes:** now, let us take different examples among oncogenes. In **Subsection 1.3.3**, we have already mentioned the different ways a proto-oncogene can be transformed into an oncogene. We will here give more details. Except in the mutation case which appears spontaneously, the activation of the proto-oncogene into an oncogene described in **Figure 1.15** can be simply explained by chromosome aberrations. For example, in neuroblastoma an amplification of *MYCN* which overproduces this protein is frequently observed. Other mechanisms involving translocation and gene fusion also play an important role (Mitelman et al., 2007). For example, if a proto-oncogene appears to be translocated near a DNA domain which normally regulates the constitutional expression of a gene at a high level, then the proto-oncogene will be highly transcribed and become an oncogene. This occurs in Burkitt's *lymphoma where *MYC* is juxtaposed with regulatory elements of the immunoglobulin heavy chain *IGH* gene: the *MYC* gene is constitutively activated because its expression is driven by immunoglobulin regulatory elements (Mitelman et al., 2007). In this case the translocation leads to an upregulation of the proto-oncogene. The last mechanism is also based on a translocation which creates a new gene by fusion of two existing genes called a chimeric gene. This phenomenon occurs in Ewing's sarcoma[22] where a translocation between chromosomes 11 and 22 merges a portion of the *EWS* gene and the *FLI1* gene into the chimeric gene *EWS-FLI1*[23]. In this tumour, gene fusion always involves *EWS* mainly with *FLI1* and less frequently with other genes. In the case of a chimeric gene the protein is either overproduced or hyperactive.

DNA copy number alteration We have seen that different types of chromosome aberrations occur in cancer. Some of them modify the copy number of entire or small portions of chromosomes while other aberrations do not modify the number of chromosomes. For situations in which the number of chromosomes is modified we speak about DNA copy number alterations. It is precisely this type of alteration which will serve as a basis for the analyses presented in **Chapter 2** and **Chapter 3**. Microarray technologies have been developed to investigate DNA copy number alterations such as array Comparative Genomic Hybridisation (aCGH) and will be presented in detail in **Subsection 1.4.2**. Briefly, this technique allows the quantification of the DNA copy number of many loci along each chromosome as shown in **Figure 1.23**: basically, the plot on the right side of **Figure 1.23** represents the DNA copy number of the different loci ordered along the genome for three different chromosomes. Typically, two DNA copy levels will have a value around 0, gains will have values shifted positively and losses will have a value shifted negatively[24]. This technology is able to detect amplifications, non-reciprocal translocations, losses and gains. Once again, we can link the importance of amplification, gain and loss of DNA copy number with the search for cancer-critical genes. Typically, we expect oncogenes to be present in gain or amplification regions while tumour-suppressor genes are supposed to be found in loss regions. Therefore, a characterisation of DNA copy number alterations should help to find new candidates for cancer-critical genes.

[22] Ewing's sarcoma is a paediatric cancer in which cells are found in the bone or in soft tissue.

[23] Special dedication to Olivier Delattre.

[24] See **Subsection 1.4.2 Page 35** for explanation how the signal is derived.

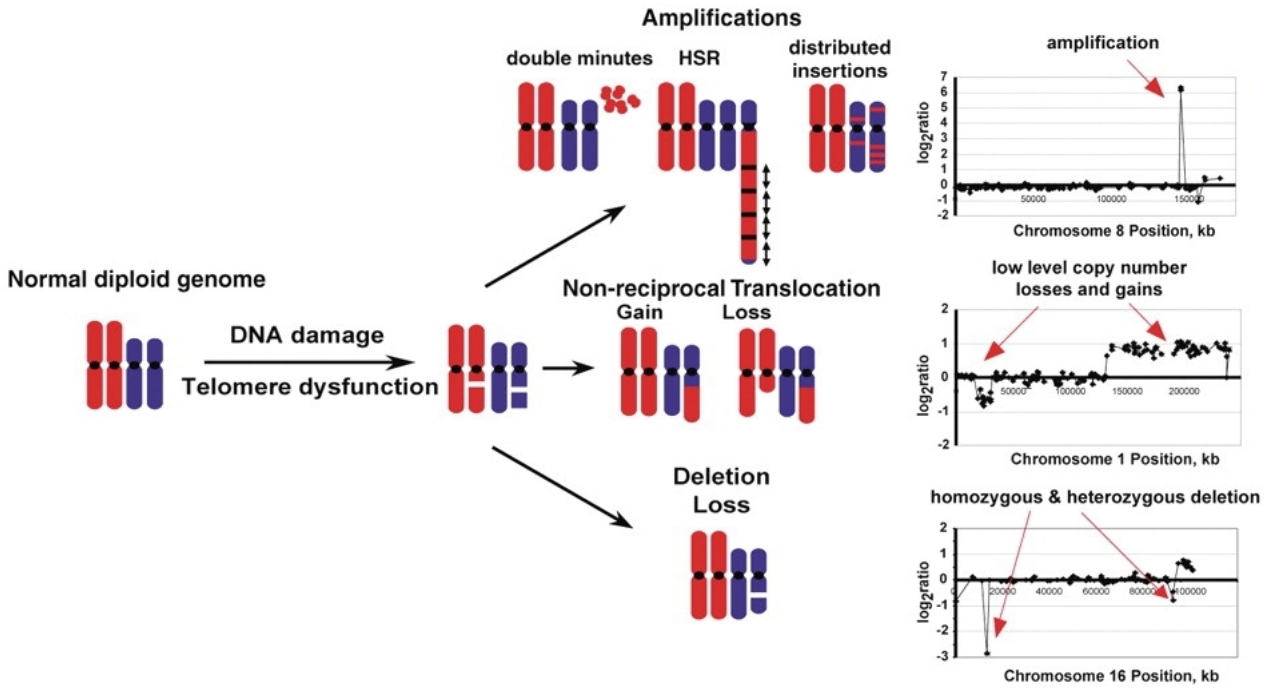


Figure 1.23: DNA copy number aberrations - Detection of copy number aberrations in tumour genomes by aCGH. Chromosomal aberrations in cancer are likely to arise following inappropriate management of DNA damage or telomere dysfunction. Common aberrations include gene amplifications, non-reciprocal translocations and interstitial deletions. Amplifications may be visible cytogenetically as double minutes, chromosomes with homogeneously staining regions (HSR) or the amplified DNA may be distributed at multiple sites. The aCGH copy number profile of the amplified *MYC* in COLO320 is shown. The amplification level is about 70 fold ($\log_2$-ratio > 6). Breakage of a chromosome or a non-reciprocal translocation event may lead to low level copy number changes, as shown in the copy number profile of chromosome 1 from 600MPE. Homozygous deletions are indicated by $\log_2$-ratio < -2 and heterozygous deletions by $\log_2$-ratio ~ -1 (image and legend from Albertson and Pinkel, 2003).

In this section, we have seen that a sequential accumulation of events has transformed the normal cell into a cancer cell which has defects in the regulation mechanisms controlling the cellular program. Those defects involve different molecular entities which play a key-role in tumoral progression. Therefore, these entities give the molecular basis both to improve the characterisation of cancer and to provide the clinician with new prognostic and predictive factors. We will see in the next section how to perform molecular profiling of tumours for these different entities using high-throughput technologies also named *omics* approaches.

1.4 Welcome to the world of *omics*!

We have seen in **Section 1.3** that when a normal cell becomes a cancer cell, a series of disregulations occurs at different molecular levels. Basically, the sequential accumulation of mutations and events occurring during tumoral progression disrupts the normal behaviour of the cell at the level of (see **Figure 1.24**):

1. DNA
    a. including mutations of the DNA sequence,
    b. changes in DNA copy number,
    c. LOH and
    d. translocations.

2. miRNA expression
3. mRNA expression
    a. including modifications in alternative splicing.
4. protein expression
    a. and particularly protein kinases which play a key-role in signal transduction.
5. epigenetic characteristics
    a. including modifications of DNA or histone methylation
    b. or modifications in histone acetylation.
6. interactions between the different molecules
    a. such as the interactions between transcription factors and DNA,
    b. or interactions between proteins.
7. phenotype characteristics of the cell

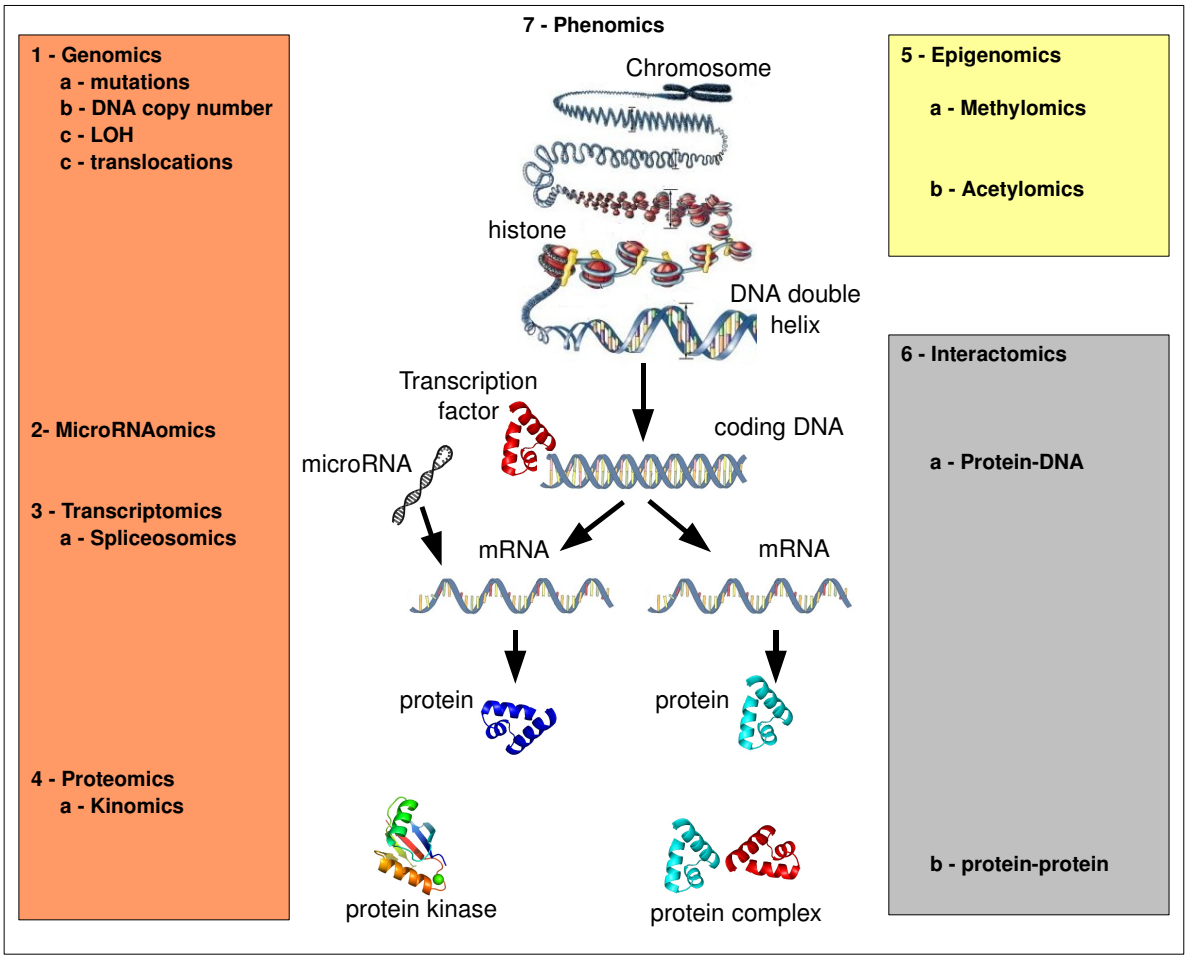


Figure 1.24: Omics technologies in oncology.

The biological knowledge regarding the biology of cancer presented in **Section 1.3** is clearly a help to know where to look in order to better characterise tumour progression and improve the classification of tumours on the basis of molecular profiles. However, the question is obviously how to carry out an investigation at the molecular level? Indeed, biological technologies are needed to accurately retrieve the molecular profiles of each tumour sample and the information retrieval must be as exhaustive as possible. For example, we would like to determine the DNA copy number of as many loci as possible over each chromosome or quantify the mRNA expression of all known genes. This exhaustive search might be reachable for some molecular profiles but in some cases, especially for proteins, it is untractable for technological reasons[25]. In the protein case, the search is restricted to interesting functional classes of proteins, such as kinases which play an important role in cancer. Since the quantification of the molecular profiles is supposed to be as exhaustive as possible, the techniques which allow the measurement are often referred to as *genome-wide* techniques whatever the type of molecular profile investigated. More generally, the name of the technology which allows us to study a particular type of molecular profile is the concatenation of the molecular entities or the biological functions under study with the *-omics* suffix. For example, as illustrated in **Figure 1.24**, *genomics* investigates the DNA, *microRNAomics* the miRNA expression, *transcriptomics* the mRNA expression, *spliceosomics* the alternative splicing, *proteomics* the protein expression, *kinomics* the phosphorylated state of protein kinases, *epigenomics* the epigenetic modifications, *interactomics* the interactions between different molecular entities and *phenomics* the observable traits of the cells. The *-omics* suffix comes from the Greek stem *omes* which stands for *all*, *every*, *whole* or *complete* reminding us of the fact that these techniques aim at achieving an exhaustive search. These techniques are also called high-throughput technologies since they produce a huge amount of information within a short time. Thus, we know where to look and now we know how to investigate. In the next section, technical details will be given regarding *omics* technologies with a particular focus on DNA copy number and mRNA expression. These two types of molecular profiles will be used in **Chapter 2** and **Chapter 3**.

1.4.1 *Microarray* era

The better understanding of biological molecular processes combined with the improvement in DNA technologies have allowed researchers to use *in vitro* some chemical reactions which happen *in vivo*. Among the main revolutions in biotechnology, let us mention the discovery of restriction enzymes and reverse transcriptase in 1970, DNA sequencing in 1977, Polymerase Chain Reaction[26] (PCR) in 1985 (see Nature Publishing Group (2007) for a history of DNA technologies). The combined improvements in chemistry, physics and molecular biology have allowed the development of new tools for genome-wide quantification: the *microarray* technology also called *biochip* has provided miniaturised sensor tools such that it is now possible to query the mRNA expression of the whole genome on a slide glass smaller than four square-centimetres (see **Figure 1.25**). Microarrays can be considered as one of the major biotechnological revolutions of the last 15 years. Originally, microarrays emerged in the field of transcriptomics. Since different molecular profiles are important in the physiology of the cell and especially in cancer study, the original microarray technology which queried mRNA has widely been transposed for all the omics approaches mentioned in **Figure 1.24**. A large variety of microarray techniques have been developed (see Hoheisel, 2006). Here, we will just give the basic principle of microarray technologies with a particular focus on the

[25] The study of proteins implies the preparation of antibodies as reporters which is a huge work to ensure their specificity. Due to the large amount of different entities (more than one million protein entities within a cell), only a few reporters have been prepared so far.

[26] Polymerase Chain Reaction is the process which allows the copying of DNA or RNA molecules.

analysis of DNA copy number and mRNA expression.

The basic principle is the following (see Southern et al., 1999): *probes* which can be DNA, RNA or protein, are tethered to a solid support (*i.e.* the chip), such as glass, plastic or silicon. They act as a specific reporter either to quantify the DNA copy number at a known locus on the genome, or the expression of a known gene or the amount of a protein. Probes are supposed to be chosen specifically in order to report the quantification of their expected target. In the case of DNA or RNA probes, the specificity is guaranteed by the choice of a unique base-paired complementarity between the probe sequence and the target sequence, and by the choice of an appropriate antibody in case of protein. Probes are amplified and deposited on a microscopic area of the chip called spot. Then, either DNA, RNA or proteins are extracted from a tumour sample and then hybridised on the chip. If present within the sample, a given DNA sequence, RNA or protein will be fixed on its matching probes. In a microarray, thousands or even millions of such spots are present which make it a very powerful tool for genome-wide screening. All these microarray technologies have been widely applied in oncology as reported by Cowell and Hawthorn (2007). The field of microarrays is evolving very quickly and new techniques are regularly described. For readers interested in epigenomics or kinomics microarray technologies we refer to the papers of Schumacher et al. (2006), Reinders et al. (2008), Schones and Zhao (2008) and Johnson and Hunter (2005). The next two subsections present in detail the microarray technologies devoted to the analysis of DNA copy number and mRNA expression.

Figure 1.25: Affymetrix GeneChip® used to quantify mRNA expression.

1.4.2 Analysis of DNA copy number

Comparative Genomic Hybridisation Originally, the study of genome-wide characterisation of DNA copy number changes was performed using Comparative Genomic Hybridisation (CGH) technique which was developed in the early 1990s. In this technique, total genomic DNA is isolated from tumour and normal control cells, labelled with different fluorochromes and hybridised to normal metaphase[27] chromosomes (Kallioniemi et al., 1992). This technique is therefore termed chromosomal CGH. Differences in the tumour fluorescence with respect to the normal fluorescence along the metaphase chromosomes are then quantified and reflect changes in the DNA copy number in the tumour genome. Subsequently, aCGH, where arrays of genomic sequences replaced the metaphase chromosomes as hybridisation reporters, was established (Solinas-Toldo et al., 1997; Pinkel et al., 1998) and solved many of the technical difficulties and problems caused by working with cytogenetic chromosome preparations. The main advantage of aCGH is the ability to perform copy number analyses with much higher resolution than was ever possible using chromosomal CGH. aCGH has already been widely used in oncology for many purposes such as global analysis

[27] Metaphase is a step in the cell cycle.

of copy number aberrations, identification of putative target genes, tumour classification or assessment of clinical significance of copy number changes (Kallioniemi, 2007). Pinkel and Albertson (2005) give details in their review about the technology and its application in oncology. We will present here only the basics (see **Figure 1.26**):

1. total genomic DNA is isolated from a tumour sample (*i.e.* the test DNA) and from a normal sample (*i.e.* the reference DNA). Genomic DNA is then digested with a restriction enzyme and the DNA fragments are differentially labelled: the tumoral DNA is labelled with a red fluorochrome and the normal DNA with a green fluorochrome.

2. equal amounts of tumoral and normal DNA are combined.

3. the mixture of both the tumoral and normal DNA is hybridised on the chip. Within each spot, there is a competitive hybridisation between the tumoral DNA target sequences and the normal DNA target sequences.

4. once hybridised, a scanning step quantifies the signal intensity for the red and green channels and outputs image files in which each pixel is given a red and green intensity.

5. an image analysis software accurately reconstructs the signal intensity at the level of each spot.

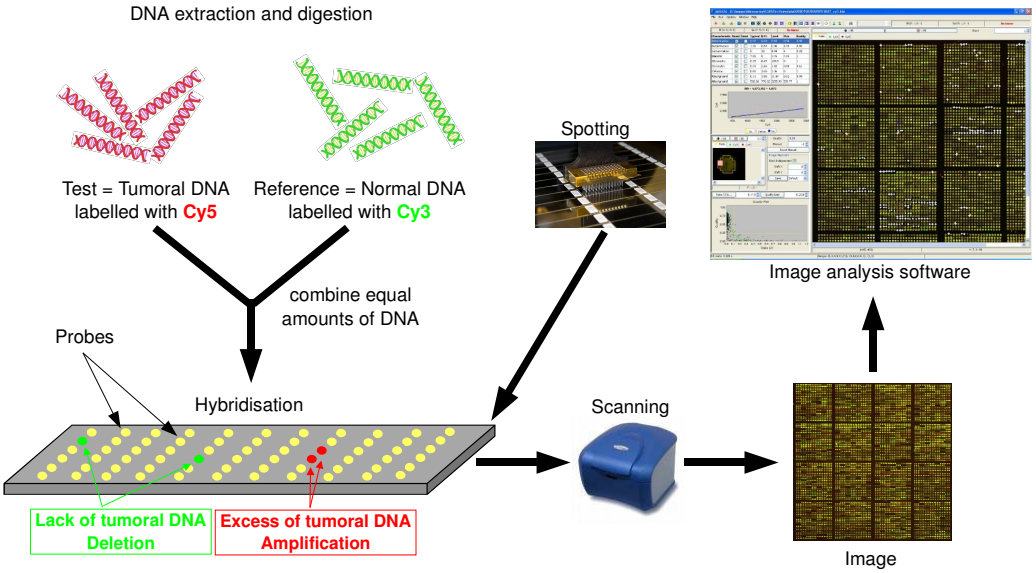


Figure 1.26: aCGH protocol - The protocol includes the extraction and labelling of the DNA, the hybridisation on the chip, the scanning and image analysis to quantify the signal.

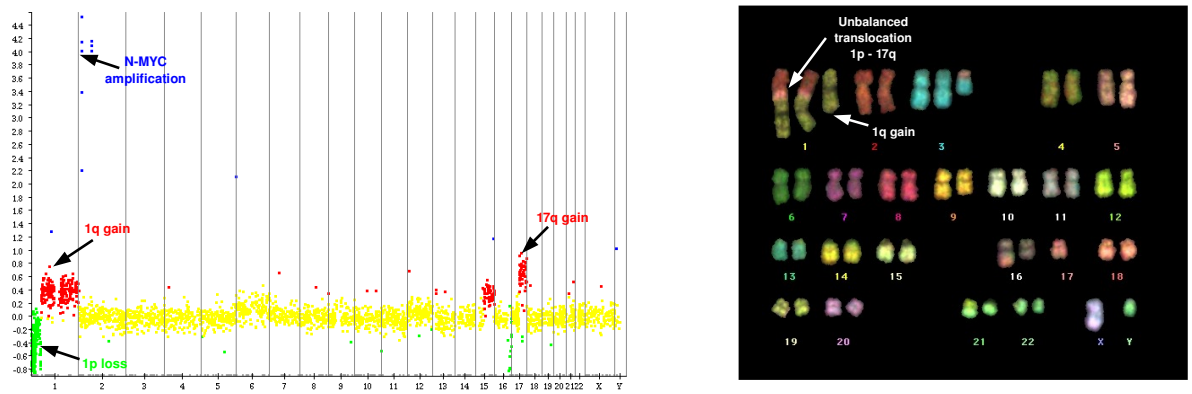
Quantification of DNA copy number How do we expect the signal to vary with respect to the DNA copy number of each sample? Let us take the simplest example in which a normal DNA is red-labelled and another normal DNA is green-labelled. Then, for each locus of each chromosome we expect two DNA copies for the two samples. For each spot competitive hybridisation takes place and we obtain half red normal DNA and half green normal DNA. The relative hybridisation intensity of the test DNA signal over the reference DNA signal equals one and the spot will be yellow following the additive colour mixing theory. And then,

what happens if the test DNA is a tumoral DNA? In this case, the relative hybridisation intensity of the test signal over the reference signal at a given location is (ideally) proportional to the relative DNA copy number of those sequences in the test and reference genomes. If the tumoral DNA copy number is greater than the normal DNA copy number, then the signal will be shifted towards red. On the contrary, if the tumoral DNA copy number is lower than the normal DNA copy number, then the signal will be shifted towards green. Therefore, the DNA copy number of the tumoral DNA is directly proportional to the red/green ratio and its theoretical value is given in **Figure 1.27**. For statistical reasons, we do not use the ratio of red/green but the $\log_2$ of this ratio therefore named $\log_2$-ratio[28]. In practice, due to technical variability there is a fluctuation of the signal around its expected value and statistical methods are necessary to retrieve the true signal. This will be the scope of **Chapter 2**. Moreover, the quantified signal is generally less than expected for three reasons. First, the quantification made with the technology is not perfect and the signal is generally less than proportional with respect to the true DNA copy number (Pinkel et al., 1998; Pollack et al., 1999). Then, the tumoral DNA generally contains contamination from normal tissue which reduces the signal. Finally, the tumour might be heterogeneous since it can derive from different clonal populations (see **Figure 1.13**) which share different patterns of DNA copy number alterations. The aCGH technology relies on the fact that the reference DNA is diploid whatever the locus. In practice, this is not the case since, even in normal individuals, DNA copy number variations exist: some parts of the DNA sequence can be present in many copies inside the genome. Such a part of the genome is called Copy Number Variant (CNV) (Iafrate et al., 2004; Freeman et al., 2006). Ideally, to avoid the identification of DNA copy number changes due to CNVs between the test DNA and the reference DNA, the two DNAs used in the aCGH protocol should come from the same patient (in this case the DNAs are termed paired). However, normal DNA from the patient is not always available and a normal reference DNA from a normal standard individual is used. Therefore, in the case of non-paired DNAs, relevant DNA copy alterations for the disease should not correspond to a CNV: the Database of Genomic Variants (Iafrate et al., 2004) available at http://projects.tcag.ca/variation/ integrates known CNVs and allows validation of relevant alterations due to the pathology.

Figure 1.27: Theoretical aCGH quantification - The theoretical ratios and $\log_2$-ratios are given for different DNA copy number alterations occurring in the tumoral DNA.

[28]The log transformation allows the distribution of the values to be closer to normality, which is a nice property in statistics.

Graphical representation of a DNA copy number molecular profile The typical representation of an aCGH molecular profile is depicted in **Figure 1.28a**: the x-axis represents the probe location ordered along the genome from chromosome 1 to 22, X and Y; the y-axis represents the $\log_2$-ratio value of the DNA copy number. In this profile of the IMR32 neuroblastoma cell line we can clearly see a loss of chromosome 1p[29], a gain of chromosome 1q and 17q due to an imbalanced translocation which is seen on the karyotype in **Figure 1.28b**. Note that the aCGH resolution allows the detection of *MYCN* amplification which cannot be seen in the karyotype.



a - IMR32 aCGH profile

b - IMR32 karyotype

Figure 1.28: IMR32 neuroblastoma cell line - (a) Array CGH profile. The imbalanced translocation 1p-17q and the 1q gain are identified by aCGH. A small alteration like *MYCN* amplification can been seen due to the increasing resolution of the aCGH technology (data from Janoueix-Lerosey et al., 2005). (b) 24-colour painting of chromosomes by Fluorescence In Situ Hybridisation (FISH). The unbalanced translocation 1p-17q and the 1q gain are highlighted (image provided by Isabelle Janoueix-Lerosey, Institut Curie).

Always more probes on the chip Different aCGH platforms are available and with the recent advances in microarray technologies we have moved from *BAC aCGH to *oligonucleotide aCGH allowing an increase in the number of loci per chip (Davies et al., 2005; Ylstra et al., 2006). BAC arrays are mainly in-house microarrays while oligonucleotide microarrays mainly come from commercial companies. Among the widely used commercial technologies, let us mention Agilent Human Genome CGH Microarray[30], Nimblegen Human Whole Genome Tiling arrays[31], Illumina BeadChip[32] and Affymetrix GeneChip®[33] (note that for Affymetrix and Illumina technologies, no normal DNA is needed in the protocol and they are named 1-colour arrays in contrast to the other technologies which use both normal and tumoral DNA and are named 2-colour arrays). At the very beginning of BAC array, the number of loci investigated was around 1000-2000 and never exceeded 32000 loci (Ishkanian et al., 2004). The use of oligonucleotide array has allowed a huge increase in the number of loci investigated on a single chip. At the time of writing the present manuscript, the highest number of loci quantified in the human genome with a single oligonucleotide array is provided by Affymetrix®

[29] p and q define the short and long chromosome arm respectively.

[30] http://www.home.agilent.com/

[31] http://www.nimblegen.com/

[32] http://www.illumina.com/

[33] http://www.affymetrix.com/

Genome-Wide Human SNP Array 6.0 and covers more than 1.8 million loci for the whole human genome. The Nimblegen company offers the possibility to use one chip of 385000 loci for each human chromosome corresponding to a total coverage of 9.24 million loci. Although the most recent chips cover more exhaustively the genome, their exact resolution does not only depend on the number of loci but also on their sensitivity. Coe et al. (2007) has proposed a definition of resolution for aCGH technology, termed *functional resolution*, which incorporates the uniformity of loci spacing on the genome, as well as the sensitivity of each platform to single-copy alteration detection. From their study, the current commercial platforms allow a single-copy detection of the order of 35-55Kb while it was 10Mb for chromosomal CGH and 1Mb for BAC aCGH (At the time of the study by Coe et al. (2007) the highest number of loci in a single chip was offered by Nimblegen Human Whole Genome Tiling arrays which allowed the quantification of 385000 loci over the whole human genome). The oligonucleotide chips making it possible to scan the genome for more than 50000 loci are often termed high-density chips.

1.4.3 Analysis of mRNA expression

The development of microarray technology was first initiated in the field of transcriptomics study using experimental protocol quite similar to the aCGH protocol described in **Figure 1.26** except that mRNA is used instead of genomic DNA. New technologies have appeared developed by the Affymetrix[33] company which changed the chip building and the protocol in such a way that no reference sample is necessary anymore. The Affymetrix GeneChip® (see **Figure 1.25**) thus provides an approach to have a semi-quantitative level of mRNA instead of a relative value with respect to a reference. As we will see for DNA copy number microarray experiments, a normalisation step which corrects systematic experimental effects is necessary to improve the quality of the data. Several methodologies have been developed for Affymetrix GeneChip® and the most widely used so far are RMA (Irizarry et al., 2003) and GC-RMA (Wu et al., 2003).

1.4.4 Emerging sequencing technologies

In 1977, a revolution in the era of genetic engineering was the development of technologies allowing the sequencing of DNA. The same year, Maxam and Gilbert, and Sanger and Coulson proposed methods to sequence genomes. Sanger's method was widely used especially for the Human Genome Project. Although the Sanger method has been improved, it does not allow sequencing of genomes either in a reasonable time or at a reasonable price[34]. To overcome these limitations, a second generation sequencing (also called next-generation sequencing) appeared in 2005 with an increase in the throughput capacity for a lower cost[35]. The main next-generation sequencers are Roche Applied Science 454[36] Genome Sequencer FLX™, Illumina's Solexa[37] and Applied Biosystems[38] SOLiD™(Rusk and Kiermer, 2008; Chi, 2008). Next-generation sequencing has many applications and especially in the field of medical science as we will see below (Schuster, 2007; Mardis, 2008). Note that high-throughput sequencing is a very competitive domain and we have already spoken about third generation sequencing (also called next-next generation sequencing). This new generation sequencing will appear in 2008 and will be based on single-molecule analysis. It should make it possible

[34]The most recent sequencers based on Sanger's method are able to sequence 2-5 million bases per day at $1 per 1000 bases.
[35]The Solexa Genome Analyzer is able to sequence 600 million bases per day at $1 per 100000 bases.
[36]http://www.454.com/
[37]http://www.illumina.com/
[38]http://www.appliedbiosystems.com/

to sequence a human genome in 24 hours with a $1000 cost. The main competitors are Helicos BioSciences[39], VisiGen Biotechnologies[40] (Blow, 2008) and Pacific Biosciences[41] .

Why is it important to mention these new generation sequencings here? We have seen in **Figure 1.24** that, among the different molecular levels investigated, many include either DNA or RNA. Thus, for molecular profiling involving nucleotide sequences, the new generation sequencing technology represents a new way of investigation in order to:

- search for mutations
- discover polymorphism
- quantify DNA copy number
- quantify LOH
- map chromosomal rearrangements at a resolution of one base (Chen et al., 2008; Campbell et al., 2008)
- quantify mRNA expression
- quantify miRNA expression
- discover ncRNAs
- identify alternative splicing
- identify protein-DNA interactions using chromatin immunoprecipitation followed by sequencing
- map nucleosome position with respect to the DNA sequence
- study epigenomic modifications
- study the spatial organisation of the chromatin

Therefore, it is very likely that these new generation sequencings will progressively replace microarray experiments in the future.

By definition, high-throughput technologies produce a huge amount of data which need to be analysed. Powerful tools and statistical methods are needed to handle such an amount of data. These tools constitute the field of bioinformatics which is briefly introduced in the following section.

1.5 From biology to bioinformatics

1.5.1 Molecular profiling of cancer: proof of concept

This subsection provides a brief overview of pioneering articles which have demonstrated molecular profiling to be a valuable tool in the field of cancer study using different statistical approaches from the field of bioinformatics.

[39] http://www.helicosbio.com/
[40] http://visigenbio.com/
[41] http://www.pacificbiosciences.com/

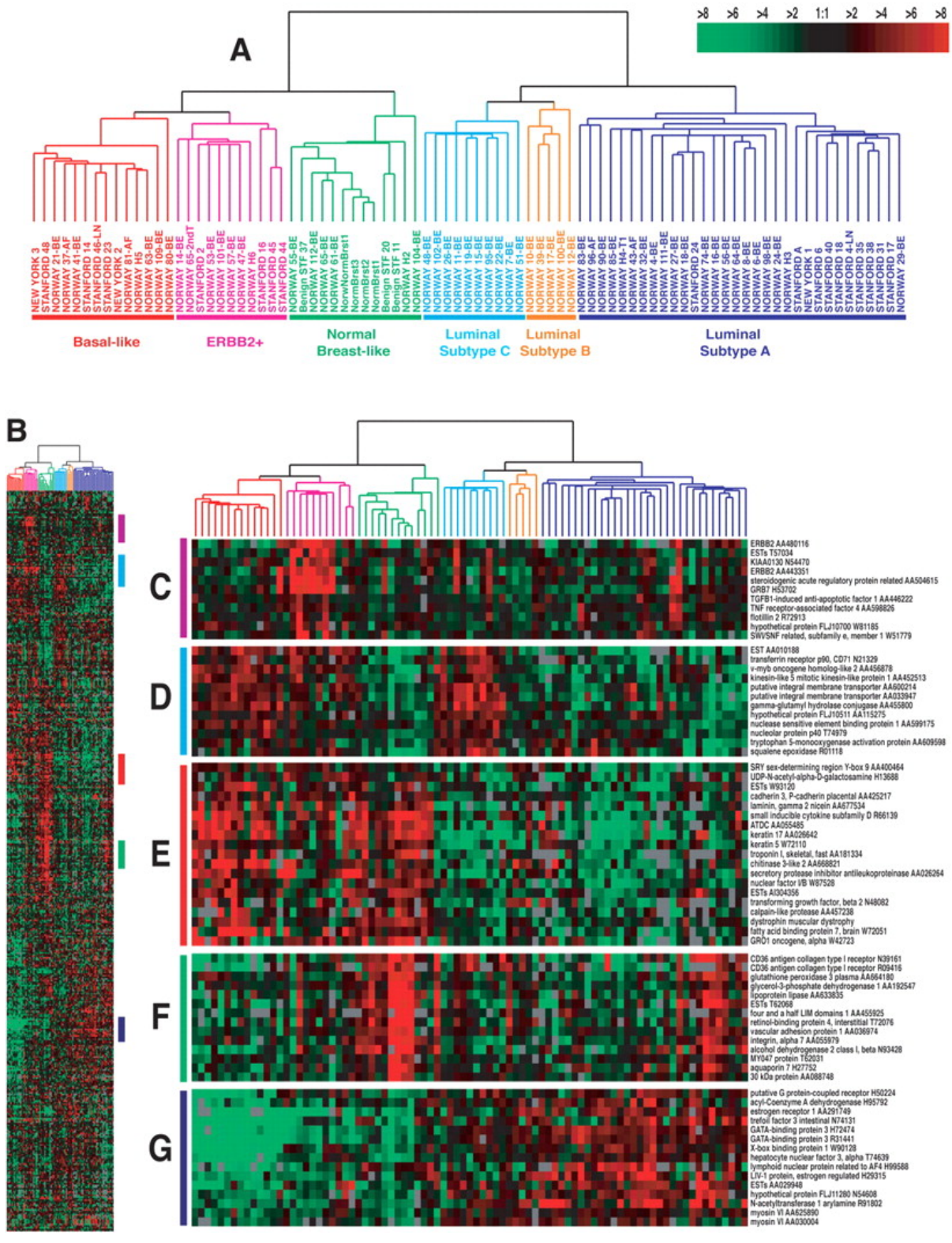


Figure 1.29: Molecular classification of breast cancer from mRNA expression profiles - Gene expression patterns of 85 experimental samples representing 78 *carcinomas, three benign tumours, and four normal tissues, analysed by hierarchical clustering using the 476 cDNA intrinsic clone set. (A) The tumour specimens were divided into five (or six) subtypes based on differences in gene expression. The cluster dendrogram showing the five (six) subtypes of tumours are coloured as: luminal subtype A, dark blue; luminal subtype B, yellow; luminal subtype C, light blue; normal breast-like, green; basal-like, red; and *ERBB2+*, pink. (B) The full cluster diagram scaled down. The coloured bars on the right represent the inserts presented in C-G. (C) *ERBB2* amplicon cluster. (D) Novel unknown cluster. (E) Basal epithelial cell-enriched cluster. (F) Normal breast-like cluster. (G) Luminal epithelial gene cluster containing ER. (image and legend from Sørlie et al., 2001).

Identification of new tumour subtypes Microarrays and especially mRNA profiling have been widely applied to tumours: gene expression patterns have been shown to efficiently retrieve biological and clinical properties of tumours as reviewed by Miller and Liu (2007) for breast cancer. For example, Perou et al. (1999) showed that the vast and complex transcriptional data generated by microarrays contained discernible subtypes of gene expression patterns which related to tumour biology and behaviour. These subtypes, termed luminal A, luminal B, normal breast-like, *ERBB2+*, and basal-like (see **Figure 1.29**) were shown to be stable and reproducible classes observable in different patient populations, and significantly associated with tumour recurrence and patient survival (Sørlie et al., 2001, 2003). Among these subtypes, a lot of attention is paid to basal-like group which has poor prognosis and for which there is no efficient therapy: understanding the mechanisms involved in the tumoral progression within this subtype should give insights in order to find new therapeutic targets. As we have already seen, protein kinases play a key-role in the biology of cancer. Recently, Finetti et al. (2008) have shown that statistical analyses restricted to the set of protein kinase coding genes were able to distinguish poor from good prognosis in luminal A subtypes. In addition to mRNA expression, miRNA expression profiling has also been shown to provide informative classification of cancers (Lu et al., 2005). Indeed, in a clustering analysis of miRNA profiles, they have been able to identify distinct patterns with respect to the different cancer types or the developmental origins of the tissues (see **Figure 1.30a**), to partition samples with respect to the mechanisms of transformation of the tumour (see **Figure 1.30b**), and to identify patterns which could not be seen based on mRNA profiling (see **Figure 1.30c**).

Identification of predictive signatures Other studies have focused on the identification of gene expression signatures in order to predict the patient outcome. One of the most famous signatures is the Amsterdam signature for breast cancer which consists of 70 genes (Van't Veer et al., 2002). It was subsequently validated on a second series of patients (Van de Vijver et al., 2002). This signature is also used in a prospective European clinical trial called MINDACT (Microarray In Node negative Disease may Avoid ChemoTherapy) (Buyse et al., 2006). miRNA expression profiling of human tumours also allows identification of signatures associated with diagnosis, staging, progression, prognosis and response to treatment (Calin and Croce, 2006b). Therefore, miRNAs represent interesting prognostic factors and even therapeutic targets (Lowery et al., 2008).

As we will see in **Chapter 2** in an application to uveal *melanoma, the DNA copy number profiling of tumours can be used for both subtype discovery and supervised classification (Trolet et al., 2008, this article is supplied as material part of the thesis in **Section 2.5**).

1.5.2 Analysis of DNA copy number: a need for new methods

Molecular profiling represents a valuable prospect to better characterise tumours and to help in diagnosis and prognosis. For mRNA expression microarray experiments, a long experience has existed for more than ten years and many methodologies have been proposed to analyse such data (Grant et al., 2007). The field of DNA copy number profiling is more recent and still evolving with new microarrays increasing the ability to detect smaller and smaller alterations. Bioinformatics tools and methods are definitely needed to analyse such data since a direct application of existing methodologies devoted to the analysis of mRNA expression microarray experiments cannot be straightforwardly applied. **Chapter 2** will present the statistical approaches which have been developed during the thesis to take into account the specificity of the DNA copy number microarray experiments so that we are able to correctly retrieve the relevant biological information from experiments.

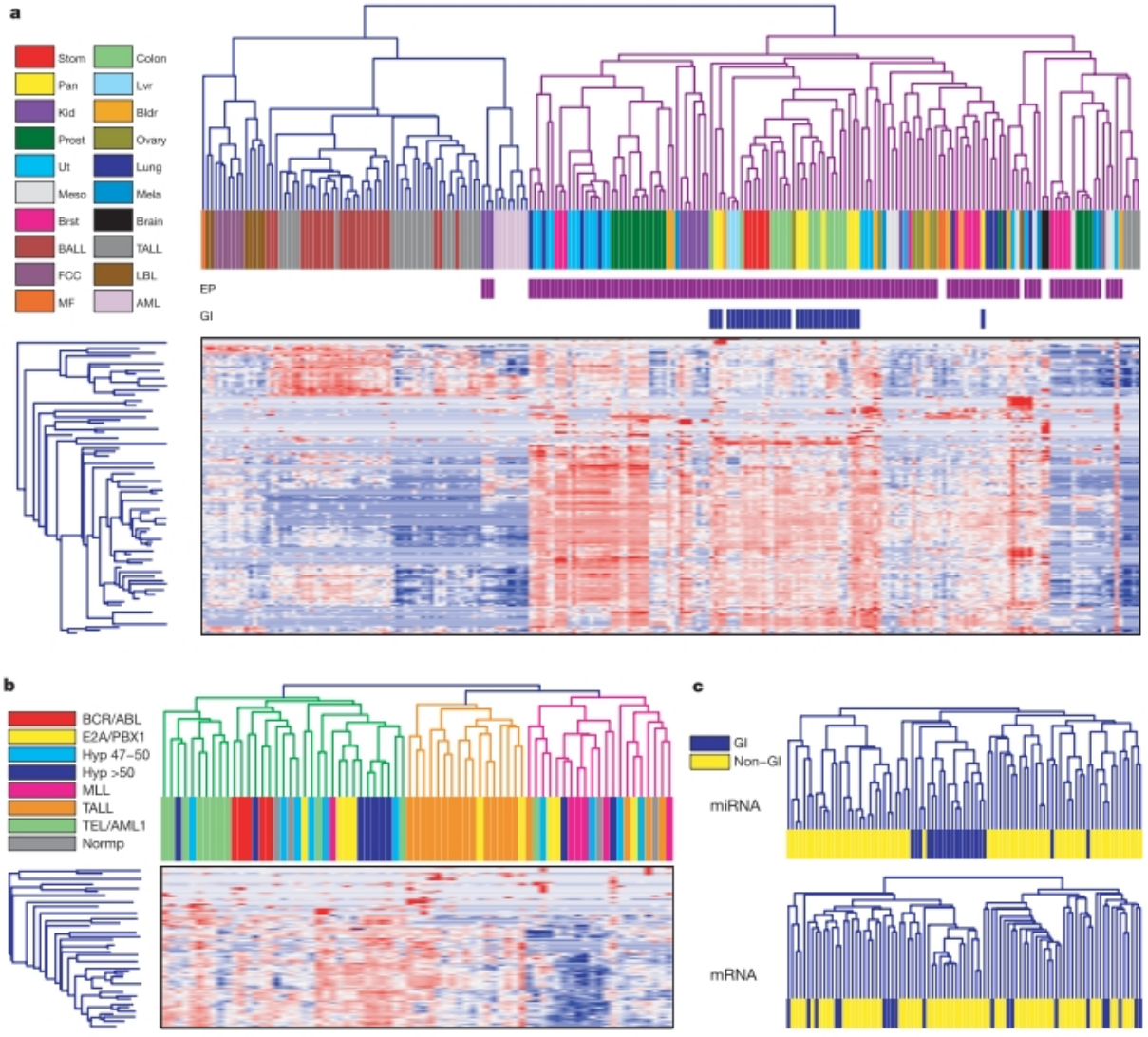


Figure 1.30: Molecular classification from miRNA expression profiles - a, miRNA profiles of 218 samples from several different tissues were clustered (average linkage, correlation similarity). Samples are in columns, miRNAs in rows. Samples of epithelial (EP) origin or derived from the gastrointestinal tract (GI) are indicated. b, Clustering of 73 bone marrow samples from patients with acute lymphoblastic leukaemia (ALL). Coloured bars indicate the different ALL subtypes. c, Comparison of miRNA data and RNA data. For 89 epithelial samples from a that had RNA expression data, hierarchical clustering was performed. Samples of GI origin are shown in blue. GI-derived samples largely cluster together in miRNA expression space, but not in RNA expression space. Abbreviations used: Bldr, bladder; Brst, breast; Fcc, follicular lymphoma; Kid, kidney; Lvr, liver; Mela, melanoma; Meso, mesothelioma; Pan, pancreas; Prost, prostate; Stom, stomach; Ut, uterus; AML, acute myelogenous leukaemia; BALL, B-cell ALL; LBL, diffuse large-B cell lymphoma; MF, mycosis fungoides; MLL, mixed lineage leukaemia; TALL, T-cell ALL; Hyper 47-50, hyperdiploid with 47-50 chromosomes; Hyper $>$ 50, hyperdiploid with over 50 chromosomes; Normp, normal ploidy (image and legend from Lu et al., 2005).

1.5.3 Issue of Knowledge integration

Up to date, most publications have used one single molecular profile at a time to characterise tumours with remarkable resolution and accuracy. As we have seen, different molecular profiles are now available and it is natural to combine them to improve the knowledge we have on cancer and also to identify new reliable and efficient biomarkers useful for clinical purposes. Indeed, the nature and strength of each biomarker, the certainty of its contribution to cancer, and therefore its translational importance, vary substantially. Some biomarkers will be strong, causal *drivers* of important cancer hallmarks. Others will be weaker but important *contributors* to the development of cancer pathophysiology. And many will be genomic *noise* (or *passengers*): that is, elements which are biologically neutral and have been accumulated by chance during the cancer's lifespan. Distinguishing the drivers and contributors from the passengers is a central challenge in genomic research. This is made more difficult by the diversity of biomarker function and the likelihood that biomarker function may depend on the tumour type (or subtype), as well as on the tumour microenvironment (Chin and Gray, 2008). Technological advances which allow the cancer genome to be examined in multiple omic dimensions are helping to focus the search for drivers and contributors, by uncovering biomarkers which tend to be dysregulated by several mechanisms. Thus, data showing that a biomarker can be dysregulated in several complementary ways in cancer, through the integration of more than one dimension of genomic information, provide strong evidence that a biomarker is likely to be pathogenetic. The current large-scale cancer genome projects that are carrying out genome-wide characterisation in a coordinated and comprehensive manner will be the most powerful at leveraging such multidimensional data for integrative analyses. Software and databases are needed to provide biologists with efficient tools to explore the huge quantity of information inside these data: such tools will be presented in **Section 2.6**. More and more studies focus on the combination of chromosome alterations and gene expression data to explain direct or indirect relation between DNA copy number and the mRNA expression level using correlation approaches (Lee et al., 2008). In **Chapter 3**, we will present a statistical approach to combine both mRNA expression and DNA copy number microarray experiments in the framework of prediction (also called supervised classification). The goal is to provide signatures using both molecular profile levels so that it can help in the choice of a tailored therapy for the patient.

1.5.4 Contributions of the thesis

As we have seen in this introductory chapter, DNA copy number alterations are a hallmark of cancer. This molecular level provides a valuable information to pinpoint new cancer-critical genes and to identify new prognostic and predictive biomarkers. High-throughput technologies have made it possible to quantify DNA copy number alterations and the contribution of the thesis is to provide statistical methods to analyse DNA copy number microarray experiments. In the next two chapters, the following contributions are presented:

- In **Chapter 2**, biostatistical methods have been developed to extract the relevant biological information from DNA copy number microarray experiments. The biostatistical methods have been applied to uveal *melanoma to identify new prognostic factors.
- In **Chapter 3**, a statistical method able to combine biological information from both DNA copy number and mRNA expression microarray experiments has been developed in order to identify new prognostic and predictive factors.

Colour study: squares with concentric circles
Wassily Kandinsky, *1913*

Necessaria est methodus ad rerum veritatem investigandam.

René Descartes, *Regulae ad directionem ingenii*

2

Extraction of the biological information from high-throughput experiments: application to DNA copy number microarray experiments

Contents

This chapter presents the biostatistical algorithms and bioinformatics tools which have been developed during the thesis to analyse DNA copy number microarray experiments. Before introducing the outline of the chapter, let us make a brief overview of the different steps to perform in a study dealing with high-throughput experiments (see **Figure 2.1**). Once the biological and/or clinical question is asked (❶), an experimental design is defined in order to efficiently answer the problem raised (❷). This step is still too often neglected while a lot of attention should be paid to it: *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of"* (Sir Ronald Aylmer Fisher). Then, the high-throughput

experiments are performed and the bioinformatics really starts from this point (❸). In almost all high-throughput experiments, there is a scanner which analyses the microarray slides and produces images. These images need to be analysed using image analysis software to quantify the raw signal (❹). This step is followed by normalisation which aims at correcting the systematic sources of variability in order to improve the signal (❺). The quality of data is checked at the level of both image analysis and normalisation steps (❻). At this stage, the information provided after normalisation is still rough and the meaningful biological information relevant for biologists must be extracted from the data (❼). Once the relevant information is extracted, the data can be used in a transversal analysis to perform clinical biostatistics, classification or systems biology modeling (❽). Finally, the results need to be validated, interpreted and can lead to new experiments (❿).

With respect to this bioinformatics framework, the outline of the chapter is the following: we first describe a normalisation method which has been developed to improve the signal from BAC aCGH experiments (steps ❺ and ❻). Then, an algorithm able to identify the DNA copy number alterations is detailed (step ❼). In the next section, we show that, for some microarray technologies, it is more powerful to combine at the same time both the normalisation and the extraction of the biological information (steps ❺, ❻ and ❼). The first three sections work at the level of single DNA copy number profile analysis. Therefore, we then illustrate how the biological information can be retrieved in a multi-profile analysis using a transversal analysis strategy (step ❼). In the fifth section, we show how the combination of the algorithms previously described can help to predict high-risk patients from DNA copy number experiments in uveal melanoma (step ❽). Finally, the tools, software and database devoted to the analysis of DNA copy number profiles which have been developed either entirely or partially are listed.

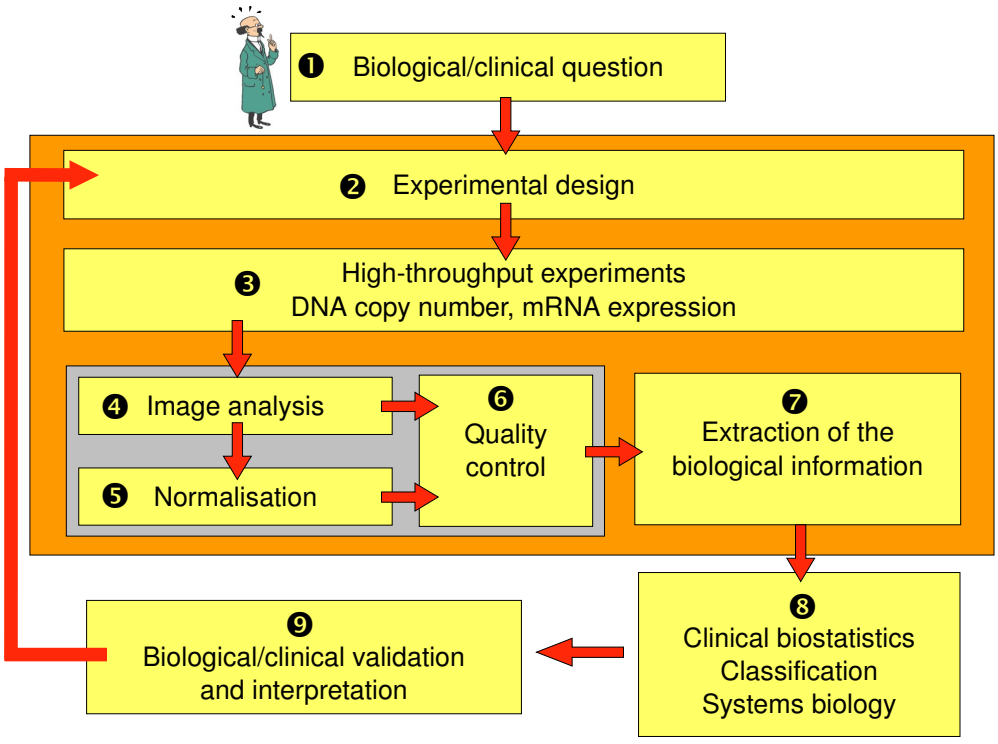


Figure 2.1: Bioinformatics approach to analyse high-throughput experiments.

2.1 Normalisation of array-CGH data

Normalisation aims at correcting the systematic sources of variability in order to improve the signal. In all microarray technologies, there are inherent sources of variability which have a direct impact on the signal quantified with image analysis software (see **Figure 2.1**). In the field of mRNA expression microarrays, many methods have been proposed (see Quackenbush, 2002; Do and Choi, 2006; Irizarry et al., 2006, for a review). They cannot be directly transposed to the analysis of BAC aCGH profiles since there are specificities which need to be taken into account. Indeed, spatial artifacts have been noticed on aCGH data but no method was appropriate for the removal of such artifacts. Therefore, we have developed a normalisation method devoted to the spatial normalisation of aCGH data. This method is called MicroArray NORmalisation (MANOR) (Neuvial et al., 2006) and the paper describing the algorithm is supplied as a material part of the thesis. Briefly, the method consists of a spatial smoothing of the data followed by a segmentation which identifies aberrant spatial areas on the chip. Koren et al. (2007) suggested that normalisation methods which correct for spatial biases, such as MANOR, should be routinely applied when analysing microarray data.

BMC Bioinformatics

BioMed Central

Research article

Open Access

Spatial normalization of array-CGH data

Pierre Neuvial*†[1], Philippe Hupé†[1,2], Isabel Brito[1], Stéphane Liva[1], Élodie Manié[3], Caroline Brennetot[3], François Radvanyi[2], Alain Aurias[3] and Emmanuel Barillot[1]

Address: [1]Institut Curie, Service de Bioinformatique, 26, rue d'Ulm, Paris, 75248 cedex 05, France, [2]Institut Curie, CNRS UMR 144, 26, rue d'Ulm, Paris, 75248 cedex 05, France and [3]Institut Curie, INSERM U509, 26, rue d'Ulm, Paris, 75248 cedex 05, France

Email: Pierre Neuvial* - pierre.neuvial@curie.fr; Philippe Hupé - philippe.hupe@curie.fr; Isabel Brito - isabel.brito@curie.fr; Stéphane Liva - stephane.liva@curie.fr; Élodie Manié - elodie.manie@curie.fr; Caroline Brennetot - caroline.brennetot@curie.fr; François Radvanyi - francois.radvanyi@curie.fr; Alain Aurias - alain.aurias@curie.fr; Emmanuel Barillot - emmanuel.barillot@curie.fr

* Corresponding author †Equal contributors

Published: 22 May 2006

BMC Bioinformatics 2006, **7**:264 doi:10.1186/1471-2105-7-264

Received: 15 September 2005

Accepted: 22 May 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/264

© 2006 Neuvial et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Array-based comparative genomic hybridization (array-CGH) is a recently developed technique for analyzing changes in DNA copy number. As in all microarray analyses, normalization is required to correct for experimental artifacts while preserving the true biological signal. We investigated various sources of systematic variation in array-CGH data and identified two distinct types of spatial effect of no biological relevance as the predominant experimental artifacts: continuous spatial gradients and local spatial bias. Local spatial bias affects a large proportion of arrays, and has not previously been considered in array-CGH experiments.

Results: We show that existing normalization techniques do not correct these spatial effects properly. We therefore developed an automatic method for the spatial normalization of array-CGH data. This method makes it possible to delineate and to eliminate and/or correct areas affected by spatial bias. It is based on the combination of a spatial segmentation algorithm called NEM (Neighborhood Expectation Maximization) and spatial trend estimation. We defined quality criteria for array-CGH data, demonstrating significant improvements in data quality with our method for three data sets coming from two different platforms (198, 175 and 26 BAC-arrays).

Conclusion: We have designed an automatic algorithm for the spatial normalization of BAC CGH-array data, preventing the misinterpretation of experimental artifacts as biologically relevant outliers in the genomic profile. This algorithm is implemented in the R package MANOR (Micro-Array NORmalization), which is described at http://bioinfo.curie.fr/projects/manor and available from the Bioconductor site http://www.bioconductor.org. It can also be tested on the CAPweb bioinformatics platform at http://bioinfo.curie.fr/CAPweb.

Background

Array-based comparative genomic hybridization (array-CGH) provides a quantitative measure of differences in copy number between two DNA samples [1]. The technique is typically applied to cancer studies because chromosome aberrations frequently occur during tumor progression [2]. Array-CGH facilitates the localization and identification of oncogenes and tumor suppressor genes,

which are likely to be present in chromosomal regions gained and lost, respectively, in cancer cells.

Recent developments in the statistical analysis of array-CGH data have focused on high-level analysis, typically the identification of breakpoints from the genomic profile [3-7], rather than normalization. Most of the normalization techniques used to date for array-CGH data analysis have therefore involved the simple transposition of methods originally designed for expression data [8,9], correcting for differences in the labeling efficiency of the two dyes, spotting effects (block, row, column, or print-tip effects), and local or global intensity dependence of the ratios [10]. As far as we are aware, Khojasteh *et al*. [11] have reported the only method specific to CGH arrays.

Investigation of the systematic sources of variation in the array-CGH data studied showed that the effects affecting expression arrays were negligible with respect to spatial effects of two types. We describe here an algorithm for spatial normalization, which can also be combined with existing normalization methods for handling non-spatial artifacts. We will define and illustrate these two types of spatial effect, and show that such effects are not properly taken into account by traditional normalization techniques.

Two distinct types of spatial artifact

The methods proposed here were originally developed for the analysis of bladder cancer data from tumors collected at Henri Mondor Hospital (Créteil, France) [12], analyzed by hybridization on CGH arrays (F. Radvanyi, D. Pinkel *et al*., unpublished results), including 2464 clones spotted at the University of California San Francisco (UCSF) [13]. They were then adapted to several data sets for CGH arrays produced and hybridized at the Institut Curie, including the breast cancer data (O. Delattre, A. Aurias *et al*., unpublished results) and the neuroblastoma data [14] (which is publicly available [15]) used to illustrate the technique.

We identified two types of spatial effect with fundamentally different natures: *local spatial bias* (Fig. 1(a)) and *continuous spatial gradients* (Fig. 2-1(a)):

Local spatial bias

The array image shows clusters of spots with a discrete signal shift, with the other spots of the array remaining unchanged. These clustered shifted spots on the array image (Fig. 1(a)) have no biological explanation, and correspond to outliers on genomic profiles (Fig. 3(e) and 6(e)). In the data sets studied here, this artifact was found to affect about half of all arrays. We describe it as *local* because it affects only limited areas of the array.

Continuous spatial gradient

The array image shows a smooth gradient in signal from one side of the slide to the other (Fig. 2-1(a)). This artifact leads to genomic profiles with high variability, even between regions with the same DNA copy number. When

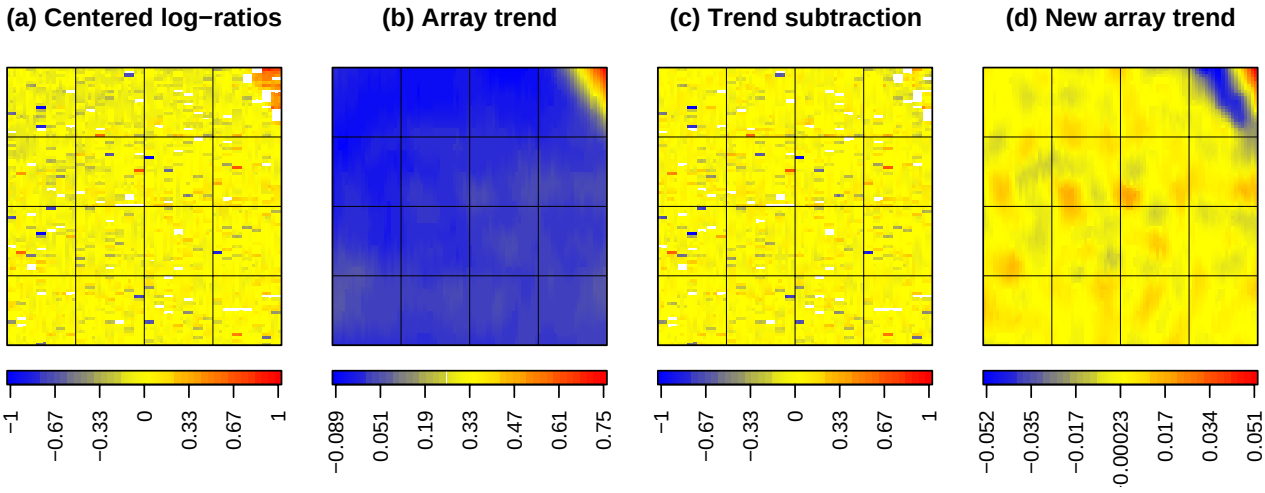


Figure 1

The need for an image segmentation method. An array with areas of local spatial bias (bladder cancer data): a straightforward trend correction method does not address the spatial effect appropriately. (a) Median-centered log-ratios; (b) spatial trend; (c) log-ratios after trend subtraction; (d) remaining spatial trend after subtraction (the color scale is not the same as in (b)). Colors are proportional to signal log-ratios; white dots correspond to missing values.

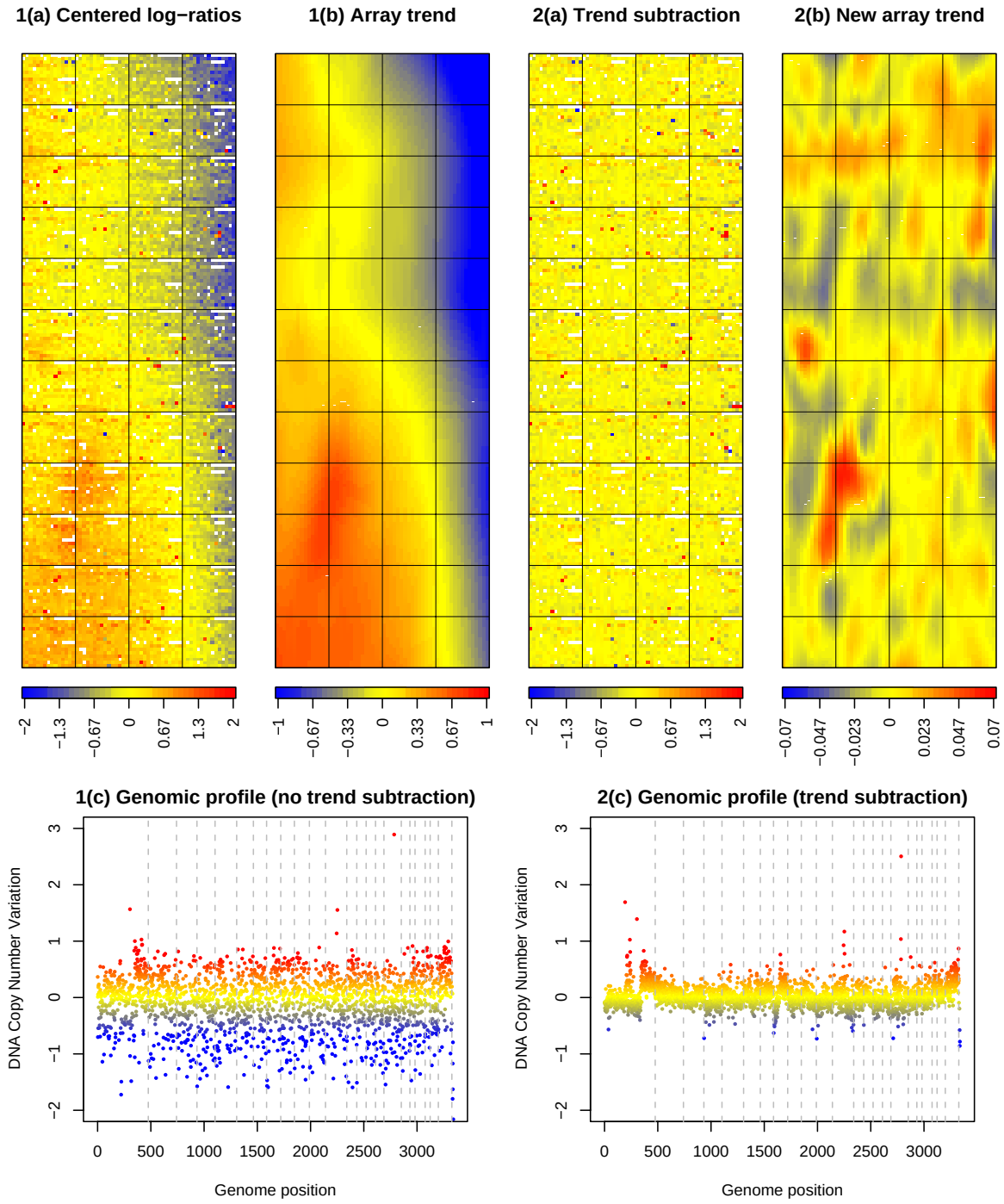


Figure 2

Results of the gradient subtraction step (2dLoess) on a breast cancer array. Correction of the spatial gradient of a breast cancer array: continuous spatial gradients are correctly taken into account by the proposed normalization method. 1(a) Median-centered log-ratios; 1(b) spatial trend; 1(c) genomic profile without spatial normalization; 2(a) corrected log-ratios; 2(b) spatial trend after correction (the color scale is not the same as in 1(b)); 2(c) genomic profile after spatial normalization. The vertical gray dashed lines indicate the separation between chromosomes.

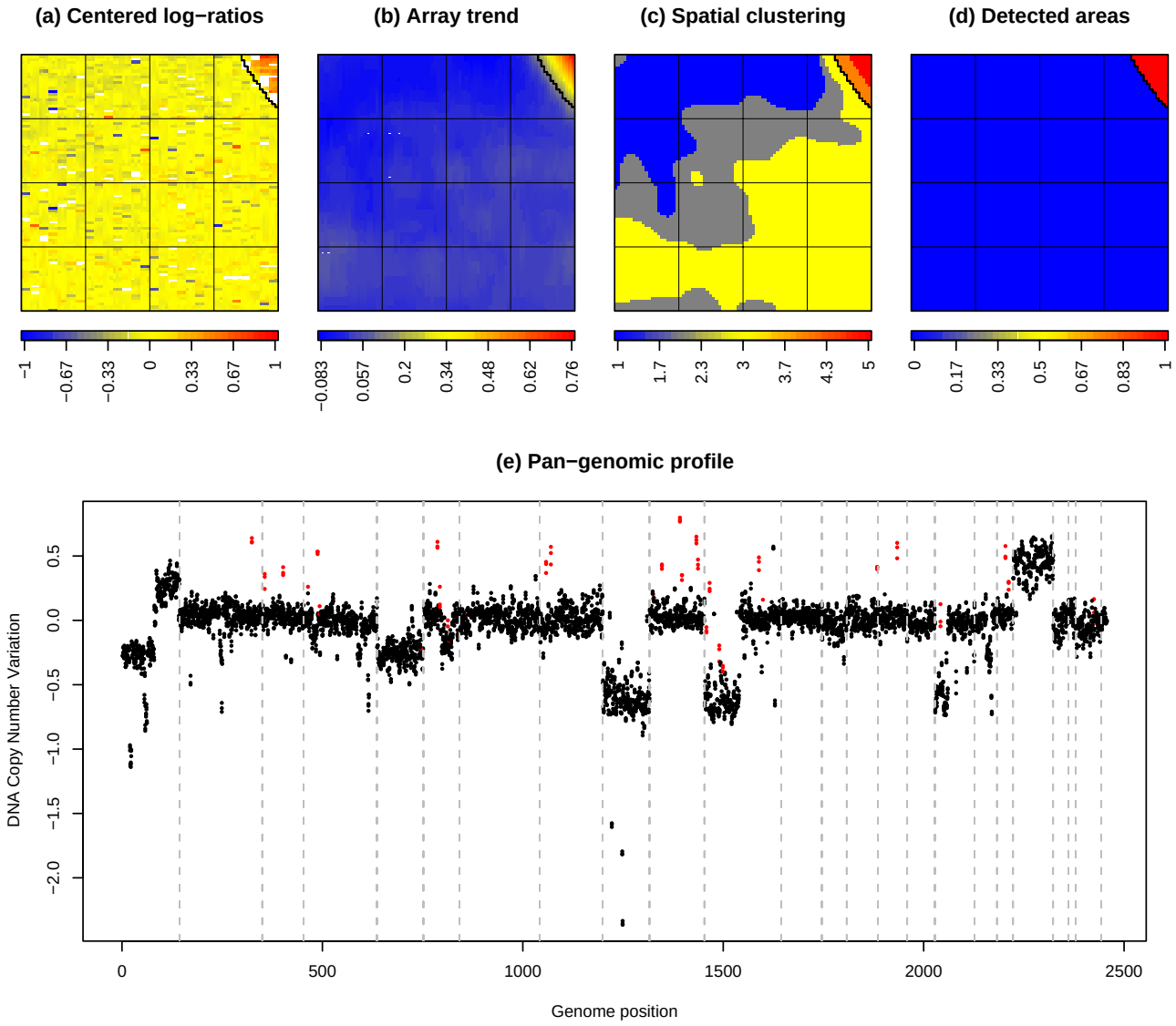


Figure 3

Results of the proposed spatial segmentation method (seg) on a bladder cancer array. Bladder cancer array with local spatial bias accurately detected by the proposed normalization method. (a) Median-centered log-ratios; (b) spatial trend; (c) spatial segmentation; (d) local spatial bias. The border of areas affected by local spatial bias that have been detected in panel (d) are reported on panels (a), (b) and (c) as a black step-function for easy interpretation; (e) genomic profile without spatial normalization (spots detected as local spatial artifacts are marked in red, and the vertical gray dashed lines indicate the separation between chromosomes).

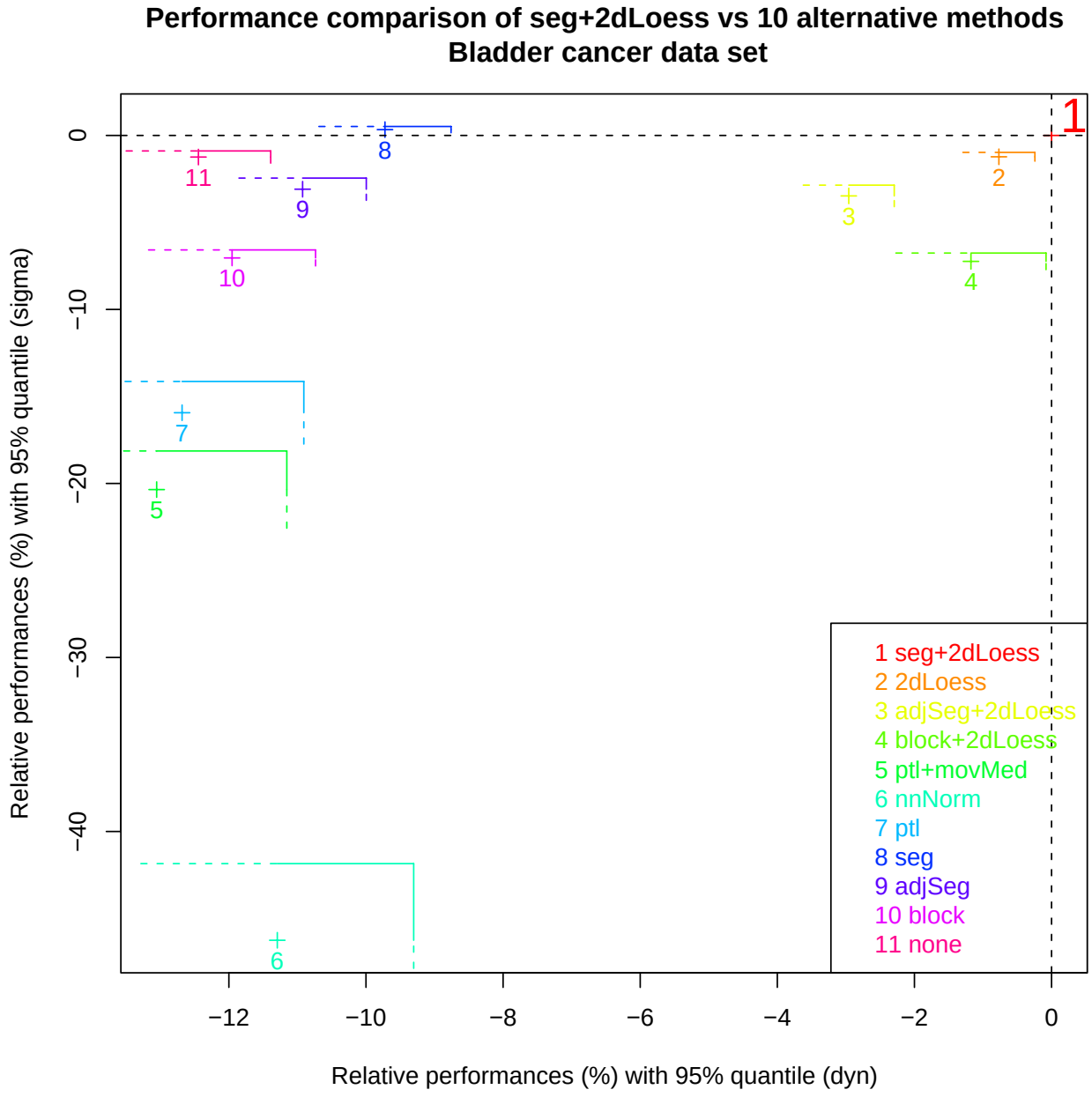


Figure 4

The proposed method (seg+2dLoess) compares favorably to all other normalization methods – bladder cancer data set. We compared the proposed method (*seg+2dLoess*) to ten methods for two quality criteria: *sigma* and *dyn*. Each color corresponds to the comparison of *seg+2dLoess* with a different method. The proposed method is taken as a reference (red point 1 at (0, 0)). For each method *i*, the cross indicates the mean relative performance (see methods section) of the data set for *dyn* (x axis) and in *sigma* (y axis), and the lines give the corresponding 95% quantile of relative performance. For *sigma* (*dyn*, respectively), the methods with a 95% quantile below (left to, respectively) the horizontal (vertical, respectively) dashed black line are significantly outperformed by our proposed method. Here *seg+2dLoess* significantly outperforms all methods for *dyn* and *sigma*, except *seg*, which performs slightly better for *sigma*. Methods 2, 3, and 4, which contain a gradient subtraction step using 2*dLoess*, perform the best against *seg+2dLoess*, as they cluster near the top-right corner of the image. However, *seg+2dLoess* still significantly outperformed these methods for both *sigma* and *dyn*.

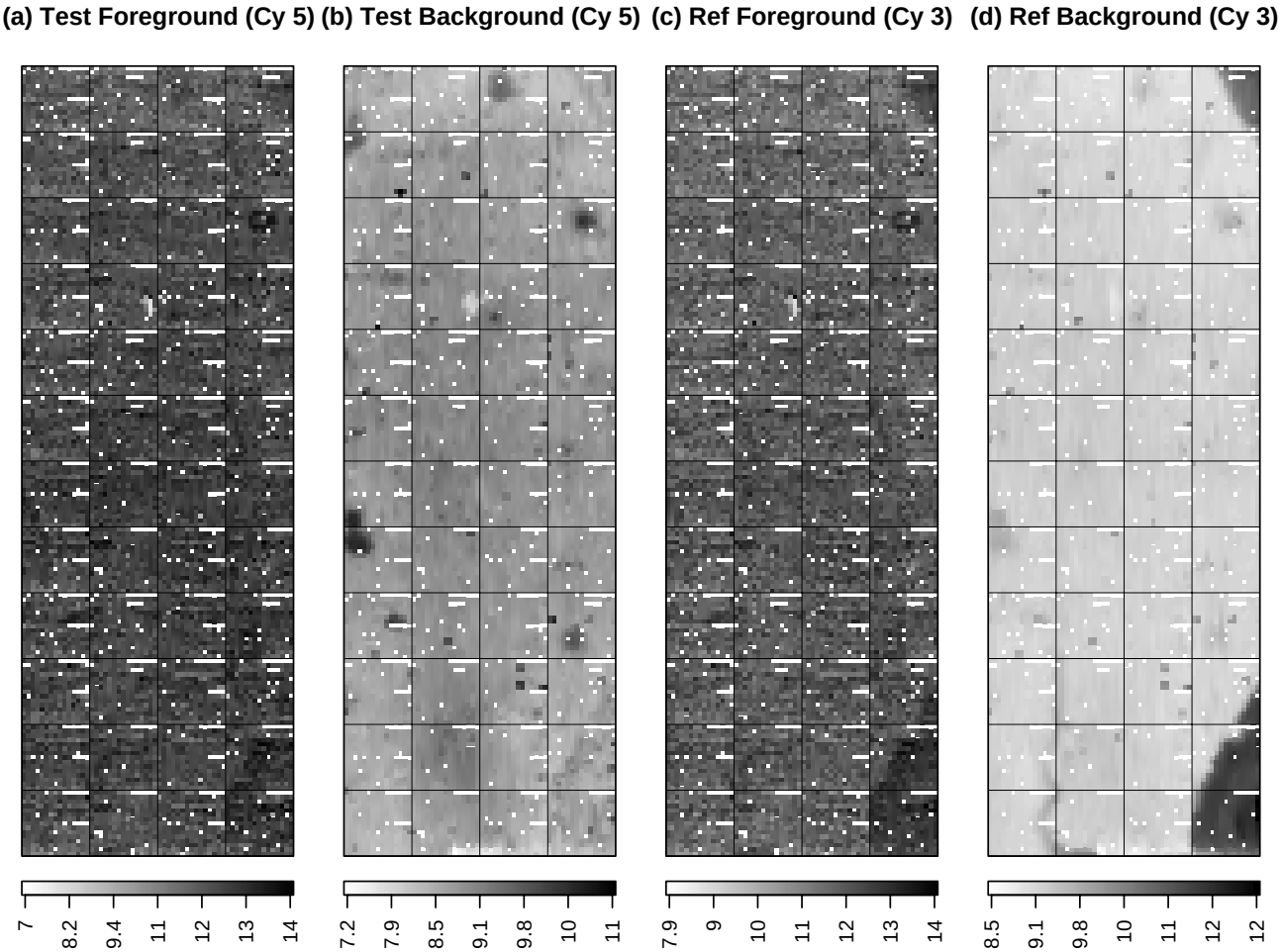


Figure 5
Evidence of local spatial bias on foreground and background raw signals on a breast cancer array. Log-ratios of the four raw signals of a breast cancer array: local spatial biases are easier to detect on a Cy3 background. (a) Test foreground; (b) test background; (c) reference foreground; (d) reference background. Gray-scale level is proportional to signal value.

this effect is observed, it affects all spots to various degrees.

These two types of effect are experimental artifacts of non-biological origin:

- They occur on arrays designed such that neighboring spots on the array correspond to non-neighboring clones in the genome, so there is no obvious biological reason for the clustering of high (or low) signals on the array;

- They are frequently observed on control (normal tissue vs normal tissue) hybridizations, and even on background signals (see Figure 5 for illustration with the breast cancer data set).

The methods proposed are designed to remove or reduce these two types of spatial effect, while preserving the true biological signal.

The need for a spatial segmentation method

The spatial effects described above cannot be attributed to spotting, for two reasons: firstly, they are not limited to array rows, columns or blocks; secondly, they are not reproducible from one array to another, even for arrays taken from batches of slides printed at the same time.

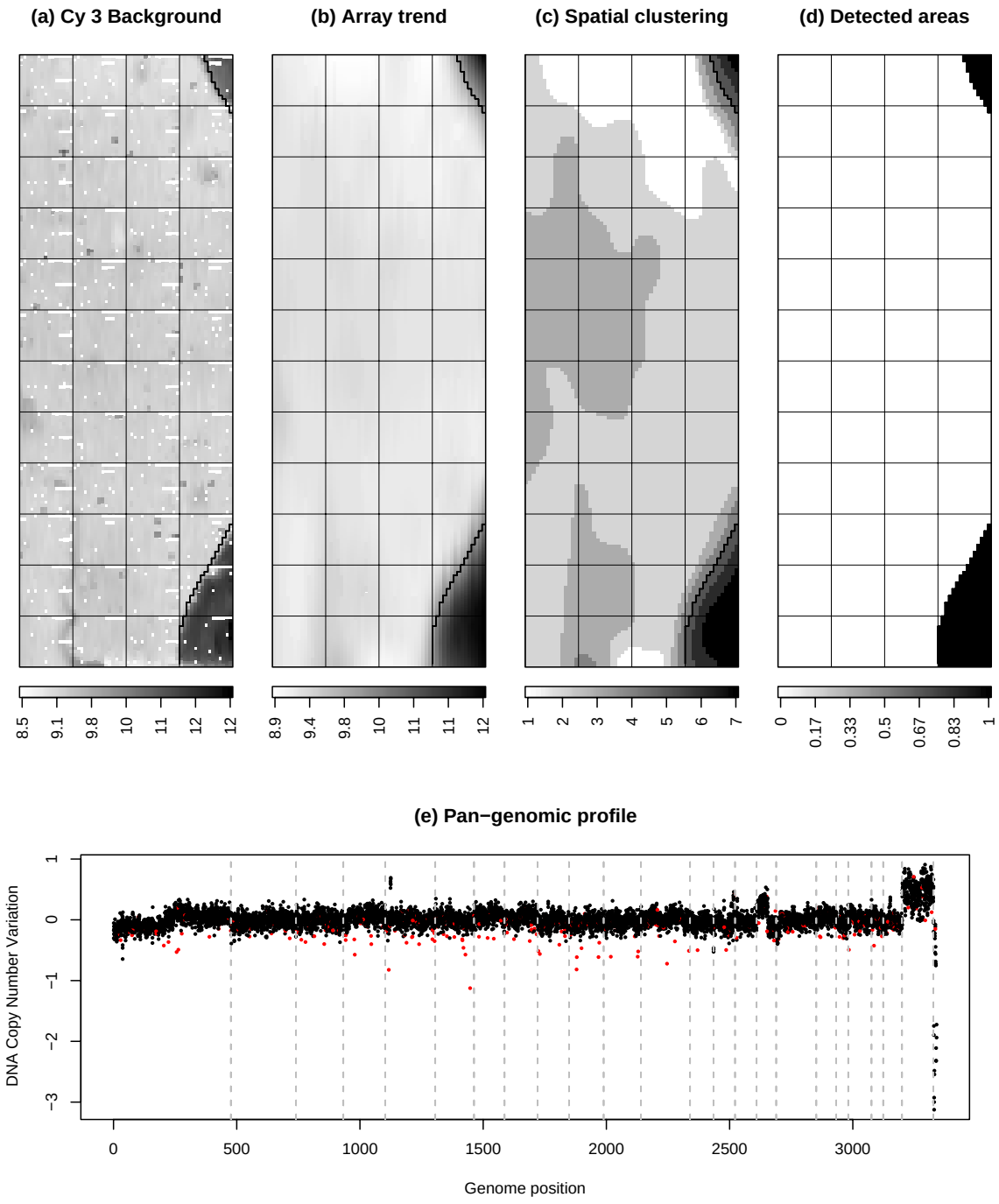


Figure 6

Results of the local spatial normalization step (seg) on a breast cancer array. Breast cancer array with local spatial bias accurately detected by the proposed normalization method. (a) Background signal log-ratios (Cy 3); (b) spatial trend; (c) spatial segmentation; (d) local spatial bias. The border of areas affected by local spatial bias that have been detected in panel (d) are reported on panels (a), (b) and (c) as a black step function for easy interpretation; (e) genomic profile without spatial normalization (spots detected as local spatial artifacts are marked in red, and the vertical gray dashed lines indicate the separation between chromosomes).

Therefore, it is not possible to correct for them properly with the normalization methods generally used for expression arrays, in which "spatial" effects are captured only by row, column, or print-tip group effects. For a method to be appropriate, it must take into account the spatial structure of the array as a whole, and the arbitrary shape of these biased areas.

Several different studies have taken into account spatial effects in expression microarray data and have provided signal correction methods. For example, Workman *et al*. [16] defined a spatial gradient normalization method using a two-dimensional Gaussian function to estimate local background bias in a probe neighborhood. Baird *et al*. [17] proposed a mixed model for cDNA array data, using splines with spatial autocorrelation, assuming the existence of a one-step correlation between adjacent spots in a row or column. Colantuoni *et al*. [18] proposed a method for normalizing the element signal intensities to a mean intensity calculated locally across the surface of a DNA microarray. Others studies have combined intensity-dependent and spatially-dependent effects. Wilson *et al*. [19] have proposed fitting a single LOESS curve on the MA plot and then spatially smoothing the residuals using a median filter to estimate the spatial trend. Tarca *et al*. [20] proposed correcting intensity-dependent and spatially-dependent effects using a feed-forward neural network. Khojasteh *et al*. [11] have compared different CGH array data normalization methods and suggested that a three-step normalization that combines print-tip LOESS with spatial correction using moving median and microplate effect correction gave the best results.

These methods may be suitable for correcting continuous spatial gradients, but they were not designed to detect abrupt changes in signal value across the array, and therefore may not adequately handle local spatial bias: Figure 1 illustrates the need for a spatial *segmentation* method to handle such local spatial effects. From the median-centered log-ratios (a) we estimate a spatial trend (b) by two-dimensional LOESS regression [21,22]; subtracting this spatial trend from the raw values partially corrects the spatial effect (c), but the array trend after correction (d) demonstrates that the spatial effect is undercorrected at the inner border of the biased area, and overcorrected at the outer border, consistent with the observation that signal disturbances vary steeply at the border of the biased area. This systematic overcorrection or undercorrection may lead to misinterpretation in the corresponding genomic profile.

A similar type of spatial effect was reported for expression microarrays by Reimers *et al* [23]. For CGH arrays, this type of effect should be easier to detect and correct, as they have a much smaller range of signal ratio variation than expression microarrays. However, this smaller range necessitates a much greater measurement precision for array-CGH data.

We describe here a spatial segmentation algorithm for the automatic *delineation* and *elimination* of unreliable areas, facilitating the exclusion of local spatial bias from array-CGH data. This algorithm consists of three steps, which are explained in detail in the Methods section:

[step 1]: Estimation of a spatial trend on the array using two-dimensional LOESS regression [21,22]

[step 2]: Segmentation of the array into spatial areas with similar trend values using NEM, an unsupervised classification algorithm including spatial constraints [24,25]

[step 3]: Identification of the areas affected by spatial bias.

A wide variety of microarray techniques based on BACs, cDNAs or oligonucleotides (see [26] for a review) may be used to quantify changes in DNA copy number. From a technical aspect, our method could be applied to any of these microarray types, although we detected local spatial bias only on BAC arrays.

Therefore, we focused on this technology, which has also been the most widely used so far. We provide examples of the implementation of this method and illustrate its performance with three data sets collected on two CGH-array platforms:

- The first data set (bladder cancer data) was produced at the UCSF. In this data set, local spatial effects were observed on 57% of 198 arrays, with a median of 229 affected spots, and no visual evidence of spatial gradients;

- The two other data sets were produced at the Institut Curie, INSERM U509. They consist of a breast cancer data set, in which local spatial effects were observed on 45% of 175 arrays, with a median of 592 affected spots, and a neuroblastoma data set [14,15], with local spatial effects on 23% of 26 arrays, and a median of 551 affected spots.

MANOR: an algorithm combining segmentation and signal correction

In addition to local spatial bias, we also frequently identified continuous spatial gradients, especially in breast cancer data set (Fig. 2-1(a)) and neuroblastoma data set. A straightforward way to correct for spatial gradients (Fig. 2-1(b)) is to subtract from the log-ratios an estimate of the spatial trend on the array (Fig. 2-2(a, b)). The first step of the spatial segmentation algorithm for detecting local spatial bias (step 1) provides such an estimate. This estimate

is calculated using two-dimensional LOESS regression as explained in detail in the Methods section.

In many cases, the CGH arrays were affected by both types of spatial effect: local spatial effects and continuous spatial gradients. In practice, we do not know in advance what type of spatial effect affects a given array. Thus, we propose the following two-step approach:

1. run the spatial segmentation algorithm (*seg*) to identify potential areas of local spatial bias

2. correct spots not excluded during the first step for continuous spatial gradients (*2dLoess*).

This algorithm, implemented in the MANOR package, will be referred to as *seg+2dLoess* in the remainder of this article. The rationale underlying this two-step approach is that arrays affected by continuous spatial gradients only will not be detected as containing local spatial bias by the step *seg*, and will therefore be properly corrected by the step *2dLoess*. This two-step approach is suitable for the spatial normalization of data sets containing both types of spatial effect.

Results and discussion

We have used our method for the spatial normalization of array-CGH data from two different platforms. In this section, we provide information about the practical implementation of the method on these two platforms, and quantitative results comparing our method to ten other normalization techniques. These compare the values of three quality criteria calculated after normalization of each array: the first, *sigma*, estimates the experimental variability between replicates, whereas the others, *smt* and *dyn*, evaluate quality in the context of the estimation of differences in DNA copy number between test and reference samples: *smt* quantifies the smoothness of the signal over the genome, and *dyn* assesses the dynamics of the signal, defined by the signal-to-noise ratio between gained and normal regions; these criteria are defined more formally and explained in detail in the Methods section.

To our knowledge, the ten normalization procedures used for the comparisons cover all the different types of approaches proposed so far and include the methods proposed by Tarca *et al*. [20], Yang *et al*. [10] and Khojasteh *et al*. [11]. These methods are detailed in the Methods section. For each normalization method, we calculated the three quality criteria for each array. When comparing two methods, we calculated a relative performance for each quality criterion, and assessed the significance of this performance using a Student's t-test, as explained in the Methods section. We show that our proposed method outperforms all previously published approaches for the three data sets.

Application to data produced at UCSF

The bladder cancer data set to which our algorithm was applied concerns 198 arrays that were spotted and hybridized at UCSF. These arrays consist of 7392 spots, corresponding to 2464 clones – all of which are BACs (Bacterial Artificial Chromosomes) – with the following design:

- Neighboring clones in the genome are dispersed on the array – a necessary condition for distinguishing between spatial artifacts and real biological information;

- Each clone is replicated three times on the array, and the three replicated spots are adjacent, so a high level of consistency for the three corresponding ratios does not prove that there are no spatial effects.

For this data set, spatial normalization is the last step in the following comprehensive normalization process. After image analysis of the arrays with SPOT 2.0 software [27], we screened for low-quality spots: spots with a foreground reference signal (and foreground DAPI signal) less than 125% of the background reference signal (reference DAPI signal) were discarded, as were clones with a log-ratio standard deviation exceeding 0.1. Clones for which only one of the three replicates was retained after these steps were then also discarded.

Finally, we applied the proposed spatial normalization method *seg+2dLoess* as follows: the spatial segmentation *seg* was applied to the log-ratios of this filtered array, with $K = 5$ and $\beta = 1$ (see Methods for a definition of these parameters and a discussion of how to choose them), followed by the correction for continuous spatial gradients *2dLoess*.

Spatial normalization step

Our segmentation algorithm detected local spatial effects on 113 of 198 bladder cancer arrays (57%); the median proportion of biased areas on these arrays was 3.1%. Figure 3 (top) illustrates the successive steps of the algorithm, from centered log-ratios to array trend, spatial segmentation of the array, and finally the delineation of biased areas. Red dots on the corresponding genomic profile (Figure 3, bottom) correspond to the spots discarded during spatial normalization (on this figure, signal log-ratios have not yet been averaged by clone: *spot-level information* is displayed).

Figure 3 (bottom) illustrates the improvement in data quality achieved with our spatial normalization method: among the apparent outliers (i.e. clones with log-ratio values significantly different from the mean log-ratio value

for the genomic region), it distinguished between experimental artifacts (red dots) and potentially biologically relevant outliers accounting for localized genomic amplifications.

Evaluation of the performance of the seg+2dLoess *method*

For each normalization method (11 methods including ours), we calculated the three quality criteria for each array and performed pairwise comparison of methods using the estimate and significance of their relative performance for each criterion, as explained in detail in the Methods section.

Figure 4 shows the results of comparison of the ten methods with *seg+2dLoess*. For the *dyn* criterion, *seg+2dLoess* significantly outperformed all methods (with all p-values $\leq$ 0.039), and most significantly methods 5 to 11, that do not include the *2dLoess* step (with all p-values below 8.5×10^{-18}). The *dyn* criterion is particularly important as it assesses the quality of copy number change detection. *seg+2dLoess* also gives significantly better results for the *sigma* criterion than all other methods (with all p-values below 1.1×10^{-8}) except one: *seg* performs significantly better ($p = 7.9 \times 10^{-4}$) but the relative improvement has a limited amplitude (only 0.36%).

For the *smt* criterion, *seg+2dLoess* also significantly outperforms all methods (with all p-values below 8.1×10^{-6}, except *block+2dLoess* for which $p = 0.048$).

Section 1 of the Additional file 1 shows similar plots to Figure 4, but for the *smt* and *dyn* criteria, and for the *smt* and *sigma* criteria. Tables 1 to 3 of the Additional files 2 and 3 summarize the results of all the pairwise comparisons of methods for the three quality criteria.

Taken together, these results show that the seg+2dLoess method outperforms its competitors for the bladder cancer data set.

Application to data produced at Institut Curie, INSERM U 509

The Institut Curie, INSERM U509 has developed its own high-density CGH array; all steps in the production of these chips are performed in Institut Curie laboratories, including array spotting, DNA preparation, hybridization, scanning and image processing. The current version of the array contains 3342 clones, each of which is spotted at least three times on the array, giving a total of 10800 to 11520 spots (including controls).

This array was designed to facilitate distinction between relevant biological effects and experimental artifacts: "empty" spots and spots of water were included as controls, clone replicates were scattered over the array, and the positions of clones on the array are not correlated with their actual positions in the genome. A reliable ratio value can therefore be calculated even if one of the three replicates is flagged. The arrays were scanned using an Axon Genepix 4000b scanner, and images were processed with Genepix Pro 5.1.

We analyzed a breast cancer data set and a neuroblastoma data set from this platform.

For this platform, we applied the proposed spatial normalization method *seg+2dLoess* as follows: the spatial segmentation *seg* was applied to the Background signal as explained in the paragraph below, and the spatial gradients were corrected by *2dLoess* calculated over the log-ratios. A post-processing step that includes spot and clone screening was then applied (allowing us, for example, to discard spots having too low a signal-to-noise ratio, or with poor replicate consistency).

Detail of the spatial segmentation step

Although we can correct the foreground signal for background intensity, a significant proportion of arrays still show localized spatial patterns that cannot be attributed to biological causes. Visual examination of spatial representations of the four signals (foreground and background intensities for test and reference signals) revealed that the bias was much clearer for the background signal of Cy3-labeled samples (Figure 5), which was not the case for bladder cancer data. We therefore applied the spatial segmentation method described above to the background signal of the Cy3 channel, with $K = 7$ and $\beta = 1$ (see Methods for a definition of these parameters and a discussion of how to choose them).

Biased areas of the CGH array are flagged and excluded from subsequent analysis. As clone replicates are not adjacent on the array, at least two of the three replicates generally remain after spatial bias correction, and a reliable ratio value can still be calculated. Figure 6 shows the results of this spatial segmentation step in the case of an array with local spatial bias but no spatial gradients.

Evaluation of the performance of the method seg+2dLoess

As for bladder cancer data, we calculated the three quality criteria for each normalization method and for each array for the breast cancer data set and the neuroblastoma data set. We then compared the methods paiwise using the estimate and significance of their relative performance for each criterion, as explained in detail in the Methods section.

Figures 7 and 8 show the results of comparing the ten methods with *seg+2dLoess* for the *dyn* and *sigma* criteria. *seg+2dLoess* significantly outperforms all other methods

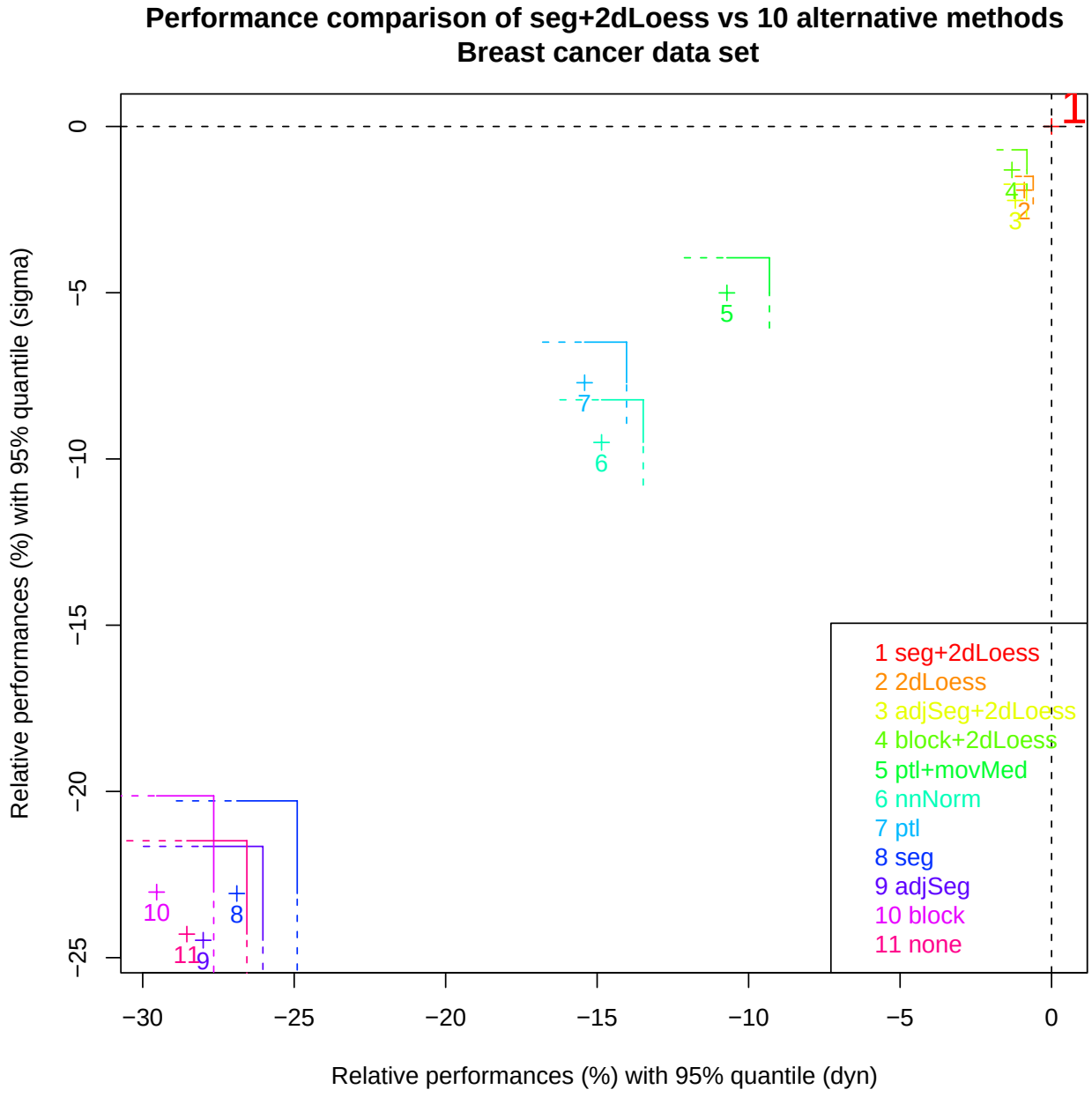


Figure 7

The proposed method (seg+2dLoess) compares favorably to all other normalization methods – breast cancer data set. We compared the proposed method (*seg+2dLoess*) to ten methods for two quality criteria: *sigma* and *dyn*. Each color corresponds to the comparison of *seg+2dLoess* with a different method. The proposed method is taken as a reference (red point 1 at (0, 0)). For each method *i*, the cross indicates the mean relative performance (see methods section) of the data set for *dyn* (x axis) and in *sigma* (y axis), and the lines give the corresponding 95% quantile of relative performance. For *sigma* (*dyn*, respectively), the methods with a 95% quantile below (left to, respectively) the horizontal (vertical, respectively) dashed black line are significantly outperformed by our proposed method. Here *seg+2dLoess* significantly outperforms all methods for *dyn* and *sigma*.

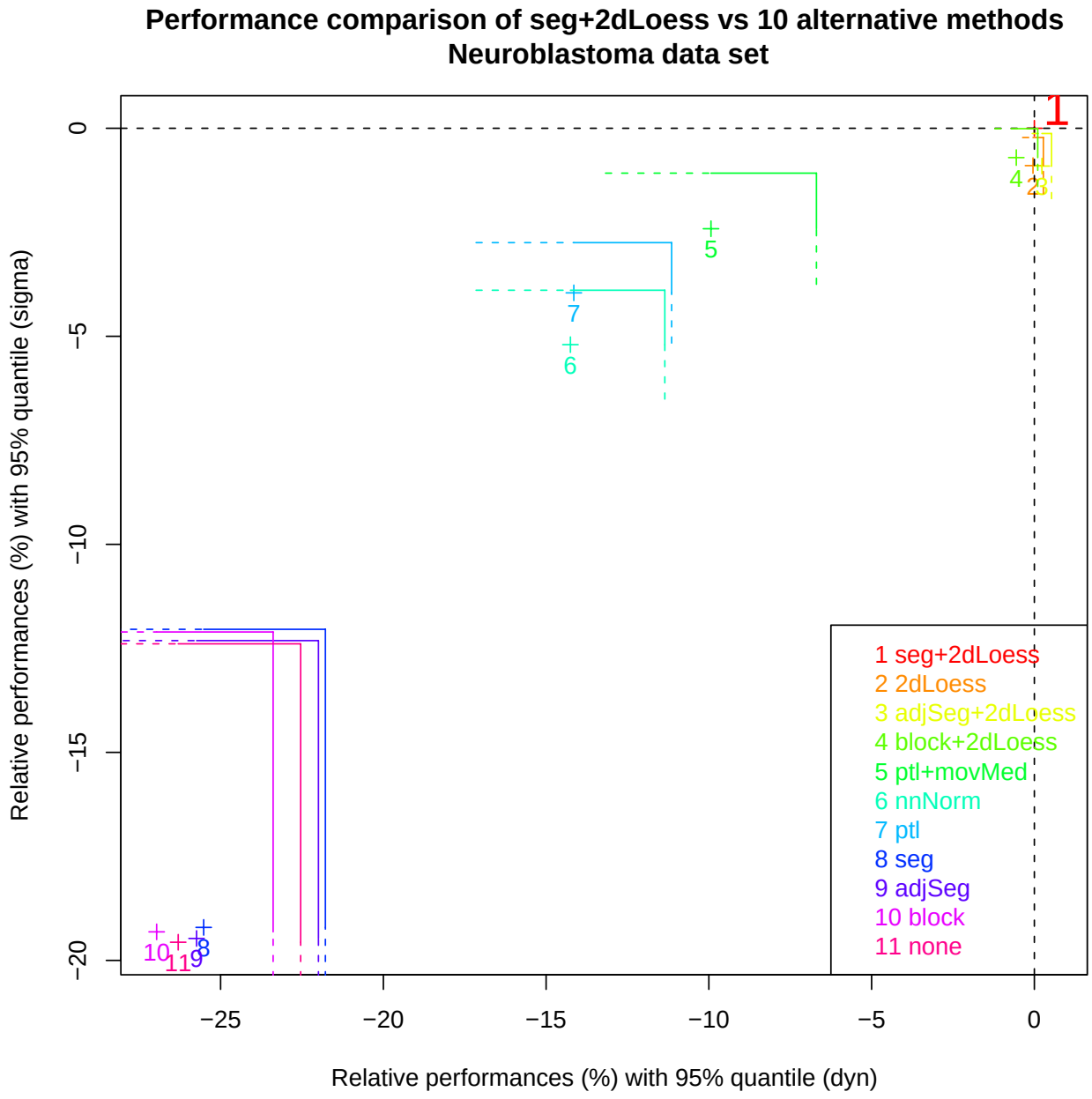


Figure 8

The proposed method (seg+2dLoess) compares favorably to all other normalization methods – neuroblastoma data set. We compared the proposed method (*seg+2dLoess*) to ten methods for two quality criteria: *sigma* and *dyn*. Each color corresponds to the comparison of *seg+2dLoess* with a different method. The proposed method is taken as a reference (red point 1 at (0,0)). For each method *i*, the cross indicates the mean relative performance (see methods section) of the data set for *dyn* (x axis) and in *sigma* (y axis), and the lines give the corresponding 95% quantile of relative performance. For *sigma* (*dyn*, respectively), the methods with a 95% quantile below (left to, respectively) the horizontal (vertical, respectively) dashed black line are significantly outperformed by our proposed method. Here *seg+2dLoess* significantly outperforms all methods for *dyn* and *sigma*, except those containing a gradient subtraction step with *2dLoess*.

for the three criteria on the breast cancer data set (with all p-values below 2.3×10^{-4}).

The neuroblastoma data set gives similar results: *seg+2dLoess* quality criteria are always better than those of the other methods, except for *dyn*, in which *adjSeg+2dLoess* is slightly better (0.22%) but not significantly so ($p = 0.1$). For *smt*, *seg+2dLoess* is only slightly better than *ptl+movMed* and the methods including the *2dLoess* step, but not significantly so for *adjSeg+2dLoess* and *ptl+movMed*. In these cases, the small size of the data set (26 arrays, 6 with local spatial bias) affects the statistical power.

Section 2 and 3 of the Additional file 1 and Tables 4 to 9 of the Additional files 2 and 3 detail and complement these results.

These results show that the *seg+2dLoess* method outperforms the other methods on the two data sets produced on the Institut Curie, INSERM U509 platform. The results also allow the methods to be ranked in terms of performance. Those methods that include a two-dimensional LOESS step are the highest ranked, with the methods proposed by [11,10] and [20], which all include some spatial processing, being next, and the other methods being the lowest ranked (see Figure 7 for example).

Conclusion

We have designed an efficient and automated algorithm for the spatial normalization of BAC array-CGH data, and defined a set of parameters for CGH array data quality assessment. We have shown that our method significantly improves the quality of data from two different BAC-array platforms and outperforms other normalization techniques on three data sets.

The proposed algorithm is particularly suitable for correcting spatial effects not related to array design (row, column, or print-tip group effects): indeed, the arrays studied show two distinct types of such spatial effect (local spatial bias and continuous spatial gradients), which can simultaneously affect any given array. In such cases, using spatial trend correction after spatial segmentation helps to remove or reduce these two types of spatial effect, while preserving the true biological signal.

This method is original in the application of a segmentation algorithm for detecting and removing local spatial bias, preventing the misinterpretation of experimental artifacts as biologically relevant outliers in the genomic profile.

This method was developed for array-CGH experiments, and gave very good results. However, it can be applied to any microarray experiment having the same types of spatial effect.

Availability and requirements

Our method is implemented in the R package MANOR (Micro-Array NORmalization) [28], which is available from the Bioconductor site [29]. It can also be tested on the CAPweb bioinformatics platform [30,31].

Methods

In this section, we provide details of the segmentation method and the other normalization techniques used for comparison, and of the quality criteria proposed. We also discuss the choice of the two parameters of the segmentation algorithm: K and β.

Description of the segmentation algorithm (seg)

The segmentation method consists of three steps:

[step 1]: Estimation of a spatial trend on the array using two-dimensional LOESS regression [21,22]

[step 2]: Segmentation of the array into spatial areas with similar trend values, using NEM, an unsupervised classification algorithm including spatial constraints [24,25]

[step 3]: Identification of the areas affected by spatial bias.

[step 1]: spatial trend estimation

We decided to carry out spatial segmentation based on an estimate of the spatial trend on the array, to optimize the robustness of segmentation. Furthermore, estimation of this trend makes it possible to replace missing values by interpolating the spatial trend.

The trend is estimated by means of a two-dimensional LOESS procedure with three iterative reweighting steps [21,22]. The local estimation is linear and the neighborhood taken into account to fit the local model corresponds to 3% of the total number of points. We use an iterative reweighting procedure to avoid outlier effects. Indeed, in the context of cancer studies, we are investigating changes in DNA copy number, and some clones displaying an amplification or a homozygous deletion may generate extreme but biologically meaningful values, which should not be interpreted as a local spatial bias.

When the spatial trend is estimated from the log-ratios, we first apply a basic correction to these log-ratios to prevent confusion between spatial artifacts and biologically relevant effects. For each chromosome arm, *centered* log-ratios are calculated as follows: the median of the corresponding log-ratio values is calculated and then subtracted from the initial values. The spatial trend is estimated from these centered log-ratios. This method helps to decrease the

impact of true genomic aberrations on the detection of spatial trends in the data, particularly for samples with many, or large genomic alterations, as most of these alterations correspond to the gain or loss of whole chromosome arms.

[step 2]: spatial segmentation

This step aims to identify K clusters corresponding to spots with similar signal levels located close together geographically. This is achieved by Neighborhood Expectation Maximization (NEM) [24,25]. We assume that the data are drawn from a mixed Gaussian density function $f(\mathbf{x}_i \mid \mathbf{\Phi}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}_i \mid \boldsymbol{\theta}_k)$ where p_k are the proportions of the mixture model, $f_k(\mathbf{x}_i \mid \theta_k)$ denotes the density function of a Gaussian distribution with parameter $\theta_k = (\mu_k, \Sigma_k)$ and $\Phi = \{p_1, ..., p_k, \theta_1, ..., \theta_K\}$ is the set of parameters to be estimated. The classical EM algorithm considers the following decomposition of the likelihood:

$$L(\mathbf{c}, \mathbf{\Phi}) = \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik} \log p_k f_k(\mathbf{x}_i \mid \boldsymbol{\theta}_k) - \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik} \log c_{ik} \qquad (1)$$

where

$$c_{ik} = \frac{p_k f_k(\mathbf{x}_i \mid \theta_k)}{f(\mathbf{x}_i)} \text{ and } \mathbf{c} = (c_{ik}) \qquad (2)$$

In the mixture model context, [32] pointed out that the EM algorithm is formally equivalent to the alternative maximization of L ($\mathbf{c}$, Φ) with respect to $\mathbf{c}$ ("E" step) and with respect to Φ ("M" step). The NEM algorithm is original in that it regularizes the likelihood by means of a term that takes into account the spatial dimension of the problem through the following adjacency matrix:

$$v_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases}$$

Here, the neighbors of a point located at coordinates (l, m) are the four points with the following coordinates: (l +1, m), (l - 1, m), (l, m - 1). We define the following quantity:

$$G(\mathbf{c}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K} c_{ik} c_{jk} v_{ij} \qquad (3)$$

Thus, instead of maximizing L ($\mathbf{c}$, Φ) in the E step, we maximize L ($\mathbf{c}$, Φ) + βG ($\mathbf{c}$). The value of β controls the weighting of the geographical context in the maximization. The M step remains unchanged.

[step 3]: elimination of local spatial bias

The basic idea is to remove from the array those spatial clusters with signal values significantly higher (or lower) than the unbiased areas of the array. We describe here the situation for positive spatial bias, but the idea can be adapted to negative bias. As local spatial biases cover a limited proportion of the array, we introduced a tuning parameter p_{max}, which corresponds to the maximum proportion of the array image corresponding to local spatial bias. In our experiment, local spatial bias typically applies to less than one quarter of the array, so we used p_{max} = 0.25.

After sorting the clusters identified by NEM by decreasing mean signal, we consider only those clusters with cumulative frequencies lower than p_{max} to be potentially biased, making it possible to define a set of candidate clusters. The mean signal value of the remaining clusters is used as a reference value for the unbiased signal. Each candidate cluster with a mean signal differing from this reference value by more than a given threshold value is considered biased. The other candidates are considered unbiased, unless their mean signal is closer to that of the biased cluster than to that of the reference: such clusters are also considered biased. This threshold was chosen based on the cross-validation of arrays analyzed by experts.

Comparison to other normalization methods

We compared the described methodology with other classical normalization methods. All these methods are listed below:

- A print-tip group method:

block (block normalization): we subtract off the row and column block median log-ratio values for each spot, and adds back the overall block median log-ratio value.

- A print-tip group with intensity dependent effect method:

ptl (print-tip loess): we apply the print-tip LOESS normalization [10] method using the marray R package (1.8.0 release, with default parameters) available from Bioconductor.

- A spatial smoothing method:

2dLoess (correction of continuous spatial gradients): a spatial trend is estimated by two-dimensional LOESS [21,22], which is then substrated from the log-ratio values.

- Two spatial segmentation methods:

seg (segmentation of local spatial bias): we apply the spatial segmentation algorithm described above to automatically eliminate the biased area.

adjSeg (correction of local spatial bias): we apply the spatial segmentation algorithm to automatically delineate the biased area. The median log-ratio value of such an area is then adjusted to the median log-ratio value of the unbiased area.

- A method combining print-tip group and spatial smoothing:

block+2dLoess (block normalization and global correction): we apply the *2dLoess* method on the normalized log-ratio values obtained with *block*.

- Two methods combining intensity dependent effect and spatial smoothing:

nnNorm (neural network normalization): we apply the normalization method described by Tarca *et al*. [20] using the nnNorm R package (1.5.1 release, with default parameters) available from Bioconductor. Briefly, this technique uses a neural network approach to correct the intensity-dependent and spatially-dependent effects.

ptl+movMed (print-tip loess and moving median filter): Khojasteh *et al*. [11] compared different normalization methods and suggested that combining the print-tip LOESS method with spatial correction (using a moving median calculated over a neighborhood of 11 rows by 11 columns) and microplate correction gave the best results. As the microplate information was not available in our data, we discarded the third step and only considered the print-tip LOESS and spatial correction.

- Two methods combining spatial segmentation and spatial smoothing:

adjSeg+2dLoess (correction of local spatial bias and continuous spatial gradients): we apply the *2dLoess* method on the normalized log-ratio values obtained with the *adjSeg* method.

seg+2dLoess (local segmentation and correction of continuous spatial gradients): we apply the *2dLoess* method on the log-ratio obtained with the *seg* method.

- Raw log-ratio values with no normalization (**none**).

Array-CGH data quality assessment

Definition of quality criteria

Evaluation of the quality of the signal ratios of an array facilitates the comparison of different image analyses or normalization algorithms, and makes it possible to quantify the improvement achieved by each step of a given normalization algorithm. We define three criteria for assessing the quality of the analyzed array: the first addresses the issue of overall quality whereas the other two provide quality evaluations for the estimation of differences in DNA copy number between test and reference samples.

sigma The first item provides an estimate of experimental noise. We isolate each clone and calculate the standard deviation of the log-ratio of the corresponding replicates. *sigma* is defined as the median of these standard deviations: the smaller the value of *sigma*, the higher the quality of the array.

The other two criteria are calculated after detection of the altered (gained or lost) regions in the test sample. We used the GLAD algorithm, developed by Hupé *et al*. [4] for this purpose:

smt Within a given DNA copy number region, the ratios of contiguous clones should not differ considerably. The second quality criterion concerns the *smoothness* of the signal log-ratios within such a chromosomal region: signal smoothness is defined as the median absolute difference between log-ratios for contiguous normal clones. If N denotes the set of clones considered normal after DNA copy number estimation, we can calculate

$$smt = \text{median}_{n \in N} |x_{(n)} - x_{(n-1)}|,$$

where $x_{(n)}$ is the value of the log-ratio at the n^{th} clone in genome order.

dyn The last criterion estimates the *dynamics* of DNA copy number variation between test and reference samples. We calculate the discrepancy between the median ratios of the regions considered "gained"(G) and "normal"(N) after DNA copy number estimation, and compare it with signal smoothness, as measured by *smt*:

$$dyn = \frac{\text{median}_{g \in G} x_g - \text{median}_{n \in N} x_n}{smt}$$

If no gained region is detected, we compare "normal" regions with "lost"(L) regions.

smt and *dyn* are not independent parameters and are anti-correlated. However, they quantify related but different ideas, as *smt* estimates the noise level after data normalization whereas *dyn* measures the ability to detect genome alterations after data normalization.

Paiwise comparison of quality criteria

These three criteria help us to decide which of two normalization methods gives the best results for a given array. In this pairwise comparison context, *smt* and *dyn* must be calculated with the same definition of *G*, *N*, and *L* regions for the two normalized arrays. We therefore define consensus *G*, *N*, and *L* regions associated with an array processed with two different normalization methods as the intersection of the two corresponding *G*, *N*, and *L* regions obtained using the two different normalization methods.

In order to test whether method *j* is better than method *i*, we defined a relative performance for each quality criterion as follows:

$$\begin{cases} RP^{sigma}(i,j) &= \dfrac{sigma(i) - sigma(j)}{sigma(i)} \\ RP^{smt}(i,j) &= \dfrac{smt(i) - smt(j)}{smt(i)} \\ RP^{dyn}(i,j) &= \dfrac{dyn(j) - dyn(i)}{dyn(i)} \end{cases}$$

We calculated this relative performance for each array, and assessed its significance by testing the hypotheses $\mathcal{H}_{i,j}: \{RP^{qc}(i,j) < 0\}$ for each quality criterion *qc*, using a Student's unilateral t-test.

In figures 4, 7, and 8, we calculated relative performances *RP*(*seg+2dLoess*, *test*) where *test* corresponds to one of the ten other methods. Hence a *negative value* for *RP* (*seg+2dLoess*, *test*) indicates that our proposed method outperforms the *test* method.

Parameter choice for the segmentation algorithm

The segmentation algorithm includes two parameters: the number *K* of clusters, and the regularization parameter β, which controls the weighting of geographic context in signal segmentation. Our experience suggests that the optimal choice of *K* and β may depend on the array-CGH technology used. We therefore provide guidelines for the choice of suitable parameters of the algorithm. We have investigated two different approaches to the choice of (*K*, β): incorporating a model selection criterion into the algorithm so that an optimal (*K*, β) can be chosen *for each array*, or developing a calibration method to help the user to find relevant sets of parameters for analyzing *a whole data set*. In this section, we discuss these two approaches and justify our choice of the second solution.

The difficulty finding optimal parameters on a per array basis

Choice of the number *K* of components in a mixture model can be addressed using model selection criteria. The basic idea is as follows: as the maximum likelihood estimator of the model increases mechanically with *K* (as model complexity increases with *K*), this method subtracts an increasing function of *K* from the likelihood of the model with *K* components, to prevent model overfitting. Many applications use the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) for this purpose. However, in our framework, *K* and β must be chosen simultaneously, because β also affects the maximum likelihood estimator. As we have no information concerning the quantitative behavior of the maximum likelihood estimator with respect to *K* and β (this complex question is beyond the scope of this paper), the choice of an appropriate penalization remains arbitrary.

We also considered an approach involving the fitting of *K* using model selection criteria and cross-validating the choice of β, but this approach has major drawbacks: first, it strongly increases the complexity of the estimation process, making this method too time-consuming for use as a routine normalization method; second, it makes the normalization method difficult to interpret, because two arrays from the same platform will not be treated with the same parameters.

Guidelines for choosing relevant parameters for analyzing a new data set

Rather than searching for optimal (*K*, β) values for each array, we provide a calibration method making it possible to choose appropriate (*K*, β) values for each data set. The basic principle of the calibration method is comparison of the output of our algorithm run on different (*K*, β) pairs, taken from a pre-defined grid (e. g. $K \in \{2,... 10\}$ and $\beta \in \{0.1,0.2,...2.0\}$).

We considered two different approaches to compare the results of the segmentations and to choose appropriate (*K*, β) values. The first approach involved choosing a (*K*, β) combination that optimizes quality criteria. The second involves expert assessment. An expert examines each array from a representative set and determines whether there is local spatial bias: he or she checks both the array image and the genomic profile to guarantee that the spatial effect is due to an experimental artifact rather than a biological effect. We then select the (*K*, β) combination that gives the best agreement between the expert decision and the algorithm decision. We call this second approach *expert assessment*. We found this second method simpler and more efficient than the first, for a number of reasons, outlined below.

In the first approach, quality criteria are calculated after normalization and DNA copy number assessment, so these three steps have to be carried out for each (*K*, β) combination. Therefore, although this method has the

obvious advantage of not relying on expert assessment, it is time-consuming, and provides only indirect evaluations of the differences between pairs of parameters, which may make the results hard to interpret. Moreover, a much lower level of variation was observed in the values of quality criteria for different (K, β) combinations for a given array than between arrays, so we were unable to identify optimal (K, β) values with this method (data not shown).

In the second approach, we considered two different ways of performing the expert assessment: either identifying arrays displaying local spatial bias (qualitative assessment), or estimating the number of spots that should be discarded (quantitative assessment). We found quantitative assessment to be very poorly reproducible, with large differences between experts, and much more time-consuming than the qualitative method. Therefore, we adopted the qualitative method, which made possible the rapid expert assessment of a larger number of arrays, thus increasing the accuracy of parameter choice.

Based on the qualitative expert assessment of an entire data set or a subset of data, we compare, for each array, the decision of our algorithm (has the algorithm detected a local spatial bias?) with that of the expert. We then calculate the proportion of false positives and false negatives for each combination of the parameters $K \in \{2,...10\}$ and $\beta \in \{0.1, 0.2,... 2.0\}$. Qualitative expert assessment remains highly variable (significant differences between experts), as a substantial proportion of arrays are difficult to classify. Nevertheless, all assessments show the same form of dependence in the error rate in (K, β), and lead to selection of the same parameters (data not shown).

For illustration, we use a subset of arrays on which two different expert assessments agree. The analysis is shown in Figure 9 for breast cancer data (134/179 arrays), and

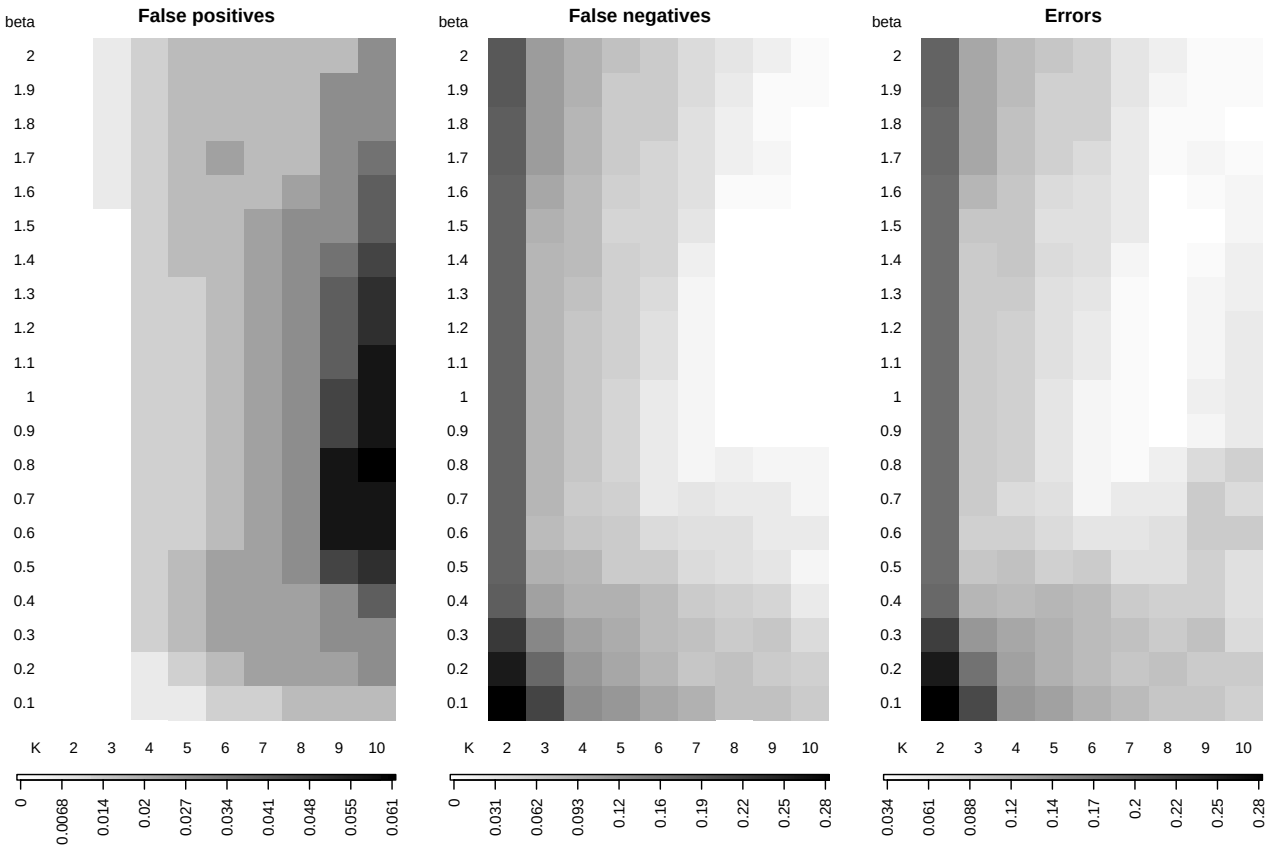


Figure 9
Comparison between qualitative assessment and segmentation results with various (K, β) –breast cancer data set. Thesegmentation algorithm is run with $K \in \{2,...10\}$ (x axis) and $\beta \in \{0.1, 0.2,...2.0\}$ (y axis) and compared with the expert assessment of the breast cancer data set. (a) False positive rate; (b) False negative rate; (c) Total error rate.

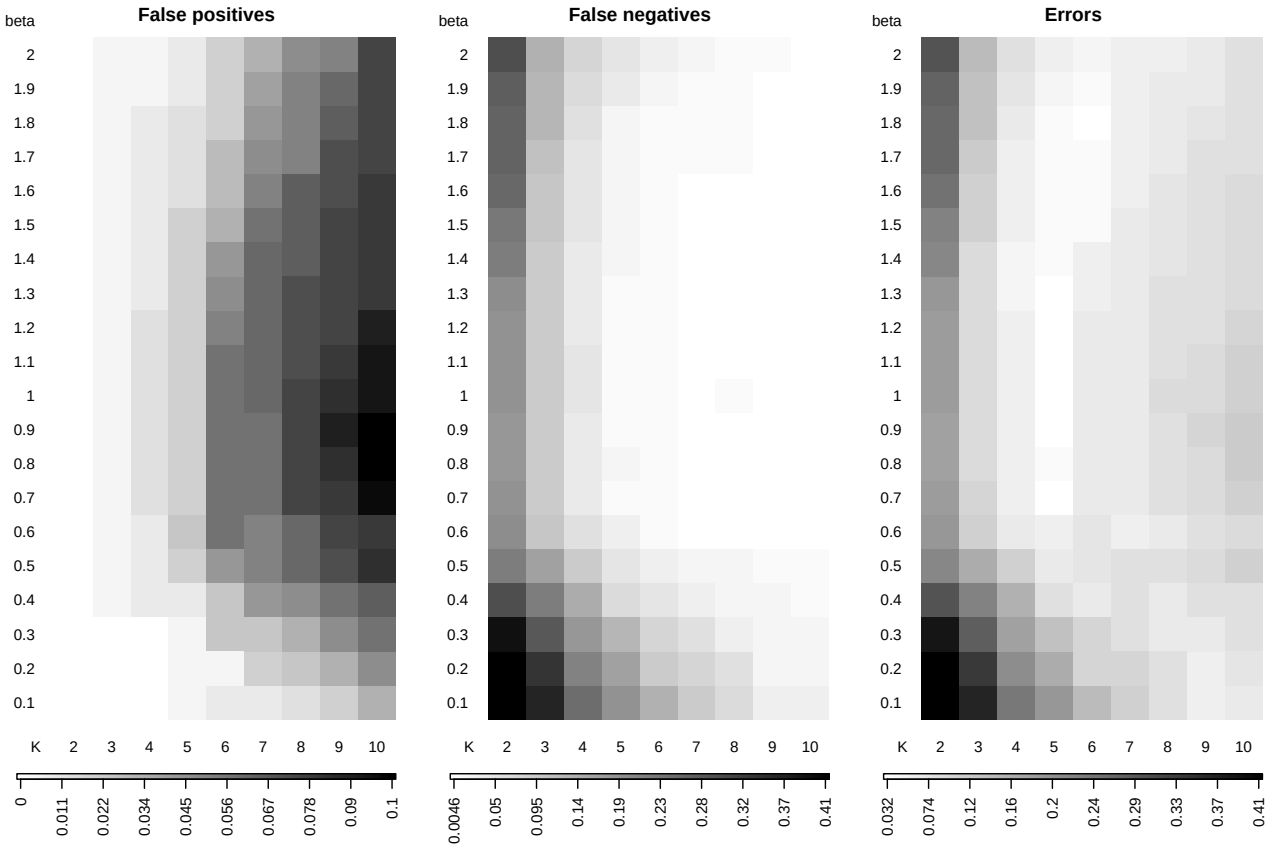


Figure 10
Comparison between qualitative assessment and segmentation results with various (K, β) – bladder cancer data set. The segmentation algorithm is run with $K \in \{2,...10\}$ (x axis) and $\beta \in \{0.1, 0.2,...2.0\}$ (y axis) and compared with the expert assessment of the breast cancer data set. (a) False positive rate; (b) False negative rate; (c) Total error rate.

Figure 10 for bladder cancer data (169/198 arrays). False positives are arrays that experts identified as having no local spatial bias, but which were identified by the algorithm as having local spatial bias. False negatives are arrays that the expert considered to contain local spatial bias, and for which no such areas were reported by the algorithm. Roughly speaking, K controls cluster size, and β influences both the size and spatial coherence of the clusters. As K increases (with fixed β), clusters tend to shrink, leading to an increase in the mean signal value of the highest cluster, making it more likely that this cluster will be identified as a local spatial bias. For fixed K, the highest cluster is slightly more likely to be detected as local spatial bias for intermediate β, corresponding to an extreme cluster with high, homogenous values: for low β this cluster is often quite large and incorporates too small signal values, whereas for very high β, the geographic context is too strong, leading to a highest cluster with heterogeneous signal values.

Drawing figures such as Figure 9 or 10 for any new data set can facilitate the identification of relevant sets of parameters for the segmentation algorithm. In our case, they suggest values of $K = 5$ and β between 0.9 and 1.3 for bladder cancer data set, and $K = 7$ or 8 and β between 0.9 and 1.3 for breast cancer data set. We used $K = 5$, $\beta = 1$ for the bladder cancer data set, and $K = 7$, $\beta = 1$ for the breast cancer data set.

Authors' contributions

PH and EB designed the study. PN and PH designed, coded and validated the spatial normalization algorithm. IB designed and coded the quality criteria. SL performed data integration. PH, PN, IB and EB drafted the manu-

script. EM, CB, FR and AA performed the microarray experiments and validated the spatial normalization algorithm. FR, AA and EB supervised the study. All authors read and approved the final manuscript.

Additional material

Additional File 1

Comparison of method seg+2dLoess with 10 alternative normalization methods. *We compared the method (*seg+2dLoess*) to ten methods for three quality criteria:* sigma, smt *and* dyn. *All images can be described as follows. Each color corresponds to the comparison of* seg+2dLoess *with a different method. The proposed method is taken as a reference (red point 1 at (0, 0)). For each method* i*, the cross indicates the mean relative performance on the data set for the two quality criteria compared, and the lines give the corresponding 95% quantile of the relative performance. The proposed method significantly outperforms, for the quality criterion shown in the y axis (at level 5%), all methods with a 95% quantile below the horizontal dashed black line. Similarly, the proposed method significantly outperformed, for the quality criterion shown in the x axis (at level 5%), all methods with a 95% quantile left of the vertical dashed black line. On most images, methods 2, 3, and 4, which contain a gradient subtraction step using 2*dLoess*, perform the best against* seg+2dLoess*, as they cluster near the top-right corner of the image. However,* seg+2dLoess *still significantly outperforms them for* sigma*,* smt *and* dyn.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-264-S1.pdf]

Additional File 2

p-*values of the relative performances of 11 normalization methods.* *We compare the results of 11 normalization methods on 3 data sets. Each table gives the significance levels of all pairwise comparisons between these 11 methods, for a given data set and a given quality measurement (*sigma, smt, dyn*). We calculated a relative performance for each array (as explained in the Methods section), and assessed its significance by testing the hypotheses* $\mathcal{H}_{i,j}^{qc}$ *:* $\{RP^{qc}(i, j) < 0\}$ *for each quality criterion* qc*, using a Student's unilateral t-test. The* p*-value associated to* $\mathcal{H}_{i,j}$ *is reported in cell (*i, j*).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-264-S2.pdf]

Additional File 3

Estimates of the relative performances of 11 normalization methods. *We compare the results of 11 normalization methods on 3 data sets. Each table gives the estimates of relative performance of all pairs of methods, for a given data set and a given quality measurement (*sigma, smt, dyn*). We calculated a relative performance for each array, and reported the mean value across all arrays of a given project in the following tables.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-264-S3.pdf]

Acknowledgements

This work was supported by the Institut Curie, the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, the IST program from the European Commission through the HKIS project (IST-2001-38153), the Cancéropole Ile de France, and the association *Courir pour la vie, courir pour Curie*.

The construction of the 3.3K BAC-array by Institut Curie, INSERM U509 was supported by grants from the Carte d'Identité des Tumeurs program of the Ligue Nationale Contre le Cancer.

We thank Isabelle Janoueix-Lerosey and Olivier Delattre (Institut Curie, INSERM U509) for making the neuroblastoma data set publicly available.

We thank Nadège Gruel, Virginie Raynal, Gaelle Pierron, Olivier Delattre (Institut Curie, INSERM U509) and Daniel Pinkel (University of California San Francisco) for fruitful discussions.

References

1. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.** *Nat Genet* 1998, **20:**207-211.
2. Albertson DG, Collins C, McCormick F, Gray JW: **Chromosome aberrations in solid tumors.** *Nat Genet* 2003, **34:**369-76.
3. Fridlyand J, Snijders A, Pinkel D, Albertson DG, Jain AN: **Application of Hidden Markov Models to the analysis of the array CGH data.** *Journal of Multivariate Analysis* 2004. Special Issue on Multivariate Methods in Genomic Data Analysis
4. Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E: **Analysis of array CGH data: from signal ratios to gain and loss of DNA regions.** *Bioinformatics* 2004, **20:**3413-3422.
5. Jong K, Marchiori E, van der Vaart A, Ylstra B, Weiss M, Meijer G: **Chromosomal Breakpoint Detection in Human Cancer.** In *Applications of Evolutionary Computing, EvoWorkshops2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, EvoSTIM, Volume 2611 of LNCS* Edited by: Raidl GR, Cagnoni S, Cardalda JJR, Corne DW, Gottlieb J, Guillot A, Hart E, Johnson CG, Marchiori E, Meyer JA, Middendorf M. University of Essex, England, UK: Springer-Verlag; 2003:54-65.
6. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data.** *Biostatistics* 2004, **5:**557-572.
7. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ: **A statistical approach for array CGH data analysis.** *BMC Bioinformatics* 2005, **6:**27.
8. Pollack JR, Sorlie T, Perou CM, Rees A, Jeffreys SS, Lonning P, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a direct role of DNA copy number alteration in the transcriptional program of breast tumors.** *PNAS* 2002.
9. Wang J, Meza-Zepeda LA, Kresse SH, Myklebost O: **M-CGH: Analysing microarray-based CGH experiments.** *BMC Bioinformatics* 2004, **5:**74.
10. Yang YH, Dudoit S, Luu P, Lin DM, Pend V, Ngai J, Speed TP: **Normalization of cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Research* 2002, **30:**e15:1-e15:11.
11. Khojasteh M, Lam WL, Ward RK, MacAulay C: **A stepwise framework for the normalization of array CGH data.** *BMC Bioinformatics* 2005, **6:**274.
12. Billerey C, Chopin D, Aubriot-Lorton MH, Ricol D, Gil S Diez de Medina, Van Rhijn B, Bralet MP, Lefrere-Belda MA, Lahaye JB, Abbou CC, Bonaventure J, Zafrani ES, van der Kwast T, Thiery JP, Radvanyi F: **Frequent FGFR3 mutations in papillary non-invasive bladder(pTa) tumors.** *Am J Pathol* 2001, **158:**955-1959.
13. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, SL S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG: **Assembly of microarrays for genome-wide measurement of DNA copy number.** *Nat Genet* 2001, **29:**263-4.
14. Janoueix-Lerosey I, Hupé P, Maciorowski Z, La Rosa P, Pierron G, Manié E, Liva S, Barillot E, Delattre O: **Preferential occurence of**

chromosome breakpoints within early replicating regions in neuroblastoma. *Cell Cycle* 2005, **4:**1842-1846.
15. **Replication timing data analysis in Neuroblastoma** [http://microarrays.curie.fr/publications/U509/reptiming]
16. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild HH, Nielsen C, Brunak S, Knudsen S: **A new non-linear normalization method for reducing variability in DNA microarray experiments.** *Genome Biology* 2002, **3(9):**research0048.1-0048.16.
17. Baird D, Johnstone P, Wilson T: **Normalization of Microarray Data Using a Spatial Mixed Model analysis which includes Splines.** *Bioinformatics* 2004, **20:**3196-3205.
18. Colantuoni C, Henry G, Zeger S, Pevsner J: **Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts.** *Biotechniques* 2002, **32:**1316-1320.
19. Wilson DL, Buckley MJ, Helliwell CA, Wilson IW: **New normalization methods for cDNA microarray data.** *Bioinformatics* 2003, **19:**1325-1332.
20. Tarca AL, Cooke JEK, Mackay J: **A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data.** *Bioinformatics* 2005, **21(11):**2674-2683.
21. Cleveland W, Devlin S, Grosse E: **Regression By Local Fitting.** *Journal of Econometrics* 1988, **37:**87-114.
22. Cleveland WS, Grosse E: **Computational Methods for Local Regression.** *Statistics and Computing* 1991, **1:**47-62.
23. Reimers M, Weinstein JN: **Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases.** *BMC Bioinformatics* 2005, **6:**166.
24. Ambroise C: **Approche probabiliste en classification automatique et contraintes de voisinage.** In *PhD thesis* Université Technique de Compiègne, France; 1996.
25. Ambroise C, Dang M, Govaert G: **Clustering of spatial data by the EM algorithm.** In *Geostatistics for Environmental Applications* Edited by: Soares A, Gomes-Hernandez J, Froidevaux R. Kluwer Academic Publisher; 1997:493-504.
26. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer.** *Nat Genet* 2005:S11-S17.
27. Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D: **Fully automatic quantification of microarray image data.** *Genome Res* 2002, **12:**325-332.
28. **MANOR: CGH Micro-Array NORmalization** [http://bioinfo.curie.fr/projects/manor]
29. **Bioconductor: Open software development for computational biology and bioinformatics** [http://www.bioconductor.org]
30. Liva S, Hupé P, Neuvial P, Brito I, Viara E, La Rosa P, Barillot E: **CAPweb : a bioinformatics CGH array Analysis Platform.** *Nucleic Acids Research* 2006 in press.
31. **CAPweb : a bioinformatics CGH array Analysis Platform** [http://bioinfo.curie.fr/CAPweb]
32. Hathaway RJ: **Another interpretation of the EM algorithm for mixture distributions.** *Journal of Statistics and Probability Letters* 1986, **4:**53-56.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

2.2 Identification of DNA copy number alterations

A need for an automatic algorithm We have seen in **Subsection 1.3.6** that some chromosome aberrations are related to DNA copy number changes. In **Subsection 1.4.2**, the aCGH technology used for the DNA copy number profiling of tumours has been described in detail. For each locus, this technology outputs a quantitative value which is ideally proportional to the DNA copy number. However, due to technical variability there is a fluctuation of the signal around its expected value and statistical methods are necessary to retrieve the true signal. Moreover, the DNA copy number of contiguous loci on the genome are very likely to have the same DNA copy number except at very particular loci, called *breakpoints*, which correspond to an abrupt change in DNA copy number. Therefore, a statistical method which takes into account the geographical proximity of the loci on the genome and has the ability to detect abrupt changes has been developed. This method is called Gain and Loss Analysis of DNA (GLAD) (Hupé et al., 2004) and the paper describing the algorithm is supplied as a material part of the thesis. Briefly, as shown in **Figure 2.2**, the GLAD algorithm allows the detection of breakpoints in the molecular profile (this step is called *segmentation*) and the assignment of a status (either loss, normal, gain or amplification) to each region identified (this step is called *calling*). The calling step provides valuable information for downstream analyses as suggested by Van Wieringen et al. (2007). The development of such an algorithm also avoid the tedious task of a manual expertise which is subject to error, non-reproducible and time-consuming (and even untractable for high-density chips).

Other methods for the DNA copy number analysis Different algorithms have been developed and their efficiency on BAC aCGH has been compared by Willenbrock and Fridlyand (2005) and Lai et al. (2005, 2008). Among the available methods, the main competitors were CBS (Olshen et al., 2004), CGHseg (Picard et al., 2005), GA (Jong et al., 2003), HMM (Fridlyand et al., 2004) and GLAD (Hupé et al., 2004). On Affymetrix GeneChip® Genome-Wide Human SNP Array chip, GLAD was also shown to be efficient (Baross et al., 2007). We give here the basic principles of the most efficient algorithms. In Olshen et al. (2004), the binary segmentation method (Sen and Srivastava, 1975) is modified to allow splits into either two or three segments. In this algorithm, termed Circular Binary Segmentation (CBS), the maximum of a likelihood ratio statistic is used recursively to detect narrower segments of aberration. Jong et al. (2003) used a Genetic Algorithm (GA) to maximise a likelihood with a penalty term containing the number of breakpoints. Fridlyand et al. (2004) used Hidden Markov Model (HMM) in which the underlying DNA copy numbers are the hidden states with certain transition probabilities. Picard et al. (2005) used a dynamic programming approach using a penalised likelihood in order to choose the most appropriate number of breakpoints.

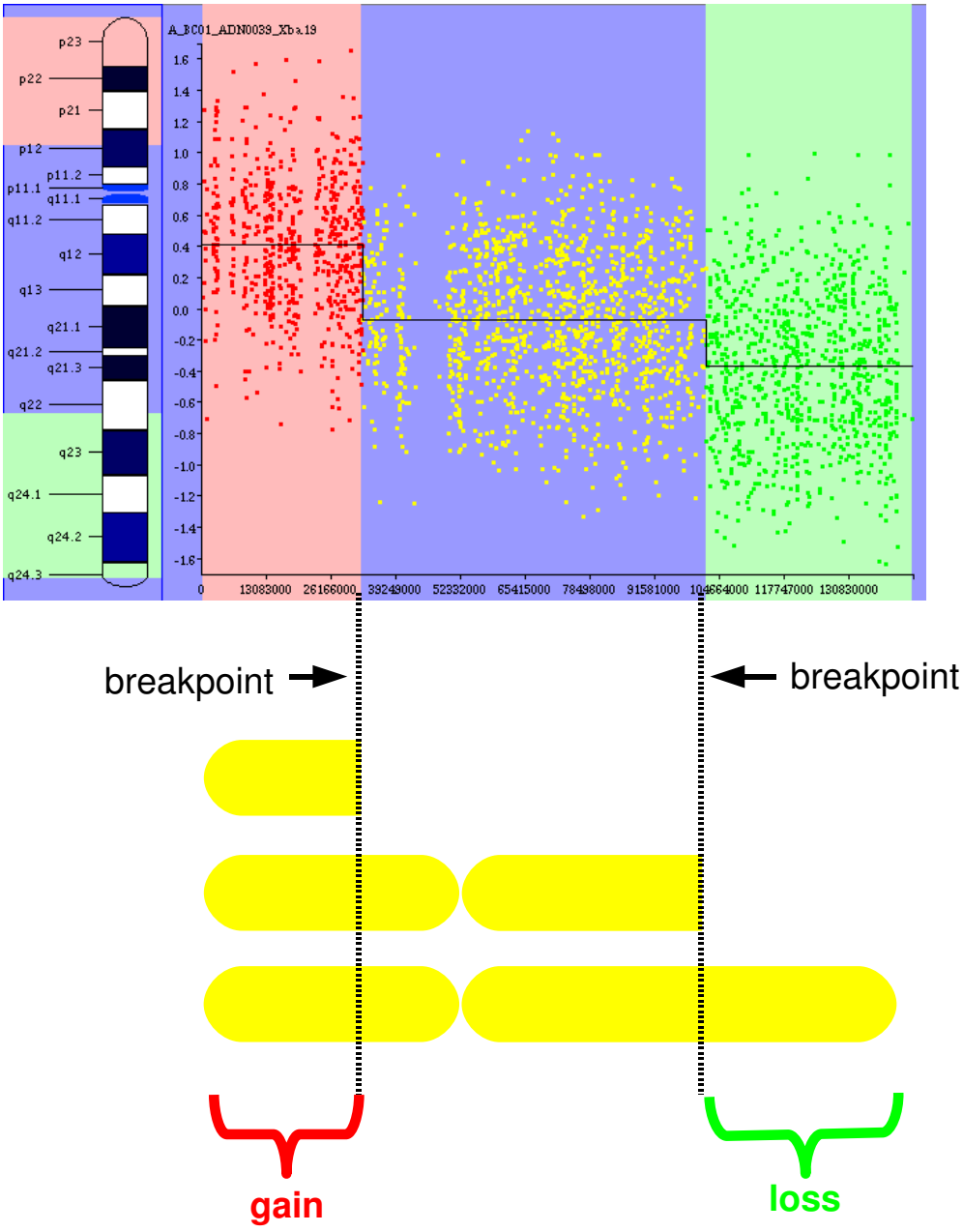


Figure 2.2: Breakpoint detection in DNA copy number profile - At the top is represented the DNA copy number value for chromosome 8 of a breast cancer tumour (data from Bollet et al., 2008). At the bottom is represented the underlying karyotype for this profile. The breakpoints correspond to a change of DNA copy number and allow the definition of a gain region and a loss region. The profile has been analysed with the GLAD algorithm (Hupé et al., 2004).

BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3413–3422
doi:10.1093/bioinformatics/bth418

Analysis of array CGH data: from signal ratio to gain and loss of DNA regions

Philippe Hupé[1,2,*], Nicolas Stransky[2], Jean-Paul Thiery[2], François Radvanyi[2] and Emmanuel Barillot[1]

[1]Service Bioinformatique and [2]UMR 144 CNRS/Institut Curie, 26, rue d'Ulm, Paris, 75248 cedex 05, France

Received on March 19, 2004; revised on June 18, 2004; accepted on July 12, 2004
Advance Access publication September 20, 2004

ABSTRACT

Motivation: Genomic DNA regions are frequently lost or gained during tumor progression. Array Comparative Genomic Hybridization (array CGH) technology makes it possible to assess these changes in DNA in cancers, by comparison with a normal reference. The identification of systematically deleted or amplified genomic regions in a set of tumors enables biologists to identify genes involved in cancer progression because tumor suppressor genes are thought to be located in lost genomic regions and oncogenes, in gained regions. Array CGH profiles should also improve the classification of tumors. The achievement of these goals requires a methodology for detecting the breakpoints delimiting altered regions in genomic patterns and assigning a status (normal, gained or lost) to each chromosomal region.

Results: We have developed a methodology for the automatic detection of breakpoints from array CGH profile, and the assignment of a status to each chromosomal region. The breakpoint detection step is based on the Adaptive Weights Smoothing (AWS) procedure and provides highly convincing results: our algorithm detects 97, 100 and 94% of breakpoints in simulated data, karyotyping results and manually analyzed profiles, respectively. The percentage of correctly assigned statuses ranges from 98.9 to 99.8% for simulated data and is 100% for karyotyping results. Our algorithm also outperforms other solutions on a public reference dataset.

Availability: The R package GLAD (Gain and Loss Analysis of DNA) is available upon request

Contact: glad@curie.fr

INTRODUCTION

Array Comparative Genome Hybridization (array CGH) is a recently developed technology based on DNA microarrays (Pinkel *et al*., 1998; Snijders *et al*., 2001; Solinas-Toldo *et al*., 1997; Ishkanian *et al*., 2004) and dedicated to the investigation and mapping of changes in DNA copy number. The array generally consists of spotted genomic sequences inserted into bacterial artificial chromosomes (BACs), e.g. (for ease of notation, we will refer to genomic sequences as BACs): each sample DNA is labeled with a fluorescent dye and the reference DNA is labeled with another fluorescent dye. This mixture is then hybridized to the array CGH. Typical applications of arrays CGH are cancer studies since chromosome aberrations frequently occur during tumor progression (Albertson *et al*., 2003) and human genetic disease research (Albertson and Pinkel, 2003; Shaw-Smith *et al*., 2004). In cancer studies, tumor DNA samples are compared with a normal reference DNA sample. The normal sample should have two copies of each genomic region (with the exception of the non-pseudo-autosomal regions of sexual chromosomes, for which a single copy is expected in males), whereas tumor DNA may present a loss or gain of DNA regions. In the simplest case, for a diploid tumor, the loss of a region will result in there being 0 or 1 copy whereas the gain of a region will result in there being three or more copies (the reality is more complex because a tumor is often not diploid). Measurement of the signal intensities of the reference and tumor samples for each BAC should make it possible to determine which regions have been gained or lost in the tumor sample.

Once a microarray has been constructed and hybridization carried out, several steps must be completed to determine which regions have been gained or lost: image acquisition, image analysis (including gridding, spot addressing, spot segmentation, spot quantification and outlier detection), signal normalization (e.g. to correct for systematic spatial or intensity biases) and duplicate treatment (each BAC is generally spotted in several copies to make possible statistical assessment of confidence). Once these steps have been completed, a synthetic value for the signal ratio is obtained, corresponding to the amount of DNA in the BAC concerned in the tumor with respect to that in the reference sample. The regions gained and lost can then be inferred from the ratio profile. Finally, correlation of the loss and gain profiles for a sufficiently high number of tumor samples should provide insight into the regions involved in tumorigenesis or tumor

*To whom correspondence should be addressed.

progression: oncogenes are likely to be present in the regions gained and tumor suppressor genes in the regions lost.

In this study, we assume that signal ratios for each BAC, such as those provided by SPOT 2.0 (Jain *et al*., 2002) or GenePix (Axon Instruments, 2003) software, are available and we focus on the problem of identifying the regions gained and lost from the ratio profile. Let us define the *status* of a homogeneous genomic region as the number of copies of the DNA of this region (here homogeneous means that all points in the region have the same DNA copy number) and Maximum Spanning Homogeneous Region (MSHR) as a region of homogeneous DNA status bordered either by a chromosome end or by another region of different status. Plots of BAC ratios (in fact, we use the ratio base 2 logarithm $\log_2$-ratio) versus BAC position (or rank) along the genome typically generate patterns in which MSHRs should be composed of spots distributed around a mean value that characterizes the status (cf. Fig. 1). Two adjacent MSHRs are separated by a breakpoint. Our approach can be broken down into two main steps: the detection of breakpoints and the assignment of a status to each MSHR. In some cases, a point deletion or amplification may affect the DNA, appearing on the ratio profile as an outlier among BACs with the same DNA status. This special case needs a particular treatment, called outlier detection.

In the absence of experimental biases, ratios should be 0 for double loss, $\frac{1}{2}$ for a single loss, 1 for the normal situation, $\frac{3}{2}$ for a single gain and more generally $\frac{n}{2}$ for a sample with n copies of DNA. In practice, microarray experiments are subject to various sources of variation, including differences in incorporation efficiency between the two fluorescent dyes, an intensity-dependent effect and a print-tip effect, as reported by Yang *et al*. (2001) for expression data. These variations create noise and bias the theoretical values. In addition, tumor biopsy samples generally contain a mixture of normal and tumor cells, and tumor cells may even present heterogeneity in terms of genome losses and gains, corresponding to different stages of tumor progression; these heterogeneities result in smaller signal gaps between regions.

To our knowledge, only two articles have dealt with the problem of breakpoint detection and none have considered the question of region assignment. Jong *et al*. (2003) used a genetic algorithm and local optimization to detect breakpoints. The algorithm developed by Olshen and Vankatraman (2002) is based on circular binary segmentation, as described by Sen and Srivastava (1975). This paper is organized as follows: we will begin by describing our breakpoint detection algorithm; we will then present the region assignment method, followed by a validation of our approach based on simulations, karyotyping results, loss of heterozygosity (LOH) (Vogelstein *et al*., 1989) and manually analyzed data. Finally, we discuss the result obtained and perspectives.

BREAKPOINT DETECTION

The problem of chromosomal breakpoint detection can be approached by estimating a piecewise constant function defining each MSHR of the chromosome. A solution to this problem of estimation has been proposed by Polzehl and Spokoiny (2002), with application in two dimensions to image segmentation. We present here the main principles of their algorithm—adaptive weights smoothing (AWS)—and describe how this algorithm should be applied to chromosomal breakpoint detection with array CGH data. The AWS procedure is an iterative, data-adaptive smoothing technique that was designed for smoothing in regression problems involving discontinuous regression function. It is assumed that the regression function can be approximated, e.g. by a simple local constant model. The regression function is estimated as a weighted maximum-likelihood estimate (MLE), with weights chosen in a completely data-adaptive way. The algorithm finds, around each point, the maximal neighborhood in which the local constant assumption holds. In our case, the maximal neighborhood of every BAC should allow us to delineate in a straightforward manner the MSHRs and the parametric estimation should provide its copy number. The procedure has a number of features of potential value for our problem: it has been shown to preserve contrasts and edges between regions (and should therefore detect breakpoints accurately), it requires very little prior information about the data to model and has a numerical complexity of nM with n the number of points (BACs) and M the size of the maximum neighborhood. The AWS is more general than simple piecewise constant function estimation, but it is straightforward to restrict it to our case.

Statistical model

Let us consider a series of N independent observations $S = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ in which each X_i is valued in a metric space $\mathcal{X}$ and determines the location (the BAC rank on the chromosome) and each Y_i is valued in another metric space $\mathcal{Y}$ and is the observation at X_i (the measured $\log_2$-ratio); the locations X_i are ordered such that $X_1 < \cdots < X_i < \cdots < X_N$. We also assume that the observation Y_i depends on the location X_i via a parameter $\theta \in \Theta$, where Θ is a subset of a finite-dimensional space $\mathbb{R}^d$. Conditionally on $X_i = x$, the random variable Y_i is distributed with the density probability function $p[y, \theta(x)]$ for some unknown $\theta(x)$ on $\mathcal{X}$ valued in Θ. Here, we consider the local constant gaussian regression model $Y_i = \theta(X_i) + \epsilon_i$, where the ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. We wish to infer the function θ such that θ is of the form $\theta(x) = \sum_{m=1}^{M} a_m \mathbf{1}(x \in \mathcal{X}_m)$ with disjoint regions $\mathcal{X}_1, \ldots, \mathcal{X}_M$ and $\mathcal{X} = \bigcup_{m=1}^{M} \mathcal{X}_m$. The regions $\mathcal{X}_m$, the values a_m and even the total number of regions M are unknown. Estimation of the parameter θ is a local estimation problem in that this parameter depends on the location.

The AWS procedure

The approach used for the local estimation of θ is based on local-likelihood modeling (Polzehl and Spokoiny, 2002) and extends the AWS procedure proposed by Polzehl and Spokoiny (2000). An iterative algorithm finds, around every location X_i, the maximal possible neighborhood in which the parameter θ is constant: a weight $w_{ij}(0 \leq w_{ij} \leq 1)$ is assigned to every observation Y_j at X_j, which depends on the previous step of the algorithm. The weighted MLE $\hat{\theta}(X_i) = \hat{\theta}_i$ is of the form:

$$\hat{\theta}_i = \operatorname*{argsup}_{\theta \in \Theta} L(W_i, \theta, \theta')$$

with

$$L(W_i, \theta, \theta') = \sum_{j=1}^{N} w_{ij} \log \frac{p(Y_j, \theta)}{p(Y_j, \theta')},$$

where θ' is an arbitrary point in Θ and $W_i = \text{diag}\{w_{i1}, \ldots, w_{iN}\}$.

At each iteration k, the geometric increase in $h^{(k-1)}$ by a growth rate $a > 1$ defines a new larger neighborhood around each X_i, which is used to calculate the MLE of θ_i. New weights are calculated by means of a location penalty kernel function K_l, which takes into account the proximity of the X_j's in the neighborhood, and a statistical penalty kernel function K_s, which takes into account the comparison of two local models. The kernels K_s and K_l are non-increasing functions and must fulfill $K_s(0) = K_l(0) = 1$. Moreover, a parameter λ controls the statistical penalty and a memory parameter $\eta(0 \leq \eta \leq 1)$ is used to stabilize the procedure. The detail of the procedure is given below (see Polzehl and Spokoiny, 2002):

(1) *Initialization*: Calculate the global MLE $\hat{\theta}^{(0)}$ of θ:

$$\hat{\theta}^{(0)} = \operatorname*{argsup}_{\theta \in \Theta} \sum_{i=1}^{N} \log p(Y_i, \theta) = \frac{1}{N} \sum_{j=1}^{N} Y_j.$$

For every $i = 1, \ldots, N$, set $\hat{\theta}_i^{(0)} = \hat{\theta}^{(0)}$ and define $W_i^{(0)}$ as the unit matrix. Set $k = 1$.

(2) *Iteration*: for every $i = 1, \ldots, N$

(a) *Calculate the adaptive weights*: For every point X_j, calculate the penalties

$$\boldsymbol{l}_{ij}^{(k)} = |\rho(X_i, X_j)/h^{(k)}|^2,$$

$$\boldsymbol{s}_{ij}^{(k)} = \lambda^{-1} \left[L\left(W_i^{(k-1)}, \hat{\theta}_i^{(k-1)}, \hat{\theta}_j^{(k-1)}\right) + L\left(W_j^{(k-1)}, \hat{\theta}_j^{(k-1)}, \hat{\theta}_I^{(k-1)}\right) \right] \Big/ 2,$$

where $\rho(x, x')$ is a metric in $\mathcal{X}$ and $h^{(k)}$ controls the size of the neighborhood of each X_i.

calculate

$$\tilde{w}_{ij}^{(k)} = K_l\left(\boldsymbol{l}_{ij}^{(k)}\right) K_s\left(\boldsymbol{s}_{ij}^{(k)}\right)$$

and define the weight $w_{ij}^{(k)}$ as

$$w_{ij}^{(k)} = \eta w_{ij}^{(k-1)} + (1 - \eta)\tilde{w}_{ij}^{(k)}.$$

Denote by W_i^k the diagonal matrix $W_i^k = \text{diag}\left\{w_{i1}^{(k)}, \ldots, w_{iN}^{(k)}\right\}$.

(b) *Estimation:* Calculate the new local MLE $\hat{\theta}_i^{(k)}$ of θ_i

$$\hat{\theta}_i^{(k)} = \operatorname*{argsup}_{\theta \in \Theta} L\left(W_i^{(k)}, \theta, \theta'\right).$$

(3) *Stopping*: Stop if $ah^{(k)} > h^*$, otherwise increase k by 1, set $h^{(k)} = ah^{(k-1)}$ and continue with step 2.

According to the assumption of our local constant gaussian model we have:

$$\hat{\theta}_i = \min_{\theta \in \Theta} \frac{1}{2\sigma^2} \sum_{j=1}^{N} w_{ij}(Y_j - \theta)^2.$$

$$L(W_i, \hat{\theta}_i, \theta') = \frac{\sum_{j=1}^{N} w_{ij}}{2\sigma^2}(\hat{\theta}_i - \theta')^2.$$

For the local constant gaussian regression model, the AWS procedure requires the parameter σ to be known. An estimate of σ is given by:

$$\frac{IQR(Z_1, \ldots, Z_{N-1})}{IQR(\mathcal{N}(0,1)) \times \sqrt{2}}, \tag{1}$$

where $Z_i = Y_{i+1} - Y_i$ and IQR defines the interquartile range.

The results of the AWS procedure provide one estimate of $\hat{\theta}_i$ for every $i = 1, \ldots, N$. Based on these estimates, we define a breakpoint as a location X_i such that $\hat{\theta}_i \notin [\hat{\theta}_{i+1} - \epsilon; \hat{\theta}_{i+1} + \epsilon]$ (in our case, $\epsilon = 10^{-2}$). Thus, a breakpoint corresponds to the last position of an MSHR. The chromosome can be split into $N' + 1$ MSHRs for a total number N' of breakpoints: $(X_1, \ldots, X_{B_1}), (X_{B_1+1}, \ldots, X_{B_2}), \ldots, (X_{B_{N'}+1}, \ldots, X_N)$. Note that we apply a particular process for singularity or outlier detection (detailed below). The procedure is run for each chromosome separately.

AWS parameters

The procedure requires the tuning of various parameters. We apply the exponential kernel $K_l(u) = K_s(u) = \exp(-u)$. For the neighborhood, we have chosen $h^{(0)} = 1$, $a = 1.2$ and $h^* = 10X_N$. The parameter λ has been set to the 0.999-quantile of the $\chi^2(1)$ distribution, to prevent there being too many breakpoints. The value of η has been set to 0.5 and corresponds to the memory parameter of the algorithm. Polzehl and Spokoiny (2002) suggested using the symmetric statistical

penalty $s_{ij}^{(k)}$ to detect fine structures, as might occur in cancer data. Nevertheless, very fine structures, such as single amplicons or deletions, may be missed and a special procedure is proposed in the next paragraph.

Outlier detection

The AWS procedure is based on the assumption that the maximal neighborhood on which parametric estimation can be carried out is large compared with the distance between two neighboring points. This procedure may therefore fail to detect very fine structures such as a BAC located in a MSHR for which the signal Y_i differs significantly from the expected values of this MSHR. Such a BAC is called an outlier (we point out that our definition of an outlier is purely statistical, and therefore an outlier corresponds either to a biological effect—local amplicon or deletion—or to an experimental artefact). To overcome this limitation in the detection of outliers, we have designed a special procedure based on median-absolute-deviation (MAD) for detecting the remaining outliers. It should be noted that when an outlier presents a large deviation, it is detected at the breakpoint detection step. This first type is called AWS outlier and is characterized by a location X_i such that $\hat{\theta}_{i-1} \in [\hat{\theta}_{i+1} - \epsilon; \hat{\theta}_{i+1} + \epsilon]$ and $\hat{\theta}_i \notin [\hat{\theta}_{i-1} - \epsilon; \hat{\theta}_{i-1} + \epsilon]$ (N.B. a special treatment is applied for starting location and ending location: if $\hat{\theta}_1 \notin [\hat{\theta}_2 - \epsilon; \hat{\theta}_2 + \epsilon]$ (respectively $\hat{\theta}_{N-1} \notin [\hat{\theta}_N - \epsilon; \hat{\theta}_N + \epsilon]$) then X_1 (respectively X_N) is considered as well as an outlier). The second type of outlier is called MAD outlier and such outliers are identified as follows: for each MSHR, we remove all the AWS outliers; based on the assumption that the observations Y_i in an MSHR are drawn from the normal distribution $\mathcal{N}(\mu_k, \sigma_k^2)$, a location X_i for which the observation Y_i lies in the $\alpha/2$-quantile upper or lower tail of the normal distribution $\mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$ is considered to be a MAD outlier (α has been set to 0.001). As we are looking for outliers, $\hat{\mu}_k$ is estimated by the median and $\hat{\sigma}_k^2$ by the square of the median-absolute-deviation for robustness considerations.

Optimization of the number of breakpoints

Our data show that despite the use of a strong statistical penalty λ, the AWS procedure may in some cases identify breakpoints which correspond to small shifts and define regions of ~10–20 BACs. This is probably due to specific local effects on the chromosome, unrelated to the biological variation we want to investigate but nevertheless real. Thus, a filtering step was added to remove these undesirable breakpoints. Before this step, all the outliers are excluded from the analysis. The likelihood of our data can be written as:

$$L = \prod_{i=1}^{B_1} \frac{1}{\sigma_1\sqrt{2\pi}} \; \mathrm{e}^{-\frac{1}{2}\left(\frac{Y_i-\mu_1}{\sigma_1}\right)^2} \cdots \prod_{i=B_{N'+1}}^{N} \frac{1}{\sigma_{N'+1}\sqrt{2\pi}} \; \mathrm{e}^{-\frac{1}{2}\left(\frac{Y_i-\mu_{N'+1}}{\sigma_{N'+1}}\right)^2}.$$

We calculate the following function:

$$f = \sum_{k=1}^{N'+1} (B_k - B_{k-1}) \log\left(\hat{\sigma}_k^2\right) + \lambda' \sum_{k=1}^{N'} K(\hat{\sigma}^{-1}|\hat{\mu}_k - \hat{\mu}_{k+1}|) \log(N)$$

with $B_0 = 0$, $B_{N'+1} = N$, $\hat{\sigma}_k^2$ and $\hat{\mu}_k$ are the usual MLE of σ_k^2 and μ_k, and $\hat{\sigma}$ is calculated from Equation (1). The function f corresponds, up to an additive constant, to a penalized form of $-\log L$. The function $K(x)$ is the tricubic kernel function and takes the value $[1-(x/6)^3]^3$ for $x \in [0;6]$ and zero elsewhere. A kernel function in the penalty term is chosen mainly to prevent the removal of true breakpoints defining a MSHR of very small cardinality. The algorithm is then very similar to the JOIN procedure of the GLSo algorithm proposed by Jong *et al*. (2003): the breakpoint for which removal leads to the largest decrease in the function f is eliminated and the procedure is iterated until the function f ceases to decrease. When a breakpoint is removed, a new larger MSHR appears and its MAD outliers are re-evaluated.

REGION ASSIGNMENT

The purpose of the region assignment is to assign a gain, loss or normal status to each MSHR. Our algorithm involves three steps:

- First, for each chromosome, MSHR are grouped in classes, each class containing MSHRs of the same expected (but unknown) DNA copy number.
- Second, the resulting classes for all chromosomes are clustered to produce superclasses, of same expected DNA copy number; these superclasses are called homogeneous chromosomal status regions (HCSR).
- Finally, each HCSR is given a label: gain, normal or loss. An evaluation of the ratios is computed and corresponds to different levels of gain or loss.

This two-step clustering (chromosome, then genome level) ensures that label assignments are consistent for all clusters within a chromosome. This refinement is necessary since the signal measured on the array may be chromosome-dependent.

MSHR clustering by chromosome

The aim of this step is to cluster the MSHR identified on a chromosome such that each cluster corresponds to a set of MSHR with identical statuses. In practice, we do not know a priori the number of clusters for a given chromosome, and we therefore propose criteria for determining the most appropriate number of clusters. We do this as follows: first, we eliminate all the outliers detected previously; then, we calculate the mean and cardinality of each MSHR; finally, we perform hierarchical clustering of the means of MSHRs with centroid criteria,

taking into account the cardinality of each MSHR. From the dendrogram produced, we then try to find the optimal number of clusters for chromosomes with more than one breakpoint (if there is only one breakpoint then the chromosome has two clusters). We successively cut the dendrogram to obtain sets S_i of clusters $C_1^i, \ldots, C_i^i$ with $i = 2, 3, \ldots, N^*_{\max}$ ($N^*_{\max}$ is less than or equal to the number of MSHR). We now use all the points belonging to each cluster (except outliers) to calculate the likelihood as follows:

$$L_i = \prod_{j \in C_1^i} \frac{1}{\sigma_1 \sqrt{2\pi}} \; \mathrm{e}^{-\frac{1}{2}\left(\frac{Y_j - \mu_1}{\sigma_1}\right)^2} \ldots$$

$$\prod_{j \in C_i^i} \frac{1}{\sigma_i \sqrt{2\pi}} \; \mathrm{e}^{-\frac{1}{2}\left(\frac{Y_j - \mu_i}{\sigma_i}\right)^2}.$$

We calculate the following function:

$$f_i^* = \sum_{k=1}^{i} \#C_k^i \log\left(\hat{\sigma}_k^2\right) + \lambda^* \sum_{k=1}^{i-1} K\left(\hat{\sigma}^{-1} |\hat{\mu}_k - \hat{\mu}_{k+1}|\right) \log(N),$$

where $\hat{\sigma}_k^2$ and $\hat{\mu}_k$ are the usual MLE of σ_k^2 and μ_k, $\hat{\sigma}$ is calculated from Equation (1) and $\#C_k^i$ is the cardinality of the cluster (N.B. the clusters are sorted in increasing order of means). This function corresponds, up to an additive constant, to a penalized form of $-\log L_i$. The optimal number of clusters is $i^* = \mathrm{arginf}_i f_i^*$. The clusters identified correspond to HCSRs. The value of λ^* has been set to 8.

HCSR clustering throughout the genome

The preceding step provides us with a set of HCSRs for each chromosome. At this stage of the analysis, we now consider globally the HCSRs of the whole genome. Based on the same principle, we cluster HCSRs according to their means, using the centroid agglomeration method and taking into account the cardinality of each HCSR. We retain the number of clusters for which the new function f_i^* is minimal. In this case, the minimal value of i is 1 (and for this value of i, the f_i^* function is calculated without the penalty term). The estimate $\hat{\sigma}$ is calculated from the data for the whole genome. For our analysis, λ^* has been set to 40.

Label assignment

We now have to decide which regions are normal, and which have been lost or gained. In array CGH experiments, as in standard microarray experiments, a bias results from differences in dye incorporation efficiency such that, even for normal/normal hybridization, the expected $\log_2$-ratios are not centered around zero. Thus, $\log_2$-ratiosare median-centered before identification of the normal DNA regions. Once clustering has been achieved for the whole genome, the cluster with the median closest to zero is considered to be normal DNA. Clusters with higher medians are considered to reflect gains and those with lower values are considered to reflect losses.

VALIDATION

Validation on simulated data

We simulated 210 genomic profiles of three types: normal profiles, profiles displaying moderate rearrangement and profiles displaying high levels of rearrangement. For each profile, we generated a series of 2457 points drawn from a normal distribution with a mean of zero and an SD of 0.079, evaluated from 12 normal/normal hybridization arrays. For moderate- and high-rearrangement profiles, a status (loss, normal or gain) was defined according to a three-state first-order Markov process with a probability transition matrix:

$$\begin{pmatrix} 0.99 & 0.008 & 0.002 \\ 0.0005 & 0.999 & 0.0005 \\ 0.002 & 0.008 & 0.99 \end{pmatrix}$$

and

$$\begin{pmatrix} 0.995 & 0.004 & 0.001 \\ 0.0025 & 0.995 & 0.0025 \\ 0.001 & 0.004 & 0.995 \end{pmatrix}$$

respectively. We added realistic values of 0.3 for gain status and -0.3 for loss status to the profile generated. We also used a Poisson process to add outliers such that the expected number of outliers in the series was 20. A value of 0.3 was either added to the value or subtracted, with a probability of 0.5. The global performance of our methodology was assessed according to the following criteria: (#correctly labeled BACs + #true positive outliers)/total number of BACs.

For both values of λ', this criterion ranges from 98.94 to 99.84%. For a total of 1195 breakpoints, 81.9% were correctly located and 15.1% were incorrectly located, with a maximum localization error of 3 BACs (cf. Table 1) for $\lambda' = 8$. For $\lambda' = 10$, no improvement was observed because the decrease in false positive rate did not counteract the increase in false negative rate. A total of 278 and 283 breakpoints were removed for λ' values of 8 and 10, respectively. We found that 66.2% of the outliers were correctly identified (cf. Table 1). The large number of false negatives may be accounted for by these points being picked up in a distribution with only a small shift (± 0.3) with respect to their neighborhood.

We have estimated the resolution of our method by simulating a chromosomal profile of 200 BACs. In the middle position, an alteration of length 1, 2, 4 or 8 has been added with a signal mean amplitude of 0.15, 0.20, 0.25 or 0.30 and a gaussian distribution. The SD is 0.079, as measured on our real data. For each combination of length and signal, 1000 simulations have been done (*nb* the HCSRclustering step has been ignored since we are working on only one chromosome). The resolution is estimated both by the percentage of correctly assigned BACs in the altered region and the number of times

Table 1. The results for the detection of breakpoints and outliers on 210 simulated genomic profiles for two values of λ'

	$\lambda' = 8$	$\lambda' = 10$
Total number of breakpoints	1195	1195
Number of breakpoints correctly identified	979	978
Number of breakpoints mislocated	181	178
Number of missed breakpoints	35	39
Number of additional breakpoints	26	25
Difference in position for mislocated breakpoints		
1	167	164
2	13	13
3	1	1
Outlier detection		
True positives	2679	2678
False negatives	1364	1365
False positives	1243	1249

We obtained 98.9–99.8% correct assignations, see text for details.

Table 2. Resolution of the method estimated on a chromosomal profile of 200 BACs depending on the length of the altered region and the signal amplitude

Signal	Length of altered region			
	1	2	4	8
Percentage of correctly labelled BACs				
0.15	$10 \pm .95$	$9 \pm .66$	$9 \pm .47$	$14 \pm .85$
0.20	23 ± 1.33	$21 \pm .92$	$26 \pm .92$	56 ± 1.33
0.25	48 ± 1.58	45 ± 1.14	56 ± 1.11	$90 \pm .66$
0.30	67 ± 1.49	69 ± 1.04	$81 \pm .85$	$97 \pm .19$
Percentage of altered regions				
0.15	$10 \pm .95$	17 ± 1.17	28 ± 1.42	43 ± 1.55
0.20	23 ± 1.33	38 ± 1.52	60 ± 1.55	81 ± 1.23
0.25	48 ± 1.58	70 ± 1.45	88 ± 1.04	$98 \pm .41$
0.30	67 ± 1.49	$90 \pm .95$	$98 \pm .47$	100 ± 0

The performance (mean $\pm$ SD) are estimated by the percentage of correctly assigned BACs in the altered region and the number of times that at least an alteration has been found in this region. SD on signal ratios was estimated on real data and set to 0.079.

Table 3. The results for breakpoint detection and label assignment on 15 human cell strains (Snijders' dataset)

Cell strain/chromosome	$\lambda' = 8$	$\lambda' = 10$
GM00143/False	8	0
GM01524/6	Yes	Yes
GM01524/False	0	0
GM01535/5	Yes	Yes
GM01535/12	Yes	Yes
GM01535/False	0	0
GM01750/9	Yes	Yes
GM01750/14	Yes	Yes
GM01750/False	0	0
GM02948/False	1	0
GM03134/8	Yes	Yes
GM03134/False	4	4
GM03563/3	Yes	Yes
GM03563/9	Yes	Yes
GM03563/False	8	4
GM03576/False	0	0
GM04435/False	2	2
GM05296/10	Yes	Yes
GM05296/11	Yes	Yes
GM05296/False	8	6
GM07081/7	Yes	Yes
GM07081/15	No	No
GM07081/False	6	6
GM07408/False	2	2
GM10315/False	3	0
GM13031/17	Yes	Yes
GM13031/False	4	4
GM13330/1	Yes	Yes
GM13330/4	Yes	Yes
GM13330/False	0	0

Following the / after the cell strain name is the number of the chromosome on which a breakpoint is present or 'False', indicating the number of false-positive breakpoints identified by the procedure in each cell strain. Yes means that breakpoints have been correctly located for the chromosome under consideration. All breakpoints were detected and all label assignments are correct except for GM07081/15 (not detected by the array CGH technology) and BAC RP11-237j07 of GM05296. In this last case, the breakpoint was located on the neighboring BAC.

that at least an alteration has been found in this region. The results are presented in Table 2 and show that a signal less or equal to 0.2 give low performance unless the length of the region is greater than 8 BACs and a signal greater or equal to 0.25 give good performance. Note that the results of our simulations depends only on the signal-to-noise ratio of the data, that should be kept higher than approximately 2.5 to avoid deterioration of performances.

Validation on the dataset from Snijders *et al*. (2001)

We present here the results obtained with our methodology applied to a public dataset (Snijders *et al*., 2001). The data correspond to 15 human cell strains with known karyotypes (12 fibroblast cell strains, 2 chorionic villus cell strains and 1 lymploblast cell strain) from the NIGMS Human Genetics Cell Repository (http://locus.umdnj.edu/nigms). Each cell strain has been hybridized with an array CGH of 2276 BACs, spotted in triplicate. The variable used for the analysis is the test over reference $\log_2$-ratio, as described by the authors. This dataset had already been analyzed with another algorithm; the results obtained are presented in Olshen and Vankatraman (2002).

Our results for breakpoint detection and label assignment are shown in Table 3 for two values of λ'. Our algorithm gave perfect detection of breakpoints: none was missed in the nine cell strains that had breakpoints. For strain GM05296, the first breakpoint of chromosome 10 was detected on BAC RP11-14i14 instead of RP11-237j07, which immediately follows it: visual checking showed that the conclusion in favor of BAC RP11-237j07 was far from clear. The number of false-positive breakpoints decreases dramatically if the

value of λ' is increased from 8 to 10. However, for some cell strains, false-positive breakpoints remain (especially for GM00143 and GM03563): such false-positive breakpoints may result from local trends on the chromosome (a BAC effect or a drift along the genome can be observed, even for normal/normal hybridizations). Similar false-positive breakpoints were reported by Olshen and Vankatraman (2002) for the cell strain GM03563, on chromosome 11. All label assignments were correct, except for the monosomic region on chromosome 15 of GM07081, which was not detected by array CGH technology (Snijders *et al*., 2001). For cell strains GM04435, GM07081 and GM07408, our algorithm identified a small monosomic region (although karyotyping did not show this region to be monosomic) of two BACs on chromosome 8 (RP11-122N11 and RP11-287P18), corresponding to the region identified in strain GM03134. If we compare our results with those obtained by Olshen and Vankatraman (2002), our algorithm gave fewer false-positive breakpoints. For cell strain GM03134, our algorithm identified the small monosomic region on chromosome 8 whereas Olshen and Vankatraman (2002) did not identify this region. For cell strain GM01535, Olshen and Vankatraman (2002) did not find the monosomic region consisting of a single BAC located at the end of chromosome 12, whereas this BAC was detected as an AWS Outlier by our algorithm.

Validation on bladder cancer data

We have applied our algorithm to bladder cancer data from tumors collected at Henri Mondor Hospital (Créteil, France) (Billerey *et al*., 2001) and hybridized on arrays CGH composed of 2464 BACs (F. Radvanyi, D. Pinkel *et al*., unpublished data). The data consist of 13 arrays CGH experiments (using DNA from 13 different bladder tumors with the following stages-grades: 1 T1G2, 1 T1aG3, 1 T2G2, 2 T3G3 and 8 T4G3) hybridized according to Pinkel's protocol (Pinkel *et al*., 1998) (Table 4). Images were analyzed with SPOT 2.0 software (Jain *et al*., 2002). A pre-processing step was used to remove poor-quality spots. Spots with a reference signal intensity (and DAPI signal intensity) below 125% of the background reference signal (DAPI signal) were discarded. Triplicates with an SD of $\log_2$-ratio >0.1 were removed from the analysis and spots located in areas of spatial bias (unpublished data) were also eliminated. The value used is the mean for each BAC of the Log2Rat variable calculated by SPOT 2.0, which corresponds to the test over reference $\log_2$-ratio (as each BAC was spotted three times on the array CGH). For our data, the karyotype is unknown. Thus, we mainly focused on breakpoint detection validation on the basis of visual expertize. Nevertheless, supporting evidence for the location of breakpoints was provided by LOH analysis.

Based on visual expertize, AWS smoothing gave an excellent fit to the CGH profile (cf. Figs 1 and 2) and this algorithm seems highly appropriate for array CGH analysis. Despite the small number of errors observed, the optimization procedure for incorrect breakpoint removal is necessary to remove false positives. A total of 108 and 116 breakpoints were removed (for $\lambda' = 8$ and 10, respectively), even though some were of biological interest. For four tumors, label assignment was highly problematic, even from visual expertize. These tumors corresponded to high-stage and high-grade tumors (1 T2G2 and 3 T4G3) with many genome rearrangements. Indeed, signal variation at breakpoint may be blurred by several biological limitations of the technology: tumor biopsy samples generally contain a mixture of normal and tumor cells, and cells within a tumor may display differences in genomic losses and gains, a phenomenon known as tumor heterogeneity. Moreover, aneuploidy may affect several chromosomes differently. These limitations make breakpoint detection and label assignment difficult. For the other nine tumors, label assignment was consistent with visual expertize.

Table 4. The results for the detection of breakpoints and outliers on 13 bladder tumor genomic profiles for two values of λ'

	$\lambda' = 8$	$\lambda' = 10$
Total number of breakpoints	267	267
Number of breakpoints correctly identified	251	245
Number of breakpoints mislocated	7	7
Number of missed breakpoints	9	15
Number of additional breakpoints	9	8
Difference in position for mislocated breakpoints:		
1	6	6
2	1	1

Performances are similar to those of a human expert.

These 13 bladder tumors had been assessed for LOH on chromosome 10, using polymorphic markers (Cappellen *et al*., 1997). Although CGH and LOH studies do not provide the same information (Albertson *et al*., 2003), the results of the two studies were consistent: the regions of gains and losses detected by array CGH correspond to regions of allelic imbalance detected with polymorphic markers. For example, Cappellen *et al*. (1997) found an allelic imbalance for polymorphic markers between D10S185 and D10S168 on chromosome 10 of tumor 1533e: these markers are located between BACs RP11-9M11 (Position 1402) and RP11-32I9 (Position 1431), which delineate the lost region detected by array CGH for the same tumor (cf. Fig. 2).

A region of amplification including the *CCND1* (cyclin D1) gene was detected on the long arm of chromosome 11 for tumor 1533e (cf. Fig. 2). Interestingly, the breakpoints defining this previously identified region on chromosome 11 of tumor 1533e were also detected in the peritumoral urothelium of the patient concerned, although the mean $\log_2$-ratio of this region was only 0.25 (data not shown), demonstrating the sensitivity of our method.

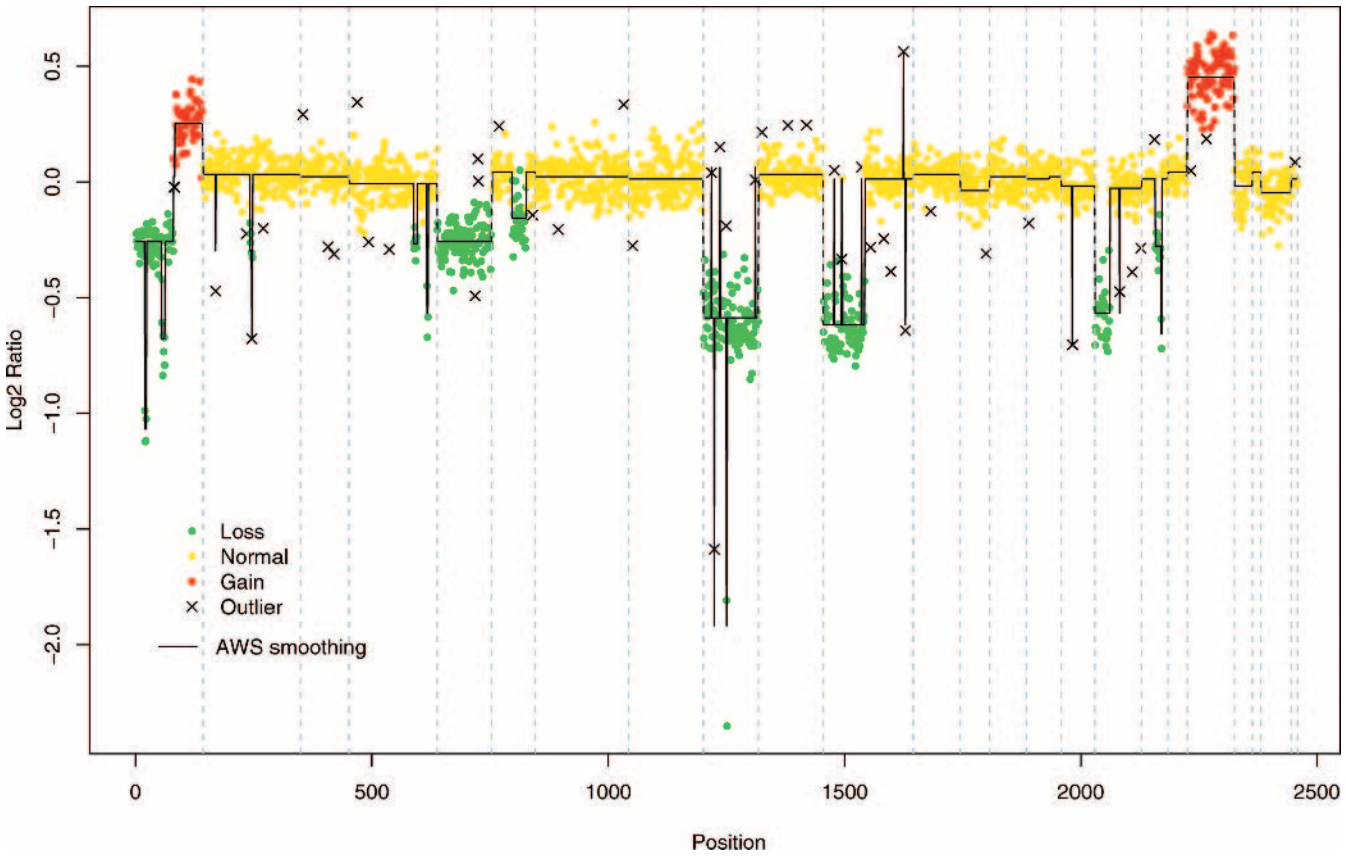


Fig. 1. Genomic profile of bladder tumor 824 (T1G2) according to our methodology: the BREAKPOINT DETECTION step makes it possible to calculate the piecewise constant function, in black, and to detect outliers; during the REGION ASSIGNMENT step, a two-step clustering process groups together regions of same status and then assigns a label (gain, normal or loss) to each region. The vertical gray dashed lines indicate the separation between chromosomes. The horizontal axis shows the rank position of each BAC along the genome and the vertical axis shows the tumor/normal $\log_2$-ratios after median centering.

DISCUSSION

We present here a new methodology for breakpoint detection and status assignment to each BAC in a array CGH experiment. Our algorithm is highly efficient with both simulated and manually analyzed data. For real data, our results are similar to those obtained by a human expert. On a public dataset, our algorithm outperformed the method described by Olshen and Vankatraman (2002). Simulated data are also correctly analyzed by our method: in most cases, missed breakpoints or outliers were not detected properly simply because the randomization procedure gave them signal ratios far from the expected ratios of their class of origin. In such cases, the available information is insufficient for the correct detection of breakpoints or outliers from the data, whatever the algorithm used.

The AWS procedure correctly detects large regions but also accurately fits data for both fine structures and small local effects. Local effects have been already reported by Olshen and Vankatraman (2002), consisting of regions of the genome showing a recurrent bias in the signal ratios confirmed on normal/normal hybridization (N/N) (data not shown): in our data, the strongest local effects were observed on chromosome 6 and chromosome 13. Both involved a shifting down of the tumor signal with respect to the normal signal. With homogeneous biological samples (e.g. cell lines), a local effect generally induces a much smaller shift than a gain or a loss of DNA. However, tumor biopsy samples are generally a mixture of normal and tumor cells and thus, heterogeneity reduces shifts, making it difficult to distinguish a biological effect from a local effect, leading to the identification of false-positive breakpoints. This suggests that an *ad hoc* procedure should be developed to eliminate such biases.

From our experience with normal/normal hybridizations, this local effect and other sources of variation, such as a BAC effect, appear to be array-dependent, rather than systematic. One solution is to flag the regions or BACs subject to such biases and to consider them with caution. More generally, this problem raises the question of array CGH data normalization and shows that breakpoint detection and label assignment are closely linked to the normalization step. Our

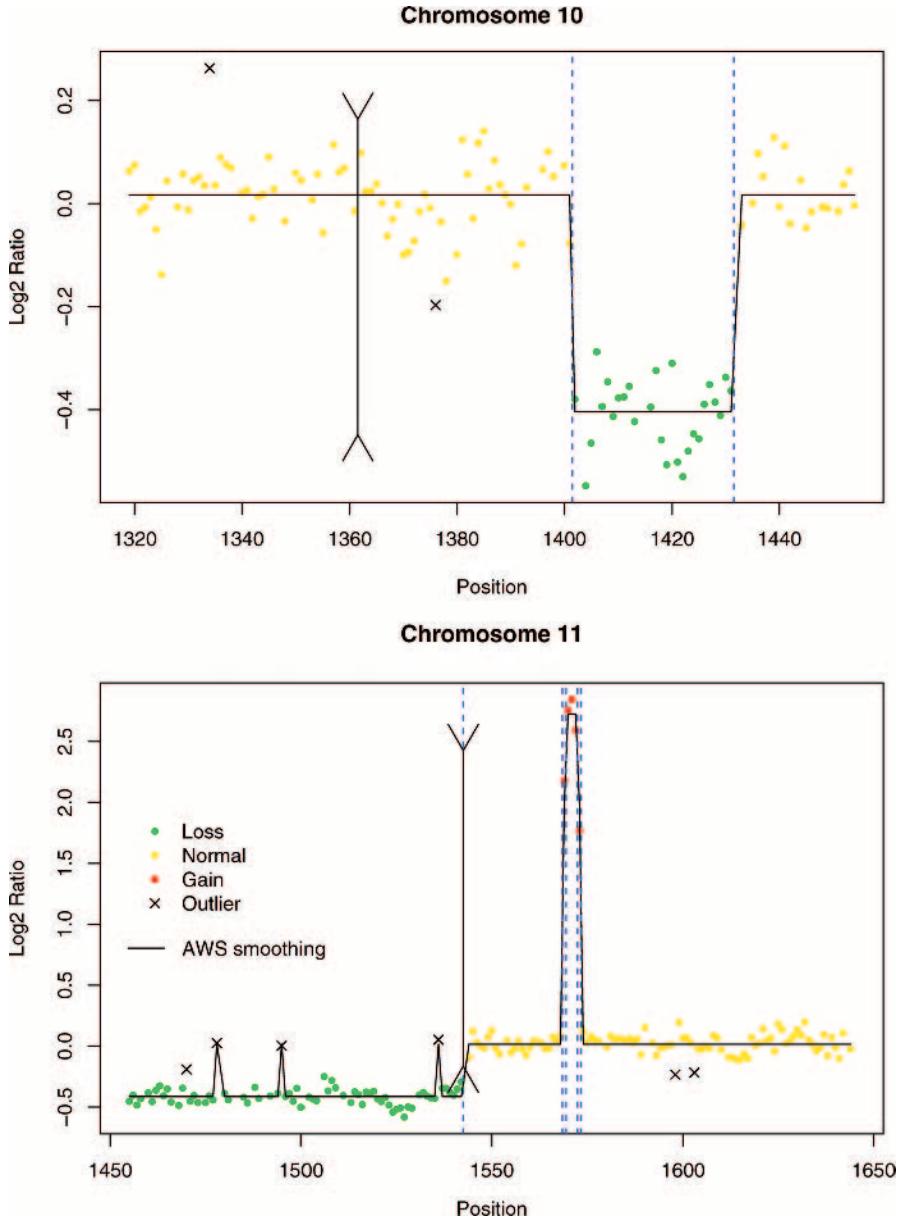


Fig. 2. Profiles for chromosomes 10 and 11 for the bladder tumor 1533e (T4G3). The vertical red dashed lines indicate the breakpoints and the vertical black double arrow indicates the centromere. The horizontal axis indicates the rank position of each BAC along the genome and the vertical axis indicates the $\log_2$-ratios after median centering.

findings also show that normalization should be carried out with an adaptive (array-dependent) algorithm. In this study, we simply applied a filter based on spot quality control criteria and removed abnormally high $\log_2$-ratios measured in some areas of the array, referred to as spatial biases (generally an edge or corner effect). Further improvements to normalization are envisaged and will form the subject of another publication. The biological significance of the outliers detected must be considered carefully for several reasons: first, natural polymorphisms may result in outliers, as shown in some cases on normal/normal hybridizations. These particular clones must therefore be flagged (such polymorphisms have been observed in our data). Second, some BACs may systematically display aberrant behavior. Finally, some BACs may have been mislocated on the genome: between two consecutive versions of the draft sequence, some BACs may be transferred from one position on a chromosome to another.

When using our algorithm, several parameters must be set: the main parameters are the statistical penalty λ for the AWS procedure, the λ' value for optimization of the number of breakpoints and λ^* in the two-step clustering step. We have set these values empirically based on our own data, but when

applying our method to arrays CGH obtained on another platform, it may be necessary to modify these parameters and a model selection step may be required (array replicates and normal/normal arrays are particularly useful at this stage).

Although breakpoint and outlier detection are entirely satisfactory with our method, label assignment is much more difficult. Several phenomena make it difficult to classify regions correctly into three classes (loss, normal and gain), not to mention to assign a number of DNA copies to a region. We have already raised the problem of sample heterogeneity. In cases of polyploidy, a single loss results in mathematically smaller shifts. In situations in which label assignment is problematic, the use of other sources of biological knowledge, such as genotyping, is likely to improve performance.

Although our methodology requires further improvement, it already provides new materials for the large-scale analysis of array CGH profiles and makes it possible to envisage further analysis. Indeed, the segmentation of CGH profiles and the assignment of statuses to BACs are required for more advanced transverse analysis in sets of patients: detection of regions recurrently lost or gained, unsupervised and supervised classification based on the CGH profile, integration of the genome and transcriptome profiles for the identification of new genes involved in tumorigenesis and/or tumor progression. This work should lead to new insight valuable for clinical research and cancer treatment. Our work was driven by and applied to cancer array CGH analysis but can also be applied to any genetic disease involving deletion or amplification in genomic DNA.

ACKNOWLEDGEMENTS

This work was supported by the Centre National de la Recherche Scientifique, the Institut Curie, the Comité de Paris Ligue Nationale contre le Cancer (Laboratoire Associé) and the IST program from the European Commission through the HKIS project (IST-2001-38153). Data processing was managed by the ™ Amadea software from ISoft (Gif sur Yvette, France).

REFERENCES

Albertson,D.G., Collins,C., McCormick,F. and Gray,J.W. (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34**, 369–376.

Albertson,D.G. and Pinkel,D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.*, **12**, R145–R152.

Axon Instruments (2003) *GenePix Pro 5.0 User's Guide*. ©Axon Instruments, Inc.

Billerey,C., Chopin,D., Aubriot-Lorton,M.H., Ricol,D., de Medina,S.G.D., Rhijn,B.V., Bralet,M.P., Lefrere-Belda,M.A., Lahaye,J.B., Abbou,C.C., *et al*. (2001) Frequent FGFR3 mutations in papillary non-invasive bladder (pTa) tumors. *Am. J. Pathol.*, **158**, 955–1959.

Cappellen,D., Gil Diez de Medina,S., Chopin,D., Thiery,J.P. and Radvanyi,F. (1997) Frequent loss of heterozygosity on chromosome 10q in muscle-invasive transitional cell carcinomas of the bladder. *Oncogene*, **14**, 3059–3066.

Ishkanian,A.S., Malloff,C.A., Watson,S.K., DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.*, **36**, 299–303.

Jain,A.N., Tokuyasu,T.A., Snijders,A.M., Segraves,R., Albertson,D.G., and Pinkel,D. (2002) Fully automatic quantification of microarray image data. *Genome Res.*, **12**, 325–332.

Jong,K., Marchiori,E., van der Vaart,A., Ylstra,B., Weiss,M. and Meijer,G. (2003). Chromosomal breakpoint detection in human cancer. In Raidl,G.R., Cagnoni,S., Cardalda,J.J.R., Corne,D.W., Gottlieb,J., Guillot,A., Hart,E., Johnson,C.G., Marchiori,E., Meyer,J.-A. and Middendorf,M. (eds), *Applications of Evolutionary Computing, EvoWorkshops2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, EvoSTIM*, Volume 2611 of *LNCS*, University of Essex, England, UK. Springer-Verlag, Berlin, pp. 54–65.

Olshen,A.B. and Vankatraman,E.S. (2002) Change-point analysis or array-based comparative genomic hybridization data. *Proceedings of the Joint Statistical Meetings*, New York, August 11–15, 2530–2535.

Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Polzehl,J. and Spokoiny,S. (2000) Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc., Ser. B*, **62**(2), 335–354.

Polzehl,J. and Spokoiny,S. (2002) Local likelihood modelling by adaptive weights smoothing. WIAS-Preprint 787.

Sen,A. and Srivastava,M.S. (1975) On tests for detecting a change in mean. *Ann. Stat.*, **3**, 98–108.

Shaw-Smith,C., Redon,R., Rickman,L., Rio,M., Willatt,L., Fiegler,H., Firth,H., Sanlaville,D., Winter,R., Colleaux,L., Bobrow,M. and Carter,N.P. (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.*, **41**, 241–248.

Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, **29**, 263–264.

Solinas-Toldo,S., Lampel,S., Stilgenbauer,S., Nickolenko,J., Benner,A., Dohner,H., Cremer,T., and Lichter,P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Vogelstein,B., Fearon,E.R., Kern,S.E., Hamilton,S.R., Preisinger,A.C., Nakamura,Y. and White,R. (1989) Allelotype of colorectal carcinomas. *Science*, **244**, 207–211.

Yang,Y., Dudoit,S. and Speed,T. (2001) Normalization for cDNA microarray data. *SPIE BiOS 2001*, San Jose, CA, January 2001.

2.3 Iterative approach for normalisation and identification of DNA copy number alterations

We have seen in **Subsection 1.4.2** that besides BAC aCGH new platforms using oligonucleotide aCGH were also available. They provide a better resolution to detect smaller alterations. Among the different platforms available, Affymetrix® Genome-Wide Human SNP Array is widely used. This kind of microarray has its specific source of variability and the MANOR method (Neuvial et al., 2006) which has been developed for BAC aCGH was not suitable for this technology. Therefore, we have developed a normalisation method devoted to the analysis of Affymetrix® Genome-Wide Human SNP Array. Besides normalisation, the proposed method has the originality to perform the identification of the DNA copy number alterations using the GLAD algorithm (Hupé et al., 2004). The algorithm alternatively identifies the DNA copy number alterations and normalises the data. Those two alternative steps are iterated to improve the signal-to-noise ratio of the data at each iteration. The normalisation step takes into account the information of the genome alterations to better estimate the sources of variability to correct during the normalisation step; this strategy was also indicated by Staaf et al. (2007). The method we have developed is named ITerative and Alternative normaLIsation and Copy number calling for affymetrix Snp arrays (ITALICS) (Rigaill et al., 2008) and the paper describing the algorithm is supplied as a material part of the thesis.

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 6 2008, pages 768–774
doi:10.1093/bioinformatics/btn048

Genome analysis

ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays

Guillem Rigaill[1,2,5,†], Philippe Hupé[1,2,3,5,*,†], Anna Almeida[4], Philippe La Rosa[1,2,5], Jean-Philippe Meyniel[4], Charles Decraene[3,4] and Emmanuel Barillot[1,2,5]

[1]Institut Curie, Service de Bioinformatique, [2]INSERM, U900, [3]CNRS UMR144, [4]Institut Curie, Translational Research Department, 26 rue d'Ulm, Paris F-75248 and [5]Ecole des Mines de Paris, ParisTech, Fontainebleau, F-77300 France

Received on August 21, 2007; revised and accepted on January 29, 2008
Advance Access publication February 5, 2008
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Affymetrix SNP arrays can be used to determine the DNA copy number measurement of 11 000–500 000 SNPs along the genome. Their high density facilitates the precise localization of genomic alterations and makes them a powerful tool for studies of cancers and copy number polymorphism. Like other microarray technologies it is influenced by non-relevant sources of variation, requiring correction. Moreover, the amplitude of variation induced by non-relevant effects is similar or greater than the biologically relevant effect (i.e. true copy number), making it difficult to estimate non-relevant effects accurately without including the biologically relevant effect.

Results: We addressed this problem by developing ITALICS, a normalization method that estimates both biological and non-relevant effects in an alternate, iterative manner, accurately eliminating irrelevant effects. We compared our normalization method with other existing and available methods, and found that ITALICS outperformed these methods for several in-house datasets and one public dataset. These results were validated biologically by quantitative PCR.

Availability: The R package ITALICS (ITerative and Alternative normaLIzation and Copy number calling for affymetrix Snp arrays) has been submitted to Bioconductor.

Contact: italics@curie.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The development of high-throughput technologies, and of microarrays in particular, has made it possible to analyze DNA copy number throughout the entire genome, with ever-increasing resolution. Various techniques for detecting DNA copy number alterations are available (for a review, see Ylstra *et al.*, 2006). Affymetrix SNP arrays, such as the Affymetrix GeneChip Human Mapping 100K Set (Kennedy *et al.*, 2003), seem to be one of the most widely used tools. These chips can be used for simultaneous genotyping and copy number determination for single nucleotide polymorphism (SNP), at high resolution. This technology has various uses, including studies of copy number variations in populations and the identification of genomic alterations in developmental genetics or cancer (for a review, see Pinkel and Albertson, 2005). In cancer studies, Affymetrix SNP arrays provide new insight into the mechanisms of tumor progression; they can be used to pinpoint new candidate genes for tumor-suppressor genes (Liu *et al.*, 2007) and oncogenes (thought to be present in loss and gain regions, respectively), and to classify tumors, improving diagnosis for new patients and the evaluation of prognosis.

Like all microarrays, Affymetrix SNP arrays are affected by systematic non-relevant sources of experimental variation. For accurate extraction of the biologically relevant effect (i.e. the true DNA copy number of each SNP in the genome, corresponding to the biological signal), the raw data must be corrected, taking these different effects into account. We present here a normalization algorithm for this purpose, which can be used for the simultaneous correction of different sources of experimental variation and biological signal estimation when trying to infer DNA copy number.

Several methods have already been developed for correcting non-relevant sources of variation. These methods include CNAG (Nannya *et al.*, 2005), GIM (Komura *et al.*, 2006) and CARAT (Huang *et al.*, 2006). However, none of these methods take into account that the range of variation due to the non-relevant effects is similar or higher than the biologically relevant effect. Therefore, the impacts of the biologically relevant effect and non-relevant effects may easily be confused. Correct estimation of the non-relevant effects also depends on the correct estimation of copy number. We therefore propose an alternative, iterative method for estimating the biologically relevant effect and non-relevant effects, to improve biological signal estimation. We will begin by briefly presenting Affymetrix SNP arrays. We will then describe our algorithm (*ITerative and Alternative normaLIzation and Copy number calling for affymetrix Snp arrays*: ITALICS) for data normalization in detail. We then discuss the results obtained with this algorithm, comparing them with those obtained with other algorithms. Finally, we discuss the advantages of ITALICS and possible improvements to this method.

*To whom correspondence should be addressed.
†The authors wish to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

 © The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

2 MATERIALS AND METHODS

2.1 Affymetrix SNP arrays

Technology: Affymetrix SNP arrays can be used to detect DNA copy number alterations at a resolution of 6–210 kb, using around 11 000–500 000 human SNPs. The Affymetrix GeneChip Human Mapping 100K and 500K Sets are comprised of two arrays. Each array is based on specific restriction enzymes: *Xba*I and *Hind*III for the 100K set and *Sty*I and *Nsp*I for the 500K set. The Affymetrix 50K *Xba*I and *Hind*III arrays contain no common SNPs and their combination provides the DNA copy numbers of more than 115 000 SNPs.

Each allele of each SNP is represented by n_i perfect match (PM) probes and n_i mismatch (MM) probes. Reverse or forward probes may be used and these probes may be centered on the SNP position or offset by −4 to +4 base pairs. Thus, all the PM probes of an SNP allele have different DNA sequences. Probes are grouped into probe quartets of four probes: one PM and one MM probe for each of alleles A and B. All four probes have the same orientation and offset.

The Affymetrix SNP arrays assay is carried out as follows. Genomic DNA is digested with a restriction endonuclease. Adaptors are ligated to all fragments. These fragments are amplified by PCR and then fragmented, labeled with biotin and hybridized with the chip. The chip is then washed and scanned to generate the cell intensity file (.CEL) which is used as input to the proposed algorithm.

Hereafter, the raw signal $Y_{i.}$ of a given SNP i is given by:

$$Y_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \text{ with } Y_{ij} = Y_{ij}^A + Y_{ij}^B$$

where Y_{ij}^A and Y_{ij}^B are the log-intensity of the PM probe A and B of the j-th probe quartet for the SNP i, and Y_{ij} is the sum of PM log-intensities for the j-th quartet. $Y_{i.}$ is the mean PM log-intensity of the n_i quartets for the SNP i. MM probes are not taken into account in our algorithm. The two PM probes defining the entity Y_{ij} are referred subsequently as $Quartet_{PM}$, the subscript i is referred to as SNP i, and the subscript j as one of the n_i quartets.

Non-relevant sources of variation: ITALICS deals with known systematic sources of variation, such as the GC-content of the $Quartets_{PM}$ (QGC_{ij}), the length of the PCR-amplified fragment (FL_i) and the GC-content of the fragment amplified by PCR (FGC_i) (Nannya *et al.*, 2005; Komura *et al.*, 2006). It also takes into account the $Quartet_{PM}$ effect (Q_{ij}), resulting from the systematically low intensity of some $Quartets_{PM}$ and the systematically high intensity of others.

We also found that some Affymetrix SNP arrays suffer from spatial artifacts, as reported by Neuvial *et al.* (2006) for CGH array data. A spatial artifact is illustrated in Figure 1A: neighboring $Quartets_{PM}$ on the chip present abnormal intensities. The corresponding SNPs which appear as outliers in the genomic profile, as shown in Figure 1C, D and E, and should be removed. We have addressed this issue using a filtering criterion, making it possible to discard bad probes, as described subsequently.

2.2 The ITALICS algorithm

Overview: In Affymetrix SNP arrays, non-relevant sources of variation ($NonRel_{ij}$) have comparable or greater influence on the raw signal variability than the biological signal ($CopyNb_i$) (see Section 3.2 to compare the type III sum of squares of the different effects in a multiple linear model). We therefore propose an iterative, alternative normalization method, making it possible to estimate the biological signal and non-relevant effects and, therefore, to eliminate most of the non-relevant effects while preserving most of the biological information. During each iteration, ITALICS:

(1) Estimates the biological signal $CopyNb_i$ using the GLAD algorithm (Hupé *et al.*, 2004),

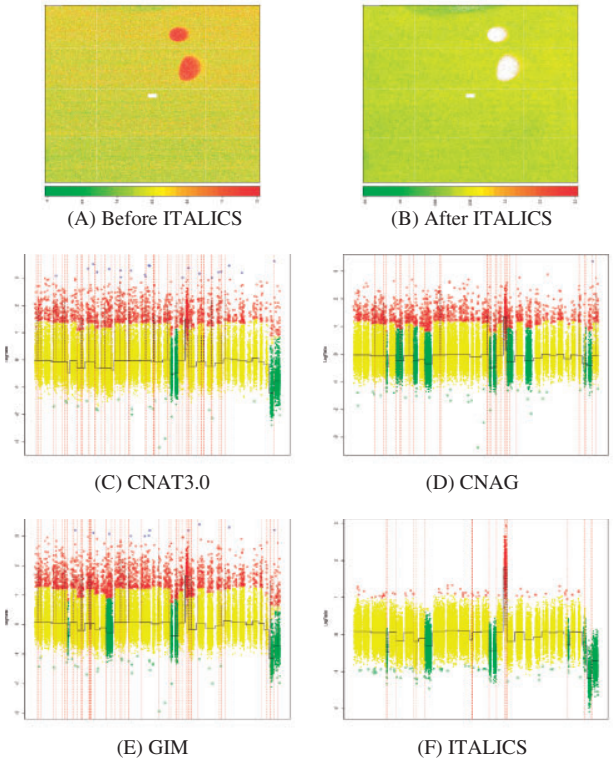


Fig. 1. Impact of spatial artifacts on genomic profiles. Image of an *Xba*I 100K Set chip (HF0844_Xba, Kotliarov *et al.* (2006)) before (**A**) and after normalization with ITALICS (**B**) (flagged $Quartets_{PM}$ in white). The Y_{ij} value of each $Quartet_{PM}$ is represented, using a gradient from green to red. (**C**), (**D**), (**E**) and (**F**) are the genomic profiles normalized with CNAT 3.0, CNAG, GIM and ITALICS. Vertical dashed red lines represent the breakpoints detected with GLAD and the assigned statuses are indicated by a color code: green for loss, yellow for normal and red for gain. Two stains of abnormally high $Quartets_{PM}$ values (in red) are visible in (A) and their corresponding SNP values correspond to outliers (colored in red) in the genomic profiles (C), (D) and (E), for which 1661, 1818 and 2331 outliers respectively, were detected. ITALICS flagged most of these $Quartets_{PM}$ (B) but evaluated the signals for their SNPs using the $Quartets_{PM}$ from the rest of the chip, resulting in the removal of only 13 of the 57 500 SNPs. ITALICS eventually identified only 88 outliers (F).

(2) Assuming the biological signal to be known, it estimates the non-relevant effects $NonRel_{ij}$ on raw data, by multiple linear regression.

After the last iteration, the $Quartets_{PM}$ for which multiple linear regression predicts the signal poorly are flagged. They correspond to $Quartets_{PM}$ with abnormal values and are excluded from the final step, in which ITALICS uses GLAD to estimate the biological effect $CopyNb_i$ on the remaining normalized $Quartets_{PM}$. The algorithm is presented in more detail below.

Biological signal estimation (CopyNb_step): ITALICS applies the GLAD algorithm to $Y_{i.}$ values to estimate the biological signal. The GLAD algorithm segments the genomic profile, defining regions of homogeneous DNA copy number. For each of these regions, it provides a smoothing value and a status (gain, normal or loss). The smoothing

value is the median of the $Y_{i.}$ values within the region concerned, and corresponds to the inferred copy number $CopyNb_i$.

Non-relevant effect estimation (NonRel_step): After estimating the biological effect $CopyNb_i$, ITALICS infers the non-relevant effects by multiple linear regression. The model used is as follows:

$$Y_{ij} = \mu + \alpha CopyNb_i + f(NonRel_{ij}) + \varepsilon_{ij}$$
$$f(NonRel_{ij}) = P_1(FL_i) + P_2(FGC_i) + P_3(QGC_{ij}) + \beta Q_{ij}$$

with:

$$i = 1, \ldots, N \text{ (the number of SNPs)}$$
$$j = 1, \ldots, n_i \text{ (the number of } Quartets_{PM} \text{ per SNP)}$$
$$P_k(x) = \sum_{l=1}^{l=3} \gamma_{kl} x^l, k = 1, \ldots, 3$$
$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

The multiple linear regression can also be expressed in classical matrix notation:

$$Y = X\theta + \varepsilon$$

with:

$$\theta = (\mu, \alpha, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{31}, \gamma_{32}, \gamma_{33}, \beta)$$

The parameter θ is estimated using the ordinary least-squares method. The degrees of the polynomial functions P_k were chosen using the BIC criterion (Schwarz, 1978) on a training data set of 128 reference diploid chips (Matsuzaki *et al*., 2004).

The $Quartet_{PM}$ effect is dealt with by calculating Q_{ij} as the mean of each $Quartet_{PM}$ on the 64 female chips of the same Affymetrix reference data set (Matsuzaki *et al*., 2004).

Once the non-relevant effects have been estimated, the Y_{ij} values are corrected as follows:

$$Y_{ij}^{\text{cor}} = Y_{ij} - \hat{f}(NonRel_{ij}),$$

where $\hat{f}(NonRel_{ij})$ corresponds to the estimate of non-relevant effects based on multiple linear regression. The corrected $Y_{i.}^{\text{cor}} = (\sum_{j=1}^{j=n_i} Y_{ij}^{\text{cor}}/n_i)$ is used in the next step of the GLAD procedure, to re-estimate the biological effect. This algorithm is repeated until the number of iterations reaches the predetermined fixed number of iterations *itermax*.

ITALICS uses GLAD and therefore we investigate if the normalization was influenced by the choice of GLAD parameters. In Supplementary information, we give guidelines for choosing parameters and expose the result of sensitivity analysis that shows a large robustness of ITALICS to parameter settings.

Elimination of poorly predicted $Quartets_{PM}$: After the last iteration, $Quartets_{PM}$ Y_{ij} poorly predicted by multiple linear regression are flagged out. This is achieved by calculating the 95% prediction interval. All Y_{ij} outside this interval are flagged. SNPs with less than three non-flagged $Quartets_{PM}$ in a total of n_i are then discarded. If more than three Y_{ij} are not flagged, $Y_{i.}^{\text{cor}}$ is recalculated as:

$$Y_{i.}^{\text{cor}} = \frac{\sum_{j \notin F_i} Y_{ij}^{\text{cor}}}{n_i - NbF_i},$$

with F_i the set of flagged $Quartets_{PM}$ for the SNP_i and NbF_i the number of flagged $Quartets_{PM}$ for the SNP_i.

Data scaling: The data are scaled to allow between-chip comparison. After the first GLAD step, the biological signal is subtracted and the standard deviation s of $(Y_{i.} - CopyNb_i)$ is calculated for each chip using all SNPs i of the chip. The data are then scaled as follows:

$$Y_{ij}^{\text{scaled}} = \frac{Y_{ij}}{s}$$

Table 1. ITALICS algorithm overview

```
iter: =0
while iter<itermax do
  CopyNb_step()
  if iter=0 then
    Data_Scaling()
  end if
  NonRel_step()
  iter: =iter+1
end while
elimination_of_poorly_predicted_quartet_PM( )
CopyNb_step()
```

The ITALICS procedure is summarized in Table 1.

2.3 Comparison with other methods

Other methods: Several other methods have already been developed. Most use linear regression to estimate and correct for non-relevant effects. They differ in the effects taken into account and in their pre- and post-processing steps.

CNAG: Copy Number Analysis for GeneChip (Nannya *et al*., 2005). CNAG corrects the raw signal intensity of a sample, by introducing the notion of *averaged best fit*, corresponding to a pseudochip constructed from the five samples most similar to the reference samples. CNAG subtracts this averaged best fit from the raw signal and then corrects for the length of the PCR-amplified fragment and GC-content effects by linear regression. This method is available within CNAG 2.0 and is also used in CNAT 4.0 (Copy Number Analysis Tool, see below).

CNAT 3.0: Chromosome Copy Number Analysis Tool 3.0. Affymetrix developed this method for the extraction of DNA copy number. No specific step for the correction of non-relevant effects is included. This method uses samples with varying chromosome X copy number for intensity calibration and transforms SNP intensity into copy number values.

CNAT 4.0: Chromosome Copy Number Analysis Tool 4.0. This tool uses CNAG to normalize the data and then smoothes the data with a user-defined window. This step artificially reduces the variance of the data and visibly improves the quality of the profile.

CARAT: Copy Number Analysis with Regression And Tree (Huang *et al*., 2006). CARAT uses a reference data set to select probes showing a high-allelic response and to remove those with no such response. For each new sample, it first standardizes the probe signal, based on mismatch probe information. It then corrects for probe GC-content and PCR fragment length effects, by linear regression. Finally, each SNP intensity is regressed against the average intensity of the reference samples with the same genotype.

GIM: Genomic Imbalance Map (Komura *et al*., 2006). GIM roughly estimates the biological effect and subtracts it from the raw signal, using a simpler version of ChARM (Myers *et al*., 2004). It removes defective probes with a high local GC-content and then re-estimates the biological effect without using the defective probes and subtracts this effect from the raw signal. It takes into account probe GC-content, the length of the PCR-amplified fragment and its GC-content, and mean SNP intensity for the reference dataset, by linear regression. GIM is implemented in Matlab and is freely available.

We compared ITALICS with CNAG, CNAT 3.0 and GIM. We did not compare ITALICS with CARAT, because no software was

available for CARAT at the time of the study, or with CNAT 4.0, which presents no improvement over CNAG. For the CNAG, CNAT 3.0 and GIM genomic profiles, copy number and the status of the genomic regions were inferred with the GLAD algorithm, using the same parameters as for the ITALICS algorithm.

Quality criteria: As described by Neuvial *et al*. (2006), we used several quality criteria to compare the various normalization algorithms.

As defined by Neuvial *et al*. (2006), the *dyn* criterion estimates the dynamics of the DNA copy number signal. Its value is:

$$dyn(a) = \frac{(median(Y_{i.}^{\mathrm{cor},a})_{i\in G} - median(Y_{i.}^{\mathrm{cor},a})_{i\in N})}{smt}$$

with G and N the regions considered to correspond to *Gain* and *Normal* and $Y_{i.}^{\mathrm{cor},a}$ the corrected signal of SNP i using the normalization method a. $smt = median(|Y_{i.}^{\mathrm{cor},a} - Y_{(i-1).}^{\mathrm{cor},a}|)$ for ordered $Y_{i.}^{\mathrm{cor},a}$ throughout the genome. *smt* quantifies the smoothness of the signal over the genome, and *dyn* assesses the dynamics of the signal, as defined by the signal-to-noise ratio (SNR). If no gain region have been identified, the *dyn* criteria is computed over loss regions. A high *dyn* should be obtained with good normalization methods.

The criterion *out* is the number of outliers detected by GLAD. GLAD defines regions of homogeneous DNA copy number and outliers are SNPs with values different from those of other SNPs in the same region. These abnormal values may be accounted for by point mutations in the genome. However, a large number of such changes is unlikely, so the total number of outliers should be relatively low and the *out* parameter close to zero.

The criterion *flag* is the number of flagged SNPs. We introduced this criterion for the comparison of methods that remove SNPs, such as GIM and ITALICS. These methods may artificially improve the quality of the signal (as measured by *dyn* and *out*), by removing SNPs with abnormal behavior. The number of flagged SNPs should, therefore, not be too high. When faced with a choice between two methods with equal SNR, the method with the lowest *flag* should be preferred.

Comparison of two normalization methods: These three criteria can be used to determine which of the two normalization methods gives the best results for a given array. In this pairwise comparison context, *dyn* must be calculated with the same definition of gain, normal and loss regions for both normalized arrays. We therefore define consensus gain, normal and loss regions associated with an array processed with two different normalization methods, as the intersection of the two corresponding gain, normal and loss regions obtained with the two different normalization methods [see also Neuvial *et al*. (2006) for details].

For the comparison of two different methods, a and b, in terms of a certain criterion, we calculate relative performances as follows:

$$RP^{dyn}(a,b) = (dyn(a) - dyn(b))/dyn(a)$$
$$RP^{out}(a,b) = -(out(a) - out(b))/out(a)$$
$$RP^{flag}(a,b) = -(flag(a) - flag(b))/flag(a)$$

RP measures the percentage improvement observed with method a, with respect to method b. The minus signs for the *out* and *flag* criteria ensure that a positive $RP^{cri}(a,b)$ always means that method a is better than method b for criterion *cri*.

2.4 Datasets

We carried out our study on two public datasets: a dataset for 128 reference diploid chips (Matsuzaki *et al*., 2004) and a glioma dataset corresponding to 356 chips (Kotliarov *et al*., 2006). We also used datasets produced by the Affymetrix platform of the Institut Curie obtained with 22 uveal melanoma samples, 40 ovarian cancer samples and 26 breast cancer samples.

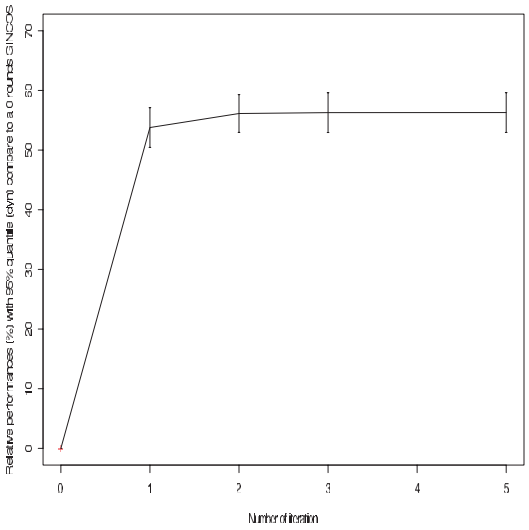


Fig. 2. Improvement in SNR with the number of ITALICS iterations The improvement in SNR obtained with each iteration was assessed by calculating the percentage improvement RP^{dyn} for 1, 2, 3, and 5 iterations with respect to no iterations. The results are summarized in this graph, showing RP^{dyn} as a function of the number of iterations. The SNR improved with the first two iterations, with no major improvement observed for subsequent iterations.

3 RESULTS

3.1 Choosing the number of iterations

We assessed the extent to which each iteration within the ITALICS algorithm improved the SNR, by calculating the *dyn* criteria for different values of *itermax* (0, 1, 2, 3 and 5) for each chip of the 356-glioma chips dataset. The percentage improvement RP^{dyn} for different values of *itermax* (1, 2, 3 and 5) with respect to no iteration was then calculated (Fig. 2). One iteration gave 53.8% improvement, two gave 56.1% improvement and three and five gave 56.3% improvement. As the third and subsequent iterations gave only a very slight improvement, we set *itermax* to two in the ITALICS algorithm.

3.2 Importance of each effect on the signal

For each chip of the glioma dataset, we calculated the type III sum of squares for each effect in our multiple linear regression model. A low type III sum of squares indicates that the difference between the full model and the model excluding the studied effect is very small. The $Quartets_{PM}$ effect gave the highest type III sum of squares, with a mean of 550×10^3 versus 10.4×10^3, 16×10^3 and 14×10^3 for $Quartets_{PM}$ GC-content, fragment length and fragment GC-content. The biological effect was the second most important effect, with a mean of 24×10^3.

3.3 ITALICS outperformed the other methods

We calculated *dyn* and *out* with ITALICS, GIM, CNAT 3.0 and CNAG, using three different cancer datasets: two in-house datasets corresponding to 22 choroidal melanoma chips and

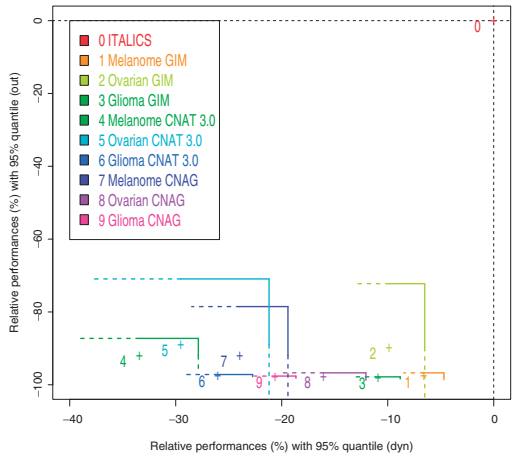


Fig. 3. Comparison of ITALICS with other normalization methods. We compared ITALICS with CNAT 3.0, CNAG and GIM for two quality criteria—*dyn* and *out*—using three different cancer datasets: two in-house data sets corresponding to 22 choroidal melanoma chips and 40 ovarian cancer chips and one public dataset corresponding to 356 glioma chips (Kotliarov *et al.*, 2006). Each color corresponds to the comparison of ITALICS with a different method or data set. ITALICS is taken as the reference [red point 0 at (0, 0)]. For each method, the cross indicates the mean relative performance on the data set concerned, for the *dyn* and *out* criteria, and the lines give the corresponding 95% quantile for relative performance. ITALICS significantly outperforms all methods for both quality criteria, *dyn* and *out*.

40 ovarian cancer chips and one public data set of 356 glioma chips. All methods were used with their default parameters.

We calculated the percentage improvement (*RP*) for CNAT 3.0, CNAG and GIM, in terms of *dyn* and *out*, with respect to ITALICS (Fig. 3). For the three competitors $RP^{cri}(competitor,ITALICS)$ is calculated and we performed *t*-tests to assess the significance of the improvement. We found that ITALICS outperformed CNAT 3.0, CNAG and GIM, in terms of *dyn* and *out*, with *t*-test *P*-values below 10^{-5} for all three data sets. For GIM, RP^{dyn} ranged from −10.9% to −6.5%, for CNAG, it ranged from −23.9% to −16.0% and for CNAT 3.0 it ranged from −33.4% to −26.0%. RP^{out} ranged from −98.1% to −89.0% for all three methods. Chip data normalized with ITALICS therefore had a significantly better SNR than those normalized with CNAT, CNAG and GIM, with fewer outliers.

Both ITALICS and GIM flag certain SNPs for elimination. The improvement in SNR obtained with these methods may therefore be partially due to the mechanical effect of this removal. We compared the number of SNPs flagged between GIM and ITALICS and found that ITALICS flagged significantly fewer SNPs than GIM, with a mean of 300 SNPs per chip for ITALICS versus 3000 for GIM. The $RP^{flag}(GIM,ITALICS)$ is −90%.

3.4 Spatial artifact correction

Some Affymetrix SNP arrays suffer from spatial artifacts. The step flagging poorly predicted $Quartets_{PM}$ removes most $Quartets_{PM}$ with abnormal intensity detected by visual inspection, as shown in Figures 1A and B. To our knowledge, ITALICS is the only method capable of doing this. Moreover, the removal of these abnormal $Quartets_{PM}$ increases the quality of the signal, by removing many outliers from the genomic profile: 1661, 1818 and 2331 outliers were detected for CNAT 3.0, CNAG and GIM (Figure 1C, D and E). With ITALICS, there were only 88 outliers (Figure 1F), but only 13 of the 56 000 SNPs were removed because they had less than three non-flagged $Quartets_{PM}$.

3.5 Biological validation

Quantitative PCR validation: We used QPCR (see Supplementary Material for more detail) to validate our method with a different technology. As a test case, we used a set of paired breast cancer samples (primary tumor and relapse, Bollet *et al.* 2008) and tried to identify a breakpoint in chromosome 20. We compared the results obtained with QPCR with those obtained with ITALICS, CNAG, GIM and CNAT, for the *Xba*I and *Hind*III arrays. We also carried out QPCR on two breast cancer tumors, each with a normal chromosome 20 (white and striped bars in Fig. 4) to assess noise for QPCR and to validate the significance of copy number change. As shown in Figure 4, ITALICS was more accurate than CNAG, GIM and CNAT 3.0 for comparisons of copy numbers, based on the estimates obtained with PCR. ITALICS, CNAG, GIM and CNAT 3.0 detected changes in copy number in this region of chromosome 20. However, ITALICS breakpoints were closer to QPCR breakpoints than CNAT breakpoints (see Fig. 4A, C and D) and CNAG and GIM breakpoints (see Figure 4A). In Figure 4A, QPCR and ITALICS breakpoints are found at identical positions (between P14 and P15). In Figure 4C and D, CNAG, GIM and ITALICS detect a copy number change between P12 and P13, close to that detected by QPCR between P14 and P15, whereas CNAT detects this breakpoint further away, between P06 and P07 in Figure 4C and between P08 and P09 in Figure 4D. In Figure 4B, QPCR, CNAT, GIM, CNAG and ITALICS found the same breakpoint.

Patients with breast cancer relapses: The problem tackled was determining whether the second cancer was a true recurrence of the first cancer or a new primary tumor, based on the two Affymetrix SNP array profiles (Bollet *et al.*, 2008). We tried to identify common breakpoints between the cancer chips for the two tumors. The breakpoints detected with CNAT 3.0 or ITALICS normalization are represented in Figure 5A and B for chromosome 6 and 9, respectively, for one patient. GIM and CNAG results are similar to ITALICS for chromosome 6 and similar to CNAT for chromosome 9 (data not shown). ITALICS identified breakpoints at identical locations for both cancers and this is true for the two chromosomes presented in Figure 5A and B. It is important to notice that this was not possible with CNAT 3.0, CNAG and GIM. The precise match between the breakpoints mapped in the two cancers with ITALICS suggests that the second cancer is a true recurrence, whereas the opposite conclusion would have been drawn with CNAT 3.0. As CNAG and GIM detect less precise matches, they lead to the same conclusion as ITALICS, but the evidences for this conclusion are weaker. Expert assessment based on clinical data also indicated that this was a true recurrence, and

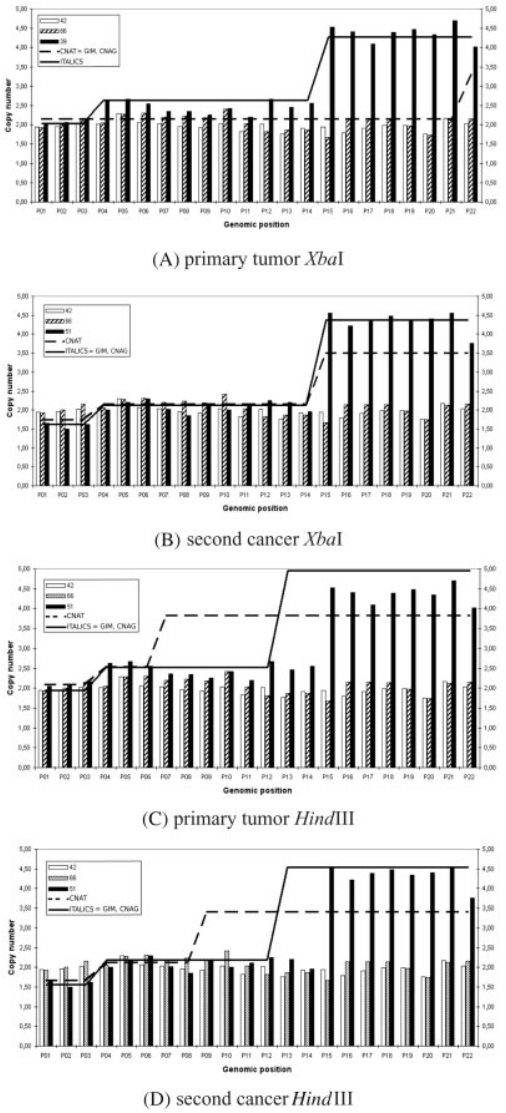


Fig. 4. Affymetrix SNP arrays and QPCR DNA copy number profiles for a patient with breast cancer relapse. CNAT 3.0 (dashed line) and ITALICS (solid line) DNA copy number determination along chromosome 20, from position 17453432 (P01) to position 49386812 (P22), for the primary tumor (**A**, **C**) and the relapse (**B**, **D**) using the *Hind*III (C, D) and *Xba*I (A, B) Affymetrix SNP arrays. CNAG and GIM results are identical to CNAT for (A) and identical to ITALICS for (B, C and D). We performed QPCR on two breast cancer tumors with a normal chromosome 20, to estimate the noise associated with QPCR and to validate the significance of copy number change. The bar charts generated show the QPCR estimation of DNA copy number in two breast cancer tissues with a normal chromosome 20 (white and striped bars, A, B, C and D), the primary breast tumor (black bars, A and C) and the corresponding relapse (black bars, B and D). In (A), both ITALICS and QPCR detect a copy number change between P14 and P15, whereas GIM, CNAG and CNAT detects a change between P21 and P22. In (C) and (D), ITALICS detects a copy number change between P12 and P13, close to that detected by QPCR between P14 and P15, whereas CNAT detects a breakpoint further away, between P06 and P07 in (**C**), and between P08 and P09 in (D). In (B) QPCR, CNAT and ITALICS found the same breakpoints.

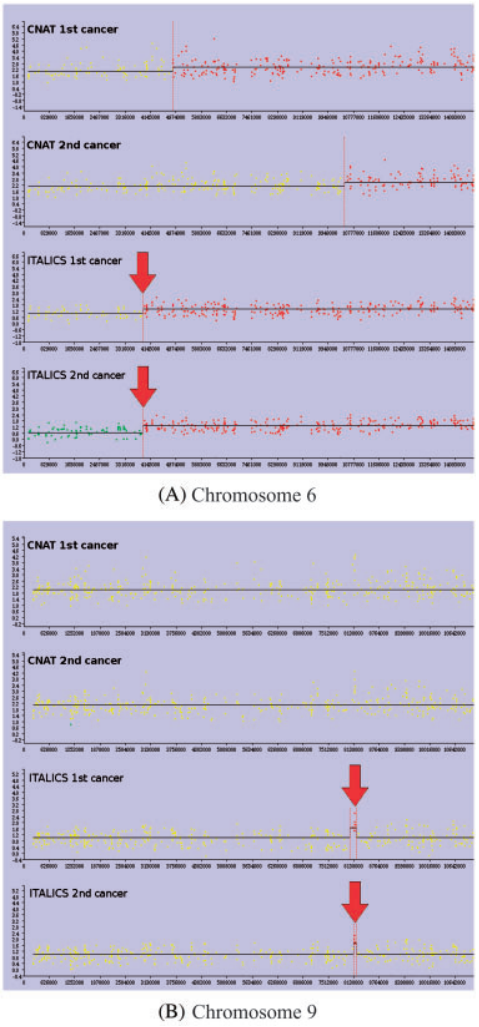


Fig. 5. Detection of breakpoints common to first and second cancers, using ITALICS. We present part of the chromosome 6 (**A**) and 9 (**B**) profiles obtained with VAMP (La Rosa *et al.*, 2006) for a patient with two breast tumors. For both (**A**) and (**B**), the first two profiles are CNAT 3.0 profiles of the first and second cancers and the last two profiles are ITALICS profiles of the first and second cancers. GIM and CNAG results are similar to ITALICS for chromosome 6 and similar to CNAT for chromosome 9 (data not shown). CNAT 3.0 identified no breakpoints (red dashed lines) common to the two cancers, whereas ITALICS did (red arrows), strongly suggesting that the second cancer was a true recurrence. Moreover, the results obtained with ITALICS are supported by an expert classification based on clinical data.

was therefore consistent with the results obtained with ITALICS. Similar conclusions were drawn for the rest of the data set (13 first and second cancer pairs). Thus, ITALICS improves the classification of true recurrences and new primary tumors.

4 DISCUSSION AND PERSPECTIVES

We present here a new method for normalizing Affymetrix SNP arrays: ITALICS. This method is highly efficient and

outperforms other normalization methods, such as CNAT 3.0, CNAG and GIM, in terms of SNR, giving a more accurate localization of breakpoints validated by QPCR. This improvement may be due to various features of the ITALICS algorithm. This algorithm estimates alternatively and iteratively both non-relevant and biologically relevant effects. The correct estimation of relevant effects depends on correct estimation of the biological signal and vice versa, as the relevant effects induce similar or higher ranges of variation than the biologically relevant effect. By estimating both the non-relevant and biologically relevant effects in an iterative manner, we avoid overestimation of the non-relevant effects and a loss of biological signal. The first estimation on raw data is necessarily rough, but improves the subsequent estimation of non-relevant effects. Each new estimation of the biological or non-relevant effects leads to a better estimation of the other effects. In practice we iterate our algorithm twice, as additional iterations were found to lead to no significant improvement in the SNR. This algorithm also includes a flagging step, making it possible to remove aberrant SNPs. Indeed, some PM intensity values are subject to spatial artifacts. The PM intensity of their $Quartets_{PM}$ is therefore abnormal, poorly predicted by the regression model and flagged. The discarding of poorly predicted $Quartets_{PM}$ does not necessarily lead to the discarding of the corresponding SNP, provided that enough $Quartets_{PM}$ remain elsewhere on the chip. As a result, very few SNPs are removed from the final genomic profile. This filtering step detects spatial artifacts only indirectly, but nevertheless gives good results in practice. Methods for the precise detection of spatial artifacts and the removal of all probes within spatial artifacts have already been developed (Neuvial *et al*., 2006). However, their direct application to SNP chips is impossible due to the very high density of these chips (more than 2 million probes per chip). Computing $Quartets_{PM}$ effect on an in-house reference dataset would certainly improve the quality of the normalization. Nevertheless, the $Quartets_{PM}$ effect is the most important effect and ignoring it would decrease the efficiency of the normalization.

We normalized *Xba*I and *Hind*III chips separately. The same major changes were detected with both chips. However, it is difficult to merge *Xba*I and *Hind*III data due to the difference in signal amplitude for consecutive alterations between the two chips. The merging of the *Xba*I and *Hind*III genomic profiles would result in a higher resolution profile, but also in a lower SNR. The ITALICS algorithm could be improved by taking into account the enzyme effect (*Xba*I and *Hind*III) to overcome this problem.

Technically, the ITALICS algorithm could be applied to higher density chips, such as the Affymetrix GeneChip Human Mapping 500K Set and even the Genome Wide SNP array 5.0 and 6.0, which do not have MM probes, as ITALICS is based solely on PM probes. Of course, we would have to check whether the non-relevant effects in our model are also observed with these higher density chips. We would also need to obtain a reference dataset for calculating the quartet effect.

5 CONCLUSION

We developed ITALICS, a new normalization algorithm for Affymetrix SNP arrays. This method was designed for the normalization and analysis of DNA copy number and significantly outperformed other methods, such as CNAT 3.0, CNAT 4.0, CNAG and GIM, in terms of SNR and can also be used to correct for experimental artifacts due to spatial effects. This method was validated by QPCR and accurately detected the breakpoints in genomic profiles. It could therefore be used to improve the characterization of samples in genomic studies.

ACKNOWLEDGEMENTS

This work was supported by the Institut Curie and the Centre National de la Recherche Scientifique. We thank Sophie Piperno-Neumann and Simon Saule, Jean-Paul Thiery and Marc Bollet, who were kind enough to provide us with access to their uveal melanoma, ovarian cancer and breast cancer datasets, respectively. We thank Marc Bollet, Nicolas Servant and Pierre Neuvial for fruitful discussions. We thank Audrey Rapinat and David Gentien for performing the Affymetrix Genechip experiments.

Conflict of Interest: none declared.

REFERENCES

Bollet,M. *et al*. (2008) High resolution mapping of breakpoints to define true recurrences among ipsilateral breast tumor recurrences. *J. Natl Cancer Inst*., **100**, 48–58.

Huang,J. *et al*. (2006) CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics*., **7**, 83.

Hupé,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Kennedy,G.C. *et al*. (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol*., **21**, 1233–1237.

Komura,D. *et al*. (2006) Noise reduction from genotyping microarrays using probe level information. *In Silico Biol*., **6**, 79–92.

Kotliarov,Y. *et al*. (2006) High resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. *Cancer Res*., **66**, 9428–9436.

La Rosa,P. *et al*. (2006) VAMP: visualization and analysis of array- CGH, transcriptome and other molecular profiles. *Bioinformatics*, **22**, 2066–2073.

Liu,W. *et al*. (2007) Deletion of a small consensus region at 6q15, including the MAP3K7 gene, is significantly associated with high-grade prostate cancers. *Clin. Cancer Res*., **13**, 5028–5033.

Matsuzaki,H. *et al*. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*., **1**, 109–111.

Myers,C.L. *et al*. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Nannya,Y. *et al*. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*, **65**, 6071–6079.

Neuvial,P. *et al*. (2006) Spatial normalization of array-CGH data. *BMC Bioinformatics*., **7**, 264.

Pinkel,D. and Albertson,D. G. (2005) Comparative genomic hybridization. *Annu Rev. Genomics Hum. Genet*., **6**, 331–354.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat*., **6**, 461–464.

Ylstra,B. *et al*. (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucl. Acids Res*., **34**, 445–450.

2.4 Extraction of informative DNA copy number alterations

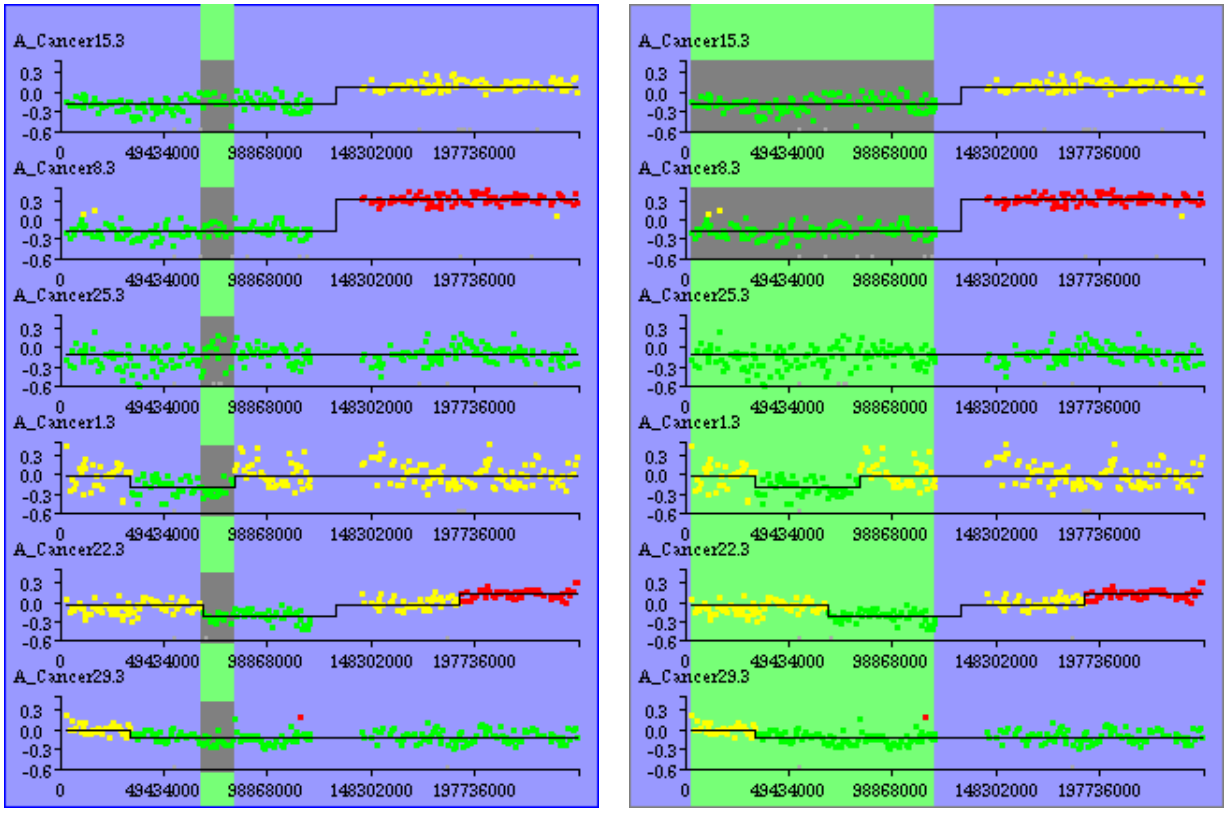
The three statistical methods, MANOR, GLAD and ITALICS deal with one genomic profile at a time. However, methods which are able to analyse many profiles simultaneously are needed in order to identify the relevant alterations for a given pathology. Indeed, alterations frequently observed in a set of tumours or at least in a subset are likely to be involved in tumoral progression. A method has been developed to extract such informative alterations from a set of DNA copy number microarray experiments (Rouveirol et al., 2006, this article is supplied in the **Annexes**). The method uses as input the calling done by GLAD or ITALICS or any algorithms able to provide a call for each probe (*i.e.* either loss, normal, gain or amplification). Two types of alterations can be detected by the algorithm:

- **minimal alteration**: this type of alteration corresponds to the smallest one intersecting a sufficient number of tumours (this parameter needs to be chosen by the user) as shown in **Figure 2.3a**. The identification of such alterations should pinpoint new cancer-critical genes (see **Subsection 1.3.3**): the tumour-suppressor genes are supposed to be present in minimal alterations of losses and oncogenes in minimal alterations of gains.

- **recurrent alteration**: this type of alteration is more restrictive than the previous definition. Indeed, in this case, a sufficient number of tumours (this parameter needs to chosen by the user) have a common alteration whose extremities are at the same location ($\pm$ tolerance) as shown in **Figure 2.3b**. This type of alteration is relevant when a precise alteration is needed for tumoral progression such as gains/losses of complete chromosomes or gains/losses of chromosome arms.

The algorithm outputs an indicator matrix as shown in **Figure 2.4**: for each tumour it is indicated whether the sample has a given alteration or not. Finding informative DNA copy number alterations presents many advantages. First, it allows the information to be reduced taking into account the redundancy of the data since contiguous probes on the genome are very likely to have the same DNA copy number. This way, the aCGH profiles are converted into a set of relevant features which leads to more powerful downstream analyses (Van de Wiel and Van Wieringen, 2007). Secondly, in downstream analyses, it allows the same weight to be given for each alteration whatever its size. Indeed, possibly very small alterations, such as amplifications, may be relevant as predictive or prognostic factors. Since there are few probes in such small alterations it is better to use the alteration as a single entity so that all the regions are weighted the same. Thirdly, it allows an easier interpretation of the data since biologists can just study a limited number of alterations rather than all the probes. Besides the proposed algorithm, other methods exist such as GISTIC (Genomic Identification of Significant Targets in Cancer) (Beroukhima et al., 2007), STAC (Significance Testing for Aberrant Copy number) (Diskin et al., 2006) and CGHregions (Van de Wiel and Van Wieringen, 2007).

These DNA copy number informative alterations are used in **Section 2.5** to identify high-risk tumours and in **Chapter 3** to combine both DNA copy number and mRNA expression microarray experiments.



a - Minimal alteration

b - Recurrent alteration

Figure 2.3: Example of informative DNA copy number alterations - The same six tumours are represented in the left and right graphics (The aCGH profiles have been retrieved from ACTuDB and correspond to colon cancer from the dataset of Douglas et al., 2004).(a) Six tumours have lost the green minimal alteration. (b) Two tumours have lost the green recurrent alteration which has the same extremities in the two samples.

	Gain Chr1	Loss Chr2	Gain Chr8	Amplification Chr8
Tumour1	0	0	1	0
Tumour2	1	0	0	1
Tumour3	1	0	1	1
Tumour4	0	1	0	0
Tumour5	0	0	0	1

Figure 2.4: Representation of DNA copy number data - Each column represents an informative DNA copy number alteration (either a minimal or recurrent alteration). A row represents a tumour sample in which it is indicated whether the sample has the corresponding alteration (1) or not (0).

2.5 Example of aCGH study: identification of high-risk tumours in uveal melanoma

In this section is presented a study to identify high-risk tumours in uveal melanoma which is the most common intra-ocular tumour (Trolet et al., 2008, under revision in *Investigative Ophthalmology and Visual Science*). The paper describing the analysis is supplied as a material part of the thesis. The particularity of this cancer is its high propensity to metastasise almost exclusively in the liver: in this case the prognosis of the patient is very poor, and therefore, patients with high-risk of metastasis need to be accurately identified. Moreover no treatment other than eye removal exists. In the study, BAC aCGH profiles were available for primary tumours from the eye and from liver metastases. The statistical methods which have been developed during the thesis to analyse DNA copy number experiments have been used in this study. Briefly, the following analyses have been performed:

- normalisation of the BAC aCGH profiles with the MANOR algorithm (Neuvial et al., 2006).
- segmentation and calling with the GLAD algorithm (Hupé et al., 2004).
- identification of informative DNA copy number alterations with the algorithm by Rouveirol et al. (2006).
- unsupervised classification (hierarchical clustering) based on the informative DNA copy number alterations (Brito et al., 2008, this article has been submitted to *Bioinformatics* and is supplied in the **Annexes**) in order to identify new subgroups of tumours.
- supervised classification (log-linear model - see **Subsection 3.2.4**) based on the informative DNA copy number alterations in order to predict high-risk tumours.

Genomic Profiling and Identification of High Risk Uveal Melanoma by array-CGH Analysis of Primary Tumors and Liver Metastases

Journal:	*Investigative Ophthalmology & Visual Science*
Manuscript ID:	IOVS-08-2296
Manuscript Type:	Article
Date Submitted by the Author:	16-May-2008
Complete List of Authors:	Trolet, Julien; Institut Curie, Dept of Bioinformatics; INSERM, U 900; Ecole des Mines, ParisTech Hupe, Philippe; Institut Curie, Dept of Bioinformatics; INSERM, U 900; Ecole des Mines, PariTech; CNRS, UMR 144 Huon, Isabelle; Institut Curie, Dpt. of Genetics Lebigot, Ingrid; Institut Curie, Centre of Biological Resources Decraene, Charles; Institut Curie, Dept of Translational research Delattre, Olivier; INSERM, U 830 Sastre-Garau, Xavier; Institut Curie, Dept of Pathology SAULE, Simon; Institut Curie, CNRS UMR146 Thiery, Jean-Paul; Institut Curie, Dept of Translational research Plancher, Corine; Institut Curie, Dpt of Biostatistics Asselain, Bernard; Institut Curie, Dpt of Biostatistics Desjardins, Laurence; Institut Curie, Dpt of Ophtalmology Mariani, Pascale; Institut Curie, Dpt of Surgery Piperno-Neumann, Sophie; Institut Curie, Dpt of Medical Oncology Barillot, Emmanuel; Institut Curie, Dept of Bioinformatics; INSERM, U 900; Ecole des Mines, ParisTech COUTURIER, Jerome; Institut Curie, Dpt of Genetics; INSERM, U 830
Keywords:	melanoma, molecular biology, gene expression
Abstract:	Purpose. About 50% of patients carrier of a uveal melanoma develop incurable metastases. The purpose of the study was to determine genomic profiles in a large series of uveal melanomas (UM) and liver metastases, in order to design a genome profile-based prognostic classifier. Methods. A series of 86 UM tumors and 66 liver metastases was analysed using BAC CGH-microarrays. A clustering was performed, and correlation with the metastatic status was sought in a subset of 78 patients (median follow-up: 54 months). A prognostic classifier was built using a log-linear model on minimal regions and leave-one-out cross-validation. Results. The clustering refines the classical classification of UM dividing the disomic 3 and the monosomic 3 groups into, respectively, two and three subgroups. Same subgroups were found in primary tumors and in metastases, but with different frequencies. Monosomy 3 was present in 70% of ocular tumors, and

	81% of metastases. Isolated monosomy 3 was present in 0% of metastatic ocular tumors and in 3% of metastases only. Highest metastatic rate in ocular tumors was in a subgroup defined by gain of 8q and losses of 3, 8p and 16q, also most represented in metastases. Position of breakpoint in 8q gains was proximal in 77% of metastatic tumors and 79% of metastases. A prognostic classifier including losses of 3, 8p, 16q, gain of 6p, and breakpoint position on 8q, lead to 82.1% of good classification. Conclusion. Genome profiling should be a reliable help for the identification of high-risk patients for future adjuvant therapy protocols.

scholarONE™ Manuscript Central

For Review Only

Genomic Profiling and Identification of High Risk Uveal Melanoma by array-CGH Analysis of Primary Tumors and Liver Metastases

Julien Trolet,[1,9,14] Philippe Hupé,[1,9,10,14] Isabelle Huon,[2] Ingrid Lebigot,[11] Charles Decraene,[3] Olivier Delattre,[12] Xavier Sastre-Garau,[4] Simon Saule,[13] Jean-Paul Thiéry,[3] Corine Plancher,[5] Bernard Asselain,[5] Laurence Desjardins,[6] Pascale Mariani,[7] Sophie Piperno-Neumann,[8] Emmanuel Barillot,[1,9,14] and Jérôme Couturier[2,12]

Departments of [1]Bioinformatics, [2]Genetics, [3]Translational Research, [4]Pathology, [5]Biostatistics, [6]Ophtalmology, [7]Surgery, and [8]Medical Oncology; [9]INSERM U900, [10]CNRS UMR144, [11]Biological Resources Centre, [12]INSERM U830, [13]CNRS UMR146, Institut Curie, Paris, France.

[14]Ecole des Mines de Paris, ParisTech, Fontainebleau, France.

For Review Only

Presented in part at the International Society of Ocular Oncology (ISOO) Meeting, Siena, Italy, June 2007.

This work was supported by a grant of the Département de Transfert, Institut Curie.

Manuscript 4 275 words.

Corresponding author : Dr. Jérôme Couturier, Department of Genetics, Institut Curie – Hôpital, 26 rue d'Ulm, F-75248 Paris Cedex 05, France.

E-mail: jerome.couturier@curie.net

ABSTRACT

Purpose. About 50% of patients carrier of a uveal melanoma develop incurable metastases. The purpose of the study was to determine genomic profiles in a large series of uveal melanomas (UM) and liver metastases, in order to design a genome profile-based prognostic classifier.

Methods. A series of 86 UM tumors and 66 liver metastases was analysed using BAC CGH-microarrays. A clustering was performed, and correlation with the metastatic status was sought in a subset of 78 patients (median follow-up: 54 months). A prognostic classifier was built using a log-linear model on minimal regions and leave-one-out cross-validation.

Results. The clustering refines the classical classification of UM dividing the disomic 3 and the monosomic 3 groups into, respectively, two and three subgroups. Same subgroups were found in primary tumors and in metastases, but with different frequencies. Monosomy 3 was present in 70% of ocular tumors, and 81% of metastases. Isolated monosomy 3 was present in 0% of metastatic ocular tumors and in 3% of metastases only. Highest metastatic rate in ocular tumors was in a subgroup defined by gain of 8q and losses of 3, 8p and 16q, also most represented in metastases. Position of breakpoint in 8q gains was proximal in 77% of metastatic tumors and 79% of metastases. A prognostic classifier including losses of 3, 8p, 16q, gain of 6p, and breakpoint position on 8q, lead to 82.1% of good classification.

Conclusion. Genome profiling should be a reliable help for the identification of high-risk patients for future adjuvant therapy protocols.

For Review Only

INTRODUCTION

Uveal melanoma is the most common intraocular malignant tumor, with an incidence of about 6 cases per million per year in the Caucasian population. It shows a high propensity to metastasize, in 90% of cases to the liver. Its prognosis is poor, with a survival of about 50% at 10-15 years, despite successful treatment of the primary tumor (1). Ophtalmologists and oncologists have recently considered the possibility of developing adjuvant systemic treatments for high-risk patients (2). This involves an accurate detection of tumors associated with a high metastatic risk at time of diagnosis to identify eligible patients. Beside clinico-pathological features (tumor size, location, histology, extrascleral invasion), certain genomic alterations of the tumor, affecting mainly chromosomes 3, 6, and 8, have been identified by karyotype analyses, then by Fluorescence In Situ Hybridization (FISH) and Comparative Genomic Hybridization (CGH) (for review, 3). Status of chromosome 3 has been shown to be strongly associated with the outcome of patients. Monosomy 3 is an early event present in 50-60% of tumors, often associated with an isochromosome 8q, and about 60% of patients having a monosomic 3 tumor experience a metastatic evolution, while disomic 3 tumors are considered as rarely leading to metastatic disease (4-7). Additionally, other recurrent chromosome alterations, such as imbalance of chromosome 6 and losses of 1p and 16q, have been described (8-12). Genome-wide techniques of genomic and expression profiling, make it now possible to analyze these tumors with combined imbalances with a much higher resolution and without the limitations of cytogenetic analyses. These approaches may improve the characterization of high-risk uveal melanoma. Recently, using gene expression profiling, two distinct molecular classes strongly associated with metastatic risk could be identified (13-15). However, DNA-based techniques are known to be more robust than expression-based methods and should be useful in reliably identifying regions of imbalance that might be of interest for a prognostic purpose, and for the search of genes involved in the development of this tumor. To date, only two pangenomic study using array-CGH, performed on 18 and 49 primary tumors, have been reported (16, 17), and little is known about the genomic profiles of uveal melanoma metastases (10). We report here the array-CGH analysis of 86 primary tumors and, for the first time, of 66 liver metastases, to attempt to identify a genomic profile associated with high-risk uveal melanoma.

MATERIALS AND METHODS

Patient and Tumor Samples

Ocular tumor samples were obtained from a series of 86 unselected patients who were treated by

For Review Only

enucleation. Informed consent was obtained. The study followed the tenets of the Declaration of Helsinki and was approved by of the Department of Translational Research and the Institutional Ethical Review Board. At first, unsupervised analysis and genomic characterization were carried out on all 86 tumors. Then, for the supervised analysis regarding the metastatic status, 8 tumors of patients with less than 24 months follow-up were removed, and a subset of 78 samples was used (median follow-up: 54 months, range: 24-96 months). Samples of 66 liver metastases were obtained from patients having undergone a carcinologic resection. Among these samples, 8 pairs of ocular tumor and the corresponding metastasis were available. All specimens included in the study were histologically confirmed, and were checked on a frozen section for showing more than 60% tumor cells, before DNA extraction.

Array-CGH

DNA extraction, labelling and hybridization were performed as previously described (18). A genome-wide DNA microarray made of about 4K BAC clones, FISH mapped, sequenced, verified for marker content, and spotted in triplicate, with a 1 Mb average resolution (CIT / INSERM U830, Institut Curie, Paris), was used. Hybridized slides were scanned using an Axon GenePix 4000B scanner (Molecular Devices, Sunnyvale, CA). Image analysis was performed with the Axon GenePix 5.1 software (Molecular Devices). In addition, one ocular tumor showing the minimal 3p loss in the series was analyzed on 250K GeneChip Array (Affymetrix , Santa Clara, CA).

Array Data Processing

Normalization

We applied the MANOR algorithm, as described in Neuvial et al. (19), to correct for local spatial bias and continuous spatial gradient. Spots showing a too low signal-to-noise ratio or poor replicate consistency were discarded.

Alteration detection and minimal regions

Each array-CGH profile is centered on the median log-ratio and then analyzed using the GLAD algorithm (20). GLAD performs a segmentation of the genomic profile, defines regions of homogeneous DNA copy number, and returns for each of these regions a smoothing value and a status (Gain, Normal or Loss). For status assignment, the following thresholds were used: smoothing values lower than -0.15 and greater than 0.15 are set to loss and gain, respectively.

For Review Only

Minimal common alterations were identified using the formalization proposed by Rouveirol et al. (21). Minimal regions supported by at least 20% of the total number of tumors of the whole dataset have been considered in this analysis.

Tumors were represented under three different kinds of genomic profiles: (i) the sequence of the log-ratio values of each clone ordered along the genome (LR profile), (ii) the sequence of the status of each clone (SC profile), or (iii) the sequence of minimal regions (MR profile).

Clustering on Minimal Regions

Hierarchical clustering was performed on MR profiles using Euclidian distance as the similarity measure, and the Ward method in order to minimize the intra-class inertia during cluster building. Separation into groups was then proposed on the basis of the structure of the dendrogram.

Differential analysis of Log-Ratio profiles

A differential analysis was performed on the LR profiles to highlight clones which have significantly different log-ratios between two user-defined groups of tumors. For each clone, a Student test was performed and the obtained *P*-values were then adjusted using the Benjamini-Hochberg algorithm (22) for multiple testing correction. Clones with an adjusted *P*-value lower than 10% were considered to be significantly different between the two groups.

Data visualization and analysis

The visualization of the data, the computation of the minimal regions and clustering were done using the VAMP software (23).

Building of a prognosis classifier

Supervised classification was based on the MR profiles. Data were represented within a multiple contingency table in which each cell contains the number of tumors for the genomic category. A log-linear model was then used to analyze the contingency table and to build the classifier. The leave-one-out procedure has been used to assess the global performance, sensitivity and specificity of the classifier. Positive and negative predictive values correspond respectively to the proportion of metastatic tumors which are predicted metastatic, and the proportion of non-metastatic tumors predicted non-metastatic. The final classifiers were computed on the whole dataset. Variable selection on the minimal regions to include in the model was performed using Akaike's information criterion (24). To build a classifier on continuous variables, we used a MANOVA model, and the location model was used to combine both categorical and continuous variables (25).

For Review Only

RESULTS

Genomic Profiles of Primary Ocular Tumors

Minimal regions were detected using the whole primary ocular tumors dataset. Partial or complete loss of chromosome 3 was found in the majority of the tumors (60/86, 70%). Among the 6 tumors showing a partial loss of the chromosome, the deletion involved the 3p only in 3 cases, the long arm only in 2 cases, and both arms in one case. These cases with partial deletions were considered as monosomic 3 tumors in the study. The smallest region of deletion was observed in a case showing a terminal 3p loss, beyond clone RP11-34L16. Result of CGH on Affymetrix 250K GeneChip showed a breakpoint in 3p25.3, between positions 8 883 800 and 8 897 506. The two deletions involving the 3q only, were large, distal to 3q11.2. A gain involving at least the distal part of the long arm of chromosome 8, band q24, was the most frequent imbalance (77/86, 90%). The frequency of gain of the individual clones has been computed for chromosome arm 8q on the On the basis of the SC profiles, the frequency of gain of individual clones of the 8q arm decreases regularly when approaching the centromere (8q11–q21.1, 52/86, 60%). Indeed, two types of 8q status can be defined, whether the tumor shows a 8q gain with a proximal breakpoint (denoted below as type A), located from centromere to 8q21.1, or not (type B). So, type A corresponds to gain of the whole or almost the whole 8q, whereas type B is related to gain of distal 8q, of a whole chromosome 8, or to no alteration of chromosome 8. Other highly recurrent regions were also found, namely 6p gain (6p25–p22, 49/86, 58%), 1p loss (1p36–p12, 39/86, 45%) and 16q loss (16q23–q24, 27/86, 31%).

The hierarchical clustering (Fig. 1a) performed using the MR profiles clearly shows, after manual verification, that chromosome 3 status defines two groups of tumors, one with disomy 3 (group 1) and the other with monosomy 3 (group 2), respectively composed of 26 and 60 cases (Table 1). Mean LR profiles were generated by computing for each clone the mean value of the log-ratio in the tumors of a given group. Group 1 mean LR profile (Fig. 2A) shows gains of 6p and of distal 8q. Mean LR profile of group 2 (Fig. 2B) is well characterized by the loss of the whole chromosome 3 associated with the gain of the entire 8q.

The observation of the clustering dendrogram (Fig. 1A) and the minimal regions shared, lead to define two disomic 3 subgroups (1a and 1b), and three monosomic 3 subgroups (2a, 2b, and 2c) (Table 1). Subgroup 1a (12 tumors) shows a 6p gain only, and subgroup 1b (14 tumors) is mainly defined by 6p gain associated with a loss of 6q (9/14, 64%), and by a gain of distal 8q (13/14, 93%; mean log-ratio of 1.5). Loss of 16q appears less often (5/14, 36%). Among the three monosomic 3 subgroups, subgroup 2a is composed

of 8 tumors without any other minimal region than monosomy 3. In subgroup 2b (27 tumors), monosomy 3 is associated with a relatively high level gain of the 8q (26/27, 96%; mean log-ratio of 2.0), mainly of the whole arm, a loss of 8p (16/27, 59%), and a loss of 16q (11/27, 42%). Subgroup 2c (25 tumors) is characterized by a set of minimal regions composed of monosomy 3, a moderate gain of 8q (21/25, 84%; mean log-ratio of 1.6), a loss of 1p (17/25, 68%), a loss of 16q (11/25, 44%), and a rare loss of 8p (6/25, 24%). Concerning the breakpoints in 8q (Table 2), most tumors of subgroup 1b belong to type B (10/14, 71%), whereas tumors of group 2b and 2c are of type A (39/52, 75%) (chi-square test, P = 3.49e-3).

Genomic Profiles of Liver Metastases

The liver metastases dataset was processed using the same procedure of clustering analysis (Fig. 1B). Minimal regions reported previously in ocular tumors were also found in liver metastases: monosomy 3 (51/66, 77%), gains of 8q (59/66, 89%), and 6p (21/66, 32%), losses of 1p (31/66, 47%), 8p (30/66, 45%) and 16q (21/66, 32%). Two new frequent imbalances, gain of 1q (23/66, 35%) and loss of 6q (42/66, 64%), were observed. The mean log-ratio of 8q gain was high (log-ratio of 1.9). Among monosomic 3 metastases, 6 cases showed a partial deletion of chromosome 3, 4 in the short arm only, and two in both arms. These 3p deletions were large, with a minimal region in 3p26-22.

Except three samples showing numerous imbalances that could not be classified, liver metastases could be separated into the same groups and subgroups as defined in ocular tumors (Table 1). The 12 disomic 3 samples are composed of subgroups 1a (2/12, 17%) and, predominantly, 1b (10/12, 83%). The 51 monosomic 3 metastases are distributed in subgroups 2a (2/51, 4%), 2b (28/51, 55%), and 2c (21/51, 41%). Thus, most of monosomic 3 metastases belong to subgroups 2b and 2c (49/51, 96%).

Regarding breakpoints on 8q, most of the liver metastases belong to type A (50/63, 79%) (Table 2).

Genomic profiles of Paired Primary Tumors and Metastases

All these cases corresponded to group 2 tumors. Some imbalances were recurrently found as additional alterations in metastases by comparison with the corresponding ocular tumors, such as gain of 1q (3/8) and loss of 6q (3/8). One metastasis shows 11 additional copy number changes, mainly gain of whole chromosomes, in comparison with the primary tumor.

Comparison of Genomic Profiles of Ocular Tumors with Respect to Metastatic Status

For Review Only

On the whole, group 2 tumors show a higher metastatic potential (38/55, 69%) than group 1 tumors (5/23, 22%; chi-square test *P* = 3.37e-4) (Table 1). Among tumors with partial loss of chromosome 3, two of the 3 tumors with 3p loss, and one of the two with 3q loss, were associated with metastasis. Metastastic tumors significantly show in addition to a chromosome 3 loss a gain of the whole 8q (Table 2), with a type A breakpoint (33/43, 77%), whereas non-metastatic tumors present type B breakpoints (30/35, 86%; *P* = 1.4e-7). We have examined separately in monosomic and disomic 3 tumors, the eventual differences in chromosome imbalances according to the metastatic status.

Monosomic 3 tumors

We compared profiles of the 38 monosomic 3 primary tumors having lead to the development of liver metastases to those of the 17 non-metastatic ones. Metastastic tumors predominantly show a gain of the whole 8q (Table 2), with a type A breakpoint (32/38, 84%), frequently associated with a loss of 8p (18/38, 47%). Conversely, non-metastatic tumors show a balanced distribution of 8q breakpoints (7 type A and 10 type B), and the loss of 8p is rare (2/17, 12%). Thus, metastatic tumors specifically exhibit type A breakpoints (chi-square *P* = 3.4e-3). A second change concerns chromosome 16, metastatic tumors showing frequent losses of 16q (22/38, 58%), which are not frequently observed in the non-metastatic ones (3/17, 18%). Finally, gain of 6p is more frequently associated with non-metastatic (7/17, 41%) than with metastatic tumors (8/38, 21%). Except gain of 6p (*P* = 2.2e-1), all other chromosome alterations, 8q gain, 8p loss, and 16q loss, are significantly associated with metastatic tumors in comparison with non metastatic ones (*P* = 3.3e-2, 2.5e-2, and 1.3e-2, respectively).

Using a differential analysis based on the LR profiles, chromosome arms 8p and 8q are detected as showing respectively lower and higher ratios in the metastatic tumors.

Disomic 3 tumors

There were only 5 metastatic disomic 3 tumors in our dataset (Table 1), and they showed no specific alterations which could separate them from the 18 disomic non-metastatic tumors.

Metastatic monosomic 3 tumors vs. monosomic 3 metastases

Using SC profiles, frequencies of alterations were compared in the 51 monosomic 3 liver metastases and in the 38 monosomic 3 metastatic ocular tumors. They shared the same imbalances, such as losses of 1p (40%) and 8p (90%), and gain of 8q (50%), with close frequency rates as shown on their respective mean LR profiles (Fig. 2C and Fig. 2D). Few differences exist for regions 1q and 6q which are respectively gained (44% vs. 18% in ocular tumors, *P* = 1.2e-3) and lost (60% vs 28% in ocular tumors, *P* = 1.0e-2) in

For Review Only

metastases. Differential analysis showed no chromosomal regions significantly highlighted, proving that the levels of gains and losses are close in these two groups.

Determination of a High-risk Profile in Ocular Tumors

In univariate analysis, each minimal regions reported in Table 3 was assessed individually and ranked according to its prediction performance. Gain of 8q and monosomy 3 are the most significant variables, with respectively 74.4% and 73.1% of good classification, with balanced sensitivity and specificity. Multivariate analysis performed by adding the best remaining variables one at a time, led to better prediction performances. The best performance, 79.5% with 81.4% of sensitivity and 77.1% of specificity, is obtained with the set of the following 5 minimal regions: monosomy 3, gains of 6p and of 8q, and losses of 8p and 16q.

We also included the breakpoints position on 8q (types A and B) as a new categorical variable in the model, as it appeared as a characteristic feature between metastatic and non-metastatic tumors. In univariate analysis, prediction performance of this new variable is better (78.2%) than any previous regions taken individually. Then, we replaced the variable gain of 8q in our previous set of 5 minimal regions by this new variable and improved the performances to 82.1% of good classification, with very balanced specificity (83.3%) and sensitivity (80.6%), and very close positive and negative predictive values (83.7% and 80.6%). We then applied variable selection on the full model, considering all minimal regions of interest and the breakpoint position type, in order to remove non-significant variables. All variables were selected. The model was tested on the metastases dataset, and a performance of 75.8% (50/66) of good classification was obtained. We found, as expected, the two cases belonging to subgroups 1a and 1b, and the 3 cases unclassified because of their numerous alterations, falsely predicted as non-metastatic.

Finally, we introduced the mean LR after breakpoint on chromosome 8q. It took the mean LR from the breakpoint on the 8q arm to the telomere, or the mean LR value of the whole chromosome 8 if there was no breakpoint on the 8q. In a univariate analysis, performance of good classification is as good as our predictive model of 5 minimal regions (79.5%), with close sensitivity and specificity. This variable was added to our best model, but performance of classification did not improved, remaining at 82.1%.

Discussion

We present a pangenomic array-CGH analysis performed on the largest series of uveal melanoma

For Review Only

tumors ever reported, and also, for the first time, on a series of liver metastases. A set of 8 paired ocular tumors and the corresponding metastases could be studied.

In ocular tumors, as known for more than a decade, the most frequent imbalances are monosomy 3 and 8q gain. Among the 86 cases analyzed, a partial chromosome 3 deletion was found in 6 cases, with a smallest region of deletion of 8.9 Mb, spanning 3p25.3-pter. This breakpoint, more distal than the one found by Parrella et al. (26), matches with the proximal one of case M16397 of Tschentscher et al. (27), which had the minimal 3p deletion in their series. Among the 6 patients carrier of a tumor with partial monosomy 3, three were metastatic, two with a loss in 3p, and one in 3q. This leads to think that genes important for the prognosis are located in both arms of chromosome 3.

Unsupervised clustering shows that chromosome 3 status is a stable variable that allows to identify two groups: disomic (group 1) and monosomic 3 (group 2) tumors. Group 1 is characterized by gain of 6p as most frequent initial imbalance. The same clustering into two main genomic groups is reported by Hughes et al. (16) and Ehlers et al. (17), from array-CGH analyses performed on 18 and 49 primary uveal melanoma, respectively. This classification is in agreement with the almost mutually exclusive relationship between monosomy 3 and gain of 6p noticed by Parella et al. (28) and Ehlers et al. (17), and their model of tumor progression. Hierarchical clustering leads to define, within these two groups, subgroups based on gain of 8q, mainly, and on other highly recurrent alterations involving 1p, 8p, and 16q losses. Ehlers et al. (17) describe a third group with a normal status for chromosomes 3 and 6p, and associated with the best prognosis.

The examination of 8q gains showed a discrete variation of breakpoints leading either to a gain of the whole arm, or to a distal gain. Type A breakpoints, located close to the centromere, leading to a gain of the whole 8q, is mostly found in monosomic 3 tumors, as observed also by Hughes et al. (16) and Ehlers et al. (17). These whole 8q gains, often associated with 8p loss, are related to the presence and the frequent duplication of isochromosomes 8q, an additional abnormality well-known in karyotypic studies (5, 12, 29). On the opposite, type B refers to a breakpoint distal to 8q21.1, or to an absence of breakpoint (gain of an entire chromosome 8). Most breakpoints of disomic 3 tumors belong to this type. This suggests that gains of 8q would mainly result from unbalanced translocations in group 1, and from isochromosome formation in group 2 tumors.

When genomic profiles of ocular tumors are compared with the status of the patients, group 2 tumors show a higher metastatic potential than group 1 tumors. However, interestingly, none of the 8 tumors with isolated monosomy 3 only, (subgroup 2a) lead to metastasis (Table 1). So, metastastatic propensity appears

For Review Only

only partially explained by chromosome 3 status. Indeed, subgroups with higher metastatic potential (2b, 2c, and to a lesser extent, 1b) can be identified. Metastatic evolution appears as associated with 8q gain, both in monosomic and disomic 3 tumors. Breakpoints in 8q preferentially belong to type A in most metastatic tumors, and to type B in non-metastatic ones (Table 2). This is probably related to the fact that isochromosome 8q is a frequent additional imbalance both in monosomic and in isodisomic 3 tumors, an anomaly that has been demonstrated in about 16 % of group 2 tumors (30, 31). Differential analysis and frequency comparison confirmed that, beside the status of chromosome 3, the main differences between metastatic and non-metastatic tumors are 8q gain and 8p loss, making them high-risk indicators. The low number of metastatic disomic 3 tumors prevented us to valuably compare their profiles to those of the non-metastatic disomic 3 tumors.

In the liver metastasis dataset, except for 3 samples showing highly altered profiles that could not be classified, all subgroups recognized in primary ocular tumors were found. Although mostly monosomic 3 profiles (group 2) are observed, and with a higher frequency than in ocular tumors, disomic 3 (group 1) profiles are also found (21% of the cases) (Table 1). Six metastases showed a partial loss of chromosome 3, 4 of them in the 3p, and two involving both arms. Most metastases belong to the two monosomy 3 subgroups with a gain of 8q (2b, 2c). These gains correspond mainly to type A breakpoints (Table 2). Two samples only show a monosomy 3 (subgroup 2a) and two a 6p gain (subgroup 1a) as isolated imbalances, confirming that these two groups are rarely metastatic. Ten metastases (15%) belong to subgroup 1b, which shows an intermediate metastatic rate. By comparison to ocular tumors, liver metastases specifically show additional gain of 1q and loss of 6q in 44% and 60% of the samples, respectively. The study of the 8 pairs of ocular tumors and their liver metastases shows very similar results, with a recurrent gain of 1q and a loss of 6q in metastases, in comparison with their respective primary tumor.

The classifier built in this work aims to predict the prognosis of any individual tumor, by examining a set of a few minimal regions of interest (Table 3). In univariate analysis, gain of 8q and monosomy 3 show better predictions for metastatic tumors, as they present higher positive predictive values than negative predictive values. In multivariate analysis, best rate of classification (79.5%) is obtained when combining a set of 5 regions (losses of chromosome 3, 8p, and 16q, and gain of 6p and 8q). In contrast with the observation of Kilic et al. (32), loss of 1p was not retained in the classifier. Using the 8q breakpoint position alone, we obtained 78.2% of good prediction, which is better than in univariate analysis of any of the regions. Moreover, when replacing gain of 8q by the breakpoint position in our set of 5 regions, we improved the

For Review Only

classification rate to 82.1% of good classification, with balanced specificity and sensitivity, and very close positive and predictive values.

Several features of our analysis suggested that the log-ratio of 8q gain could be a pertinent prognostic indicator. First, it was apparent that high level 8q gain was present in the subgroups of ocular tumors with the highest metastatic rates. Secondly, differential analysis result showed that monosomic metastatic tumors presented a higher 8q gain than monosomic non-metastatic tumors. Finally, in our liver metastases analysis, gain of 8q with a high log-ratio was reported as the major alteration. However, although good results were obtained in univariate analysis (79.5%), introducing the mean LR after breakpoint on chromosome 8q did not improved the performance of classification (82.1%).

The present study should provide useful information, in addition to clinico-pathological features, for designing an optimal strategy for identifying high-risk ocular tumors in a clinical setting. Given the relatively high frequency of partial deletions of chromosome 3 disclosed by pangenomic and microsatellite analyses, FISH tests should be performed with a probe located in the most recurrent minimum region of loss in chromosome 3, 3p25.1-p25.2 (24, 25, and present work), and at least, a 8q probe of region q11-q21.1, in order to differentiate types A and B breakpoints. However, with the advent of fine-needle biopsies, it is likely that DNA or RNA-based techniques will be more reliable and more adaptable to the analysis of a large number of small samples. It has been shown that expression profiling is more sensitive and specific than genomic profiling for the prognostic evaluation of tumors (15, 33), but it is foreseeable that some samples will yield RNA of insufficient quality for the analysis, and its applicability on individual tumors in the clinical setting remains to be determined. The reason of the superiority of expression profiling is probably in part its ability to better classify tumors with isodisomy 3. This leads to recommend genome profiling and assessment of chromosome 3 allelic status as diagnostic strategy for an optimal prognostic evaluation of uveal melanoma tumors for future clinical trials.

For Review Only

REFERENCES

1. Diener-West M, Hauwkins BS, Markowitz JA, et al. A review of mortality from choroidal melanoma. II. A meta-analysis of 5-year mortality rates following enucleation, 1966 through 1988. Arch Ophthalmol 1992;110:245-50.
2. Damato B. Developments in the management of uveal melanoma. Clin Experiment Ophtalmol 2004;32:639-47.
3. Mudhar HS, Parsons MA, Sisley K, et al. A critical appraisal of the prognostic and predictive factors for uveal malignant melanoma. Histopathology 2004;45:1-12
4. Prescher G, Bornfeld N, Hirche H, et al. Prognostic implications of monosomy 3 in uveal melanoma. Lancet 1996 347:1222-5.
5. Sisley K, Rennie I, Parsons M, et al. Abnormalities of chromosomes 3 and 8 in posterior uveal melanoma correlate with prognosis. Genes Chromosomes Cancer 1997;19:22-8.
6. White VA, Chambers JD, Courtright PD, et al. Correlation of cytogenetic abnormalities with the outcome of patients with uveal melanoma. Cancer. 1998;83:354-9.
7. Damato B, Duke C, Coupland SE, et al. Cytogenetics of uveal melanoma: a 7-year clinical experience.Ophthalmology. 2007;114:1925-31.
8. Gordon K, Thompson C, Char D, et al. Comparative Genomic Hybridization in the detection of DNA copy number abnormalities in uveal melanoma. Cancer Res 1994;54: 4764-8.
9. Speicher M, Prescher G, du Manoir S, et al. Chromosomal gains and losses in uveal melanomas detected by Comparative Genomic Hybridization. Cancer Res 1994;54:3817-23.
10. Aalto Y, Eriksson L, Seregard S, et al. Concomitant loss of chromosome 3 and whole arm losses and gains of chromosome 1, 6, or 8 in metastasizing primary uveal melanoma. Invest Ophthalmol Vis Sci 2001;42:313-7.
11. Sisley K, Parsons MA, Garnham J, et al. Association of specific chromosome alterations with tumour phenotype in posterior uveal melanoma. Br J Cancer 2000;82:330-8.
12. Kilic E, van Gils W, Lodder E, et al. Clinical and cytogenetic analyses in uveal melanoma. Invest Ophthalmol Vis Sci 2006;47: 3703-7.
13. Tschentscher F, Husing J, Holter, et al. Tumor classification based on gene expression profiling shows that uveal melanomas with and without monosomy 3 represent two distinct entities. Cancer Res 2003;63:2578-84.

For Review Only

14. Onken M, Worley L, Ehlers J, et al. Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death. Cancer Res 2004;64:7205-9.
15. Petrausch U, Martus P, Tönnies H, et al. Significance of gene expression analysis in uveal melanoma in comparison to standard risk factors for risk assessment of subsequent metastases. Eye 2007; 1-11.
16. Hughes S, Damato BE, Giddings I, et al. Microarray comparative genomic hybridisation analysis of intraocular uveal melanomas identifies distinctive imbalances associated with loss of chromosome 3. Br J Cancer 2005;93:1191-6.
17. Ehlers JP, Worley L, Onken MD, et al. Integrative genomic analysis of aneuploidy in uveal melanoma. Clin Cancer Res 2008;14:115-22.
18. Idbaih A, Marie Y, Lucchesi C, et al. BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas. Int J Cancer 2008;122:1778-86.
19. Neuvial P, Hupé P, Brito I, et al. Spatial normalization of array-CGH data. BMC Bioinformatics 2006;7:264.
20. Hupé P, Stransky N, Thiery JP, et al. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 2004;20:3413-22.
21. Rouveirol C, Stransky N, Hupé P, et al. Computation of recurrent minimal genomic alterations from array-CGH data. Bioinformatics 2006;22:849-56.
22. Benjamini Y and Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 1995;57:289-300.
23. La Rosa P, Viara E, Hupé P, et al. VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. Bioinformatics 2006;22:2066-73.
24. Sakamoto Y: Efficient use of Akaike's information criterion for model selection in high dimensional contingency table analysis. Metron 1982;40:257-75.
25. Daudin JJ: Selection of variables in mixed-variable discriminant analysis. Biometrics 1986;42:473-481.
26. Parrella P, Fazio VM, Gallo AP, et al. Fine mapping of chromosome 3 in uveal melanoma: identification of a minimal region of deletion on chromosomal arm 3p25.1-p25.2. Cancer Res 2003;63:8507-10.

For Review Only

27. Tschentscher F, Prescher G, Horsman DE, et al. Partial deletions of the long and short arm of chromosome 3 point to two tumor suppressor genes in uveal melanoma. Cancer Res 2001;61:3439-42.

28. Parrella P, Sidransky D and Merbs S: Allelotype of posterior uveal melanoma: implications for a bifurcated tumor progression pathway. Cancer Res 1999;59:3032-7.

29. Prescher G, Bornfeld N, Friedrichs W, et al. Cytogenetics of twelve cases of uveal melanoma and patterns of nonrandom anomalies and isochromosome formation. Cancer Genet Cytogenet 1995;80:40-46.

30. White VA, McNeil BK and Horsman DE: Acquired homozygosity (isodisomy) of chromosome 3 in uveal melanoma. Cancer Genet and Cytogenet 1998;102:40-45.

31. Onken MD, Worley LA, Person E, et al. Loss of heterozygosity of chromosome 3 detected with single nucleotide polymorphisms is superior to monosomy 3 for predicting metastasis in uveal melanoma. Clinical Cancer Research 2007;13:2923-7.

32. Kilic E, Naus C, van Gils W, et al. Concurrent loss of chromosome arm 1p and chromosome 3 predicts a decreased disease-free survival in uveal melanoma. Invest Ophthalmol Vis Sci 2005;46:2253-7

33. Worley LA, Onken MD, Person E, et al. Transcriptomic versus chromosomal prognostic markers and clinical outcome in uveal melanoma. Clin Cancer Res 2007;13:1466-71.

For Review Only

Table 1. Tumor subgroups in ocular tumors and liver metastases defined from the clustering on 7 minimal regions, and rate of metastatic patients in the different subgroups of primary tumors.

		Ocular tumors		Liver metastases
Subgroups	**Imbalances**	**Frequency** *n = 86*	**Metastases rate** *n = 78**	**Frequency** *n = 63*
1a	G6p	14% (12)	0% (0/10)	3% (2)
1b	G6p, L6q, G8q, L16q	16% (14)	38% (5/13)	16% (10)
2a	L3	9% (8)	0% (0/7)	3% (2)
2b	L3, L8p, G8q, L16q	31% (27)	84% (21/25)	44% (28)
2c	L1p, L3, L8p, G8q, L16q	29% (25)	74% (17/23)	33% (21)

*sample of patients with a minimum follow-up of 24 months.
G: gain, L: loss.

Table 2. Frequency of 8q gains with a proximal breakpoint (type A, q11.1-q21.1) in each subgroup of ocular tumors and liver metastases.

Ocular tumors *(n = 78)*	
Metastatic tumors	77% (33/43)
Non metastatic tumors	14% (5/35)
Metastatic monosomic 3 tumors	84% (32/38)
Non metastatic monosomic 3 tumors	41% (7/17)
1a	0% (0/12)
1b	29% (4/14)
2a	0% (0/8)
2b	81% (22/27)
2c	68% (17/25)

Metastases *(n = 63)*	
1a	0% (0/2)
1b	70% (7/10)
2a	0% (0/2)
2b	93% (26/28)
2c	81% (17/21)

For Review Only

Table 3. Supervised prognostic classification of ocular tumors.
Classifiers built using a log-linear model and prediction performances assessed using leave-one-out cross-validation.
Variables used in the models are the minimal regions, 8q breakpoint position type, and 8q log-ratio after breakpoint.

Variables	Performance of classification	Specificity	Sensitivity	Positive predictive value	Negative predictive value
Minimal regions					
Univariate analysis					
L1p (p36-p12)	55.1%	55.1%	0%	100%	0%
L6q	55.1%	55.1%	0%	100%	0%
L16q (q23-q24)	64.1%	80.0%	56.6%	46.5%	85.7%
G6p (p25-p22)	67.9%	70.5%	64.7%	72.1%	62.9%
L8p (p23-p11)	66.7%	90.5%	57.9%	44.2%	94.3%
L3	73.1%	69.6%	81.8%	90.7%	51.4%
G8q (8q11.1-q21.1)	74.4%	73.5%	75.9%	83.7%	62.9%
Minimal regions					
Multivariate analysis					
G8q, L3	76.9%	79.1%	74.3%	79.1%	74.3%
G8q, L3, G6p	75.6 %	73.1%	80.8%	88.4%	60.0%
G8q, L3, G6p, L8p	78.2%	79.5%	76.5%	81.4%	74.3%
G8q, L3, G6p, L8p, L16q	79.5%	81.4%	77.1%	81.4%	77.1%
8q breakpoint					
and minimal regions					
8q breakpoint type A	78.2%	79.5%	76.5%	81.4%	74.3%
L3, L8p, L16q, G6p, 8q breakpoint type A	82.1%	83.3%	80.6%	83.3%	80.6%
8q log-ratio after breakpoint	79.5%	84.6%	74.4%	76.7%	82.9%
L3, L8p, L16q, G6p, 8q breakpoint type A, 8q log-ratio after breakpoint	82.1%	83.3%	80.6%	83.3%	80.6%

G: gain, L: loss.

For Review Only

FIGURES

Figure 1. Hierarchical clustering (Ward method with Euclidean distance) of 86 ocular primary tumors (A) and 66 liver metastases (B).

A- Two groups (1 and 2) of tumors, characterized by the status of chromosome 3, and 5 subgroups (1a, 1b, 2a, 2b and 2c) can be defined, on the basis of imbalances of minimal regions, mainly gains of 6p and 8q, and losses of 1p, 8p, and 16q. Each tumor corresponds to a row and abscissa corresponds to the chromosomes lined up from 1 to Y. Regions gained, highly represented (log-ratios >3), lost, or normal, are in red, blue, green, and yellow, respectively. The dendrogram resulting from the clustering is shown on the right.

B- The same groups and subgroups are recognized, but with different frequencies and, on average, more altered profiles (see text and Table 1).

Figure 2. Mean Log-Ratio profiles of the 26 ocular disomic 3 tumors (A), the 60 monosomic 3 tumors (B), the 38 monosomic 3 metastatic ocular tumors (C) and the 53 monosomic 3 liver metastases (D). Abscissa corresponds to the chromosomes lined up from 1 to Y, and ordinate is the log-ratio between tumor and control DNAs. Major chromosomal alterations are pointed by arrows (G: gain, L: loss).

For ReView Only

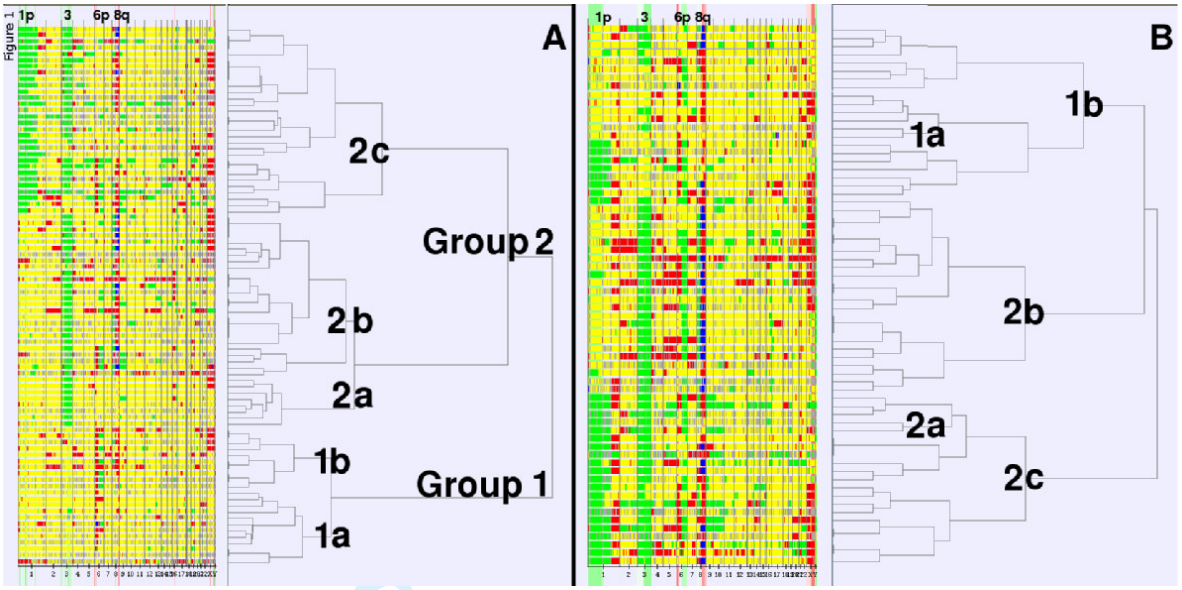


Figure 1

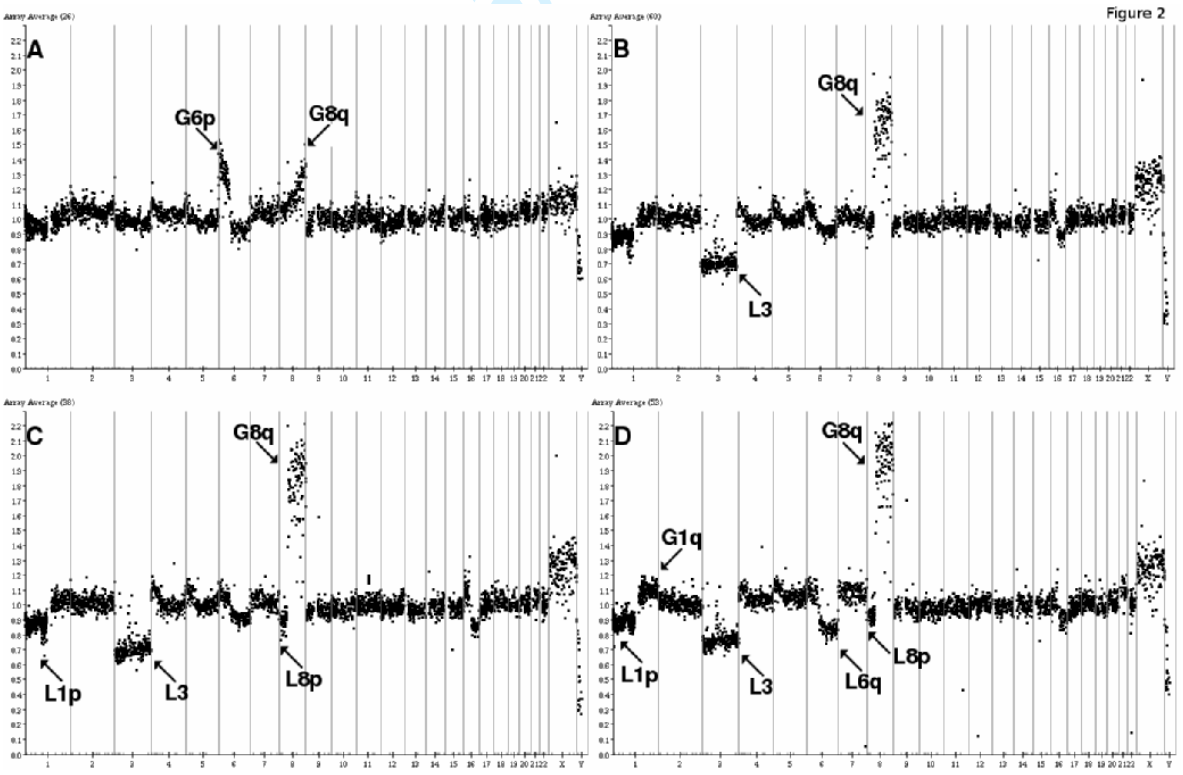


Figure 2

2.6 Tools, software and database for DNA copy number microarray experiments

This section presents the work which has be done to provide the statistical methods which have been developed during the thesis with an enhanced value.

2.6.1 R packages

R[1] (R Development Core Team, 2008) is an open source statistical language which has been used to implement the statistical methods previously described. Packages have been built and are available from Bioconductor[2] (Gentleman et al., 2004). Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data; it is based on R. The three R packages are the following:

MANOR (Neuvial et al., 2006) The normalisation method described in **Section 2.1** is available in the MicroArray NORmalisation (MANOR) R package. MANOR is integrated into CAPweb which is an analysis platform developed at Institut Curie (see **Subsection 2.6.3**). MANOR has also been integrated by a bioinformatics team into their toolkit PerlMAT devoted to the management and analysis of microarray data (Morris et al., 2008).

GLAD (Hupé et al., 2004) The method devoted to the analysis of aCGH profile described in **Section 2.2** is available in the Gain and Loss Analysis of DNA (GLAD) R package. GLAD is also integrated into CAPweb. GLAD has also been integrated by other bioinformatics teams into their web interfaces such as ADaCGH (Díaz-Uriarte and Rueda, 2007) and ISACGH (Conde et al., 2007).

ITALICS (Rigaill et al., 2008) The method devoted to the analysis of Affymetrix array described in **Section 2.3** is available in the ITerative and Alternative normaLIsation and Copy number calling for affymetrix Snp arrays (ITALICS) R package. ITALICS is integrated into CAPweb.

2.6.2 VAMP software

A graphical user interface named Visualisation and Analysis of Molecular Profiles (VAMP) has been developed for the visualisation and first level analysis of molecular profiles (DNA copy number, mRNA, LOH, *etc.*) (La Rosa et al., 2006, this article is supplied in the **Annexes**). VAMP is integrated into CAPweb and has been distributed in several academic institutions.

2.6.3 CAPweb platform

CGH Array Pipeline on the web (CAPweb) is a user-friendly tool enabling biologists to analyse aCGH from raw data to visualisation and biological interpretation (Liva et al., 2006, this article is supplied in the **Annexes**). With CAPweb it is possible to manage the data, to normalise the aCGH data with MANOR, to detect breakpoints with GLAD, to analyse Affymetrix data with ITALICS, to visualise and analyse the genomic profiles with VAMP. CAPweb is used at Institut Curie via the intranet server (more than 10000 DNA copy number experiments have been analysed) and is used by the scientific community via the internet

[1] http://www.r-project.org

[2] http://www.bioconductor.org

server (more than 3000 DNA copy number experiments have been analysed). CAPweb has been distributed in several academic research labs (about 15) and sold to a private company.

CAPweb has been used in different clinical publications: Vermeer et al. (2008), Unger et al. (2008), Fuhrmann et al. (2008), Idbaih et al. (2008) and Idbaih et al. (2007).

2.6.4 ACTuDB database

Array-CGH Tumour DataBase (ACTuDB) (Hupé et al., 2007, this article is supplied in the **Annexes**) compiles DNA copy number microarray experiments from articles which have made their data publicly available. The data have been integrated into a unified bioinformatics environment using GLAD and VAMP. ACTuDB provides biologists with the possibility to compare their findings to the existing sets of DNA copy number data for validation.

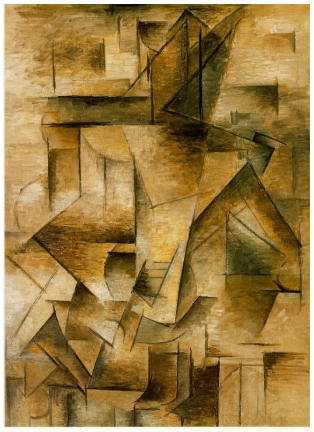
2.6.5 Clinical applications of the tools and software

The tools and software which have been developed during the thesis have first been used for research purposes. Now, they are routinely used in clinical practise at Institut Curie by the Département de Biologie des Tumeurs (for sarcomas, neuroblastoma and uveal melanoma) and at Hôpital de la Pitié Salpétrière (for gliomas).

2.7 Conclusion

In this chapter, the set of biostatistical algorithms and bioinformatics tools developed during the thesis have been described. They are devoted to the analysis of DNA copy number microarray experiments. The work has been carried out as a close collaboration between biologists, clinicians and bioinformaticians within Institut Curie, and as a close collaboration of the bioinformaticians within the Service de Bioinformatique as well. A lot of attention has been paid to packaging each method so that it can be used by the bioinformatics community. Our methods are available as R packages from the Bioconductor project website. However, some packages still require computer skills which are not suitable for biologist end-users. This is the reason why our methods have been integrated into a user-friendly unified bioinformatics environment named CAPweb so that every biologist can use it without particular skills. As a result, many biological or clinical publications from teams belonging or not to Institut Curie refer to our environment for data preprocessing and analysis. We have also successfully applied our statistical methods to predict high-risk tumours in uveal melanoma (Trolet et al., 2008, under revision in *Investigative Ophthalmology and Visual Science*).

The field of microarray technology is evolving quickly. In the first chapter we have seen that new chips are available allowing a better resolution to identify more precise DNA copy number alterations. Then, the biostatistical algorithms and bioinformatics tools we have developed potentially need to be improved to take into account new issues due to this increasing resolution. Moreover, new biostatistical methods will be needed to cope with emerging biotechnologies. This will be the case for the ultra-high throughput sequencing which is very likely to replace microarray technologies within the next five years. Typically, DNA copy number will be quantified using high-throughput sequencing. This new sequencing technology generates unprecedented amounts of data which definitely imply huge computer and bioinformatics skills to provide the biologists with a comprehensible information. Huge amount of works will be needed and new biostatistical algorithms will have to be developed to take into account the specificities of this technology. At the time of writing the manuscript, Institut Curie has just received an ultra-high throughput sequencer which will make it possible to query still unexplored biological territories. This definitely starts a new era in biology and bioinformatics and will give new insights to further improve the knowledge in oncology.



The guitar player
Pablo Picasso, *1910*

For every complex question there is a simple and wrong solution.

Albert Einstein

3

Prediction of the clinical phenotype based on both mRNA expression and DNA copy number microarray experiments

Contents

We have seen in **Chapter 1** that a large variety of molecular profiles can be available for each patient. They represent valuable information to identify new reliable and efficient biomarkers useful for clinical purposes. Among the molecular profiles, mRNA expression and DNA copy number have widely been used so far. When available for the same patient, it is natural to combine both profiles to improve the identification of new predictive and prognostic factors. The biological rational behind the combination is very simple: there are direct or indirect relations between DNA copy number and mRNA expression. Indeed, the more copies of a gene there are, the more likely the gene is expressed at a high level. Methods based on correlation analysis have been developed to identify such relations (Lee et al., 2008). Chin et al. (2006) have combined both levels of information in survival analysis

but, to our knowledge, no statistical method has been proposed in the context of supervised classification. We therefore propose the following approach:

- The informative DNA copy number alterations are extracted using the methodology described in **Section 2.4** (Rouveirol et al., 2006). This way, each DNA copy number molecular profile is summarised as shown in **Figure 3.1**: the set of profiles is represented within an indicator matrix which gives whether or not a tumour sample has a given informative alteration.

- Regarding the mRNA data, the expression values of each gene are used after normalisation by a classical method such as RMA (Irizarry et al., 2003) or CG-RMA (Wu et al., 2003) (see **Figure 3.2**).

The two issues of using both mRNA expression and DNA copy number microarray experiments in supervised classification are the following. Firstly, the DNA copy number data in **Figure 3.1** are discrete predictor variables while the mRNA expression data in **Figure 3.2** are continuous predictor variables. Therefore, the predictor variables are called mixed since they consist of two different variable types. A supervised classification method able to handle such mixed predictor variables is needed. Secondly, the mRNA data are high-dimensional data and therefore, the supervised classification method must also be able to handle high-dimensionality. These two issues are raised in this chapter. The outline is the following. In the first sections, the principles of supervised classification, the methods able to handle mixed variables and the issue of high-dimensionality are introduced. Then, a first contribution of the thesis describes a method able to handle mixed variables in high-dimensionality contexts. The problem of model selection is then introduced. A second contribution of the thesis describes a statistical criterion able to test the efficiency of the proposed supervised method.

	Gain Chr1	Loss Chr2	Gain Chr8	Amplification Chr8
Tumour1	0	0	1	0
Tumour2	1	0	0	1
Tumour3	1	0	1	1
Tumour4	0	1	0	0
Tumour5	0	0	0	1

Figure 3.1: Representation of DNA copy number data - Each column represents an informative DNA copy number alteration (either a minimal or recurrent alteration). A row represents a tumour sample in which it is indicated whether the sample has the corresponding alteration (1) or not (0).

	Gene1	Gene2	Gene3	Gene4
Tumour1	2.5	12.1	8.9	4.6
Tumour2	4.6	13.6	9.8	6.5
Tumour3	7.7	10.7	6.3	3.8
Tumour4	8.1	9.8	7.2	6.0
Tumour5	5.3	14.7	5.4	5.1

Figure 3.2: Representation of mRNA expression data - Each column represents a gene and the corresponding mRNA expression value is given for each tumour sample represented in rows.

3.1 Back to basics: supervised classification

Supervised classification also named *discriminant analysis* was first described in 1936 in the famous *Fisher's iris* example. It is a statistical technique which allows the study of the differences between two or more classes (for example, a clinical phenotype to predict can be the low- and high-risk metastasis classes of patients) based on the observation of different variables (such as the mRNA expression and DNA copy number data). The goal is therefore to predict the class label of a new sample for which the different variables have been observed. The different variables are often referred to as the *predictors* and the class label (also termed dependent variable or class variable) as the *outcome*. Predictor variables can be either discrete or continuous data. Supervised classification builds a *prediction rule* or *classifier* which is used for future predictions. It can be compared to multiple regression except that the outcome to predict is discrete. In practice, supervised classification methodology proceeds in two steps: firstly, the classifier is built over a *training set* and its prediction performance is computed over a *test set*. It is therefore a learning procedure.

In statistical decision theory, the supervised classification problem has been formalised as follows. We assume that for each observation i a set of predictor variables $\mathbf{S}_i$ in some space $\mathcal{S}$ has been observed. Moreover, the class label $Y_i \in \mathcal{G}$ is known where $\mathcal{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_K\}$ is the set of possible K classes. The classifier is defined as a function $\mathcal{C} : \mathcal{S} \rightarrow \mathcal{G}$. The prediction will be wrong if $\mathcal{C}(\mathbf{S}_i) \neq Y_i$. The function $\mathcal{C}$ is built so that it ensures the highest prediction performance or equivalently so that it minimises the error rate. The performance of the classifier is quantified with a *loss function* $L(Y_i, \mathcal{C}(\mathbf{S}_i))$ for penalising errors in prediction. In supervised classification, the loss function can be represented by a $K \times K$ matrix $\mathbf{L}$ where $K = card(\mathcal{G})$. $\mathbf{L}$ will be zero on the diagonal and nonnegative elsewhere, where $L(k, l)$ is the price paid for classifying an observation belonging to class $\mathcal{G}_k$ as $\mathcal{G}_l$. Most often the *zero-one* loss function is used where all misclassifications are charged a single unit. Each observation i can be viewed as random variable $(\mathbf{S}_i, Y_i)$ from the joint distribution $P(\mathbf{S}, Y)$. The *risk function* R for a classifier $\mathcal{C}$ is the expected loss under $P(\mathbf{S}, Y)$:

$$R(\mathcal{C}, L) = E[L(Y, \mathcal{C}(\mathbf{S}))]$$

When the zero-one loss function L_{0-1} is used then the risk of the classifier $R(\mathcal{C}, L_{0-1})$ is simply the *misclassification rate* $P(\mathcal{C}(\mathbf{S}) \neq Y)$. The supervised classification aims at finding the $\mathcal{C}$ function so that $R(\mathcal{C}, L_{0-1})$ is minimal. Such a function $\mathcal{C}^\star$ always exists. It is named the *Bayes classifier*:

$$\mathcal{C}^\star(\mathbf{S}_i) = \arg\min_{g \in \mathcal{G}} [1 - P(g|\mathbf{S} = \mathbf{S}_i)]$$

The function $\mathcal{C}^\star$ can be derived from the Bayes theorem:

$$P(Y_i = \mathcal{G}_k|\mathbf{S}_i) = \frac{p_k f_k(\mathbf{S}_i)}{\sum_{l=1}^{K} p_l f_l(\mathbf{S}_i)} \tag{3.1}$$

where f_k is the probability density function of $\mathbf{S}$ given the class label $\mathcal{G}_k$ and p_k is the prior probability of class $\mathcal{G}_k$.

In practice, the probability density functions f_k and the prior probabilities p_k are unknown. Therefore, they need to be estimated. The data within the training set which are used for this task are required to be independent and identically distributed. The goal of

the supervised classification is to build a classifier $\mathcal{C}$ which is as close as possible to the optimal Bayes classifier. Many supervised classification techniques are available and they all aim at approximating the Bayes classifier. Among the most widely used techniques let us mention *linear and quadratic discriminant analysis*, *nearest neighbours*, *support vector machine*, *classification trees*, *etc.* (for a review see Dudoit and Fridlyand, 2003; Hastie et al., 2003; Larrañaga et al., 2006).

An important issue in supervised classification is the prediction performance assessment. Indeed, the misclassification rate of the classifier is unknown and an estimation must be provided. An efficient prediction performance is defined as a low misclassification rate or a correct classification rate. In the manuscript, the prediction performance will be defined as the correct classification rate. Different methods have been proposed to assess the prediction performance. The main ones are v-fold cross-validation, Monte-Carlo cross-validation and *bootstrap (see MacLachlan, 1992; Boulesteix et al., 2008b). A commonly used form of cross-validation is Leave-One-Out (LOO) cross-validation where v is set at the number of observations. LOO often results in low bias but high variance estimators of the misclassification rate. In genomic studies where sample size is often small, Molinaro et al. (2005) have shown that, in this case, LOO, 10-fold cross-validation and bootstrap gave efficient estimations.

Probabilistic discrimination with a Gaussian model This paragraph describes a probabilistic method which is a standard in supervised classification. When the predictor variables are continuous data it is often assumed that they are normally distributed with mean row-vector $\boldsymbol{\mu}_k$ and variance-covariance matrix $\boldsymbol{\Sigma}_k$, conditionally on the class k. The probability density function f_k is:

$$f_k(\mathbf{S}_i) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{S}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{S}_i - \boldsymbol{\mu}_k)'\right]$$

where p is the number of continuous predictor variables.

As already mentioned, the classifier is derived from the Bayes theorem (see **Equation 3.1**) in which the denominator value is the same whatever the class k. Therefore, in the case of binary classification which is the most frequent situation, the prediction rule can be expressed from the logarithm of the numerator of **Equation 3.1**. A new observation i will be predicted to belong to class 1 if:

$$2\log p_1 - \log|\boldsymbol{\Sigma}_1| - (\mathbf{S}_i - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}_1^{-1}(\mathbf{S}_i - \boldsymbol{\mu}_1)' > 2\log p_2 - \log|\boldsymbol{\Sigma}_2| - (\mathbf{S}_i - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}_2^{-1}(\mathbf{S}_i - \boldsymbol{\mu}_2)' \quad (3.2)$$

In the case where the variance-covariance matrices are supposed to be identical for the two classes ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$), the prediction rule can be simplified and a new observation i will be predicted to belong to class 1 if:

$$\mathbf{S}_i\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' - \log\frac{p_2}{p_1} > 0 \quad (3.3)$$

In the general case, the prediction rule is *quadratic* with respect to $\mathbf{S}_i$ (see **Equation 3.2**) whereas in the case where the variance-covariance matrices are identical, the prediction rule is *linear* with respect to $\mathbf{S}_i$ (see **Equation 3.3**). We speak about *Quadratic Discriminant Analysis (QDA)* and *Linear Discriminant Analysis (LDA)* respectively. In practice, the mean vectors and variance-covariance matrices are unknown and need to be estimated. As it can be seen in the prediction rule, the variance-covariance matrices need to be inverted, which is problematic when the matrices are singular. This happens when the number of predictors is greater than the number of observations which is always the case with genomic data. In such cases, Regularised Discriminant Analysis (RDA) can be used as an alternative (see Hastie et al., 2003) which makes it possible to modulate the heterogeneity of the class variance-covariance matrices and to modulate the dependency between variables adding a constant to the diagonal of the matrices.

While the standard probabilistic discrimination with a Gaussian model is only able to handle continuous predictor variables, other methods have been developed to take into account both continuous and discrete predictor variables. They are described in the next section.

3.2 Supervised classification with mixed variables

This section describes supervised classification methods which are able to handle both continuous and discrete predictor variables.

3.2.1 Classification trees

A classification tree is a non *parametric method in which the classifier is based on a series of successive binary questions. In oncology, this approach is widely used by clinicians to define the most appropriate treatment for each patient or to predict the prognosis (Nadal et al., 1988): the decision is chosen based on the answers to several questions. Each question, or node, splits the space of predictors $\mathcal{S}$ into two descendant subsets, starting with $\mathcal{S}$ itself. Each terminal subset is assigned a class label and the resulting partition of $\mathcal{S}$ corresponds to the classifier. Different classification approaches exist and here we will describe Classification and Regression Tree (CART) (Breiman et al., 1984). The tree construction consists of two main steps:

1 - node splitting rule: for each node, a question needs to be defined. Therefore, a predictor variable and a threshold have to be chosen if the variable is continuous, or if the variable is discrete, a set of modalities has to be chosen. For example, Nadal et al. (1988) proposed a classification tree to predict the prognosis of women with breast cancer: the first question is based on a discrete variable which asks *Is there a metastasis in the liver?* and subsequent questions are based on thresholds over continuous variables such as *Is the albumin value greater than 30?* (see **Figure 3.3**). The splitting rule is defined so that the data in the descendant subsets are *purer* than the data in the parent subset. This is quantified with an *impurity function* I which depends on the proportion of each class $(p_1, \cdots, p_K)$ within each subset. The function I must be maximal when the p_k are equal and zero when there is only one class within the subset. The most widely used measure of impurity is the *Gini index*: for a node N, $I(N) = 1 - \sum_k p_k^2$. In each node N, the p_k are estimated by the relative proportion of class k. For a given splitting rule, a proportion p_R of the observations are sent to the right daughter node N_R and a proportion p_L to the left daughter node N_L. The optimal split will be chosen so that the *decrease in impurity* $\Delta(N) = I(N) - p_R I(N_R) - p_L I(N_L)$ is optimal.

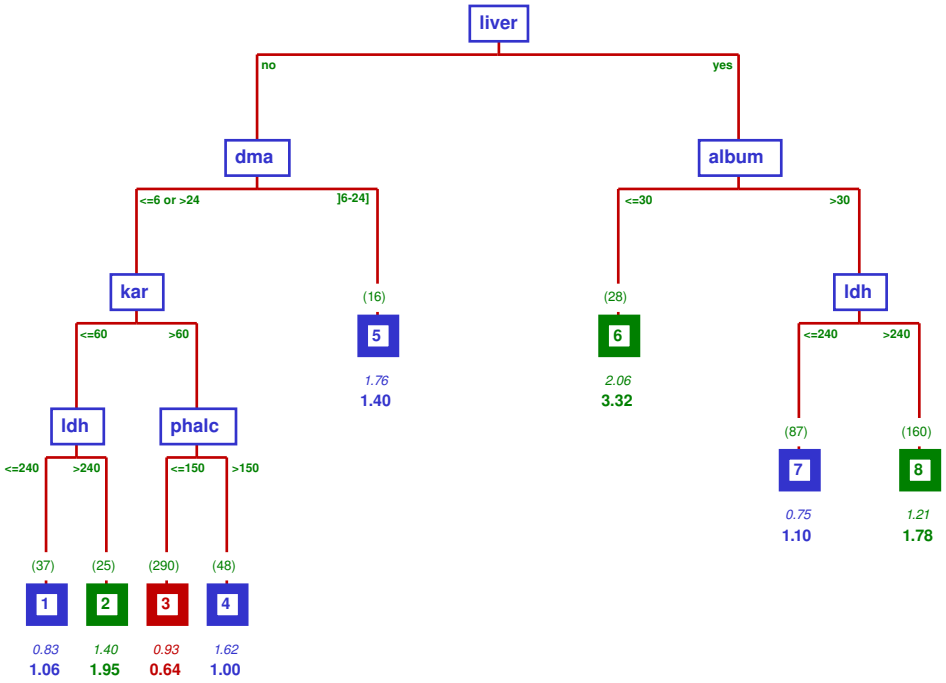


Figure 3.3: Application of classification trees to predict prognosis in breast cancer - Three prognosis classes are defined: good prognosis, bad prognosis and very bad prognosis. The prediction rule is defined from series of questions based on the presence or absence of metastasis in the liver (discrete variable) and some thresholds on biological parameters (album, ldh, phalc) or clinical parameters (dma, kar) (continuous variables). *liver*: presence of liver metastasis - *dma*: delay of metastasis appearance - *kar*: Karnofsky index - *album*: albumin level - *ldh*: lactate deshygrogenase level - *phalc*: alkaline phosphatase level (from Nadal et al., 1988) (Image provided by Yann De Rycke and Bernard Asselain, Institut Curie).

2 - tree pruning: an optimal subtree is then selected from the complete tree T_{max} to ensure efficient prediction performance. Indeed, the complete tree is able to correctly predict all the observations within the training set but these correct classifications correspond to *overfitting. The tree needs to be pruned to increase its generalisation performance on an independent dataset. The tree is therefore pruned using a penalty criterion which takes into account the complexity of the tree (*i.e.* the number of nodes) and the misclassification rate. The penalty criterion is:

$$L_{0-1}(T_i) + \alpha Card(T_i), \text{ with } Card(T_i) \text{ the number of nodes in } T_i$$

During this procedure the series of nested subtrees is considered $T_1 \supset T_2 \supset \cdots \supset T_{max}$. The subtree which optimises the penalty criterion is selected as the best classifier. In practice, the α value needs to be tuned by cross-validation. If $\alpha = 0$ the maximal tree T_{max} will be selected and if α is very large the subtree T_1 with only one node will be selected.

Random forest Breiman (2001) proposed to combine different classification trees to improve the prediction performance. The way the set of trees is built is based on a random selection of both the training samples and the predictor variables. One classification tree is built from a subsample bootstrap from the whole training set similarly to *bagging technique. Then, the tree is built using the CART methodology with slight modifications. Instead of choosing the best node splitting rule from the whole set of predictor variables, only a random subset of predictor variables is used. For each tree, no pruning is performed. The number of trees to build and the size of the random subset of predictor variables need to be chosen by

the user. A new observation is assigned to the class which is the most frequently designed over all the classification trees. In this approach Breiman (2001) has shown that the generalisation error for forests converges to a limit as the number of trees in the forest becomes large. In the case of high-dimensional data, this algorithm is well adapted since the search of the node splitting rule is reduced over a small subset of variables.

3.2.2 Logistic regression

Logistic regression belongs to the category of Generalised Linear Model (GLM) which is a generalisation of least-squares regression. In a GLM, each outcome of the dependent variables $\mathbf{Y}$ is assumed to be generated from a particular distribution function in the exponential family. The mean $\boldsymbol{\mu}$ of the distribution depends on the predictor variables $\mathbf{X}$, through the relation $E(\mathbf{Y}) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ where g is called the *link function*. In this section, $\mathbf{X}$ denotes the design matrix in which a continuous predictor variable is represented by its values and a discrete predictor variable by a dummy submatrix of zero or one. Therefore, the design matrix $\mathbf{X}$ can combine continuous and discrete predictor variables as in usual linear models. The vector $\boldsymbol{\beta}$ corresponds to the parameters of the model. In binary classification, the vector $\mathbf{Y}$ is simply the vector of class label for each observation and each outcome is assumed to follow a Bernoulli distribution. In this case the natural link function is the *logit* function:

$$g(p) = \frac{p}{1-p}$$

For the observation i its probability p_i of belonging to the class 1 is expressed as:

$$p_i = \frac{e^{\mathbf{X}^i\boldsymbol{\beta}}}{1 + e^{\mathbf{X}^i\boldsymbol{\beta}}}$$

where $\mathbf{X}^i$ is the i-th row of the design matrix $\mathbf{X}$ and represents the predictor variables for the observation i.

The value of the parameters inside the vector $\boldsymbol{\beta}$ have a direct interpretation in terms of *odds-ratio* which is widely used in medical science. An odds-ratio close to one means that there is no association between the outcome and the predictor variable while a smaller or larger value than one indicates a relation between the outcome and the predictor variable.

3.2.3 DISQUAL

The DISQUAL method (DIScriminant analysis with QUAlitative variables) has been proposed by Saporta (1990) to build classifiers from discrete predictor variables. As we will see, this method can be used as a preliminary step to combine both discrete and continuous predictor variables. The procedure consists of two steps: a Multiple Correspondence Analysis (MCA) is followed by a Linear Discriminant Analysis (LDA).

MCA is an extension of correspondence analysis which allows the analysis of relationships between discrete variables. As such, it can also be seen as a generalisation of Principal Component Analysis (PCA). The output of the MCA are factorial axes which are continuous descriptors and therefore these axes can be used as inputs for the LDA. In the case where both continuous and discrete predictor variables are available it is possible to compute the factorial axes from the discrete variables and combine these axes as well as the continuous variables within a LDA.

3.2.4 Location model (LM)

The Location Model (LM) is a statistical method which has been designed to handle data with mixed variables. Daudin (1986) proposed a general Multivariate Analysis of Variance (MANOVA) log-linear formulation of the LM. This model makes it possible to take into account the additivity which is due to the discrete predictor variables: for a given class, the discrete predictor variables define subclasses in which the continuous predictor variables have a specific mean value. For example, in the context of a breast cancer study let us assume that there is a *HER2/neu* amplification in some tumours (see **Chapter 1 - Page 26**). Then, the amplification modifies the expression of some genes (and especially *HER2/neu* which is overexpressed) whatever the class label is. According to our notations, the amplification is the discrete predictor variable (and has two modalities: *amplification* or *no amplification*) and the genes are the continuous predictor variables. Moreover, some interactions might exist between one class and some discrete predictor variables. The LM is able to capture both additivity and interactions. This is illustrated in **Figure 3.4** in which the expression values of gene1 versus gene2 are plotted for class 1 tumours (•) and class 2 tumours (×) having either an amplification (blue) or no amplification (red). The following parameters have been used to simulate the data in **Figure 3.4**:

- Inside each subclass the variance-covariance matrix has the following value:

$$\begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

- **In the additive case**, the means of each subclass are (1,1), (2,2), (3,3) and (4,4) for •, •, × and × respectively. The gene expression values are shifted by the same translation $\mathbf{a}$ whatever the class label due to the amplification effect (**Figure 3.4**, left graphic).

- **In the interaction case**, the means of each subclass are (1,1), (2,2), (3,3) and (7,5) for •, •, × and × respectively. The gene expression values are shifted both in class 1 and class 2 due to the amplification effect but the translation is not the same for the two classes: it is $\mathbf{a}$ in class 1 and $\mathbf{a}+\mathbf{i}$ in class 2 (**Figure 3.4**, right graphic). This interaction between subclasses is also named *reversal*.

Besides additivity and interaction, the LM is able to take into account the frequency of each subclass. Typically, if the frequency of amplification is higher in class 1 than in class 2, then it represents valuable information for the prediction.

Mathematical formulation of the LM

Let $\mathcal{G}_1, \cdots, \mathcal{G}_K$ be K classes, $\mathbf{Y}$ the vector of class labels, $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_q)$ be q discrete predictor variables and $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_p)$ be p continuous predictor variables. The LM consists of two distinct parts:

Part 1: the MANOVA model The conditional distribution of $\mathbf{X}$ in class k and for $\mathbf{Z} = \mathbf{z}$ is assumed to be normal with the following mean vector and variance-covariance matrix:

$$E(\mathbf{X}|\mathcal{G}_k, \mathbf{Z} = \mathbf{z}) = \boldsymbol{\mu}_{zk} \tag{3.4}$$
$$V(\mathbf{X}|\mathcal{G}_k, \mathbf{Z} = \mathbf{z}) = \boldsymbol{\Sigma} \tag{3.5}$$

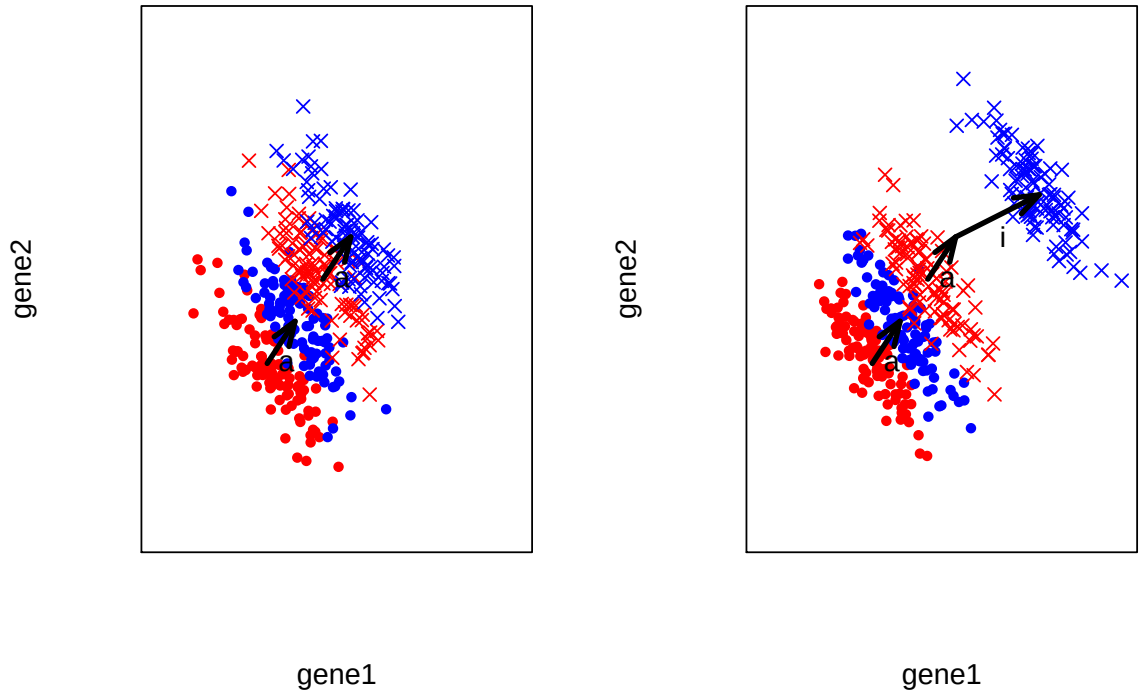


Figure 3.4: Artificial example for the LM - • represent class 1 and × represent class 2. There are 100 observations in each subclass. In red are represented the tumours without the amplification and in blue the tumours with the amplification. The expression values of gene1 versus gene2 are plotted for each tumours. In the left graphic is illustrated an example of additivity: the gene expression values are shifted by the same translation $\mathbf{a}$ whatever the class label is due to the amplification effect. In the right graphic is illustrated an example of interaction: the gene expression values are shifted both in class 1 and class 2 due to the amplification effect but the translation is not the same for the two classes (it is $\mathbf{a}$ in class 1 and $\mathbf{a}+\mathbf{i}$ in class 2).

On the random continuous predictor variables $\mathbf{X}$, a MANOVA model is assumed which takes into account the class effect and the effects associated with the discrete predictor variables $\mathbf{Z}$. Moreover, the model can take into account interactions between the class effect and the discrete predictor variables as already mentioned. The class variable and the discrete predictor variables are transformed into an indicator design matrix $\mathbf{D}$ (as it is classically done in analysis of variance) which includes at least the main effects (*i.e.* the class and the discrete predictor variables) and their interactions if necessary. The parameters of the MANOVA model are obtained by usual maximum likelihood estimates.

Part 2: the log-linear model The discrete predictor variables and the class variable can be represented within a contingency table. The probability $P(\mathbf{Z}=\mathbf{z}|\mathcal{G}_k) = p(\mathbf{z}|k)$ is estimated using a log-linear model. The parameters of the log-linear model are obtained by usual maximum likelihood estimates.

Complete formulation of the LM Combining the MANOVA and the log-linear model gives the complete formulation of the LM. In this complete formulation, the probability density function f_k in class k is:

$$f_k(\mathbf{x},\mathbf{z}) \quad = \quad P(\mathbf{Z}=\mathbf{z}|\mathcal{G}_k)\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{zk})\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{zk})'\right]$$

where $\mathbf{x}$ and $\mathbf{z}$ are the observed continuous and discrete predictor variables respectively.

The prediction rule is the following:

$$\begin{aligned} \mathcal{C}(\mathbf{x}, \mathbf{z}) &= \arg\max_{\mathcal{G}_k \in \mathcal{G}} (\mathbf{x} - \boldsymbol{\mu}_{zk})' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{zk} + \log p(\mathbf{z}|k) \quad (3.6) \\ &= \mathrm{LM}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \quad (3.7) \end{aligned}$$

The LM can be viewed as a generalisation of the LDA in the sense that the information of the discrete predictor variables are taken into account to improve the prediction.

LM versus the other methods

The present manuscript focuses on the LM to build classifiers from mixed variables. Although it is somewhat arbitrary, we can nevertheless motivate our choice with the following reasons:

- the LM is a generalisation of the linear discriminant analysis which has been widely used in a large variety of classification tasks with high efficiency.
- the normality of the continuous variables which is a reasonable assumption for mRNA expression data after log-transformation makes the LM optimal with respect to other methods which are non-parametric.
- in the favourable case in which the data can be perfectly separated the logistic regression fails to converge.

Application of the LM on mRNA expression and DNA copy number data

We propose to apply the LM combining the data from mRNA expression and DNA copy number molecular profiles. Following the previous notations, $\mathbf{Z}$ is given by the data in **Figure 3.1** and $\mathbf{X}$ by the data in **Figure 3.2**.

As already mentioned for LDA and QDA (see **Equation 3.3** and **Equation 3.2**), the prediction rule defined by the LM requires a variance-covariance matrix to be inverted (see **Equation 3.6**). Due to the high-dimensionality of the data, the variance-covariance matrix $\boldsymbol{\Sigma}$ is singular since the number of continuous predictor variables is greater than the number of observations. Therefore, the methodology cannot be applied straight forward. The data need to be preprocessed in order to reduce their dimensionality. This issue is raised in the following section.

3.3 *Curse of dimensionality*: a need for dimension reduction

The *curse of dimensionality* is a term coined by Richard Bellman to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to a mathematical space (Donoho, 2000). In classical statistical methodology, a convenient situation appears when there are many observations (n) and few variables (p). The trend today is towards more observations and especially to radically larger numbers of variables. This is especially the case in the field of genomic studies in which a classical situation is a hundred of observations and several thousands or even tens or hundreds of thousands of

predictor variables. This situation is often referred to as the *small n, large p* issue in the field of statistical learning. Why is this situation an issue? Firstly, the classical methods such as LDA have not been designed to cope with this high-dimensionality problem. Secondly, although we might think that the increase in variable number may help in the discriminating task, this is in practice not the case for the following reasons:

- among the available predictor variables, many of them are irrelevant to distinguish between the different classes and introduce noise in the classifier building procedure. Hence the prediction performance of the model is decreased.
- the risk of overfitting the data is high and especially in the case where the sample size is small. This is the rule so far in cancer study and this aspect needs to be taken into account seriously as pointed out by Ransohoff (2004) in its *Rules of evidence*.

In microarray data analysis, the high-dimensional data spaces we work on is a real issue and different approaches have been proposed to handle this problem as reviewed by Wang et al. (2008). Indeed, in statistical learning tasks, a rule of thumb is to have at least 10 training samples per feature dimension whereas in microarrays this ratio is often closer to 0.01 samples per dimension (Wang et al., 2008). To overcome this issue of high dimensionality of the data, different techniques can be applied to reduce their complexity. The different approaches can be split into three categories which are described in the following subsection.

3.3.1 Techniques to reduce the complexity of the data

1 - variable selection identifies a small subset from the original predictor variables which is used to build the classifier. This technique has widely been used in the machine learning community and Guyon and Elisseeff (2003) give an overview of this issue. Briefly, the variable selection techniques can be split into three categories: *(i) filters* select subsets of variables as a pre-processing step, independently of the supervised classification method chosen, *(ii) wrappers* utilise the supervised classification method of interest as a black box to score subsets of variable according to their predictive power, *(iii) embedded methods* perform variable selection in the process of training and are usually specific to a given supervised classification method. Saeys et al. (2007) give a review of the application of variable selection in the field of bioinformatics such as sequence analysis, microarray analysis, mass spectra analysis. Due to the large number of predictor variables, an exhaustive search of the best subset of predictor variables is impossible for computational reasons. Therefore, simple ranking of the variables or heuristics like forward, backward or stepwise selection are often applied.

Another important point to raise in predictor variable selection is the confrontation between the *univariate paradigm* and the *multivariate paradigm*. In univariate approach, each variable is scored individually without taking into account the other variables while in multivariate selection all the variables are considered together and therefore the relation between variables is used. It is often believed that if a variable alone has no discriminating power then it should be removed. This is totally false and Guyon and Elisseeff (2003) give a nice example (the *xor* example) in which two variables taken alone have no discriminating power while taken together they separate the classes of interest perfectly. This is due to the existence of interactions or correlations between variables which is the rule in mRNA expression data. Therefore, we think that methods which fully benefit the correlation structure of the data should be adopted to select the most relevant predictor variables.

Among the motivations for using variable selection as a way to reduce the complexity of the data, some authors mention the fact that high-dimensional data often contain many

redundant variables which also affect the prediction performance of supervised classification methods. They claim that these redundant variables should be eliminated. This is illustrated in the paper by Yu and Liu (2004) who have developed a variable selection strategy based on both *relevance* and *redundancy* concepts showing that discarding irrelevant and redundant variables can improve prediction performance. At this stage, two schools of thought can be opposed regarding redundancy: one school which considers redundancy as a drawback and one school which considers redundancy as an advantage. In this manuscript, we think that methods which fully exploit redundancy lead to efficient prediction performance. We can illustrate our point of view with a simple simulated dataset. Two classes have been simulated according to the following model: in class 1, four independent continuous predictor variables have been randomly generated from a normal distribution $\mathcal{N}(0, -1)$ in class 1 and $\mathcal{N}(0, 1)$ in class 2. The four continuous predictor variables can be seen as a realisation of the same event and therefore share the same information. The number of observations is 1000 in each class. The classifier has been computed with a probabilistic discrimination with a Gaussian model using the four variables (*allvar*), only one of the four variables (*onevar*), the mean vector of the four variables (*meanvar*) and the first component of the PLS algorithm (*plscomp*) (see **Subsection 3.3.2** for details). The prediction performance assessed by LOO is reported in **Table 3.1** for the four methods. The results show that taking only one variable leads to a lower prediction performance (onevar - 85.0%) than taking the mean vector (meanvar - 97.9%) or a linear combination of the four variables (plscomp - 97.9%). This can be easily explained since averaging the four variables or building a linear combination of the four variables reduces the noise and improves the separation between the two classes as it can be seen on **Figure 3.5**. Therefore, methods which are able to handle redundancy in a clever way should lead to improve prediction performance.

2 - variable grouping aims at defining groups of predictor variables which behave similarly. Traditionally, unsupervised classification methods such as hierarchical clustering or k-means are widely used to perform this task. In the cluster building procedure, the standard clustering techniques do not incorporate the class variable to predict. Therefore, this might not be suitable for prediction purposes. Hence, Dettling (2003) has proposed supervised classification methods which group predictor variables by incorporating class variable. Convincing results have been obtained on gene expression microarray data. A first algorithm *Wilma* (Dettling and Bühlmann, 2002) has been developed which partitions the predictor variables into non-overlapping groups. A second algorithm *Pelora* (Dettling and Bühlmann, 2004) improves the previous one allowing overlapping groups of variables. In the framework of gene expression studies, this improvement is biologically motivated since some genes operate in multiple pathways (see **Figure 1.11**). Although both algorithms rely on two different statistical models, the idea behind them remains the same. Briefly, the supervised classification method is a one-step procedure for variable selection, variable grouping and formation of new features by averaging the variable values within the same group of predictor variables. The identification of the groups of variables followed by the averaging of these variables renders the discrimination between the different classes easier as already illustrated in **Figure 3.5**.

Zou and Hastie (2005) proposed a regression method which can be straightforwardly applied to the context of supervised classification. Their method has the intrinsic ability to perform variable grouping. It is a combination of lasso and ridge regression (for details about these two techniques see Hastie et al., 2003) since the regression coefficients are shrunk using both L_1 and L_2 penalties. In cases where there is a group of variables among which the pairwise correlations are very high, the two penalties force all the variables from one group to be selected as soon as a variable of the group is selected. Otherwise, without these

method	performance
allvar	98.0%
meanvar	97.9%
plscomp	97.9%
onevar	85.0%

Table 3.1: Redundancy in simulated data and prediction performance - Prediction performance assessed by LOO on simulated data for four different classifiers. The classifier has been computed with a probabilistic discrimination with a Gaussian model using the four variables (*allvar*), only one of the four variables (*onevar*), the mean vector of the four variables (*meanvar*) and the first component of the Partial Least Squares (PLS) algorithm (*plscomp*).

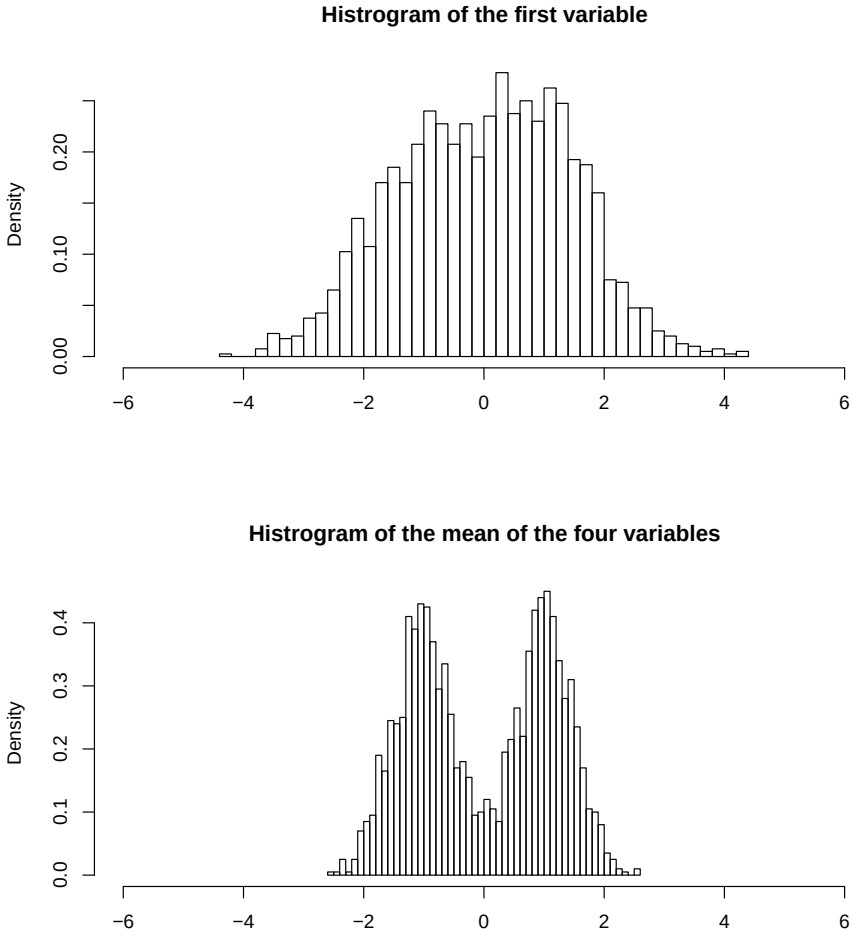


Figure 3.5: Redundancy in simulated data - Distribution of the simulated data for one of the four continuous predictor variables (top) or the mean vector of the four continuous predictor variables (bottom). The four independent continuous predictor variables have been randomly generated from a normal distribution $\mathcal{N}(0,-1)$ and $\mathcal{N}(0,1)$ in class 2.

two penalties, only one variable from the group would be selected by chance and the others discarded. Indeed, as mentioned by Zou and Hastie (2005), the ideal gene selection method should be able to do two things: eliminate the trivial genes and automatically include all the genes from a given group in the model once one gene among the group is selected. The proposed regularisation method named *elastic net* simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables which share identical regression coefficients.

3 - projection of the data onto a subspace is a technique which aims at building new variables named components (or supergenes or metagenes in the framework of microarray studies) from a combination of the original variables. The most popular method is PCA which has widely been used in microarray studies. Díaz-Uriarte (2003) used this technique combined with a variable grouping approach. Another famous method is Independent Component Analysis (ICA) (Comon, 1994): it finds components which are independent and not necessarily orthogonal contrary to the PCA. In the context of supervised classification, these methods may not be suitable since they do not take into account the class variable to predict when building the components. Recently Bair et al. (2006) proposed a supervised version of PCA which can be used for supervised classification purposes. Another popular method which takes into account the outcome variable is Partial Least Squares (PLS) which will be detailed in **Subsection 3.3.2** since this method has been retained to reduce the dimension of the data in the model we propose. These methods are by definition multivariate as they take all the variables at the same time to build the components. Therefore, they are perfectly adapted in the context of high-dimensionality and multicolinearity problems.

Which dimension reduction technique to choose? The different strategies have all shown improving prediction performance. Therefore, there is no clear evidence that one strategy is better than another. Nevertheless, we can motivate our choice with biological arguments rather than statistical ones. Indeed, in biological processes, genes cooperate in modules or pathways (see **Figure 1.11**) and taking into account the genes of the same pathways should improve the ability to separate the different classes and improve robustness of the signatures (*i.e.* the classifier): in different publications in which lists of genes were identified from the same basic scientific question and using similar patient cohorts, very few genes (and possibly none) were found in common between the different signatures (Miller and Liu, 2007). Ein-Dor et al. (2005) and Michiels et al. (2005) have shown that the signatures strongly depended on the selection of patients in the training sets. The main reason for this discrepancy and instability is the correlation between genes and the fact that the selection procedure randomly selects few genes among all the possible relevant genes. Nevertheless, if one looks beyond the genes to the pathways they represent, multiple pathways can be found in common between the signatures, indicating that the signatures and their predictive powers may come from the same underlying biological mechanism. Therefore, in order to fully exploit the fact that there is redundancy in the data we can choose either *variable grouping* or *projection* strategies. These strategies should be able to use all the genes which are relevant for class prediction and involved in the same underlying biological mechanism. Moreover, we also believe that these strategies should ensure efficient prediction performance on independent dataset that will demonstrate their generalisation ability. Once again, the choice is arbitrary and our choice has also been motivated by simplicity. Projection methods are simple algorithms which do not require extensive computation as compared with variable grouping which is more time-consuming. This is why we have decided to focus on the PLS algorithm. The first PLS components will be used as new continuous predictor variables in the LM. The PLS algorithm is described is the next subsection.

3.3.2 Partial Least Squares (PLS)

Multivariate projection methods represent powerful tools to reduce the complexity of the data especially when the dimensionality is extremely high. Among these methods, PCA has been widely used as an exploratory analysis tool to assess the data structure of a matrix $\mathbf{X}$. This technique is not well adapted when the goal is to assess the data structure of the matrix $\mathbf{X}$ (denoted as the predictor variables) with respect to a matrix $\mathbf{Y}$ (denoted as the outcome variable) and particularly in the context of class prediction (Nguyen and Rocke, 2002; Barker and Rayens, 2003). The PLS algorithm has been designed especially to retrieve the relation between $\mathbf{X}$ and $\mathbf{Y}$ using regression by means of projections to latent structures. It was originally developed in the field of econometrics by Hermann Wold (Wold, 1966) and his son Svante made this tool very popular in the field of chemometrics (Wold et al., 1982). PLS derives its usefulness from its ability to analyse data with many, noisy and multicolinear variables and therefore is particularly suitable in the framework of gene expression microarray data analysis. The PLS algorithm was originally designed to analyse data where both $\mathbf{X}$ and $\mathbf{Y}$ are continuous variables. However, it can be directly applied in the context of binary class prediction provided that $\mathbf{Y}$ is encoded by zeros and ones indicating the class label. In supervised classification, the PLS performs efficiently (Barker and Rayens, 2003) and especially in the field of gene expression microarray data analysis (Boulesteix, 2004b, 2006; Boulesteix and Strimmer, 2007; Bøvelstad et al., 2007; Nguyen and Rocke, 2002). Different PLS implementations exist but we will describe the PLS1 algorithm (Wold, 1966; Wold et al., 1982) (details about this algorithm and other implementations can be found in Tenenhaus (1998)).

Description of the algorithm

The goal of the PLS algorithm is to compute K new components $\mathbf{T}_1, \ldots, \mathbf{T}_K$ so that $COV_n(\mathbf{T}_k, \mathbf{Y})_{k=1,\ldots,K}$ are maximal. The algorithm is iterative and is summarised in **Algorithm 1**. Since the PLS algorithm maximises a covariance, it is sensitive to the scaling of the variables. Therefore, in order to give the same weight to each variable, we use the matrices $\mathbf{X}$ and $\mathbf{Y}$ which have been centered and scaled.

Algorithm 1 Partial Least Squares algorithm

K, the number of components
$k = 0$
${}^k\mathbf{X} = \mathbf{X}$
${}^k\mathbf{Y} = \mathbf{Y}$
while $k < K$ **do**
  $\mathbf{W}_{k+1} = \arg\max_{\mathbf{W} \in \mathbb{R}^p, \|\mathbf{W}\|^2 = 1} COV_n({}^k\mathbf{X}\mathbf{W}, {}^k\mathbf{Y})$
  $\mathbf{T}_{k+1} = {}^k\mathbf{X}\mathbf{W}_{k+1}$
  perform the two following regression models:
    ${}^k\mathbf{X} = \mathbf{T}_{k+1}\mathbf{P}'_{k+1} + {}^{k+1}\mathbf{X}$
    ${}^k\mathbf{Y} = \mathbf{T}_{k+1}\mathbf{R}'_{k+1} + {}^{k+1}\mathbf{Y}$
  $k := k + 1$
end while
$\mathbf{T} = (\mathbf{T}_1 \cdots \mathbf{T}_K)$
$\mathbf{W} = (\mathbf{W}_1 \cdots \mathbf{W}_K)$
$\mathbf{P} = (\mathbf{P}_1 \cdots \mathbf{P}_K)$

At each step k of the procedure, the algorithm finds the component $\mathbf{T}_{k+1}$ which is expressed as a linear combination from the columns of the matrix $\mathbf{X}$. $\mathbf{T}_{k+1}$ can be expressed as $\mathbf{T}_{k+1} = {}^k\mathbf{X}\mathbf{W}_{k+1}$. The goal is to find the vector $\mathbf{W}_{k+1}$ so that $COV_n({}^k\mathbf{X}\mathbf{W}_{k+1}, {}^k\mathbf{Y})$ is maximal subject to $\|\mathbf{W}_{k+1}\|^2 = 1$. The maximisation of $COV_n({}^k\mathbf{X}\mathbf{W}_{k+1}, {}^k\mathbf{Y})$ is equivalent to the maximisation of $COV_n^2({}^k\mathbf{X}\mathbf{W}_{k+1}, {}^k\mathbf{Y})$. The data ${}^k\mathbf{X}$ and ${}^k\mathbf{Y}$ being centered, we have $n^2 COV_n^2({}^k\mathbf{X}\mathbf{W}_{k+1}, {}^k\mathbf{Y}) = \mathbf{W}'_{k+1}{}^k\mathbf{X}'^k\mathbf{Y}^k\mathbf{Y}'^k\mathbf{X}\mathbf{W}_{k+1}$. To solve this optimisation problem the Lagrange multiplier method is used:

$$\text{Let } s = \mathbf{W}'_{k+1}{}^k\mathbf{X}'^k\mathbf{Y}^k\mathbf{Y}'^k\mathbf{X}\mathbf{W}_{k+1} - \lambda(\mathbf{W}'_{k+1}\mathbf{W}_{k+1} - 1)$$

Let it be required that the partial derivatives of s with respect to $\mathbf{W}_{k+1}$ and λ vanish:

$$\begin{aligned} \frac{\partial s}{\partial \lambda} &= (\mathbf{W}'_{k+1}\mathbf{W}_{k+1} - 1) = 0 \\ \frac{\partial s}{\partial \mathbf{W}_{k+1}} &= {}^k\mathbf{X}'^k\mathbf{Y}^k\mathbf{Y}'^k\mathbf{X}\mathbf{W}_{k+1} - \lambda\mathbf{W}_{k+1} = 0 \end{aligned}$$

At the vanishing condition the following relation holds:

$${}^k\mathbf{X}'^k\mathbf{Y}^k\mathbf{Y}'^k\mathbf{X}\mathbf{W}_{k+1} = \lambda\mathbf{W}_{k+1}$$

Therefore, $\mathbf{W}_{k+1}$ is the eigenvector associated to the highest eigenvalue λ of ${}^k\mathbf{X}'^k\mathbf{Y}^k\mathbf{Y}'^k\mathbf{X}$. It follows that:

$$COV_n^2(\mathbf{T}_{k+1}, {}^k\mathbf{Y}) = \frac{\lambda}{n^2} \tag{3.8}$$

From **Algorithm 1**, the output of the PLS can be expressed as follows:

$$(\mathbf{T}, \mathbf{W}, \mathbf{P}) = \text{PLS}(\mathbf{X}, \mathbf{Y}, K) \tag{3.9}$$

Mathematical properties of the PLS algorithm

We have seen before that $\mathbf{W}_{k+1}$ is the eigenvector associated to the largest eigenvalue of the matrix ${}^k\mathbf{X}'^k\mathbf{Y}^k\mathbf{Y}'^k\mathbf{X}$. The eigenvalue $\mathbf{W}_{k+1}$ can also be computed from the Singular Value Decomposition (SVD) of the matrix ${}^k\mathbf{X}'^k\mathbf{Y}$. Let $\mathbf{M} = {}^k\mathbf{X}'^k\mathbf{Y}$ in **Theorem 1 (singular value decomposition)**. Since ${}^k\mathbf{X}'^k\mathbf{Y}$ is a column-vector its SVD is:

$$\begin{aligned} \mathbf{U} &= \frac{{}^k\mathbf{X}'^k\mathbf{Y}}{\|{}^k\mathbf{X}'^k\mathbf{Y}\|} \\ \mathbf{S} &= \left\| {}^k\mathbf{X}'^k\mathbf{Y} \right\| \\ \mathbf{V} &= 1 \end{aligned}$$

and therefore we have the following properties:

$$\lambda = \left\| {}^{k}\mathbf{X}'^{k}\mathbf{Y} \right\|^{2} \quad (3.10)$$

$$\mathbf{W}_k = \frac{{}^{k}\mathbf{X}'^{k}\mathbf{Y}}{\|{}^{k}\mathbf{X}'^{k}\mathbf{Y}\|} \quad (3.11)$$

$$\mathbf{T}'_k\mathbf{T}_l = 0, l > k \quad (3.12)$$

$$\mathbf{W}'_k\mathbf{P}_k = 1 \quad (3.13)$$

$$\mathbf{W}'_k\mathbf{X}'_l = \mathbf{0}, l \geq k \quad (3.14)$$

$$\mathbf{W}'_k\mathbf{P}_l = 0, l > k \quad (3.15)$$

$$\mathbf{W}'_k\mathbf{W}_l = 0, l > k \quad (3.16)$$

$$\mathbf{T}'_k\mathbf{X}_l = \mathbf{0}, l \geq k \quad (3.17)$$

$$\mathbf{X}_k = \prod_{i=1}^{k}(\mathbf{I} - \mathbf{W}_i P'_i), k \geq 1 \quad (3.18)$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \quad (3.19)$$

The proofs of equations 3.12 to 3.19 can be found in Tenenhaus (1998). In **Algorithm 1**, the regression model ${}^{k}\mathbf{Y} = \mathbf{T}_{k+1}\mathbf{R}'_{k+1} + {}^{k+1}\mathbf{Y}$ can be ignored and ${}^{k}\mathbf{Y}$ can be left unchanged during the procedure without changing the results (Tenenhaus, 1998).

Connections between PLS and other methods

There are many connections between PLS and other methods. They are the following:

- **Connection between PLS and between-class PCA:** between-class PCA can be viewed as a supervised version of PCA to identify principal components which are able to explain the class separation. Instead of finding the eigenvectors of the total variance-covariance matrix $\mathbf{\Sigma}$, the between-class PCA finds the eigenvectors of the between-class variance-covariance matrix $\mathbf{\Sigma}_B$ (see **Theorem 3 (variance decomposition)**). Boulesteix (2004a) has shown that the first PLS component equals the first between-class principal component.

- **Connection between PLS and LDA:** Boulesteix (2004a) has shown that if $\mathbf{\Sigma}$ is assumed to be of the form $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, then the **Equation 3.3** can be expressed as a function of the first between-class principal component. Therefore, PLS being linked to between-class PCA, PLS is also linked to LDA. Barker and Rayens (2003) have shown that in the general case a relation exists between PLS and LDA.

All these connections between PLS and supervised methods support the use of PLS in the context of supervised classification. In the next section, we present a supervised classification method based on PLS and able to handle mixed variables in the context of high-dimensionality. This is the first contribution of this chapter.

3.4 Contribution 1: the Partial Least Squares Location Model (PLS-LM)

We have seen in **Subsection 3.2.4** that the LM provides a framework to handle both continuous and discrete predictor variables. The limitation to using this model in the context of microarray data is its inefficiency to handle high-dimensionality. In **Subsection 3.3.2**, the PLS method has been shown to be a powerful tool to reduce the dimensionality of the data since it fully exploits the redundancy inside the continuous predictor variables. Therefore, in the case where there are only two classes, we propose the following model to apply the LM in high-dimensionality situations:

1. compute $(\mathbf{W}, \mathbf{T}, \mathbf{P}) = \text{PLS}(\mathbf{X}, \mathbf{Y}, K)$ for a fixed K (see **Equation 3.9**)
2. as a supervised classification method apply $\text{LM}(\mathbf{T}, \mathbf{Y}, \mathbf{Z})$ (see **Equation 3.7**)

where:

$\mathbf{X}$	is the centered and scaled matrix of the continuous predictor variables
$\mathbf{Z}$	is the matrix of discrete predictor variables
$\mathbf{Y}$	is the centered and scaled vector of class labels

This algorithm is termed the Partial Least Squares Location Model (PLS-LM) and its efficiency has been evaluated in both simulated and real data. This is the scope of the two following subsections.

3.4.1 Prediction performance of the PLS-LM on simulated data

In this subsection, prediction performance of the PLS-LM on different simulated datasets are provided. The data have been simulated as described in **Section 3.2.4** and **Figure 3.4**. The data can be viewed as a toy example in which the expression values of gene1 and gene2 (the continuous predictor variables) allow the separation of class 1 tumours (•) with respect to class 2 tumours (×). A discrete predictor variable is introduced into the simulation model: tumours have either an amplification (blue) or no amplification (red). The amplification effect on the continuous predictor variables has been simulated to produce either the additive case or the interaction case. In what follows, C denotes the effect due to the class of the tumour and A denotes the effect due to the amplification. The prediction performance has been estimated using LOO. To compare prediction performance of the different classifiers we used the McNemar test (Dietterich, 1998) (See **Theorem 2 (McNemar's test)**).

Same number of observations in each subclass

In this simulation study, there is the same number of observations in each subclass so that only the MANOVA part of the LM is relevant for the prediction. The parameters of the simulation model are the following:

- Inside each of the four subclasses, gene1 and gene2 are normally distributed with the following variance-covariance matrix:

$$\begin{bmatrix} 1 & -0.2 \\ -0.2 & 1 \end{bmatrix}$$

- The number of observations is 5000 in each subclass.
- **In the additive case**, the means of each subclass are (0.5,0.5), (2,2), (2,2) and (3.5,3.5) for $\bullet$, $\bullet$, $\times$ and $\times$ respectively.
- **In the interaction case**, the means of each subclass are (0.5,0.5), (2,2), (2,2) and (5,2) for $\bullet$, $\bullet$, $\times$ and $\times$ respectively.
- A number of 100 independent and identically distributed random variables $\mathcal{N}(0,1)$ have been added as continuous predictor variables.

Different models have been used in the MANOVA part of the LM. They are the following:

C	only the main effect C is considered
C + A	C and A are considered as main effects
C×A	C and A are considered as main effects and their interaction is added

Results The prediction performance for the LM and the PLS-LM is provided in **Table 3.2(a)** for the additive case and **Table 3.2(b)** for the interaction case. The pairwise comparisons of the prediction performance between the different classifiers are available in **Annexes - Table A.1** and **Annexes - Table A.2**. The contribution of gene1 and gene2 in the PLS components are given in **Table 3.3(a)** and **Table 3.3(b)**. The results are the following:

- In the additive case, LM and PLS-LM give the same prediction performance when only the two discriminative genes are used. When all the continuous predictor variables are considered, the prediction performance tends to be lower than when only the two discriminative genes are considered. In all cases, taking into account the main effect A significantly improves the prediction performance (at least 13.1% improvement). For the PLS-LM classifier, the models based on two components do not improve the prediction performance. This is explained by the fact that the contribution of gene1 and gene2 in the second component is of the same order as the random variables (see **Table 3.3(a)**): all the signal is already captured in the first component.
- In the interaction case, LM and PLS-LM give the same prediction performance when only the two discriminative genes are used and when two components are used in the PLS-LM. This is explained by the fact that the second component still captures signal: the contribution of gene1 and gene2 are greater than the contribution of the random variables for the second component (see **Table 3.3(b)**). The prediction performance is significantly higher in the PLS-LM model with two components than in the PLS-LM model with one component for the reasons we have just mentioned. As in the additive case, taking into account all the continuous predictor variables tends to lower the prediction performance, which is a bit higher in the case of PLS-LM. In all cases, taking into account the main effect A significantly improves the prediction performance (at least 13.9% improvement). Besides the main effect A, adding the interaction C×A still significantly improves the prediction performance (at least 1.2%).

(a) Additive case with the same number of observations in each subclass

predictors	MANOVA	log-linear	LM	PLS-LM (1)	PLS-LM (2)
g	C	-	74.7	74.7	74.7
g	C + A	-	88.1	88.1	88.1
a	C	-	74.6	74.7	74.5
a	C + A	-	88.0	87.8	87.9

(b) Interaction case with the same number of observations in each subclass

predictors	MANOVA	log-linear	LM	PLS-LM (1)	PLS-LM (2)
g	C	-	74.7	74.8	74.7
g	C + A	-	89.7	88.9	89.7
g	C×A	-	91.0	88.9	91.0
a	C	-	74.6	74.6	74.5
a	C + A	-	89.5	88.5	89.2
a	C×A	-	91.0	88.5	90.4

(c) Additive case with a different number of observations in each subclass

predictors	MANOVA	log-linear	LM	PLS-LM (1)	PLS-LM (2)
g	C	-	84.8	84.8	84.8
g	C + A	-	88.5	88.5	88.5
g	-	C×A	70.0	70.0	70.0
g	C×A	C×A	89.6	89.6	89.6
a	C	-	84.7	84.8	84.8
a	C + A	-	88.3	88.1	88.2
a	-	C×A	70.0	70.0	70.0
a	C + A	C×A	89.5	89.4	89.3

(d) Interaction case with a different number of observations in each subclass

predictors	MANOVA	log-linear	LM	PLS-LM (1)	PLS-LM (2)
g	C	-	85.0	84.8	85.0
g	C + A	-	89.9	88.2	89.9
g	C×A	-	91.1	88.2	91.1
g	-	C×A	70.0	70.0	70.0
g	C + A	C×A	91.0	89.4	91.0
g	C×A	C×A	92.0	89.4	92.0
a	C	-	85.0	84.7	85.0
a	C + A	-	89.8	87.9	89.6
a	C×A	-	91.0	87.8	90.6
a	-	C×A	70.0	70.0	70.0
a	C + A	C×A	90.8	89.1	90.7
a	C×A	C×A	91.8	89.1	91.5

Table 3.2: PLS-LM prediction performance - The prediction performance (%) is given for the different models for LM and PLS-LM. Either the two genes (g) or the two genes with the 100 random variables (a) have been used as continuous predictor variables. The number in brackets indicates the number of components used in the PLS-LM. The prediction performance is given for the additive case (a,c) and interaction case (b,d) with the same (a,b) or a different (c,d) number of observations in each subclass.

(a) Additive case with the same number of observations in each subclass

predictor	variable	PLS (1)	PLS (2)
g	gene1	$7.1\ 10^{-1}$	$7.1\ 10^{-1}$
g	gene2	$7.1\ 10^{-1}$	$7.1\ 10^{-1}$
a	gene1	$7.0\ 10^{-1}$	$2.1\ 10^{-2}$
a	gene2	$7.1\ 10^{-1}$	$6.3\ 10^{-2}$
a	min.rand	$3.6\ 10^{-4}$	$1.0\ 10^{-3}$
a	max.rand	$2.4\ 10^{-2}$	$3.1\ 10^{-1}$
a	mean.rand	$7.4\ 10^{-3}$	$7.9\ 10^{-2}$

(b) Interaction case with the same number of observations in each subclass

predictor	variable	PLS (1)	PLS (2)
g	gene1	$8.8\ 10^{-1}$	$5.7\ 10^{-1}$
g	gene2	$4.8\ 10^{-1}$	$8.3\ 10^{-1}$
a	gene1	$8.8\ 10^{-1}$	$5.3\ 10^{-1}$
a	gene2	$4.7\ 10^{-1}$	$7.0\ 10^{-1}$
a	min.rand	$3.9\ 10^{-4}$	$1.0\ 10^{-3}$
a	max.rand	$2.6\ 10^{-2}$	$1.3\ 10^{-1}$
a	mean.rand	$8.1\ 10^{-3}$	$3.8\ 10^{-2}$

(c) Additive case with a different number of observations in each subclass

predictor	variable	PLS (1)	PLS (2)
g	gene1	$7.0\ 10^{-1}$	$7.1\ 10^{-1}$
g	gene2	$7.1\ 10^{-1}$	$7.1\ 10^{-1}$
a	gene1	$7.0\ 10^{-1}$	$1.1\ 10^{-2}$
a	gene2	$7.1\ 10^{-1}$	$1.1\ 10^{-1}$
a	min.rand	$2.4\ 10^{-4}$	$2.9\ 10^{-4}$
a	max.rand	$2.5\ 10^{-2}$	$2.4\ 10^{-1}$
a	mean.rand	$6.1\ 10^{-3}$	$8.3\ 10^{-2}$

(d) Interaction case with a different number of observations in each subclass

predictor	variable	PLS (1)	PLS (2)
g	gene1	$8.6\ 10^{-1}$	$6.1\ 10^{-1}$
g	gene2	$5.1\ 10^{-1}$	$8.0\ 10^{-1}$
a	gene1	$8.6\ 10^{-1}$	$6.0\ 10^{-1}$
a	gene2	$5.1\ 10^{-1}$	$7.1\ 10^{-1}$
a	min.rand	$2.6\ 10^{-4}$	$6.8\ 10^{-4}$
a	max.rand	$2.8\ 10^{-2}$	$9.8\ 10^{-2}$
a	mean.rand	$6.7\ 10^{-3}$	$3.3\ 10^{-2}$

Table 3.3: Variable weights in the PLS - The contribution in absolute value of the different variables is given for either the PLS model computed over two genes (predictor=g) or the PLS model computed over the two genes with the 100 random variables (predictor=a). The minimum (min.rand), maximum (max.rand) and mean (mean.rand) absolute value contribution is given for the random variables. The number in brackets indicates the component number of the PLS. The weights are given for the additive case (a,c) and interaction case (b,d) with the same (a,b) or a different (c,d) number of observations in each subclass.

To sum up the results we can say that:

- Taking into account the main effect A and the interaction between C and A significantly improves the prediction performance.
- The addition of noise within the data tends to lower the prediction performance. This effect is a bit higher in PLS-LM than in LM.
- In the case of interaction, taking into account two components in PLS-LM improves the prediction performance.
- The PLS-LM allows efficient prediction performance.

Different number of observations in each subclass

In this simulation study, there is a different number of observations in each subclass so that both the MANOVA and log-linear parts of the LM are relevant for the prediction. The parameters of the simulation model are the following:

- Inside each of the four subclasses, gene1 and gene2 are normally distributed with the following variance-covariance matrix:

$$\begin{bmatrix} 1 & -0.2 \\ -0.2 & 1 \end{bmatrix}$$

- The number of observations in each subclass is 7000, 3000, 3000 and 7000 in $\bullet$, $\bullet$, $\times$ and $\times$ respectively. Therefore, there is 30% amplification in class 1 tumours and 70% in class 2 tumours.
- **In the additive case**, the mean of each subclass is (0.5,0.5), (2,2), (2,2) and (3.5,3.5) for $\bullet$, $\bullet$, $\times$ and $\times$ respectively.
- **In the interaction case**, the mean of each subclass is (0.5,0.5), (2,2), (2,2) and (5,2) for $\bullet$, $\bullet$, $\times$ and $\times$ respectively.
- A number of 100 independent and identically distributed random variables $\mathcal{N}(0,1)$ have been added as continuous predictor variables.

Different models have been used in the MANOVA part of the LM. They are the following:

C	only the main effect C is considered
C + A	C and A are considered as main effects
C×A	C and A are considered as main effects and their interaction is added

The log-linear model used to estimate the probability of belonging to each subclass is denoted C×A: it takes into account the main effects C and A and their interaction. The log-linear model has been used or not in combination with the MANOVA part of the LM.

Results The prediction performance for the LM and the PLS-LM is provided in **Table 3.2(c)** for the additive case and **Table 3.2(d)** for the interaction case. The pairwise comparisons of the prediction performance between the different classifier are available in **Annexes - Table A.3** and **Annexes - Table A.4**. The contributions of gene1 and gene2 in the PLS components are given in **Table 3.3(c)** and **Table 3.3(d)**. For both the additive and interaction cases, and in classifiers in which no log-linear model is involved, the conclusions are exactly the same as previously. The log-linear model alone gives 70% which is the worst prediction performance (note that the log-linear classifier is necessarily the same whatever the MANOVA model considered). In all cases, adding the log-linear model in the classifier gives significantly higher prediction performance although the improvement is moderate (from 0.8% up to 1.8%).

3.4.2 Prediction performance of the PLS-LM on real data

Breast cancer data from Chin et al. (2006)

The data have been retrieved from the ACTuDB. For 89 patients we have the following data: the outcome (recurrence / no recurrence), the aCGH (BAC array with 2000 probes) and mRNA (Affymetrix HG-U133A chip) profiles. A total of 31 patients out of 89 had a recurrence. The probes whose genes expression value is lower than 5.5 are considered as non-expressed: those probes which are non-expressed for all the patients have been discarded from the analysis (a total of 20244 probes have been kept in the analysis). The minimal alterations of amplification have been computed as described in **Section 2.4**: amplifications have been found on chromosomes 8p11-12, 8q21-24, 11q13, 17q21-22 and 20q13 as reported by Chin et al. (2006). The 8p11-12 amplification was shown by Chin et al. (2006) to be associated with recurrence and therefore for simplification we use only this alteration in what follows. The molecular subtypes defined from the classification by Sørlie et al. (2001) have also been provided by the authors (normal-like, basal-like, erbb2, luminal A and luminal B) and this information has been included in our model. In what follows, C denotes the class effects (recurrence / no recurrence), A denotes the effect due to the amplification on chromosome 8p11-1 and S denotes the effect due to the molecular subtypes.

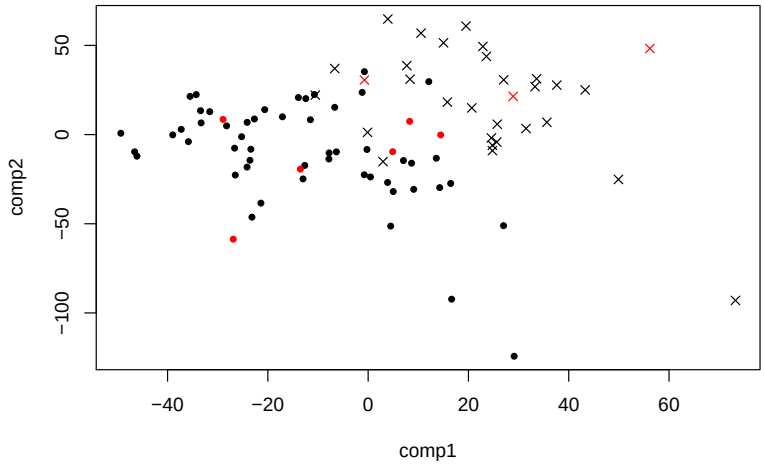
Results The prediction performance of different PLS-LMs is provided in **Table 3.4**. The representation of the data in the first two PLS components is provided in **Figure 3.6**. Whatever the model, including the A effect does not improve the prediction performance. This can be visually explained from **Figure 3.6(a)** where no obvious additive effect due to the A effect is observed. Moreover, very few observations are available to estimate the A effect accurately. Taking into account the S effect improves the prediction performance when only the first PLS component is considered: the prediction performance increases from 57.3% to 61.8%. This can be visually explained from **Figure 3.6(b)** where an additive effect can be observed especially for the basal-like subtype. Combining the MANOVA and log-linear models does not allow improvement of the prediction performance. Note that due to the small amount of data no interaction was taken into account in the MANOVA model.

Bladder cancer data from Stransky et al. (2006)

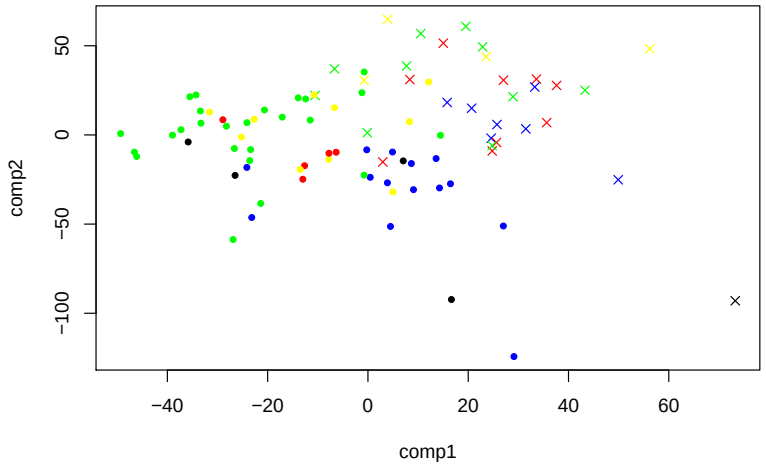
The data have been retrieved from ACTuDB. For 48 patients we have the following data: the stage of the tumour (Ta / T2-T4), the aCGH (BAC array with 2464 probes) and mRNA (Affymetrix HG-U95A/Av2) profiles. A total of 16 patients out of 48 are Ta tumours. All the 8111 probes have been used in the analysis since the non-expressed probes had already been filtered out by the authors. The minimal alterations of amplification have been computed

MANOVA	log-linear	PLS-LM (1)		PLS-LM (2)	
C	-	57.3	*14*	61.8	*16*
C + A	-	57.3	*14*	61.8	*17*
C + S	-	61.8	*14*	62.9	*16*
C + A + S	-	61.8	*14*	62.9	*16*
-	C	65.2	*0*	65.2	*0*
-	C×A	65.2	*0*	65.2	*0*
-	C×S	69.7	*9*	69.7	*9*
-	C×A×S	69.7	*13*	69.7	*13*
C + A	C×A	55.1	*8*	61.8	*14*
C + S	C×S	61.8	*12*	65.2	*14*
C + A + S	C×A×S	62.9	*9*	61.8	*14*

Table 3.4: Prediction performance of the PLS-LM on Chin et al. (2006) - The prediction performance (%) is given for the different models. For each model, the number of correctly predicted recurrences is provided in italic. The number in brackets indicates the component number of the PLS.



(a) Amplification of 8p11-12



(b) Molecular subtypes

Figure 3.6: PLS components for Chin et al. (2006) - • represents patients without recurrence (58) and × represents patients with recurrence (31). The patients are represented in the first two components of the PLS. (a) Patients with an amplification of 8p11-12 are shown in red. (b) Molecular subtypes defined with the classification by Sørlie et al. (2001) are represented: normal-like (black), erbb2 (red), basal-like (blue), luminal A (green) and luminal B (yellow).

MANOVA	log-linear	PLS-LM (1)		PLS-LM (2)	
C	-	85.4	*12*	89.6	*14*
C + A	-	89.6	*13*	89.6	*14*
-	C	66.7	*0*	66.7	*0*
-	C×A	66.7	*0*	66.7	*0*
C + A	C×A	89.6	*13*	89.6	*14*

Table 3.5: Prediction performance of the PLS-LM on Stransky et al. (2006) - The prediction performance (%) is given for the different models. For each model, the number of correctly predicted Ta tumour is provided in italic. The number in brackets indicates the component number of the PLS.

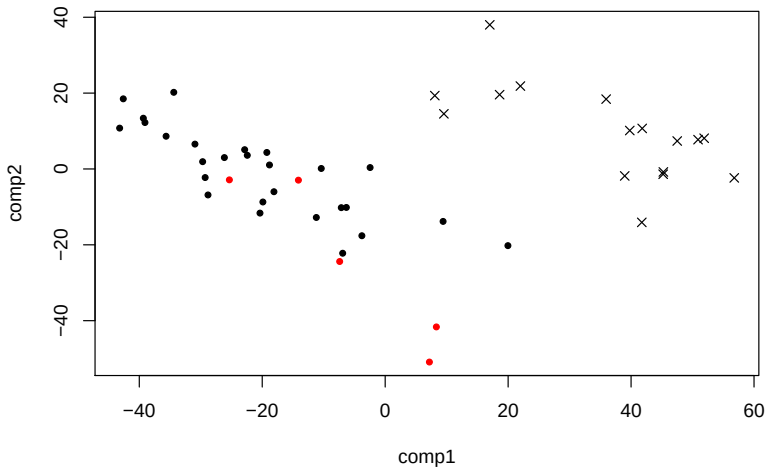


Figure 3.7: PLS components for Stransky et al. (2006) - • represents patients with T2-T4 tumour (32) and × represents patients with Ta tumour (16). The patients are represented in the first two components of the PLS. Patients with an amplification of 8q22-23 are shown in red.

as described in **Section 2.4**: amplifications have been found on chromosomes 6p22, 8q22-23 and 11q13. For simplification, only the amplification on chromosome 8q22-23 was considered. In what follows, C denotes the class effect (Ta / T2-T4), A denotes the effect due to the amplification on chromosome 8q22-23.

Results The prediction performance of different PLS-LMs is provided in **Table 3.5**. The representation of the data in the first two PLS components is provided in **Figure 3.7**. Taking into account the A effect improves the prediction performance when only the first PLS component is considered: the prediction performance increases from 85.4% to 89.6%. The log-linear model alone classifies all the patients in the same class T2-T4 tumour. Combining the MANOVA and log-linear models does not allow improvement of the prediction performance. Note that it was not possible to take interaction into account in the MANOVA model due to lack of degrees of freedom.

3.4.3 Discussion and perspectives

We have developed a supervised classification method named PLS-LM to combine data from both mRNA expression and DNA copy number experiments. By using PLS, the method handle the high-dimensionality of the data. The aim of the study was also to check the ability of the method to improve the prediction performance using both levels of information rather than only the mRNA expression data. The results on simulated data have clearly demonstrated that the PLS-LM allows efficient prediction performance. However, no striking improvement has been noticed on real data. Indeed, the prediction performance for a classifier with both mRNA expression and DNA copy number data is not clearly better than a classifier with only the mRNA expression data. This issue can be explained by two reasons. The first one is a statistical reason. When we consider both mRNA expression and DNA copy number data, the PLS-LM requires many parameters to be estimated but the number of observations available in real datasets still remains very low. Therefore, the model does not efficiently estimate the effects even if they are biologically relevant. The fact that we have few observations with many model parameters can also leads to a situation near overfitting although we have observed no degradation of the prediction performance with respect to classifiers based on mRNA expression alone. The second reason is biological. The information carried by the mRNA expression data might be sufficient to render the differences between the class to predict. Nevertheless, heterogeneity exists in mRNA expression since DNA copy number alterations directly or indirectly impacts gene expression (Lee et al., 2008). In addition to DNA copy number alterations, other sources of heterogeneity exist such as molecular subtypes (as defined by Sørlie et al. (2001) in breast cancer), mutation status of critical-cancer genes or any clinico-histopathological variables. Then, besides DNA copy number data, these other sources of variability could be included in the PLS-LM. However, the ability of the PLS-LM to improve the prediction performance could certainly be demonstrated on real datasets provided that many observations (more than several hundreds) are available in order to efficiently estimate the model parameters. Such datasets with sufficient observations are not available yet. At the time of the manuscript writing, Boulesteix et al. (2008a) has proposed a method to handle mixed data. It is based on combining PLS with random forests. Their conclusions are similar to ours. In simulated, data they have shown that the method significantly improves the prediction performance but on real cancer data, no improvement was observed. Different improvements of the PLS-LM can be suggested:

- We have illustrated the supervised classification using binary classification problems. Although binary classification is the most frequent problem there are situations with

more than two classes. Therefore, our method could be extended to mutli-class problems.

- Besides classical PLS, there is a non-linear algorithm which uses the kernel trick to capture non-linearity within the data (Rosipal and Trejo, 2001). Such an algorithm named kernel PLS could also be applied. However, adding non-linearity into the model increases its complexity that might not be suitable in situation with few observations.

- The model could take into account a prior biological knowledge to build the PLS components. Taking into account a prior biological knowledge to build a prediction rule has been proposed by Rapaport et al. (2007). They considered the topology of the gene network in their classifier. In our algorithm, one might suggest building one PLS component for each pathway which plays a key-role in cancer. However, the knowledge we have about the pathways involved in cancer is far from being exhaustive. Using pathways presents the advantage of having a direct biological interpretation while it is not direct when the components are built over all the genes. Nevertheless, in such a case a functional interpretation can be performed. For example, one can apply Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) procedure on the absolute value of gene weight to each PLS component.

3.4.4 Implementation

A program which implements the LM with variable selection and the dimension reduction using the PLS-LM has been written in C/C++. The program offers the possibility to assess the prediction performance using LOO procedure. The GNU Scientific Library (http://www.gnu.org/software/gsl/) was used for matrix and vector manipulation. The valgrind suite (http://www.valgrind.org/) was used for debugging and profiling the code. The development has been performed on PPC32/Linux architecture. The program can be used either as a command line or within a R package (R Development Core Team, 2008). On a simulated dataset with 300 observations, 20000 continuous predictor variables, the LOO procedure using the PLS-LM with 3 components takes 6 minutes. The architecture used was a SunFire X4600 with 8 dual-core AMD Opteron 885 processors (2.6 GHz) and 32 Go RAM memory running Solaris 5.10. The details of the implementation are the following:

- The MANOVA and PLS algorithms are our own implementations.

- The implementation provided by the R software has been used for the log-linear model.

- The variable selection and LOO procedures are our own implementations.

3.5 How many components to choose in the PLS-LM?

In the PLS algorithm, a number of $\min(n, p)$ components can be computed where n is the number of observations and p is the number of variables. Intuitively, the low order components will capture less and less relevant signal and will even capture the noise of the data and should be excluded from the model. Therefore, to remove the noise, a criterion is required for selecting the number of components to include in the model. Model selection is an even more general issue in statistical modelling and still remains an open question. It aims at choosing the *best model* among a collection of possible models. An efficient model selection technique must balance *goodness of fit* and *complexity* since more complex models will be better able to fit the data but the additional parameters may not represent a relevant effect. This is *the principle of parsimony* introduced by William Ockham in the 14th-century: *"entia non sunt multiplicanda praeter necessitatem"* (entities should not be multiplied beyond necessity). This principle leads to a model with the smallest number of parameters for an adequate representation of the data. In statistics, the principle of parsimony is often referred to as the *bias-variance tradeoff*: in general, bias decreases and variance increases as the dimension of the model increases (see **Figure 3.8**). Simple models may be biased but will have low variance. More complex models have greater representation power (low bias) but overfit to the particular training set (high variance). Thus, the large variance associated with using many features (including those with modest discrimination power) defeats any possible classification benefit derived from these features (Wang et al., 2008). The complexity is generally measured by counting the number of degrees of freedom in the model. Finding the best model is always the grail quest but *"all models are wrong but some are useful"* (George E.P. Box) and *"for every complex question there is a simple and wrong solution"* (A. Einstein). All model selection methods are based to some extent on the principle of parsimony. In the context of class prediction, the model selection method should give the right compromise between bias and variance in order to ensure efficient generalisation. In the following section are presented criteria for model selection.

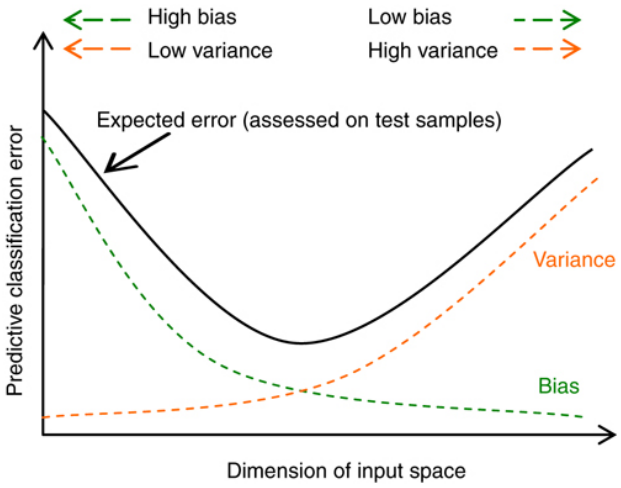


Figure 3.8: Bias-variance tradeoff - A demonstration of the bias-variance dilemma in predictive classification. Specifically, the error of model fitting can be decomposed into two components, bias (approximation error) and variance (estimation error). Added dimensions can degrade the prediction performance if the sample size is small relative to the dimensionality. For a fixed sample size in the high-dimensional data space, there is a tradeoff between the decreased approximation error and the increased estimation error (image and legend from Wang et al., 2008).

3.5.1 Model selection criteria

Theoretical criteria

An abundant literature has been published on the theoretical aspects of model selection and many criteria are available. Burnham and Anderson (2002) give elements of model selection in the framework of statistical inference. We will not give an exhaustive list of the different criteria but we will only mention the two most widely used which are based on the *likelihood theory*. They are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both of them are based on a penalised version of the *likelihood function* $\mathcal{L}(M_\theta|data)$ which represents the probability to have observed the data under the model M_θ where θ is the set of parameters of the model. The penalisation function pen of the likelihood is an increasing function of the number of degrees of freedom K of the model. Indeed, the likelihood will mechanically increase as the number of parameters of the model increases and the penalisation counteracts this phenomenon. The general form of the model selection criteria can be expressed as $2\mathcal{L}(M_\theta|data) - pen(K)$. For AIC and BIC, $pen(K)$ is $2K$ and $\log(n)K$ respectively where n is the number of observations. In practice, for a number of observations greater than 7, the BIC penalty is higher than the AIC penalty and therefore the model selected will tend to have fewer parameters. Depending on the goal of the study, AIC or BIC should be preferred. In the case of explicative model BIC should be preferred to ensure that only the relevant effects have been kept whereas for predictive purposes AIC should be preferred in order not to discard an important discriminative predictor variable. However, no clear consensus exists on that issue.

Empirical criteria

Other criteria which are based on empirical considerations have also been proposed. The criteria are based on cross-validation procedures in which the data are split into a training and a test set. The test set is used to compute the ability of the model to predict the outcome. In the context of regression Prediction Residuals Sum of Squares (PRESS) can be used. In the supervised classification context the prediction performance is used.

Application of model selection criteria in PLS regression model

In the context of PLS regression, Li et al. (2002) compared different criteria to select the number of components of the model. A widely used criterion for selecting the most appropriate number of components is named the *Wold's R criterion* which is based on the PRESS. It is defined as the ratio of the PRESS for the first $k+1$ components and the first k components. The inclusion of a new component in the model terminates at k when the ratio exceeds unity. Adjusted Wold's R criteria have been proposed using either 0.90 or 0.95 as a threshold instead of unity. On simulated data, these two adjusted criteria have been shown to ensure satisfactory performance in terms of the number of times, the known true model is selected while neither Wold's R criterion nor AIC gave satisfactory performance. An F-test based criterion using the PRESS for the first $k+1$ components and the first k components has also been proposed (Osten, 1988).

To assess the prediction performance of the method, it would be a methodological error to select the appropriate number of components using cross-validation and then to compute the prediction performance over the complete training set. Indeed, the cross-validation for model selection is part of the supervised classification method itself. Therefore, a rigorous prediction performance assessment must include a double cross-validation otherwise the prediction performance could be over-optimistic. This pitfall has already been pointed out by

Simon et al. (2003) when gene selection procedure is used as a preprocessing step in the supervised classification method. The main drawback of double cross-validation is that it can greatly increase the computation time. Therefore, it is more suitable to use a theoretical criterion which can be derived directly from the data rather than a criterion based on cross-validation although this second strategy gives in practice excellent results and even often outperforms theoretical criteria.

3.5.2 Model selection criteria and LM

In this subsection, we present the application of model selection criteria to choose the best model regarding the two parts of the LM.

Selection of the parameters in the MANOVA model

The MANOVA model can be written as $\mathbf{X} = \mathbf{D}\mathbf{\Theta} + \mathbf{E}$ where $\mathbf{D}$ is the design matrix of the model, $\mathbf{\Theta}$ is the set of parameters to be estimated and $\mathbf{E}$ is the matrix of residuals. The design matrix can be split into two sub-matrices $\mathbf{D} = (\mathbf{D}_c, \mathbf{D}_{\bar{c}})$ so that $\mathbf{D}_c$ corresponds to the effects related to the class (*i.e.* the class effect itself and the interactions between the class effect and the discrete predictor variables) and $\mathbf{D}_{\bar{c}}$ corresponds to the other effects. Therefore, there are three questions regarding the selection of parameters in the MANOVA model: *(i)* what are the continuous predictor variables to select in $\mathbf{X}$, *(ii)* what are the parameters to keep in $\mathbf{D}_c$ and *(iii)* what are the parameters to keep in $\mathbf{D}_{\bar{c}}$. Questions *(i)* and *(ii)* are related to the discrimination power of the model: the selection of the parameters will be raised in the next paragraph. Question *(iii)* is related to the improvement of the model due to the main effect of the discrete predictor variables and their possible interactions: the selection of the parameters will also be detailed.

Selection of the continuous predictor variables and the discriminant MANOVA terms This paragraph raises questions *(i)* and *(ii)*. Daudin (1986) proposed a modified AIC in order to answer these questions which are treated within a single procedure. First, the discriminative power of a model needs to be defined: it is quantified as the difference between the AIC of the full model (*i.e.* the model with the design matrix $\mathbf{D} = (\mathbf{D}_c, \mathbf{D}_{\bar{c}})$) and the AIC of the sub-model without the effects related to the class effect (*i.e.* the model with the design matrix $\mathbf{D} = (\mathbf{0}, \mathbf{D}_{\bar{c}})$). This difference is named Discriminant AIC (DAIC) and is expressed as follows:

$$DAIC(\mathbf{X}, \mathbf{D}_c, \mathbf{D}_{\bar{c}}) \;=\; 2[AIC(\mathbf{X}, \mathbf{D}_c, \mathbf{D}_{\bar{c}}) - AIC(\mathbf{X}, \mathbf{0}, \mathbf{D}_{\bar{c}})]$$

The goal of the selection procedure is to select both the optimal subset of continuous predictor variables $\mathbf{X}_{opt}$ and to keep the most relevant discriminant terms $\mathbf{D}_{c.opt}$ based on the DAIC:

$$DAIC(\mathbf{X}_{opt}, \mathbf{D}_{c.opt}, \mathbf{D}_{\bar{c}}) \;=\; \sup_{\mathbf{X}, \mathbf{D}_c} DAIC(\mathbf{X}, \mathbf{D}_c, \mathbf{D}_{\bar{c}})$$

The exhaustive search over $\mathbf{X}$ and $\mathbf{D}_c$ can be computationally very expensive and in practice backward selection procedure is used to reduce the complexity of the selection procedure. At each step of the procedure, the DAIC indicates whether a continuous predictor variable must be removed from $\mathbf{X}$ or a term from the $\mathbf{D}_c$. The procedure stops when no removal is required.

Selection of the non-discriminant MANOVA terms This paragraph raises the question *(iii)*. In the MANOVA model, the terms in $\mathbf{D}_{\bar{c}}$ are selected using AIC.

Computation of the log-likelihood The parameters of the MANOVA model are estimated using the usual maximum likelihood estimators and the log-likelihood of the data is expressed as follows:

$$\log(L) \quad = \quad -\frac{1}{2}\Big[n\log|\hat{\mathbf{\Sigma}}| + np\log(2\pi) + np\Big] \tag{3.20}$$

where $\hat{\mathbf{\Sigma}}$ is the standard maximum likelihood estimator of $\mathbf{\Sigma}$, p is the number of continuous predictor variables and n is the number of observations.

The number of degrees of freedom of the model is $K = rk(\mathbf{D})p + 0.5p(p+1)$ (*i.e.* the number associated with the mean parameters plus the number of parameters associated with the variance-covariance matrix).

Selection of the log-linear model

In the log-linear model, the terms are selected using AIC as described by Sakamoto (1982). The log-likelihood of the model is computed as follows:

$$\log(L) \quad = \quad \sum_t \hat{m}_t \log \hat{m}_t$$

where $\hat{m}_t$ is the maximum likelihood estimator of the mean count of cell t of the contingency table. If we denote C_j the number of modality of the j effect then the number of degrees of freedom K of the model is the sum over the number of effect of $C_j - 1$ plus the sum over all the interactions i, j $(C_j - 1)(C_i - 1)$. There is also the constant term of the model but since the parameters are estimated under the constraint $\sum_t \hat{m}_t = n$ where n is the total number of observations there is nothing to add anymore. The terms of the log-linear model are selected according to a backward selection procedure.

Application to the PLS-LM

The application of the DAIC to select the PLS components fails due to the introduction of covariance between the continuous predictor variables. Indeed, even on simulated data in which there is no class effect, the trend is to keep all the components included as continuous predictor variables. To illustrate this phenomenon, let us take for example the following simple model: in the MANOVA model, only the class effect is considered in the design matrix $\mathbf{D}$. There are only two classes with n_1 and n_2 observations respectively. The design matrix can be expressed as follows:

$$\mathbf{D} \quad = \quad (\mathbf{D}_c, \mathbf{D}_{\bar{c}}) \text{ with } \mathbf{D}_c = (\underbrace{1\ldots1}_{n_1}\underbrace{0\ldots0}_{n_2})' \text{ and } \mathbf{D}_{\bar{c}} = (1\ldots1)'$$

The MANOVA model can be written as $\mathbf{T} = \mathbf{D\Theta} + \mathbf{E}$ where $\mathbf{T}$ represents the PLS components. From **Equation 3.12**, the variance-covariance matrix of $\mathbf{T}$ is diagonal since the covariance is zero between two components by construction. Let us call this variance-covariance matrix $\mathbf{\Sigma} = diag\{v_1, \cdots, v_K\}$. From **Theorem 3 (variance decomposition)**

we have $\mathbf{\Sigma} = \mathbf{\Sigma}_W + \mathbf{\Sigma}_B$ where the two classes are used to perform the decomposition. In the MANOVA model, the variance-covariance matrix used to compute the log-likelihood in **Equation 3.20** is the estimation of $\mathbf{\Sigma}_W$. This matrix has necessarily non-zero entries due to the variance decomposition and a covariance is mechanically introduced between the components. This leads to the increase in the log-likelihood of the data.

Another explanation to the fact that DAIC fails to select the PLS components can be explained as follows. Each PLS component accounts only for one free parameter in the likelihood penalty. However, one component is the aggregation of information from all the original continuous predictor variables. Thus, the number of degrees of freedom associated with each PLS component is very likely to be higher as pointed out by Frank and Friedman (1993). Moreover, by using the class variable to compute the components, PLS uses even more degrees of freedom per component and can fit the data to a high degree of accuracy. The estimation of the number of degrees of freedom of a model is an important issue to have efficient model selection criteria. This issue has been raised by Zou et al. (2007) in the case of the lasso method. We think that the DAIC could be applied with PLS components provided that the true (but unknown) degrees of freedom of each component is used in the likelihood penalty. The estimation of the number of degrees of freedom in the PLS will not be raised in the present manuscript. We will only present a statistical criterion to test whether there is a significant signal on the first PLS component. This is the second contribution of this chapter which is detailed in the next section.

3.6 Contribution 2: Statistical criterion to test the significance of the first PLS component

This section presents the statistical criterion we propose to test the significance of the first PLS component. The criterion is based on asymptotic statistics on random matrices. Geman (1980) proposed a general framework to establish the expected value of the maximum eigenvalue of symmetric random matrices. However, the application conditions of his theorem did not hold to obtain the expected value of the maximum eigenvalue in the PLS. This is the reason why we established a suitable criterion for the PLS using normality assumption of the data. In this section, we assume that there is absolutely no relation between the matrix $\tilde{\mathbf{X}}$ which contains the original continuous predictor variables and the matrix $\tilde{\mathbf{Y}}$ which contains the class variable. Therefore, the $\tilde{\mathbf{X}}$ matrix only contains random entries. We first present the properties of the statistical criterion whose efficiency is then assessed on both simulated and real data.

3.6.1 Asymptotic distribution of the statistical criterion

In this subsection, the statistical criterion and its properties are presented. First, we describe how the original data are pre-processed before computing the PLS. As we already mentioned, since the PLS algorithm maximises a covariance, it is sensitive to the scaling of the variables. Therefore, in order to give the same weight to each variable we use the matrices which have been centered and scaled as follows:

- matrix $\tilde{\mathbf{X}}$:

$$\begin{aligned} X_{ij} &= \frac{\tilde{X}_{ij} - \frac{1}{n}\sum_{i\leq n}\tilde{X}_{ij}}{\hat{\sigma}_j} \\ \text{with } \hat{\sigma}_j^2 &= \frac{\sum_{i\leq n}(\tilde{X}_{ij} - \frac{1}{n}\sum_{i\leq n}\tilde{X}_{ij})^2}{n-1} \end{aligned}$$

- matrix $\tilde{\mathbf{Y}}$:

  Let $\tilde{\mathbf{Y}}' = (\underbrace{1\cdots 1}_{n_1}\underbrace{0\cdots 0}_{n_2})$ with $n = n_1 + n_2$.

  Therefore, the centered and scaled (with the biased version of the standard-deviation estimator) vector $\tilde{\mathbf{Y}}$ is:

  $\mathbf{Y} = (Y_i) = (\underbrace{c_1\cdots c_1}_{n_1}\underbrace{c_2\cdots c_2}_{n_2})$ with $c_1 = \sqrt{\frac{n_2}{n_1}}$ and $c_2 = -\sqrt{\frac{n_1}{n_2}}$.

  Note that for $\tilde{\mathbf{Y}}$, the scaling is done with the biased version of the standard-deviation estimator to simplify the calculation.

We will demonstrate in what follows how we can derive the law of the eigenvalue associated with the first PLS component computed over the centered and scaled matrix $\tilde{\mathbf{X}}$. We assume that the $\tilde{X}_{ij}$'s are independently distributed with $\tilde{X}_{ij} \sim \mathcal{N}(0, \sigma_j^2)$.

Proposition 1 *Let $\tilde{X}_1, \tilde{X}_2, ...\tilde{X}_n$ be n independent and identically distributed random variables with $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. Let $Y_1, Y_2, ...Y_n$ be n fixed binary (non random) variables with n_1 values $c_1 = \sqrt{\frac{n_2}{n_1}}$ and $n_2 = n - n_1$ values $c_2 = -\sqrt{\frac{n_1}{n_2}}$ such that $\sum_i Y_i^2 = n$. Let $\alpha = n_1/n$ which is held fixed as $n \to \infty$. Let us define the following quantities:*

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-1}\sum(\tilde{X}_i - \frac{1}{n}\sum_{i\leq n}\tilde{X}_i)^2 \\
X_i &= \frac{\tilde{X}_i - \frac{1}{n}\sum_{i\leq n}\tilde{X}_i}{\hat{\sigma}} \\
A &= \sum_{i=1}^{n} X_i Y_i
\end{aligned}$$

Then, $n^{-1/2}A \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. Let us assume that $Y_i = c_1$ for $i \leq n_1$ and $Y_i = c_2$ for $n_1 < i \leq n$. $A = B_1 + B_2$ where $B_1 = c_1 \sum_{i\leq n_1} X_i$ and $B_2 = c_2 \sum_{n_1<i\leq n} X_i$.

$$\begin{aligned}
B_1 &= \frac{c_1}{\hat{\sigma}}\left(\sum_{i\leq n_1}\tilde{X}_i - \frac{n_1}{n}\sum_{i\leq n}\tilde{X}_i\right) \\
&= \frac{c_1}{\hat{\sigma}}\left[(1-\alpha)\sum_{i\leq n_1}\tilde{X}_i - \alpha\sum_{n_1<i\leq n}\tilde{X}_i\right] \\
&= \frac{c_1}{\hat{\sigma}}\left[(1-\alpha)\sum_{i\leq n_1}(\tilde{X}_i - \mu) - \alpha\sum_{n_1<i\leq n}(\tilde{X}_i - \mu)\right] \\
&= \frac{c_1(1-\alpha)\sigma}{\hat{\sigma}}\left(\sum_{i\leq n_1}\frac{\tilde{X}_i - \mu}{\sigma}\right) - \frac{c_1\alpha\sigma}{\hat{\sigma}}\left(\sum_{n_1<i\leq n}\frac{\tilde{X}_i - \mu}{\sigma}\right) \\
n^{-1/2}B_1 &= \frac{c_1(1-\alpha)\sigma}{\hat{\sigma}}\alpha^{1/2}n_1^{-1/2}\left(\sum_{i\leq n_1}\frac{\tilde{X}_i - \mu}{\sigma}\right) - \frac{c_1\alpha\sigma}{\hat{\sigma}}(1-\alpha)^{1/2}n_2^{-1/2}\left(\sum_{n_1<i\leq n}\frac{\tilde{X}_i - \mu}{\sigma}\right)
\end{aligned}$$

The same computation for B_2 leads to:

$$\begin{aligned}
B_2 &= \frac{c_2}{\hat{\sigma}}\left(\sum_{n_1<i\leq n}\tilde{X}_i - \frac{n_2}{n}\sum_{i\leq n}\tilde{X}_i\right) \\
&= \frac{c_2}{\hat{\sigma}}\left[\alpha\sum_{n_1<i\leq n}\tilde{X}_i - (1-\alpha)\sum_{i\leq n_1}\tilde{X}_i\right] \\
&= \frac{c_2}{\hat{\sigma}}\left[\alpha\sum_{n_1<i\leq n}(\tilde{X}_i - \mu) - (1-\alpha)\sum_{i\leq n_1}(\tilde{X}_i - \mu)\right] \\
&= \frac{c_2\alpha\sigma}{\hat{\sigma}}\left(\sum_{n_1<i\leq n}\frac{\tilde{X}_i - \mu}{\sigma}\right) - \frac{c_2(1-\alpha)\sigma}{\hat{\sigma}}\left(\sum_{i\leq n_1}\frac{\tilde{X}_i - \mu}{\sigma}\right) \\
n^{-1/2}B_2 &= \frac{c_2\alpha\sigma}{\hat{\sigma}}(1-\alpha)^{1/2}n_2^{-1/2}\left(\sum_{n_1<i\leq n}\frac{\tilde{X}_i - \mu}{\sigma}\right) - \frac{c_2(1-\alpha)\sigma}{\hat{\sigma}}\alpha^{1/2}n_1^{-1/2}\left(\sum_{i\leq n_1}\frac{\tilde{X}_i - \mu}{\sigma}\right)
\end{aligned}$$

Finally we have:

$$\begin{aligned} n^{-1/2}A &= \left[\frac{c_1(1-\alpha)\sigma}{\hat{\sigma}}\alpha^{1/2} - \frac{c_2(1-\alpha)\sigma}{\hat{\sigma}}\alpha^{1/2}\right]\left(n_1^{-1/2}\sum_{i\leq n_1}\frac{\tilde{X}_i-\mu}{\sigma}\right) + \\ &\quad \left[\frac{c_2\alpha\sigma}{\hat{\sigma}}(1-\alpha)^{1/2} - \frac{c_1\alpha\sigma}{\hat{\sigma}}(1-\alpha)^{1/2}\right]\left(n_2^{-1/2}\sum_{n_1<i\leq n}\frac{\tilde{X}_i-\mu}{\sigma}\right) \\ &= \frac{\sigma}{\hat{\sigma}}\left[(1-\alpha)^{1/2}\left(n_1^{-1/2}\sum_{i\leq n_1}\frac{\tilde{X}_i-\mu}{\sigma}\right) - \alpha^{1/2}\left(n_2^{-1/2}\sum_{n_1<i\leq n}\frac{\tilde{X}_i-\mu}{\sigma}\right)\right] \end{aligned}$$

Let $n^{-1/2}A = S_n\left[(1-\alpha)^{1/2}U_{n_1} - \alpha^{1/2}U_{n_2}\right]$ where:

$$\begin{aligned} S_n &= \frac{\sigma}{\hat{\sigma}} \\ U_{n_1} &= n_1^{-1/2}\sum_{i\leq n_1}\frac{\tilde{X}_i-\mu}{\sigma} \\ U_{n_2} &= n_2^{-1/2}\sum_{n_1<i\leq n}\frac{\tilde{X}_i-\mu}{\sigma} \end{aligned}$$

Let us assume that (i) $S_n \xrightarrow{p} 1$ and (ii) $(1-\alpha)^{1/2}U_{n_1} - \alpha^{1/2}U_{n_2} \xrightarrow{d} \mathcal{N}(0,1)$. Then, from **Theorem 7.ii (Slutsky's theorem)**, $n^{-1/2}A \xrightarrow{d} \mathcal{N}(0,1)$. Let us prove that (i) and (ii) hold:

(i) : From **Theorem 4.ii (Cochran's theorem)** we have $E(1/S_n^2) = 1$ and $V(1/S_n^2) = 2/(n-1)$. The **Theorem 5 (Bienaymé-Chebyshev inequality)** implies that $1/S_n^2 \xrightarrow{p} 1$. Applying **Theorem 8i (continuous mapping)** with $g(x) = \sqrt{1/x}$ implies that $S_n \xrightarrow{p} 1$.

(ii) : From **Theorem 9 (central limit theorem)**, $U_{n_1} \xrightarrow{d} \mathcal{N}(0,1)$ and $U_{n_2} \xrightarrow{d} \mathcal{N}(0,1)$. U_{n_1} and U_{n_2} are independent and therefore it follows from **Theorem 6 (gaussian vector)** that $(1-\alpha)^{1/2}\mathcal{N}(0,1) - \alpha^{1/2}\mathcal{N}(0,1) \sim \mathcal{N}(0,1)$. ■

Proposition 2 *Let $\tilde{\mathbf{X}}$ be a (n,p) matrix of continuous predictor variables whose entries $\tilde{X}_{ij}$ are independent and identically distributed with $\tilde{X}_{ij} \sim \mathcal{N}(\mu_j, \sigma_j^2)$. Note $\mathbf{X}$ the centered and scaled $\tilde{\mathbf{X}}$ matrix in which each column $\tilde{\mathbf{X}}_j$ has been scaled by the classical unbiased estimator of the standard-deviation. Let $\mathbf{Y} = (Y_1, Y_2, ...Y_n)'$ be n fixed binary (non random) variables with n_1 values $c_1 = \sqrt{\frac{n_2}{n_1}}$ and $n_2 = n - n_1$ values $c_2 = -\sqrt{\frac{n_1}{n_2}}$ such that $\sum_i Y_i^2 = n$. Let $\alpha = n_1/n$ which is held fixed as $n \to \infty$. Then:*

$$nCOV_n^2(\mathbf{T}_1, \mathbf{Y}) \xrightarrow{d} \chi^2(p)$$

where $\mathbf{T}_1$ is the first PLS component computed with $\mathbf{X}$ and $\mathbf{Y}$.

Proof. Let $\mathbf{X'Y} = \mathbf{A} = (A_j)_{j=1...p}$ with $A_j = \mathbf{Y'X}_j$. From **Proposition 1** we have:

$$n^{-1/2}A_j \xrightarrow{d} \mathcal{N}(0,1)$$

Let derive the distribution of $\sum_{j=1}^{p}(n^{-1/2}A_j)^2$. From the notation of **Theorem 8.ii (continuous mapping)** let:

$$\begin{aligned} &g : \mathbb{R}^p \to \mathbb{R} \\ &g : (x_1, \cdots, x_p) \to \sum_{j=1}^{p} x_j^2 \\ &X_n = (B_1, \cdots, B_p) \end{aligned}$$

Since the random vectors $\mathbf{X}_j$ are independent the $n^{-1/2}A_j$ are independent too and we have:

$$\begin{aligned} \sum_{j=1}^{p}(n^{-1/2}A_j)^2 &\xrightarrow{d} \chi^2(p) \\ \frac{\|\mathbf{X'Y}\|^2}{n} &= \sum_{j=1}^{p}(n^{-1/2}A_j)^2 \\ \frac{\|\mathbf{X'Y}\|^2}{n} &\xrightarrow{d} \chi^2(p) \end{aligned}$$

Equation 3.8 and **Equation 3.10** imply that:

$$nCOV_n^2(\mathbf{T}_1, \mathbf{Y}) = \frac{\|\mathbf{X'Y}\|^2}{n}$$

Note that if the $\tilde{X}_{ij}$ are independently distributed with $\tilde{X}_{ij} \sim \mathcal{N}(0,1)$ and if no scaling is performed on $\tilde{\mathbf{X}}$ (*i.e.* $\tilde{\mathbf{X}} = \mathbf{X}$) then $nCOV_n^2(\mathbf{T}_1, \mathbf{Y}) \sim \chi^2(p)$ (the proof is straightforward). ■

3.6.2 Assessment of the statistical criterion on simulated data

This section presents simulation studies to investigate the ability of the statistical criterion proposed in **Proposition 2** to identify signal within the data. In the following subsections, we will use the following notations:

$\mathbf{H_0}$ This notation will refer to the situation in which there is no class effect.

$\mathbf{H_1}$ This notation will refer to the situation in which there is a difference between the two classes.

without correlation This notation will refer to the situation in which there is no correlation between the continuous predictor variables.

with correlation This notation will refer to the situation in which correlation has been introduced between a subset of the continuous predictor variables.

For each situation, the matrix $\tilde{\mathbf{X}}$ has been simulated and the details of the simulation model are provided within each subsection. The number of observations n has been set at different values (10, 20, 50 and 100) and half observations belong to class 1 and class 2. The total number of continuous predictor variables p has always been set at 100. Once simulated, each column $\tilde{\mathbf{X}}_j$ of the matrix $\tilde{\mathbf{X}}$ is centered and scaled by the classical unbiased estimator of the standard-deviation. For each simulation, the values $nCOV_n^2(\mathbf{T}_1, \mathbf{Y})$ of the statistical criterion proposed in **Proposition 2** have been computed. The mean, variance and the percentage of rejected hypotheses at the level of 5% under $\mathbf{H_0}$ will be presented. For simulation conditions which statisfy the hypotheses of **Proposition 2**, the statistical criterion follows a $\chi^2(100)$ distribution.

$\mathbf{H_0}$ without correlation

The $\tilde{X}_{ij}$ are independently distributed and have been simulated according to four distributions:

1. **normal** distribution $\mathcal{N}(0, 1)$
2. **uniform** distribution $\mathcal{U}(0, 1)$
3. **log-normal** distribution $\mathcal{LN}(0, 1)$
4. **student** distribution $\mathcal{T}(1)$

Once simulated, the matrix $\tilde{\mathbf{X}}$ is then scaled and centered as already described at the beginning of **Subsection 3.6.1**. In the **normal** distribution case, the hypotheses of **Proposition 2** hold and the proposition can be applied. The **uniform**, **log-normal** and **student** have also been used for simulation to test the robustness of **Proposition 2** with respect to the normality hypothesis.

Results The complete set of histograms of the statistical criterion is presented in **Annexes - Figure A.1** but only two typical histograms are available in **Figure 3.9**. The mean (E), variance (V) and the percentage of rejected hypotheses (R) at the level of 5% under $\mathbf{H_0}$ over the 20000 simulations are available in **Table 3.6**. The **normal** and **uniform** distributions show similar behaviour. Even for small sample size, the empirical mean is close to the expected mean value of 100. The variance converges to its expected value of 200 as the sample size increases. The smaller variance than expected for small sample size leads to a number of rejected hypotheses lower than expected (around 3% instead of 5% expected). Both for the **log-normal** and **student** distributions, the empirical mean is also close to the expected mean value of 100. In contrast, the empirical variance is much smaller than expected and especially in the **student** distribution case. This results in fewer rejected hypotheses (between 1.37% and 3.42% for the **log-normal** distribution and between 0.52% and 1.06% for the **student** distribution). This different behaviour can be explained by the asymmetry of the **log-normal** distribution and the heavy tail of the **student** distribution. The statistical criterion proposed shows robustness with respect to the normality hypothesis.

		normal	uniform	log-normal	student
$n = 10$	E	100.18	99.95	99.94	100.05
	V	148.72	158.23	112.25	83.47
	R	3.04	3.12	1.37	0.52
$n = 20$	E	100.14	100.16	99.93	99.93
	V	171.80	178.57	128.30	92.09
	R	3.64	4.08	1.95	0.87
$n = 50$	E	99.98	99.95	100.10	99.98
	V	191.89	193.18	151.76	96.35
	R	4.69	4.71	3.01	0.92
$n = 100$	E	99.98	100.11	100.00	100.03
	V	195.11	197.79	163.30	97.53
	R	4.86	5.00	3.42	1.06

Table 3.6: Simulations under $\mathbf{H_0}$ without correlation - E is the empirical mean, V is the unbiased empirical variance and R is the percentage of rejected hypotheses under $\mathbf{H_0}$ at the 5% level computed over 20000 simulations. The simulations have been performed for the normal, uniform, log-normal and student distributions. n corresponds to the number of observations. The number of variables has been set at 100.

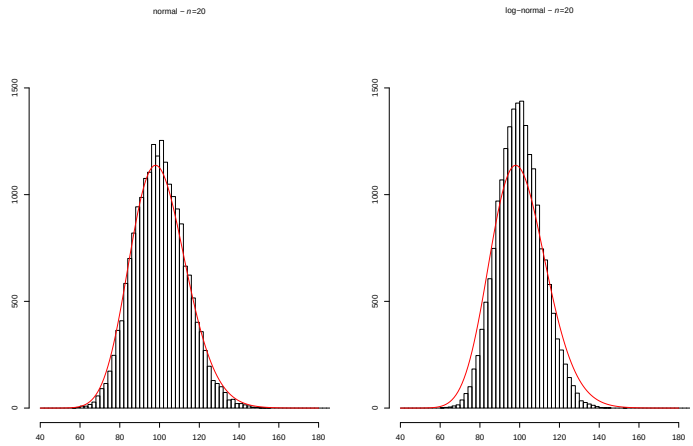


Figure 3.9: Typical histograms for simulations under $\mathbf{H_0}$ without correlation - Two typical histograms of the statistical criterion over 20000 simulations. The data have been simulated using the normal and log-normal distributions. n corresponds to the number of observations. The number of variables has been set at 100. The $\chi^2(100)$ probability density function is displayed as a red line.

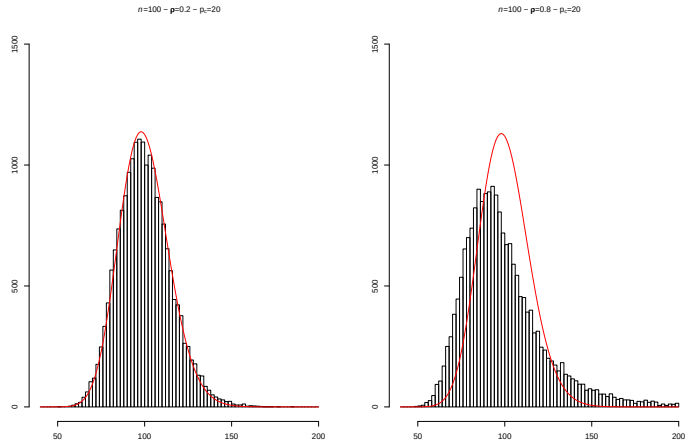


Figure 3.10: Typical histograms for simulations under $\mathbf{H_0}$ with correlation - Two typical histograms of the statistical criterion over 20000 simulations. The data have been simulated using the normal distribution. n corresponds to the number of observations, ρ is the correlation value, and p_c is the number of correlated variables. The number of variables has been set at 100. The $\chi^2(100)$ probability density function is displayed as a red line.

H_0 with correlation

In this subsection, correlation has been introduced between a subset of p_c variables using the following method:

1. The $\tilde{\mathbf{X}}_j$ columns of the matrix $\tilde{\mathbf{X}}$ are independently distributed and have been simulated according to a $\mathcal{N}(0,1)$ distribution.

2. A variable subset of size p_c (taking the values 2, 5, 10 and 20) is chosen from the matrix $\tilde{\mathbf{X}}$.

3. The pairwise correlation between the p_c variables is set at ρ (taking the values 0.2, 0.5 and 0.8). The correlation matrix $\mathbf{\Delta}$ between the p_c variables is noted:

$$\mathbf{\Delta} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} = \mathbf{\Lambda}'\mathbf{\Lambda}$$

4. Using the Cholesky decomposition, the matrix $\mathbf{\Delta}$ is re-written as the matrix-product $\mathbf{\Delta} = \mathbf{\Lambda}'\mathbf{\Lambda}$.

5. The $\tilde{\mathbf{X}}_c$ corresponding to the p_c variables from the matrix $\tilde{\mathbf{X}}$ is replaced with $\tilde{\mathbf{X}}_c\mathbf{\Lambda}$. Therefore, the matrix $\tilde{\mathbf{X}}$ has p_c variables following a normal distribution with pairwise correlation of ρ and $p - p_c$ independent variables following a normal distribution.

6. The new matrix $\tilde{\mathbf{X}}$ is then scaled and centered as already described at the beginning of **Subsection 3.6.1**.

Results The complete set of histograms of the statistical criterion is presented in **Annexes - Figure A.2** but only two typical histograms are available in **Figure 3.10**. The mean (E), variance (V) and the percentage of rejected hypotheses (R) at the level of 5% under $\mathbf{H_0}$ over the 20000 simulations are available in **Annexes - Figure A.3**. For all the simulations, the mean of the statistical criterion fluctuates around its expected value of 100. The variance increases with p_c, ρ and n: the statistical criterion distribution becomes asymmetric and this leads to an icrease in the percentage of rejected hypotheses. For low correlation ($\rho = 0.2$), the increase in the percentage of rejected hypotheses is visible only for the highest number of correlated variables $p_c = 20$ ($R = 6.22\%$ for $n = 100$). For intermediate correlation ($\rho = 0.5$), the increase in the percentage of rejected hypotheses is already visible for a number of correlated variables $p_c = 10$ ($R = 6.53\%$ for $n = 100$). For high correlation ($\rho = 0.8$), the increase in the percentage of rejected hypotheses starts being visible even with a small number of correlated variables $p_c = 5$ ($R = 5.64\%$ for $n = 100$); for the highest number of correlated variables $p_c = 20$ the percentage reaches values greater than 12.00% ($R = 12.81\%$ for $n = 100$). The increase in the correlation within the data makes the statistical criterion less conservative with respect to the hypothesis $\mathbf{H_0}$.

$\mathbf{H_1}$ without correlation

In this subsection a class effect has been introduced for a subset of p_c variables using the following method:

1. The $\tilde{\mathbf{X}}_j$ columns of the matrix $\tilde{\mathbf{X}}$ are independently distributed and have been simulated according to a $\mathcal{N}(0,1)$ distribution.

2. A variable subset of size p_c (taking the values 2, 5, 10 and 20) is chosen from the matrix $\tilde{\mathbf{X}}$.

3. Each value $\tilde{X}_{ij}$ of $\tilde{\mathbf{X}}_c$ corresponding to the p_c variables from the matrix $\tilde{\mathbf{X}}$ is replaced with $\tilde{X}_{ij} + 1$ if the observation i belongs to class 1. Therefore, the matrix $\tilde{\mathbf{X}}$ has p_c variables with a class effect (the within-class distribution being normal) and $p - p_c$ independent variables following a normal distribution.

4. The new matrix $\tilde{\mathbf{X}}$ is then scaled and centered as already described at the beginning of **Subsection 3.6.1**.

Results The complete set of histograms of the statistical criterion is presented in **Annexes - Figure A.4** but only two typical histograms are available in **Figure 3.11**. The mean (E), variance (V) and the percentage of rejected hypotheses (R) at the level of 5% under $\mathbf{H_0}$ over the 20000 simulations are available in **Table 3.7**. E, V and R increase with the number of observations n and the number of variables with a class effect p_c. With only $p_c = 5$ the statistical criterion is able to reject all the $\mathbf{H_0}$ hypotheses for a number of 100 observations. For small sample size like $n = 20$, at least $p_c = 20$ variables are needed to reach almost 100% of rejected hypotheses (99.78% exactly). In order to efficiently identify the signal in the data with the proposed statistical criterion, the following requirements are needed: either the number of observations or the number of variables with a class effect must be high.

$\mathbf{H_1}$ with correlation

In this subsection, both a pairwise correlation and a class effect have been introduced for a subset of the same p_c variables using the following method:

1. The $\tilde{\mathbf{X}}_j$ columns of the matrix $\tilde{\mathbf{X}}$ are independently distributed and have been simulated according to a $\mathcal{N}(0,1)$ distribution.

2. A variable subset of size p_c (taking the values 2, 5, 10 and 20) is chosen from the matrix $\tilde{\mathbf{X}}$.

3. The correlation has been simulated for the p_c variables according to the method described in **$\mathbf{H_0}$ with correlation**.

4. The class effect has been added for the same subset of p_c variables according to the method described in **$\mathbf{H_0}$ without correlation**.

Results The complete sets of histograms of the statistical criterion are represented in **Annexes - Figure A.5** but only two typical histograms are available in **Figure 3.12**. The mean (E), variance (V) and the percentage of rejected hypotheses (R) at the level of 5% under $\mathbf{H_0}$ over the 20000 simulations are available in **Annexes - Figure A.6**. The qualitative behaviour for E, V and R is similar to what was described in the situation where

		$n = 10$	$n = 20$	$n = 50$	$n = 100$
$p_c = 2$	E	103.17	107.22	119.21	139.09
	V	151.58	181.74	231.00	275.91
	R	4.62	10.66	35.76	81.14
$p_c = 5$	E	107.65	117.27	147.47	197.73
	V	158.14	205.66	284.96	413.78
	R	9.46	30.28	91.68	100.00
$p_c = 10$	E	115.20	135.04	194.90	294.79
	V	164.94	234.44	388.43	618.37
	R	23.67	75.50	100.00	100.00
$p_c = 20$	E	130.27	170.25	290.20	489.81
	V	185.73	294.58	578.26	1054.64
	R	66.14	99.78	100.00	100.00

Table 3.7: Simulations under $\mathbf{H_1}$ without correlation - E is the empirical mean, V is the unbiased empirical variance and R is the percentage of rejected hypotheses under $\mathbf{H_0}$ at the 5% level computed over 20000 simulations.The number of variables has been set at 100.

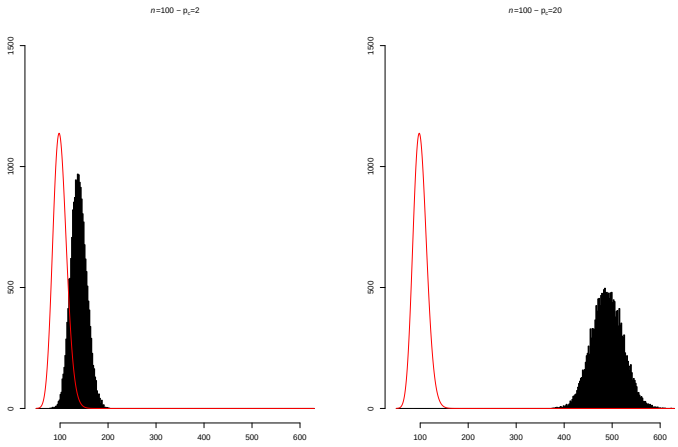


Figure 3.11: Typical histograms for simulations under $\mathbf{H_1}$ without correlation - Histograms of the statistical criterion over 20000 simulations. The data have been simulated using the normal distribution. n corresponds to the number of observations and p_c is the number of variables with a class effect. The number of variables has been set at 100. The $\chi^2(100)$ probability density function is displayed as a red line.

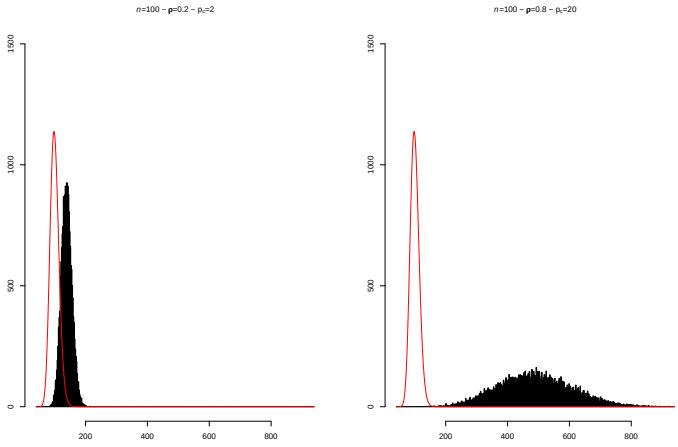


Figure 3.12: Typical histograms for simulations under $\mathbf{H_1}$ with correlation - Histograms of the statistical criterion over 20000 simulations. The data have been simulated using the normal distribution. n corresponds to the number of observations, ρ is the correlation value, and p_c is the number of variables with a class effect and correlated. The number of variables has been set at 100. The $\chi^2(100)$ probability density function is displayed as a red line.

there is no correlation. However, the most striking characteristic is the huge increase in V as the correlation value ρ increases. As a consequence, R approaches 100% in some cases while it was equal to 100% in the situation without correlation. With only $p_c = 5$, the statistical criterion is able to reject 99.98% ($\rho = 0.2$), 99.64% ($\rho = 0.5$) and 99.2% ($\rho = 0.8$) of $\mathbf{H_0}$ hypotheses for a number of 100 observations. For small sample size like $n = 20$, at least $p_c = 20$ variables are needed to reach more than 79% of rejected hypotheses while in the situation without correlation the statistical criterion was able to reject 99.78% of the hypotheses. R also increases as the correlation decreases (79.17% with $\rho = 0.8$, 86.83% with $\rho = 0.5$ and 94.92% with $\rho = 0.2$). The increase in correlation (increase of the ρ value and/or the p_c value) tends to reduce the ability to identify the signal in the data due to an increase in variance; this is especially the case for small sample size.

3.6.3 Assessment of the statistical criterion on real data

The efficiency of the statistical criterion we proposed has been evaluated using the two real datasets from Chin et al. (2006) and Stransky et al. (2006). We remind the reader that the dataset from Chin et al. (2006) consists of 20244 probes and 89 observations. The dataset from Stransky et al. (2006) consists of 8111 probes and 48 observations.

Results In both datasets the criteria are significant (p-value$< 10^{-10}$). It is 24860 and 30572 for Chin et al. (2006) and Stransky et al. (2006) respectively. We have estimated the percentage of rejected $\mathbf{H_0}$ hypotheses under 20000 permutations (the class label have been resampled). The percentage is 36.7% and 37.1% for Chin et al. (2006) and Stransky et al. (2006) respectively. This can be explained by an increase in the variance of the statistical criteria while the means remain close to their expected values (data not shown). The histograms of the distribution are similar to what is observed in **Figure 3.10** in the case of high correlation except that the asymmetry of the distribution is higher. Indeed, in real datasets we expect that a lot of probes are very likely to be correlated (the average pairewise absolute value correlation is 0.174 and 0.164 for Chin et al. (2006) and Stransky et al. (2006) respectively). In the original data we used, a correlation mechanically exists due to the class effect. Then, whatever the resampling, the correlation structure due to the between-class effect will always remain which can be unfair for our assessment. Therefore, we restricted the permutations for a subset of the data corresponding to only one class. Labels have been randomly assigned to create two balanced classes. In this case, the conclusions still hold the same which confirms the effect induced by the correlation on the statistical criterion. This suggests a limited interest of the criterion for real data.

3.6.4 Discussion and perspectives

We have developed a statistical criterion to test the significance of the first PLS component. The criterion has been shown to be efficient on simulated data for independently and identically distributed variables. In the case of correlated variables, the test based on the statistical criterion tends to be less conservative leading to reject the $\mathbf{H_0}$ hypotheses while it should not. On real data, there is a lot of correlation between genes. As a result, the statistical criterion also tends to reject the $\mathbf{H_0}$ hypothesis in situation where there is no signal in the data. Therefore, both the results on simulated and real data suggest to improve the statistical criterion in order to take into account the correlation between the variables. Besides this, we can also mention the following improvements:

- We have derived the asymptotic distribution of the statistical criterion as n, the number

of observations, tends to infinity. The asymptotic distribution could also be evaluated as p, the number of variables, tends to infinity.

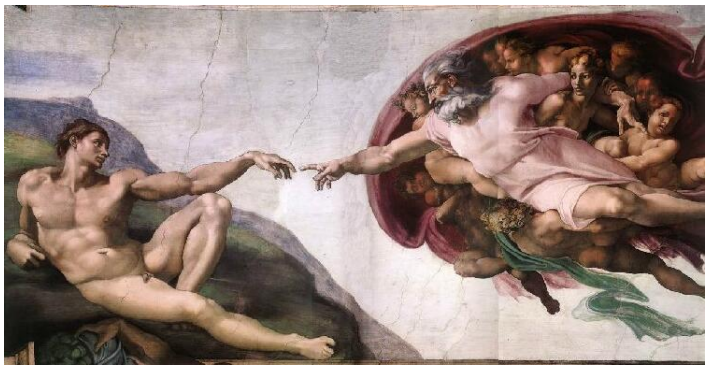
- In the PLS-LM we require that the continuous variables are scaled in order to have the same weight. The scaling is done without taking into account the class information. Then, the scaling could be done using an intra-class standard-deviation. In this case, the distribution of the statistical criterion should be derived.

- The last improvement concerns the other components. Indeed, we have only focused here on the first component. Due to the complexity of the PLS algorithm, it is not straightforward to derive the distribution of the statistical criterion for the next component. One way to answer this issue could be the use of residuals theory (Ellenberg, 1973).

3.7 Conclusion

In this chapter, two contributions related to the class prediction problem have been presented. The first one concerns a supervised classification method which is able to combine both continuous and discrete predictor variables in the context of high-dimensionality. This method has been named PLS-LM since its combines a Partial Least Squares dimension reduction approach with a Location Model. The model has been applied using both mRNA expression and DNA copy number data. In the second contribution, a statistical criterion has been defined to test the significance of the signal captured by the first PLS component. The results have shown both the supervised classification method and the statistical criterion to be efficient on simulated data. However, in the case of real microarray data, the results are not yet so convincing. Indeed, the proposed classifiers which combine both mRNA expression and DNA copy number data do not allow the improvement of prediction performance with respect to classifiers which only consider mRNA expression data. Moreover, the test based on the proposed statistical criterion is not enough conservative when there is correlation between the variables which is the rule in microarray experiments. It suggests to improve the statistical criterion in order to use it with microarray experiments.

High-throughput microarray experiments are very complex data and deciphering their complexity remains a difficult task. As a result, simple and naive classifiers often produce the best results when a small amount of data is available, beating sophisticated models, such as the PLS-LM. For example, in the framework of survival analysis in breast cancer, Haibe-Kains et al. (2008) found that models using a single gene or a small set of biologically driven selected genes yielded similar or even better performance than models fitted from genome-wide data. Therefore, we believe that increasing the sample size would be valuable to correctly estimate on real data the parameters needed by our model on one hand, and to better capture the complexity of the data on another hand. Nevertheless, we have offered an original approach to handle mixed data in genomic study in the context of high-dimensionality. Perspectives of improvements have been proposed to increase the efficiency of the PLS-LM in real situations.



Ceiling of the Sistine Chapel
Michelangelo, *1508-1512*

Tant que tu n'atteindras pas la vérité,
tu ne pourras la corriger. Toutefois,
si tu ne peux pas la corriger, tu ne l'atteindras pas.
En attendant, ne te résigne pas.

Livre des Conseils

Conclusion

In the present manuscript, we have described the biostatistical algorithms and bioinformatics tools we have developed during the thesis. They are devoted to the analysis of DNA copy number profiles from high-throughput microarray experiments. The algorithms deal with the different steps which are necessary to analyse such data. We have developed MANOR to normalise aCGH data, GLAD to identify DNA copy number alterations from aCGH profiles and ITALICS to both normalise and identify alterations from Affymetrix data. These algorithms are available as R packages from the Bioconductor project website. They allow the extraction of relevant biological information from DNA copy number profiles. All the algorithms are integrated into CAPweb which is a web-plaform allowing the biologists to easily use our different algorithms without any particular bioinformatics skills and visualise the data with the VAMP software. In a clinical application to uveal melanoma, our biostatistical algorithms have allowed the identification of informative alterations which efficiently predict the high-risk tumours.

Besides DNA copy number profiles, many other molecular profiles are available. They represent a valuable additional information such as the widely used mRNA expression profiles. Integrative analyses are needed in order to combine these different levels of information. We have therefore proposed an original supervised classification method which combines both mRNA expression and DNA copy number profiles to build signatures for prediction of the clinical phenotype of patients. Although not completly convincing yet on real data, the proposed method has been demonstrated to improve the prediction performance in simulated data. Therefore, it is very likely that the method could be valuable for prediction purposes in some real situtations close to simulation configurations. We also conjecture that increasing sample size would improve the learning capicity of the supervised classification method. Indeed, the underlying biological mechanisms responsible for the determination of patient's clinical phenotype are heterogenous and very complex. Therefore, collecting more information is clearly needed. Typically, both more samples than we have so far and other types of molecular profiles are necessary in order to unravel efficiently the complexity of the data. Morever, we know that the genes, proteins and other biological entities cooperates with precise relationships within a huge network which is far from being understood. The method we have developed in the context of supervised classification does not incorporate information regarding this biological network. Therefore, we believe that incorporating new predictor variables which take into account the relationships between the different biologi-

cal entities would improve the prediction performance since the prediction rule will better exploit the biological knowledge. Basically, the network could be split into modules of biological entities where each module represents a biological function and renders the underlying biological mechanisms. Then, the new predictor variable could be a linear combination of the biological entities involved in the module or any other clever summarisation of the data. However, building such new predictor variables so that they are biologically relevant still remains a difficult task. Indeed, very few information is known about the biological network so that these new variables can be built accurately. Therefore, lots of effort must be paid to improve the knowledge of the network we have. This is precisely the goal of systems biology which integrates inside mathematical approaches different levels of molecular information to build models able to figure out the underlying biological processes. These models allow the decomposition of the biological network into modules which are needed to build new predictor variables.

The challenges of biostatistics and bioinformatics are definitely to provide integrative analysis methods combining information from different types of molecular profiles but also from different technologies such as microarrays and high-throughput sequencers. All these data represent highly valuable prospects and their integrative analyses will give new insigths in order to decipher the complexity of cancer.

List of publications

The exhaustive list of publications I have contributed to during my thesis is listed. The publications have been separated into four categories depending on the topics raised. For each publication, the number of citations is provided from ISI Web of Knowledge at the date of 6/11/2008.

Biostatistics publications

Brito et al. (2008): Stability-based comparison of class discovery methods for array-CGH profiles, submitted to *Bioinformatics*.
Second author. Participation in the design of the study and in the writing of the manuscript. Preprocessing of the data.
This article is supplied in the **Annexes**.
number of citations: -

Rigaill et al. (2008): ITALICS: an algorithm for normalization and DNA copy number calling for affymetrix SNP arrays, *Bioinformatics*.
First co-author with Guillem Rigaill. Design of the study. Validation of the algorithm. Writing of the manuscript.
This article is supplied as a material part of the thesis.
number of citations: 1

Neuvial et al. (2006): Spatial normalization of array-CGH data, *BMC Bioinformatics*.
First co-author with Pierre Neuvial. Design of the study, coding and validation of the spatial normalization algorithm. Writing of the manuscript.
This article is supplied as a material part of the thesis.
number of citations: 18

Rouveirol et al. (2006): Computation of recurrent minimal genomic alterations from CGH data. *Bioinformatics*.
Third author. Participation in the definition of the concept of minimal and recurrent alterations. Preprocessing of the data. Writing of some parts of the manuscript (description of the data preprocessing).
This article is supplied in the **Annexes**.
number of citations: 14

Hupé et al. (2004): Analysis of array CGH data: from signal ratio to gain and loss of DNA regions, *Bioinformatics*.
First author. Design of the study, coding and validation of the algorithm. Writing of the manuscript.
This article is supplied as a material part of the thesis.
number of citations: 97

Software and database publications

Hupé et al. (2007): ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors, *Oncogene.*
First co-author with Philippe La Rosa. Design of the study. Integration of the data. Writing of the manuscript.
This article is supplied in the **Annexes**.
number of citations: 1

La Rosa et al. (2006): VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles, *Bioinformatics.*
Third author. Participation in the definition of the software specifications, in the implementation of some analysis tools and in the writing of the manuscript.
This article is supplied in the **Annexes**.
number of citations: 21

Liva et al. (2006): CAPweb: a bioinformatics CGH array Analysis Platform, *Nucleic Acids Research.*
Second author. Participation in the design of the study. Coding of the statistical analysis part of the platform. Participation in the writing of the manuscript.
This article is supplied in the **Annexes**.
number of citations: 8

Clinical and biological publications

Trolet et al. (2008) Genomic profiling and identification of high risk tumors in uveal melanoma by array-CGH analysis of primary tumors and liver metastases, submitted to *Investigative Ophthalmology and Visual Science.*
Second author. Participation in the design of the study, in the statistical analysis of the data and in the writing of the manuscript (interpretation of the statistical analysis results, description of the statistical methods).
This article is supplied as a material part of the thesis.
number of citations: -

Volpe et al. (2008): A critical function for transforming growth factor-β, interleukin 23 and proinflammatory cytokines in driving and modulating human T_H-17 responses, *Nature Immunology.*
Multivariate statistical analysis of the data. Participation in the writing of the manuscript (interpretation of the multivariate statistical analysis results, description of the statistical methods).
This article is supplied in the **Annexes**.
number of citations: 5

Bollet et al. (2008): High resolution mapping of breakpoints to define true recurrences among ipsilateral breast tumor recurrences, *Journal of the National Cancer Institute.*
Preprocessing of the data. Participation in the statistical analysis of the data. Rereading of the manuscript.
This article is supplied in the **Annexes**.
number of citations: 1

Janoueix-Lerosey et al. (2005): Preferential occurrence of chromosome breakpoints within early replicating regions in neuroblastoma, *Cell Cycle*.
Second author. Statistical analysis of the data. Participation in the writing of the manuscript (description of the statistical methods).
This article is supplied in the **Annexes**.
number of citations: 9

Review publication

Thiery et al. (2006): Challenges in the stratification of breast tumors for tailored therapies, *Bulletin du Cancer*.
Rereading of the manuscript.
This article is supplied in the **Annexes**.
number of citations: -

Glossary

BAC

A BAC (Bacterial Artificial Chromosome) is a plasmid vector in which a DNA sequence has been inserted (from 100 up to 300 kilobases).

bagging

Bootstrap aggregating (bagging) is an algorithm to improve machine learning of classification and regression models in terms of stability and classification accuracy.

bootstrap

Bootstrap is the practice of estimating properties of an estimator by measuring those properties when sampling from an approximating distribution. This is performed by constructing a number of resamplings with replacement from the observed dataset.

carcinogenesis

see **tumoral progression**.

carcinoma

A carcinoma is a cancer which arises from epithelial cells.

chromatin

Chromatin is the complex of DNA, RNA and proteins which constitutes a chromosome. See also **nucleosome**.

CpG islands

CpG islands are regions in DNA which contain many adjacent cytosine and guanine nucleotides. The p in CpG refers to the phosphodiester bond between the C and G. These islands occur in 40% of the promoters of human genes (from Esteller, 2008).

genomic imprinting

Genomic imprinting is a genetic phenomenon by which certain genes are expressed in a parent-of-origin-specific manner. Imprinted genes are either expressed only from the allele inherited from the mother, or from the allele inherited from the father.

histone

see **nucleosome**.

leukemia

A leukemia is a malignancy of any variety of hematopoetic cell types, including the lineages leading to lymphocytes and granulocytes, in which the tumour cells are non-pigmented and dispersed throughout the circulation.

lymphoma

A lymphoma is a cancer which originates in lymphocytes (a type of white blood cell in the vertebrate immune system). There exist many types of lymphomas.

melanoma

A melanoma is a tumour arising from melanocytes, the pigmented cells of the skin and iris.

metastasis

A metastasis is a tumour growth forming at one site in the body, the cells of which derive from another tumour located elsewhere in the body.

neuroblastoma

A neuroblastoma is a paediatric extra-cranial solid tumour arising from a sympathetic nervous system tissue.

nucleosome

A nucleosome is a protein octamer composed of two types of histones among H2A, H2B, H3 and H4, and around which DNA is wrapped in **chromatin**.

oligonucleotide

An oligonucleotide is a DNA sequence (typically about 25-60 nucleotides).

oncology

Oncology is the science which studies tumours including their development, diagnosis, treatment, and prevention.

overfitting

Overfitting is fitting a statistical model which has too many parameters. As a consequence, the model has a very poor ability to generalise beyond the fitting data.

parametric

A parametric statistic is a statistic where the data are assumed to follow a given probability density function. This is in opposition to non-parametric statistics where no assumption is needed for the probability density function followed by the data.

pathologist

A pathologist is a physician who studies and diagnoses diseases through examination of organs, tissues and cells.

sarcoma

A sarcoma is a cancer of the connective or supportive tissue (bone, cartilage, fat, muscle, blood vessels) and soft tissue.

stem cell

Stem cells are cells which retain the ability to renew themselves through mitotic cell division and can differentiate into a diverse range of specialised cell types.

tumoral progression

Tumoral progression is the process of multi-step evolution of a normal cell into a tumor cell. It is also termed tumorigenesis, oncogenesis or carcinogenesis.

Bibliography

Aguilera, A. and Gómez-González, B. (2008). **Genome instability: a mechanistic view of its causes and consequences**. *Nature Reviews Genetics*, **9**:204–217.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). **Molecular Biology of the Cell**. Garland Science, Taylor & Francis Group, LLC. Fourth Edition.

Albertson, D. G., Collins, C., McCormick, F., and Gray, J. W. (2003). **Chromosome aberrations in solid tumors**. *Nature Genetics*, **34**:369–76.

Albertson, D. G. and Pinkel, D. (2003). **Genomic microarrays in human genetic disease and cancer**. *Human Molecular Genetics*, **12**:R145–R152.

Aouba, A., Péquignot, F., Toullec, A. L., and Jougla, E. (2007). **Les causes médicales de décès en France en 2004 et leur évolution 1980-2004**. *Bulletin épidémiologique hebdomadaire*, **35-36**.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). **Prediction by supervised principal components**. *Journal of The American Statistical Association*, **101**:119–137.

Barker, M. and Rayens, W. (2003). **Partial least squares for discrimination**. *Journal of Chemometrics*, **17**:166–173.

Baross, A., Delaney, A., Li, H., Nayar, T., Flibotte, S., Qian, H., Chan, S., Asano, J., Ally, A., Cao, M., Birch, P., Brown-John, M., Fernandes, N., Go, A., Kennedy, G., Langlois, S., Eydoux, P., Friedman, J., and Marra, M. (2007). **Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data**. *BMC Bioinformatics*, **8**:368.

Beroukhima, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S., Du, J., Kau, T., Thomas, R. K., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R. M., Demichelis, F., Hatton, C., Rubin, M. A., Garraway, L. A., Nelson, S. F., Liau, L., Mischel, P. S., Cloughesy, T. F., Meyerson, M., Golub, T. A., Lander, E. S., Mellinghoff, I. K., and Sellers, W. R. (2007). **Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma**. *Proceedings of the National Academy of Sciences*, **104**:20007–20012.

Billerey, C. and Boccon-Gibod, L. (1996). **Etude des variations inter-pathologistes dans l'évaluation du grade et du stade des tumeurs vésicales**. *Progrès en Urologie*, **6**:49–57.

Blow, N. (2008). **DNA sequencing: generation next-next**. *Nature Methods*, **5**:267–274.

Bollet, M. A., Servant, N., Neuvial, P., Decraene, C., Lebigot, I., Meyniel, J.-P., De Rycke, Y., Savignoni, A., Rigaill, G., Hupé, P., Fourquet, A., Sigal-Zafrani, B., Barillot, E., and Thiery, J.-P. (2008). **High resolution mapping of breakpoints to define true recurrences among ipsilateral breast tumor recurrences**. *Journal of the National Cancer Institute*, **100**:48–58.

Boulesteix, A. L. (2004a). **Dimension reduction and classification with high-dimensional microarray data**. Ph.D. thesis, Ludwig-Maximilian-Universität München.

Boulesteix, A. L. (2004b). **PLS dimension reduction for classification with microarray data**. *Statistical Applications in Genetics and Molecular Biology*, **3**:Article33.

Boulesteix, A. L. (2006). **Reader's reaction to "Dimension reduction for classification with gene expression microarray data" by Dai et al (2006)**. *Statistical Applications in Genetics and Molecular Biology*, **5**.

Boulesteix, A.-L., Porzelius, C., and Daumer, M. (2008a). **Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value**. *Bioinformatics*, **24**:1698–706.

Boulesteix, A. L. and Strimmer, K. (2007). **Partial least squares: a versatile tool for the analysis of high-dimensional genomic data**. *Briefings in bioinformatics*, **8**:32–44.

Boulesteix, A.-L., Strobl, C., Augustin, T., and Daumer, M. (2008b). **Evaluating microarray-based classifiers: an overview**. *Cancer Informatics*, **4**:77–97.

Breiman, L. (2001). **Random forests**. *Machine Learning*, **45**:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). **Classification and regression trees**. *Wadsworth International, Belmont, CA*.

Brito, I., Hupé, P., Neuvial, P., and Barillot, E. (2008). **Stability-based comparison of class discovery methods for array-CGH profiles**. *Bioinformatics*. Submitted.

Burnham, K. P. and Anderson, D. R. (2002). **Model selection and multimodel inference: a practical-theoretic approach**. Springer-Verlag.

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., and Piccart, M. J. (2006). **Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer**. *Journal of the National Cancer Institute*, **98**:1183–1192.

Bøvelstad, H., Nygård, S., Størvold, H., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjaerde, O. (2007). **Predicting survival from microarray data - a comparative study**. *Bioinformatics*, **23**:2080–2087.

Calin, G. A. and Croce, C. M. (2006a). **MicroRNA-cancer connection: the beginning of a new tale**. *Cancer Research*, **66**:7390–7394.

Calin, G. A. and Croce, C. M. (2006b). **MicroRNA signatures in human cancers**. *Nature Reviews Cancer*, **6**:857–866.

Calzone, L., Gelay, A., Zinovyev, A., Radvanyi, F., and Barillot, E. (2008). **A comprehensive modular map of molecular interactions in RB/E2F pathway**. *Molecular Systems Biology*, **4**:174.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., Teague, J. W., Menzies, A., Goodhead, I., Turner, D. J., Clee, C. M., Quail, M. A., Cox, A., Brown, C., Durbin, R., Hurles,

M. E., Edwards, P. A. W., Bignell, G. R., Stratton, M. R., and Futreal, P. A. (2008). **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing**. *Nature Genetics*, **40**:722–729.

Chen, W., Kalscheu, V., Tzschach, A., Menzel, C., Ullmann, R., Schulz, M., Erdogan, F., Li, N., Kijas, Z., Arkesteijn, G., Pajares, I. L., Goetz-Sothmann, M., Heinrich, U., Rost, I., Dufke, A., Grasshoff, U., Glaeser, B. G., Vingron, M., and Ropers, H. H. (2008). **Mapping translocation breakpoints by next-generation sequencing**. *Genome Research*, **18**:1143–1149.

Chi, K. R. (2008). **The year of sequencing**. *Nature Methods*, **5**:11–14.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., , and Gray, J. W. (2006). **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies**. *Cancer Cell*, **10**:529–541.

Chin, L. and Gray, J. W. (2008). **Translating insights from the cancer genome into clinical practice**. *Nature*, **452**:553–563.

Cianfrocca, M. and Goldstein, L. J. (2004). **Prognostic and predictive factors in early-stage breast cancer**. *The Oncologist*, **9**:606–616.

Coe, B. P., Ylstra, B., Carvalho, B., Meijer, G. A., MacAulay, C., and Lam, W. L. (2007). **Resolving the resolution of array CGH**. *Genomics*, **89**:647–653.

Comon, P. (1994). **Independent component analysis, a new concept?** *Signal Processing*, **36**:287–314.

Conde, L., Montaner, D., Burguet-Castell, J., Tarraga, J., Medina, I., Al-Shahrour, F., and Dopazo, J. (2007). **ISACGH: a web-based environment for the analysis of array CGH and gene expression which includes functional profiling**. *Nucleic Acids Research*, **35**:W81–W85.

Cowell, J. and Hawthorn, L. (2007). **The application of microarray technology to the analysis of the cancer genome**. *Current Molecular Medicine*, **7**:103–120.

Crick, F. (1970). **Central dogma of molecular biology**. *Nature*, **227**:561–563.

Croce, C. (2008). **Oncogenes and cancer**. *New England Journal of Medecine*, **358**:502–511.

Daudin, J.-J. (1986). **Selection of variables in mixed-variable discriminant analysis**. *Biometrics*, **42**:473–481.

Davies, J. J., Wilson, I. M., and Lam, W. L. (2005). **Array CGH technologies and their applications to cancer genomes**. *Chromosome Research*, **13**:237–48.

Dettling, M. (2003). **Revealing predictive gene groups with supervised algorithms**. In Kurt Hornik, F. L. and Zeileis, A., editors, *Proceedings of the Conference in Distributed Statistical Computing 2003, Vienna*.

Dettling, M. and Bühlmann, P. (2002). **Supervised clustering of genes**. *Genome Biology*, **3**:research0069.1–0069.15.

Dettling, M. and Bühlmann, P. (2004). **Finding predictive gene groups from microarray data**. *Journal of Multivariate Analysis*, **90**:106–131.

Dietterich, T. G. (1998). **Approximate statistical tests for comparing supervised classification learning algorithms**. *Neural Computation*, **10**:1895–1924.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert Jr., C. J., Weber, B. L., Maris, J. M., and Grant, G. R. (2006). **STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments**. *Genome Research*, **16**:1149–1158.

Do, J. H. and Choi, D. K. (2006). **Normalization of microarray data: single-labeled and dual-labeled arrays**. *Molecules and Cells*, **22**:254–261.

Donoho, D. (2000). **High-dimensional data analysis: The curses and blessings of dimensionality**. American Math. Society "Math Challenges of the 21st Century" (2000).

Douglas, E. J., Fiegler, H., Rowan, A., Halford, S., Bicknell, D. C., Bodmer, W., Tomlinson, I. P. M., and Carter, N. P. (2004). **Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas**. *Cancer Research*, **64**:4817–4825.

Dudoit, S. and Fridlyand, J. (2003). **Classification of microarray experiments**. In Speed, T., editor, *Statistical Analysis of Gene Expression Microarray Data*. CHAPMAN & HALL/CRC.

Díaz-Uriarte, R. (2003). **A simple method for finding molecular signatures from gene expression data**. *ESF Workshop on genomic approaches to microarray data analysis, Madrid, Spain, October 30-31, 2003*.

Díaz-Uriarte, R. and Rueda, O. M. (2007). **ADaCGH: a parallelized web-based application and R package for the analysis of aCGH data**. *PLoS ONE*, **2**:e737.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics*, **21**:171–178.

Ellenberg, J. H. (1973). **The joint distribution of the standardized least squares residuals from a general linear regression**. *Journal of the American Statistical Association*, **68**:941–943.

Ellis, I. O., Galea, M., Broughton, N., Locker, A., Blamey, R. W., and Elston, C. W. (1992). **Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up**. *Histopathology*, **20**:479–89.

Elston, C. W. and Ellis, I. O. (1991). **Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up**. *Histopathology*, **19**:403–10.

Esquela-Kerscher, A. and Slack, F. J. (2006). **Oncomirs? microRNAs with a role in cancer**. *Nature Reviews Cancer*, **6**:259–269.

Esteller, M. (2007). **Cancer epigenomics: DNA methylomes and histone-modification maps**. *Nature Reviews Genetics*, **8**:286–298.

Esteller, M. (2008). **Epigenetics in cancer**. *New England Journal of Medecine*, **358**:1148–1159.

Fabbri, M., Croce, C. M., and Calin, G. A. (2008). **MicroRNAs**. *The Cancer Journal*, **14**:1–6.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). **Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?** *Nature Reviews Genetics*, **9**:102–114.

Finetti, P., Cervera, N., Charafe-Jauffret, E., Chabannon, C., Charpin, C., Chaffanet, M., Jacquemier, J., Viens, P., Birnbaum, D., and Bertucci, F. (2008). **Sixteen-kinase gene expression identifies luminal breast cancers with poor prognosis**. *Cancer Research*, **68**:767–776.

Frank, I. and Friedman, J. (1993). **A statistical view of some chemometrics regression tools**. *Technometrics*, **35**:109–148.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W., and Lee, C. (2006). **Copy number variation: new insights in genome diversity**. *Genome Research*, **16**:949–961.

Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). **Application of Hidden Markov Models to the analysis of the array CGH data**. *Journal of Multivariate Analysis*, **90**:132–153. Special Issue on Multivariate Methods in Genomic Data Analysis.

Frigola, J., Song, J., Stirzaker, C., Hinshelwood, R. A., Peinado, M. A., and Clark, S. J. (2006). **Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band**. *Nature Genetics*, **38**:540–549.

Fuhrmann, C., Schmidt-Kittler, O., Stoecklein, N. H., Petat-Dutter, K., Vay, C., Bockler, K., Reinhardt, R., Ragg, T., and Klein, C. A. (2008). **High-resolution array comparative genomic hybridization of single micrometastatic tumor cells**. *Nucleic Acids Research*, **36**:e39.

Geman, S. (1980). **A limit theorem for the norm of random matrices**. *The Annals of Probability*, **2**:252–261.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). **Bioconductor: Open software development for computational biology and bioinformatics**. *Genome Biology*, **5**:R80.

Grant, G. R., Manduchi, E., and Stoeckert, C. J. (2007). **Analysis and management of microarray gene expression data**. *Current Protocols in Molecular Biology*, **Supplement 77**:UNIT 19.6.

Guyon, I. and Elisseeff, A. (2003). **An introduction to variable and feature selection**. In Guyon, I. and Elisseeff, A., editors, *Special issue on variable and feature selection*, Journal of Machine Learning Research, pages 1157–1182.

Haibe-Kains, B., Desmedt, C., Sotiriou, C., and Bontempi, G. (2008). **A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?** *Bioinformatics*.

Hanahan, D. and Weinberg, R. A. (2000). **The hallmarks of cancer**. *Cell*, **100**:57–70.

Hastie, T., Tibshirani, R., and Friedman, J. (2003). **The Elements of Statistical Learning - Data Mining, Inference, and Prediction**. Springer Series in Statistics. Springer.

Hoheisel, J. D. (2006). **Microarray technology: beyond transcript profiling and genotype analysis**. *Nature Reviews Genetics*, **7**:200–210.

Hudis, C. A. (2007). **Trastuzumab - mechanism of action and use in clinical practice**. *New England Journal of Medecine*, **357**:39–41.

Hupé, P., La Rosa, P., Liva, S., Lair, S., Servant, N., and Barillot, E. (2007). **ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors**. *Oncogene*, **26**:6641–52.

Hupé, P., Stransky, N., Thiery, J. P., Radvanyi, F., and Barillot, E. (2004). **Analysis of array CGH data: from signal ratio to gain and loss of DNA regions**. *Bioinformatics*, **20**:3413–3422.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). **Detection of large-scale variation in the human genome**. *Nature Genetics*, **36**:949–951.

Idbaih, A., Boisselier, B., Sanson, M., Crinière, E., Liva, S., Marie, Y., Carpentier, C., Paris, S., Laigle-Donadey, F., and Mokhtari, K. (2007). **Tumor genomic profiling and TP53 germline mutation analysis of first-degree relative familial gliomas**. *Cancer Genetics and Cytogenetics*, **176**:121–126.

Idbaih, A., Kouwenhoven, M., Jeuken, J., Carpentier, C., Gorlia, T., Kros, J. M., French, P., Teepen, J. L., Delattre, O., Delattre, J.-Y., van den Bent, M., and Hoang-Xuan, K. (2008). **Chromosome 1p loss evaluation in anaplastic oligodendrogliomas**. *Neuropathology*, **28**:440–443.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics*, **4**:249–264.

Irizarry, R. A., Wu, Z., and Jaffee, H. A. (2006). **Comparison of affymetrix genechip expression measures**. *Bioinformatics*, **22**:789–794.

Ishkanian, A. S., Malloff, C. A., Watson, S. K., DeLeeuw, R. J., Chi, B., Coe, B. P., Snijders, A., Albertson, D. G., Pinkel, D., Marra, M. A., Ling, V., MacAulay, C., and Lam, W. L. (2004). **A tiling resolution DNA microarray with complete coverage of the human genome**. *Nature Genetics*, **36**:299–303.

Jablonka, E. and Lamb, M. J. (2002). **The changing concept of epigenetics**. *Annals of the New York Academy of Sciences*, **981**:82–96.

Janoueix-Lerosey, I., Hupé, P., Maciorowski, Z., La Rosa, P., Schleiermacher, G., Pierron, G., Liva, S., Barillot, E., and Delattre, O. (2005). **Preferential occurrence of chromosome breakpoints within early replicating regions in neuroblastoma**. *Cell Cycle*, **4**:1842–1846.

Johnson, S. A. and Hunter, T. (2005). **Kinomics: methods for deciphering the kinome**. *Nature Methods*, **2**:17–25.

Jong, K., Marchiori, E., Van der Vaart, A., Ylstra, B., Weiss, M., and Meijer, G. (2003). **Chromosomal breakpoint detection in human cancer**. In Raidl, G. R., Cagnoni, S., Cardalda, J. J. R., Corne, D. W., Gottlieb, J., Guillot, A., Hart, E., Johnson, C. G., Marchiori, E., Meyer, J.-A., and Middendorf, M., editors, *Applications of Evolutionary Computing, EvoWorkshops2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, EvoSTIM*, volume 2611 of *LNCS*, pages 54–65. Springer-Verlag, University of Essex, England, UK.

Kallioniemi, A. (2007). **CGH microarrays and cancer**. *Current Opinion in Biotechnology*, **18**:1–5.

Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F. M., and Pinkel, D. (1992). **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors**. *Science*, **258**:818–821.

Kim, E., Goren, A., and Ast, G. (2008). **Insights into the connection between cancer and alternative splicing**. *Trends in Genetics*, **24**:7–10.

Knudson, A. G. (1971). **Mutation and cancer: Statistical study of retinoblastoma**. *Proceedings of the National Academy of Sciences*, **68**:820–823.

Koren, A., Tirosh, I., and Barkai, N. (2007). **Autocorrelation analysis reveals widespread spatial biases in microarray experiments**. *BMC Genomics*, **8**:164.

La Rosa, P., Viara, E., Hupé, P., Pierron, G., Liva, S., Neuvial, P., Brito, I., Lair, S., Servant, N., Robine, N., Manié, E., Brennetot, C., Jannoueix-Lerosey, I., Raynal, V., Gruel, N., Rouveirol, C., Stransky, N., Stern, M.-H., Delattre, O., Aurias, A., Radvanyi, F., and Barillot, E. (2006). **VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles**. *Bioinformatics*, **22**:2066–2073.

Lai, W., Choudhary, V., and Park, P. J. (2008). **CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms**. *Bioinformatics*, **24**:1014–1015.

Lai, W., Johnson, M., Kucherlapati, R., and Park, P. (2005). **Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data**. *Bioinformatics*, **21**:3763–3770.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., and Robles, V. (2006). **Machine learning in bioinformatics**. *Briefings in Bioinformatics*, **7**:86–112.

Lee, H., Kong, S. W., and Park, P. J. (2008). **Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes**. *Bioinformatics*, **24**:889–896.

Li, B., Morris, J., and Martin, E. B. (2002). **Model selection for partial least squares regression**. *Chemometrics and Intelligent Laboratory Systems*, **64**:79–89.

Liva, S., Hupé, P., Neuvial, P., Brito, I., Viara, E., La Rosa, P., and Barillot, E. (2006). **CAPweb: a bioinformatics CGH array Analysis Platform**. *Nucleic Acids Research*, **34**:477–481.

Lowery, A. J., Miller, N., McNeill, R. E., and Kerin, M. J. (2008). **MicroRNAs as prognostic indicators and therapeutic targets: Potential effect on breast cancer management**. *Clinical Cancer Research*, **14**:360–365.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R., and Golub, T. R. (2005). **MicroRNA expression profiles classify human cancers**. *Nature*, **435**:834–838.

Lønning, P. E. (2007). **Breast cancer prognostication and prediction: are we making progress?** *Annals of Oncology*, **18**:viii3–viii7.

MacLachlan (1992). **Discriminant analysis and statistical pattern recognition**. New York: John Wiley & Sons.

Mardis, E. R. (2008). **The impact of next-generation sequencing technology on genetics**. *Trends in Genetics*, **24**:133–141.

Mattick, J. S. (2003). **Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms**. *BioEssays*, **25**:930–939.

Mattick, J. S. and Makunin, I. V. (2006). **Non-coding RNA**. *Human Molecular Genetics*, **15**:R17–R29.

Mello, C. C. and Conte Jr, D. (2004). **Revealing the world of RNA interference**. *Nature*, **43**:338–342.

Michiels, S., Koscielny, S., and Hill, C. (2005). **Prediction of cancer outcome with microarrays: a multiple random validation strategy**. *The Lancet*, **365**:488–492.

Miller, L. and Liu, E. (2007). **Expression genomics in breast cancer research: microarrays at the crossroads of biology and medicine**. *Breast Cancer Research*, **9**:206.

Mitelman, F., Johansson, B., and Mertens, F. (2007). **The impact of translocations and gene fusions on cancer causation**. *Nature Reviews Cancer*, **7**:233–245.

Molinaro, A., Simon, R., and Pfeiffer, R. (2005). **Prediction error estimation: a comparison of resampling methods**. *Bioinformatics*, **21**:3301–3307.

Morris, J. A., Gayther, S. A., Jacobs, I. J., and Jones, C. (2008). **A suite of perl modules for handling microarray data**. *Bioinformatics*, **24**:1102–1103.

Nadal, J. M., Jouve, M., Mosseri, V., Asselain, B., and Pouillart, P. (1988). **Cancer métastasé du sein traité par polychimiothérapie: une nouvelle approche du pronostic.** *Bulletin du Cancer*, **75**:757–769.

Nakamura, Y. (1998). **ATM: the p53 booster**. *Nature Medicine*, **4**:1231–1232.

Nature Publishing Group (2007). **DNA Technologies - Milestones timeline**. *Nature Milestones*. http://www.nature.com/milestones/miledna/timeline.html.

Neuvial, P., Hupé, P., Brito, I., Liva, S., Manié, E., Brennetot, C., Radvanyi, F., Aurias, A., and Barillot, E. (2006). **Spatial normalization of array-CGH data**. *BMC Bioinformatics*, **7**:264.

Nguyen, D. and Rocke, D. (2002). **Tumor classification by partial least squares using microarray gene expression data**. *Bioinformatics*, **18**:39–50.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). **Circular binary segmentation for the analysis of array-based DNA copy number data**. *Biostatistics*, **5**:557–572.

Osten, D. W. (1988). **Selection of optimal regression models via cross-validation**. *Journal of chemometrics*, **2**:39–48.

Perkins, D. O., Jeffries, C., and Sullivan, P. (2005). **Expanding the 'central dogma': the regulatory role of nonprotein coding genes and implications for the genetic liability to schizophrenia**. *Molecular Psychiatry*, **10**:69–78.

Perou, C., Jeffrey, S., Van de Rijn, M., Rees, C., Eisen, M., Ross, D., Pergamenschikov, A., Williams, C., Zhu, S., Lee, J., Lashkari, D., Shalon, D., Brown, P., and Botstein, D. (1999). **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers**. *Proceedings of the National Academy of Sciences*, **96**:9212–9217.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005). **A statistical approach for array CGH data analysis**. *BMC Bioinformatics*, **6**:27.

Pinkel, D. and Albertson, D. G. (2005). **Array comparative genomic hybridization and its applications in cancer**. *Nature Genetics*, **37 Suppl**:11–17.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W., and Albertson, D. G. (1998). **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays**. *Nature Genetics*, **20**:207–211.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). **Genome-wide analysis of DNA copy-number changes using cDNA microarrays**. *Nature Genetics*, **23**:41–46.

Quackenbush, J. (2002). **Microarray data normalization and transformation**. *Nature Genetics*, **32**:496–501.

R Development Core Team (2008). **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ransohoff, D. (2004). **Rules of evidence for cancer molecular-marker discovery and validation**. *Nature Reviews Cancer*, **4**:309–314.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.-P. (2007). **Classification of microarray data using gene networks**. *BMC Bioinformatics*, **8**:35.

Reinders, J., Vivier, C. D., Theiler, G., Chollet, D., Descombes, P., and Paszkowski, J. (2008). **Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion**. *Genome Research*, **18**:469–76.

Rigaill, G., Hupé, P., Almeida, A., La Rosa, P., Meyniel, J.-P., Decraene, C., and Barillot, E. (2008). **ITALICS: an algorithm for normalization and DNA copy number calling for affymetrix SNP arrays**. *Bioinformatics*, **24**:768–774.

Rosipal, R. and Trejo, L. J. (2001). **Kernel partial least squares regression in reproducing kernel Hilbert space**. *Journal of Machine Learning Research*, **2**:97–123.

Rouveirol, C., Stransky, N., Hupé, P., La Rosa, P., Viara, E., Barillot, E., and Radvanyi, F. (2006). **Computation of recurrent minimal genomic alterations from CGH data**. *Bioinformatics*, **22**:849–56.

Rusk, N. and Kiermer, V. (2008). **Primer: Sequencing - the next generation**. *Nature Methods*, **5**:15.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). **A review of feature selection techniques in bioinformatics**. *Bioinformatics*, **23**:2507–2517.

Sakamoto, Y. (1982). **Efficient use of Akaike's information criterion for model selection in high dimensional contingency table analysis**. *Metron*, **15**:257–276.

Saporta, G. (1990). **Probabilités, analyse des données et statistique**. Editions TECHNIP.

Satzinger, H. (2008). **Theodor and Marcella Boveri: chromosomes and cytoplasm in heredity and development**. *Nature Reviews Genetics*, **9**:231–238.

Sawyers, C. L. (2008). **The cancer biomarker problem**. *Nature*, **452**:548–552.

Schones, D. E. and Zhao, K. (2008). **Genome-wide approaches to studying chromatin modifications**. *Nature Reviews Genetics*, **9**:179–191.

Schrödinger, E. (1944). **Qu'est-ce que la vie?** Christian Bourgois Editeur, 1986.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T., and Petronis, A. (2006). **Microarray-based DNA methylation profiling: technology and applications**. *Nucleic Acids Research*, **34**:528–542.

Schuster, S. C. (2007). **Next-generation sequencing transforms today's biology**. *Nature Methods*, **5**:16–18.

Sen, A. and Srivastava, M. S. (1975). **On tests for detecting a change in mean**. *Ann. Statist.*, **3**:98–108.

Shann, Y., Cheng, C., Chiao, C., Chen, D., Li, P., and Hsu, M. (2008). **Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines**. *Genome Research*, **18**:791–801.

Sherr, C. J. (2004). **Principles of tumor suppression**. *Cell*, **116**:235–246.

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003). **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification**. *Journal of the National Cancer Institute*, **95**:14–18.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T., and Lichter, P. (1997). **Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances**. *Genes Chromosomes Cancer*, **20**:399–407.

Southern, E., Mir, K., and Shchepinov, M. (1999). **Molecular interactions on microarrays**. *Nature Genetics*, **21**:5–9.

Speybroeck, L. V. (2002). **From epigenesis to epigenetics: The case of C. H. Waddington**. *Annals of the New York Academy of Sciences*, **981**:61–81.

Srebrow, A. and Kornblihtt, A. R. (2006). **The connection between splicing and cancer**. *Journal of Cell Science*, **119**:2635–2641.

Staaf, J., Jönsson, G., Ringnér, M., and Vallon-Christersson, J. (2007). **Normalization of array-CGH data: influence of copy number imbalances**. *BMC Genomics*, **8**:382.

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S. G. D., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., Graham, A., Southgate, J., Asselain, B., Allory, Y., Abbou, C. C., Albertson, D. G., Thiery, J. P., Chopin, D. K., Pinkel, D., and Radvanyi, F. (2006). **Regional copy number-independent deregulation of transcription in cancer**. *Nature Genetics*, **38**:1386–1396.

Stratton, M. and Rahman, N. (2008). **The emerging landscape of breast cancer susceptibility**. *Nature Genetics*, **40**:17–22.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences*, **102**:15545–15550.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., de Rijn, M. V., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Børresen-Dale, A.-L. (2001). **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proceedings of the National Academy of Sciences*, **98**:10869–10874.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C., Lønning, P., Brown, P., Børresen-Dale, A., and Botstein, D. (2003). **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proceedings of the National Academy of Sciences*, **100**:8418–8423.

Tenenhaus, M. (1998). **La régression PLS - Théorie et pratique**. Editions TECHNIP.

Thiery, J.-P., Sastre-Garau, X., Vincent-Salomon, B., Sigal-Zafrani, X., Pierga, J., Decraene, C., Meyniel, J.-P., Gravier, E., Asselain, B., Rycke, Y. D., Hupé, P., Barillot, E., Ajaz, S., Faraldo, M., Deugnier, M., Glukhova, M., and Medina, D. (2006). **Challenges in the stratification of breast tumors for tailored therapies**. *Bulletin du Cancer*, **93**:E81–9.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., Helden, J. V., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). **Assessing computational tools for the discovery of transcription factor binding sites**. *Nature Biotechnology*, **23**:137–144.

Trolet, J., Hupé, P., Huon, I., Lebigot, I., Mariani, P., Plancher, C., Asselain, B., Desjardins, L., Delattre, O., Sastre-Garau, X., Thiery, J.-P., Saule, S., Piperno-Neumann, S., Barillot, E., and Couturier, J. (2008). **Genomic profiling and identification of high risk tumors in uveal melanoma by array-CGH analysis of primary tumors and liver metastases**. *Investigative Ophthalmology and Visual Science*. Under revision.

Turner, B. M. (2002). **Cellular memory and the histone code**. *Cell*, **111**:285–291.

Unger, K., Malisch, E., Thomas, G., Braselmann, H., Walch, A., Jackl, G., Lewis, P., Lengfelder, E., Bogdanova, T., Wienberg, J., and Zitzelsberger, H. (2008). **Array CGH demonstrates characteristic aberration signatures in human papillary thyroid carcinomas governed by RET/PTC**. *Oncogene*, **27**:4592–4602.

Vaidya, J. S. (2007). **Breast cancer: an artistic view**. *The Lancet Oncology*, **8**:583–585.

Van de Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., Van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). **A gene-expression signature as a predictor of survival in breast cancer**. *New England Journal of Medecine*, **347**:1999–2009.

Van de Wiel, M. A. and Van Wieringen, W. M. (2007). **CGHregions: dimension reduction for array CGH data with minimal information loss**. *Cancer Informatics*, **2**:55–63.

Van der Vaart, A. W. (1998). **Asymptotic Statistics**. Cambridge University Press.

Van Wieringen, W. M., Van De Wiel, M. A., and Ylstra, B. (2007). **Normalized, segmented or called aCGH data?** *Cancer Informatics*, **3**:331–337.

Van't Veer, L. J. and Bernards, R. (2008). **Enabling personalized cancer medicine through analysis of gene-expression patterns**. *Nature*, **452**:564–570.

Van't Veer, L. J., Dai, H., Van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., Van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature*, **415**:530–536.

Venables, J. P. (2004). **Aberrant and alternative splicing in cancer**. *Cancer Research*, **64**:7647–7654.

Vermeer, M. H., van Doorn, R., Dijkman, R., Mao, X., Whittaker, S., van Voorst Vader, P. C., Gerritsen, M.-J. P., Geerts, M.-L., Gellrich, S., Söderberg, O., Leuchowius, K.-J., Landegren, U., Out-Luiting, J. J., Knijnenburg, J., IJszenga, M., Szuhai, K., Willemze, R., and Tensen, C. P. (2008). **Novel and highly recurrent chromosomal alterations in sézary syndrome**. *Cancer Research*, **68**:2689–2698.

Vermeulen, L., Sprick, M., Kemper, K., Stassi, G., and Medema, J. (2008). **Cancer stem cells - old concepts, new insights**. *Cell Death and Differentiation*, **15**:947–58.

Vogelstein, B. and Kinzler, K. W. (2004). **Cancer genes and the pathways they control**. *Nature Medicine*, **10**:789–799.

Volpe, E., Servant, N., Zollinger, R., Bogiatzi, S. I., Hupé, P., Barillot, E., and Soumelis, V. (2008). **A critical function for transforming growth factor-β, interleukin 23 and proinflammatory cytokines in driving and modulating human T_H-17 responses**. *Nature Immunology*, **9**:650–7.

Wang, Y., Miller, D. J., and Clarke, R. (2008). **Approaches to working in high-dimensional data spaces: gene expression microarrays**. *British Journal of Cancer*, **98**:1023–1028.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., Surani, M., Sakaki, Y., and Sasaki, H. (2008). **Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes**. *Nature*, **453**:539–543.

Weinberg, R. A. (2007). **The biology of cancer**. Garland Science, Taylor & Francis Group, LLC.

Willenbrock, H. and Fridlyand, J. (2005). **A comparison study: applying segmentation to array CGH data for downstream analyses**. *Bioinformatics*, **21**:4084–4091.

Wold, H. (1966). **Estimation of principal component and related models by iterative least squares**. In Krishnaiah, P. R., editor, *Multivariate Analysis*, pages 391–420. New-York: Academic Press.

Wold, S., Martens, H., and Wold, H. (1982). **The multivariate calibration problem in chemistry solved by the PLS method**. In Ruhe, A. and Kåstrøm, B., editors, *Matrix Pencils*, Lecture Notes in Mathematics, pages 286–293. Springer Berlin / Heidelberg.

Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M., and Spencer, F. (2003). **A model based background adjustment for oligonucleotide expression arrays**. Technical report, John Hopkins University, Department of Biostatistics Working Papers, Baltimore, MD.

Ylstra, B., Van den Ijssel, P., Carvalho, B., Brakenhoff, R. H., and Meijer, G. A. (2006). **BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH)**. *Nucleic Acids Research*, **34**:445–450.

Yu, L. and Liu, H. (2004). **Efficient feature selection via analysis of relevance and redundancy**. *Journal of Machine Learning Research*, **5**:1205–1224.

Zou, H. and Hastie, T. (2005). **Regularization and variable selection via the elastic net**. *Journal of the Royal Statistical Society, Series B*, **67**:301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). **On the "degrees of freedom" of the lasso**. *Annals of Statistics*, **35**:2173–2192.

Annexes

A.1 Publications supplied as supplementary materials

The following publications are supplied:

- Brito et al. (2008)
- Bollet et al. (2008)
- Volpe et al. (2008)
- Hupé et al. (2007)
- Liva et al. (2006)
- La Rosa et al. (2006)
- Thiery et al. (2006)
- Rouveirol et al. (2006)
- Janoueix-Lerosey et al. (2005)

Stability-based comparison of class discovery methods for array-CGH profiles

Journal:	*Bioinformatics*
Manuscript ID:	BIOINF-2008-1477
Category:	Original Paper
Date Submitted by the Author:	19-Sep-2008
Complete List of Authors:	Brito, Isabel; Institut Curie, Service de Bioinformatique; INSERM, U900; Mines Paris Tech, Mines Paris Tech Hupé, Philippe; Institut Curie, Servive Bioinformatique Neuvial, Pierre; Institut Curie, Service de Bioinformatique; INSERM, U900; Mines Paris Tech, Mines Paris Tech Barillot, Emmanuel; Institut Curie, Service de Bioinformatique; INSERM, U900; Mines Paris Tech, Mines Paris Tech
Keywords:	class discovery, Cancer, array-CGH, cluster stability

scholarONE™
Manuscript Central

BIOINFORMATICS

Vol. 00 no. 00
Pages 1–7

Stability-based comparison of class discovery methods for array-CGH profiles

Isabel Brito [a,b,c,*], Philippe Hupé [a,b,c,d], Pierre Neuvial [a,b,c], Emmanuel Barillot [a,b,c]

[a]Institut Curie, 26 rue d'Ulm, Paris Cedex 05, F-75248 France,
[b]INSERM, U900, Paris, F-75248 France,
[c]Mines ParisTech, Fontainebleau, F-77300 France,
[d]CNRS UMR144, Paris, F-75248 France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

For Peer Review

ABSTRACT

Motivation: Array-CGH can be used to determine DNA copy number, imbalances in which are a fundamental factor in the genesis and progression of tumors. The discovery of classes with similar patterns of array-CGH profiles therefore adds to our understanding of cancer and the treatment of patients. Various input data representations for array-CGH, dissimilarity measures between tumor samples and clustering algorithms may be used for this purpose. The choice between procedures is often difficult. An evaluation procedure is therefore required to select the best class discovery method (combination of one input data representation, one dissimilarity measure and one clustering algorithm) for array-CGH. Robustness of the resulting classes is a common requirement, but no stability-based comparison of class discovery methods for array-CGH profiles has ever been reported.

Results: We applied several class discovery methods and evaluated the stability of their solutions, with a modified version of Bertoni's χ^2-based test (Bertoni and Valentini, 2007). We conclude that `Minimal Regions` of alteration (a concept introduced by Rouveirol *et al.*, 2006) for input data representation, `sim` (Liu *et al.*, 2006) or `agree` (van Wieringen *et al.*, 2008) for dissimilarity measure and the use of `average` group distance in the clustering algorithm produce the most robust classes of array-CGH profiles.

Availability: The software is available from `http://bioinfo.curie.fr/projects/cgh-clustering`. It has also been partly integrated into VAMP (La Rosa *et al.*, 2006). The data sets used are publicly available from ACTuDB (Hupé *et al.*, 2007).

Supplementary Material: Certain mathematical definitions and tables of results may be obtained from *Bioinformatics* online.

Contact: isabel.brito@curie.fr

1 INTRODUCTION

Recurrent non random genomic alterations, including changes in DNA copy number in particular, are hallmarks of cancer. The characterization of these imbalances is critical to our understanding of tumorigenesis and cancer progression (Albertson *et al.*, 2003, Mitelman, 2005).

Comparative Genomic Hybridization (CGH) is a molecular cytogenetics technique for the efficient characterization of chromosomal gains and losses. Two differently labeled tumoral (test) and healthy (reference) DNA samples are hybridized with normal metaphase chromosome. The relative intensity of the test signal over the reference signal (the signal ratio) reflects the imbalance in copy number between the two samples at a given location (for statistical reasons, ratio are log-transformed and the signal will be termed log-ratio hereafter). The initial resolution of the technique (about 10 Mbp) improved considerably with the advent of array-based Comparative Genomic Hybridization (array-CGH) in the late 1990s (Solinas-Toldo *et al.*, 1997; Pinkel *et al.*, 1998). Array-CGH initially used BAC clone arrays (Snijders *et al.*, 2001) or cDNA arrays (Pollack *et al.*, 1999). More recently, the use of oligonucleotide arrays (Lucito *et al.*, 2003; Carvalho *et al.*, 2004) or tiling-resolution arrays (Ishkanian *et al.*, 2004) has further improved the sensitivity and resolution of the technique (typically 20-80 bp for oligonucleotide arrays and about 100 kbp for BAC arrays).

The identification of tumor classes is an important step in cancer research. A class is defined as a family of tumors with similar biological traits and similar clinical features. Class discovery methods have been extensively used for expression data (Quackenbush, 2006 or Thalamuthu *et al.*, 2006), particularly for tumor classification (e.g. Sørlie *et al.*, 2003). In this respect, DNA copy number is as crucial as mRNA expression, and biologists and clinicians routinely make use of information concerning genome alterations to investigate tumor biology and to treat patients. For example, chromosome 3 monosomy is used as an indicator of high metastatic risk in uveal melanoma, whereas EGFR amplification is an indication for trastuzumab treatment in breast cancer (Vogel *et al.*, 2002). However, array-CGH data have specific features differentiating them from expression array data. First, the log-ratio signals calculated have a small range, which may be discretized into different classes: loss, normal, gain and amplification. Second, neighboring genomic segments are likely to be altered in the same way. Due to these particular features, class discovery for array-CGH data merits a separate analysis, and this constitutes the scope of our work.

*to whom correspondence should be addressed

© Oxford University Press .

Only a few studies dedicated to class discovery for CGH or array-CGH data have been published. Mattfeldt *et al.* (2001), Liu *et al.* (2006) and Liu *et al.* (2007) examined chromosomal CGH data whereas van Wieringen *et al.* (2008) explored array-CGH data.

Liu *et al.* (2006) stressed the unusual nature of CGH data and recommended the use of particular dissimilarity measures. They proposed several different dissimilarity measures, the most original of which is `sim`, which measures the number of contiguous genomic intervals of alterations of the same type overlapping between pairs of samples. Liu *et al.* (2007) presented an algorithm for identifying small sets of important genomic intervals called markers. They showed that markers distinguished effectively between different histological cancer types, thereby improving the quality of clustering.

van Wieringen *et al.* (2008) proposed the WECCA algorithm (weighted clustering of called array-CGH data), a method including a dissimilarity measure and a clustering algorithm devoted to array-CGH data. They defined two dissimilarity measures based on the concepts of agreement (`agree`) and concordance (`conc`). `Agree` is defined as the probability of alterations being identical at the same location in two different samples, under a null model. `Conc` reflects the similarity in ordering of the types of alteration in two different samples. The clustering algorithm functions as an agglomerative linkage adapted to these two dissimilarity measures and is called `total`. van Wieringen *et al.* (2008) demonstrated that `total` linkage is likely to produce tight clusters. Moreover, WECCA produces clusters strongly associated with survival.

Continuing on from these studies, we compared several class discovery methods with a view to identifying the method most appropriate for array-CGH data. We define a class discovery method as the combination of an input data representation, a dissimilarity measure and a clustering algorithm. In many fields, biology and cancer research in particular, it is important for the classes identified to be statistically stable. However, the stability of the classes obtained has never before been estimated for array-CGH data. We therefore tried to determine the best way to obtain stable classes of tumors. Stability is defined as follows: if the class discovery method is applied repeatedly to independent samples and generates similar solutions in each case, then it may be considered statistically stable.

This paper is structured as follows. In Section 2, we discuss several possibilities for representing the input data of an array-CGH experiment. Section 3 provides a description of array-CGH data preprocessing. In Section 4, we present the dissimilarity measures and clustering algorithms used in this article, and the stability-based validation method applied. Section 5 shows results for several public data sets. In Section 6 we present and discuss our results. The mathematical definitions used throughout this article and some results tables are provided as **Supplementary Material**.

2 INPUT DATA REPRESENTATION STRATEGIES

In an array-CGH experiment, a signal intensity is measured for each probe, for the tumor sample and the reference. The log-ratio of the signal for the sample to the signal for the reference is calculated and denoted signal log-ratio. These log-ratios may be used directly or further processed before their use as input data for classification. It remains unclear which input data representation is optimal for class discovery. Below, we consider several strategies for input data representation for array-CGH classification.

2.1 Strategies using "`All probes`"

These strategies are straightforward, as they make use of all probes. The input data representation for each probe may be:

- `log-ratio` - data are expressed on the base 2 logarithmic scale. This representation is the most common in array-CGH data analysis.
- `smoothed log-ratio` - the log-ratio of the probe is smoothed using its neighbors in the genome. In algorithms such as GLAD (Hupé *et al.*, 2004), the smoothed log-ratio values are calculated by estimating a piecewise constant function of the raw log-ratios, using a segmentation procedure. GLAD uses an adaptive weight smoothing algorithm, ensuring that only neighboring probes with similar DNA copy numbers are smoothed together. Several algorithms for the segmentation of array-CGH data have been described (see Lai *et al.*, 2005, for a review).
- `calls` - the data are encoded as discrete and ordinal variables: the calls may be -1 for a probe corresponding to a region of loss, 0 for a normal region, 1 for a region of gain and 2 for a region of amplification.

2.2 Strategies using "`Data compression`"

In array-CGH data, some probes may be redundant because neighboring genomic segments are likely to be altered in the same way. Data compression strategies involve reducing the number of dimensions so that only a few relevant variables are handled.

`Statistical compression` The number of dimensions is reduced by Principal Components Analysis (PCA). `PCA` computes a linear combination of probes that jointly account for most of the variability in the data. `PCA` is carried out on log-ratio values and the first components identified constitute the input data representation associated with this strategy.

`Biological compression` Variable compression is based on the concept of *Minimal Regions* (Rouveirol *et al.*, 2006). A *Minimal Region* (`MR`) is defined as the largest sequence of altered probes (contiguous probes with identical, and not normal, calls) common to a subset of array-CGH profiles, called *support*. Each `MR` is coded as 1 if the sample belongs to the *support* and as 0 if it does not. Other concepts similar to `MR` have been proposed, such as *markers* (Liu *et al.*, 2007), *SIRAC* (Lai *et al.*, 2007) and *CGHregions* (van de Wiel and van Wieringen, 2007).

3 DATA PRE-PROCESSING

3.1 Data sets

We used five array-CGH data sets publicly available from ACTuDB (Hupé *et al.*, 2007). The following table provides a brief description of each data set, with all datasets identified by the name of the first author.

For Peer Review

data set	no. of arrays	no. of probes	platform	tumor tissue
Blaveri *et al.*, 2005	98	2146	HumArray 2.0	bladder
Gysin *et al.*, 2005	25	2415	HumArray 2.0	pancreas
Patil *et al.*, 2005	49	2385	HumArray 1.14	liver
Douglas *et al.*, 2004	85	3127	BAC/PAC	colon
Veltman *et al.*, 2003	49	1741	HumArray 1.11	bladder

3.2 Missing values

Array-CGH experiments, like most microarray experiments, often generate missing values, due to poor hybridization, high levels of heterogeneity between replicates, image corruption or scratches on the slide.

Several methods have been used to impute missing values for expression data (Brock *et al.*, 2008). We propose a novel missing value imputation method more appropriate for array-CGH data. Our method is applied to each sample independently and is based on the *genome metric*. Each probe is assigned a chromosomal position, deduced from its distance in base pairs from the *p*-telomere.

Missing values are imputed as follows. Let us assume that probe i has a missing value in one sample. Denote by a and b the two probes closest (on the left and right, respectively) to i in the sample.

1. If a and b have the same call, they probably belong to the same genomic region and i naturally belongs to that region, with i given by:
   - The log-ratio average of a and b as the log-ratio of i,
   - The interpolation value of the smoothing points of a and b as the *smoothed log-ratio* of i,
   - The call of a and b as a *call*,
2. If a and b have different calls,
   2.1. If one is normal, preference is given to the alteration and i is given the call, log-ratio or smoothed log-ratio of the altered probe.
   2.2. If neither is normal, i is given the call, log-ratio or smoothed log-ratio of the probe closest to i.
3. If a has a missing value, then we look for the first probe before a without a missing value and proceed as in 1. or 2. The same procedure is applied if b has a missing value except that we look for the first probe after b.

For all data sets, `log-ratios`, `smoothed log-ratios` and `calls` were downloaded from ACTuDB. The sex chromosomes were excluded from the analysis. All data sets presented missing values (between 3 and 13 % of the data), which were imputed with our procedure. We performed `PCA` on log-ratios and retained the principal components jointly accounting for at least 90% of data variability. `MR` were obtained with VAMP (La Rosa *et al.*, 2006), with support ranging from 5 to 50% of the tumors, using increments of 5%.

4 CLASS DISCOVERY PROCEDURES

Mathematical definitions for the items marked $*$ in this section may be found in the **Supplementary Material**. Once the input data representation has been chosen, the class discovery procedure requires the choice of a dissimilarity measure and a clustering algorithm.

4.1 Dissimilarity measures

The objects studied here are tumor samples. As it is not possible to devise a general formula for identifying the best dissimilarity measure for each individual situation, we consider some of the most frequently used methods (Legendre and Legendre, 1998).

We use the general notation *dissimilarity measure* to refer to a distance or a similarity or a dissimilarity. To convert a distance or dissimilarity measure into a similarity measure, or *vice versa*, the value is simply subtracted from the maximum value obtained.

For each input data representation strategy, we calculated different pairwise dissimilarity measures: `Euclidean`, `Manhattan` and `Pearson correlation`. We also calculated the dissimilarity measures proposed by Liu, and by van Wieringen: `sim`, `agree` and `conc`. All three were applied only to `calls` and `biological compression` strategies.

`sim` accounts for the number of contiguous genomic intervals of alterations of the same type overlapping in pairs of samples. In some circumstances, the similarity between one sample and itself may be smaller than that between two different samples (see **Supplementary Material** for an example). To prevent this situation, we made a minor correction: let $\mathbf{S}$ be the similarity `sim` matrix between pairs of samples with generic element s_{jl}, $j, l = 1, \ldots, p$, then assign $s_{jj} = max\, s_{jl}$.

The `agree` measure is defined as the probability of measurements for an arbitrary probe in two different samples being identical and `conc` is the probability of measurements of an arbitrary probe in two different samples being concordant (i.e. with the same order in terms of magnitude; see van Wieringen *et al.*, 2008 for details). These measures are based on the assumption that samples are independent and probes are distributed according to a mixture model.

4.2 Clustering algorithms

Many different clustering algorithms have been described (see Jain *et al.*, 1999 for a review).

Hierarchical algorithms are widely used because of their appealing tree representation. Hierarchical agglomerative or bottom-up clustering is a process beginning with the joining of the two most similar objects, with iterative merging of objects or groups of objects until all are included in a single set. By contrast, hierarchical divisive or top-down clustering algorithms begin with the whole set of objects, dividing this set successively in two until each group comprises only one object. Both agglomerative and divisive procedures may be carried out with several linkage methods. In this paper we applied the agglomerative linkages `complete`*, `average`*, `weighted`* and `Ward`* and the divisive linkages `diana`*, `tsvq`* and `hybrid`*.

We also applied partitioning clustering algorithms, which produce flat, non imbricated, clusters. The most common partitioning algorithm is the k`-means` algorithm, which was designed for use with Euclidean distance. We also used a variant, k`-centroids`,

For Peer Review

which adapts the k-means algorithm to other dissimilarity measures. Cluster centroids are defined such that the average dissimilarity of the object of a cluster to all the objects in the cluster is minimal. Finally, we included the algorithm proposed by van Wieringen *et al.* (2008) and called total, which is associated only with agree and conc.

4.3 Evaluation stage

4.3.1 Stability It is difficult to evaluate class discovery solutions, particularly as no class labels are known and so no error rate can be estimated. However, a panoply of criteria for the validation of class discovery solutions has been proposed (Handl *et al.*, 2005). External indices assess class discovery solutions according to object labeling, which may be provided by an expert, whereas internal indices evaluate a particular notion of class discovery quality, such as the separation of clusters.

We validated class discovery solutions in terms of their stability. Stability is an internal index because it assesses the preservation of class discovery solutions across perturbations of the original data. We compared solutions emerging from two perturbations of the original data, using the Jaccard coefficient*. This coefficient requires a partition to be calculated. In the case of hierarchical algorithms, the Jaccard coefficient is calculated for each partition considered.

Several ways of perturbing the data have been proposed. We decided to resample the data by repeatedly drawing overlapping subsets of samples from the same dataset without replacement (Levine and Domany, 2001 and Ben-Hur *et al.*, 2002).

4.3.2 Assessing the significance of solutions We assessed the statistical significance of the stability of the structure discovered by the class discovery method, using the χ^2-based test proposed by Bertoni and Valentini (2007). This test was initially designed to determine the number of clusters in a stability framework, but can easily be transposed for class discovery method selection in the same framework, as described below.

A perturbation procedure was applied $2M$ times to the data set $\mathbf{X}$, building M pairs of subsets of $\mathbf{X}$. Let $\mathcal{C}$ be a set of R class discovery methods $\mathcal{C} = \{C_1, \ldots, C_r, \ldots, C_R\}$. R methods are then applied to the M pairs of subsets and the number k of clusters for each solution is fixed. The similarity of each pair of solutions is then calculated s_{rm}, $r = 1, \ldots, R$; $m = 1, \ldots, M$. The (s_{rm}) values are the realizations of the random variable S_r.

Bertoni and Valentini (2007) concluded that $E[S_r]$ can be used as an index of the reliability of class discovery solutions: if $E[S_r] \simeq 1$ the solution is stable. The stability of the solution is considered to decrease with increasing distance of $E[S_r]$ from 1.
This result was demonstrated by Bertoni and Valentini (2007) in the model selection framework, but it also applies *mutatis mutandis* to this context. As we tested a number of methods, we incorporated a multiple testing correction step into the stability analysis.

$E[S_r]$ may be estimated by its empirical mean ξ_r, defined as $\xi_r = 1/M \sum_{m=1}^{M} s_{rm}$. ξ_r is then sorted in descending order, $\mathcal{G} = (\xi_{p(1)}, \xi_{p(2)}, \ldots, \xi_{p(R)})$ where p is a permutation index such that $(\ldots\ \xi_{p(r_1)} \geq \xi_{p(r_2)}\ \ldots)$. Class discovery solutions are then ordered from the most to the least stable.

Let us consider the Bernoulli random variable $B_r = I(S_r > s^0)$ where $s^0 \in [0, 1]$ is a fixed threshold, $\theta_r = P(S_r > s^0)$ and I is the indicator function. The sum $X_r = \sum_{m=1}^{M} B_{rm}$ of M independent and identically distributed (i.i.d.) B_r follows a binomial distribution $\mathcal{B}(M, \theta_r)$. For a sufficiently large M,

$$Z_r = \frac{X_r - M\theta_r}{\sqrt{M\theta_r(1-\theta_r)}} \sim \mathcal{N}(0, 1).$$

Assuming Z_r i.i.d., $r = 1, \ldots, R$ and θ estimated by its pooled estimate

$$\hat{\theta} = \frac{\sum_{r=1}^{R} X_r}{RM}.$$

Then,

$$Y = \sum_{r=1}^{R} \frac{(X_r - M\hat{\theta})^2}{M\hat{\theta}(1-\hat{\theta})} \sim \chi^2(R-1).$$

The null hypothesis "H_0: all the θ_r are equal to θ" is tested against the alternative hypothesis "H_1: not all θ_k are equal", with Y used as the test statistic. If the null hypothesis is rejected, we exclude the least stable method, according to the sorting of $\mathcal{G}$, and repeat the test. *P*-values were adjusted for multiple testing by Bonferroni-Holm correction (Holm, 1979).

This χ^2-based test is repeated until no significant difference is detected or until only one class discovery method is left. The set of methods remaining represents the set of stable methods discovered.

For all data sets, resampling was performed by establishing $M = 100$ pairs of subsets of each data set. For each subsample, we randomly picked 80 % of the data set. A dissimilarity measure and a clustering algorithm were applied to each subsample. We considered partitions in $k = 2$ to 10 clusters. For each partition, the Jaccard index was used to compare pairs of solutions from pairs of subsets.

Finally, the χ^2-based test was applied iteratively for the detection of stable class discovery solutions for a Bonferroni-Holm-corrected significance level of 5%. The threshold s^0 was set at 0.98.

All methods are implemented within the R programming language (http://www.r-project.org). We used cluster and hybridHclust R packages available from http://www.r-project.org, clusterv and mosclust R packages available from http://homes.dsi.unimi.it/valenti/software.html and WECCA available from http://www.few.vu.nl/~wvanwie/software/WECCA/WECCA.html.

5 RESULTS

We intensively compared the stability performances of class discovery methods (combinations of an input data representation, a dissimilarity measure and a clustering algorithm). We considered five strategies for input data representation: all versions of All probes (log-ratio, smoothed log-ratio and calls), statistical and biological compressions. We considered six dissimilarity measures: Euclidean and Manhattan distances, Pearson correlation and sim, conc and agree similarities. We applied ten clustering algorithms: complete, average, weighted and ward linkages, diana, tsvq, hybrid, *k*-means and *k*-centroids and total. The χ^2-based test described was applied iteratively to detect stable class discovery methods.

For Peer Review

For each data set and each partition, the extensive list of class discovery methods declared stable by the above-described χ^2-based test is provided in Table 1 of **Supplementary Material**.

Figure 1 indicates, for each data set, the frequency of each input data representation, each dissimilarity measure and each clustering algorithm in the list of class discovery methods declared stable, all partitions taken together. For all data sets, `MR` clearly outperformed the other input data representations, and the hierarchical agglomerative linkage `average` outperformed the other clustering algorithms. The situation is less clear for dissimilarity measures: `sim` in three cases, `agree` in one case and `Euclidean`, `Manhattan` and `Pearson correlation` equally outperformed the other dissimilarity measures in one case.

We also calculated the frequency of each input data representation, each dissimilarity measure and each clustering algorithm in the class discovery methods declared stable for each partition from 2 to 10, all data sets taken together (see Figure 2). `MR` and hierarchical `average` were again identified as the input data representation and clustering algorithm most frequently leading to stable solutions. For dissimilarity measures, `Pearson correlation` performed well in the case of two clusters and `agree` performed well with six clusters. For 3, 4, 5, 7, 8, 9 and 10 clusters, `sim` outperformed the other dissimilarity measures.

The most stable input data representation, dissimilarity measure and clustering algorithm depended little on the data set or number of clusters considered. Figure 3 shows the frequency of class discovery methods declared stable over all possible data sets and partitions. The most stable combinations were (`MR`, `agree`, `average`) and (`MR`, `sim`, `average`). By contrast, the `hybrid` and `total` algorithms gave no stable solutions.

6 DISCUSSION AND CONCLUSION

We investigated the application of several input data representations, dissimilarity measures and clustering algorithms for array-CGH data. We compared the resulting class discovery methods in terms of the stability of their solutions.

We conclude that the characterization of array-CGH data by `MR` (Rouveirol *et al.*, 2006) is a good choice for class discovery purposes, as our experiments demonstrate that stable partitions are generally achieved with this method. As these solutions are reached by reducing the number of data dimensions, the data are characterized in a parsimonious manner.

The use of `MR` presents other advantages in addition to its parsimony. Firstly, it allows the same weight to be assigned to each alteration, regardless of its size. Indeed, potentially very small alterations, such as amplifications, may be relevant as predictive or prognostic factors. As few probes are found in such small alterations, it may be better to use the alteration as a single entity so that all regions are weighted equally. Secondly, this method facilitates data interpretation because it allows biologists to study a limited number of alterations rather than having to study all the probes to account for differences. Finally, data representation based on `MR` reduces the amount of data required for class discovery. This feature is particularly useful for high-density array-CGH technologies.

We recommend the use of hierarchical agglomerative `average` linkage associated with `sim` or `agree` similarity measures for a stable class discovery framework.

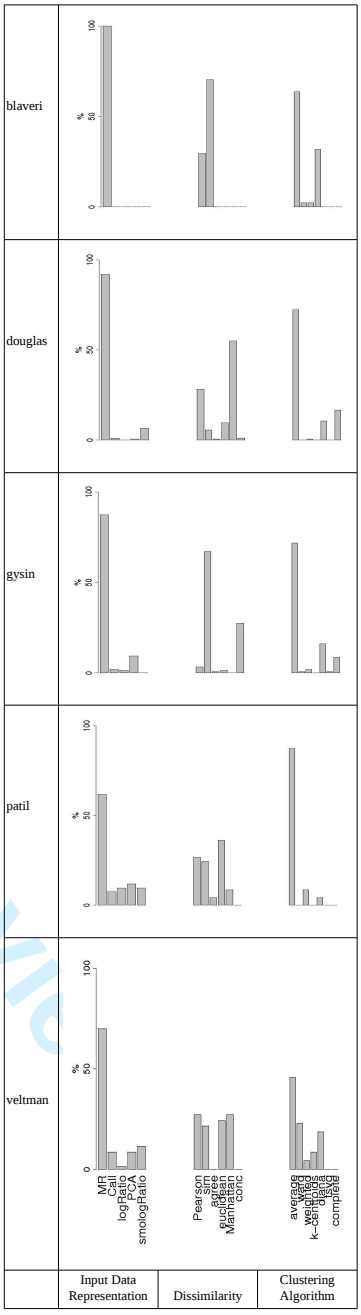


Fig. 1. Frequency of input data representation, dissimilarity measure and clustering algorithm among the class discovery methods declared stable for each data set.

For Peer Review

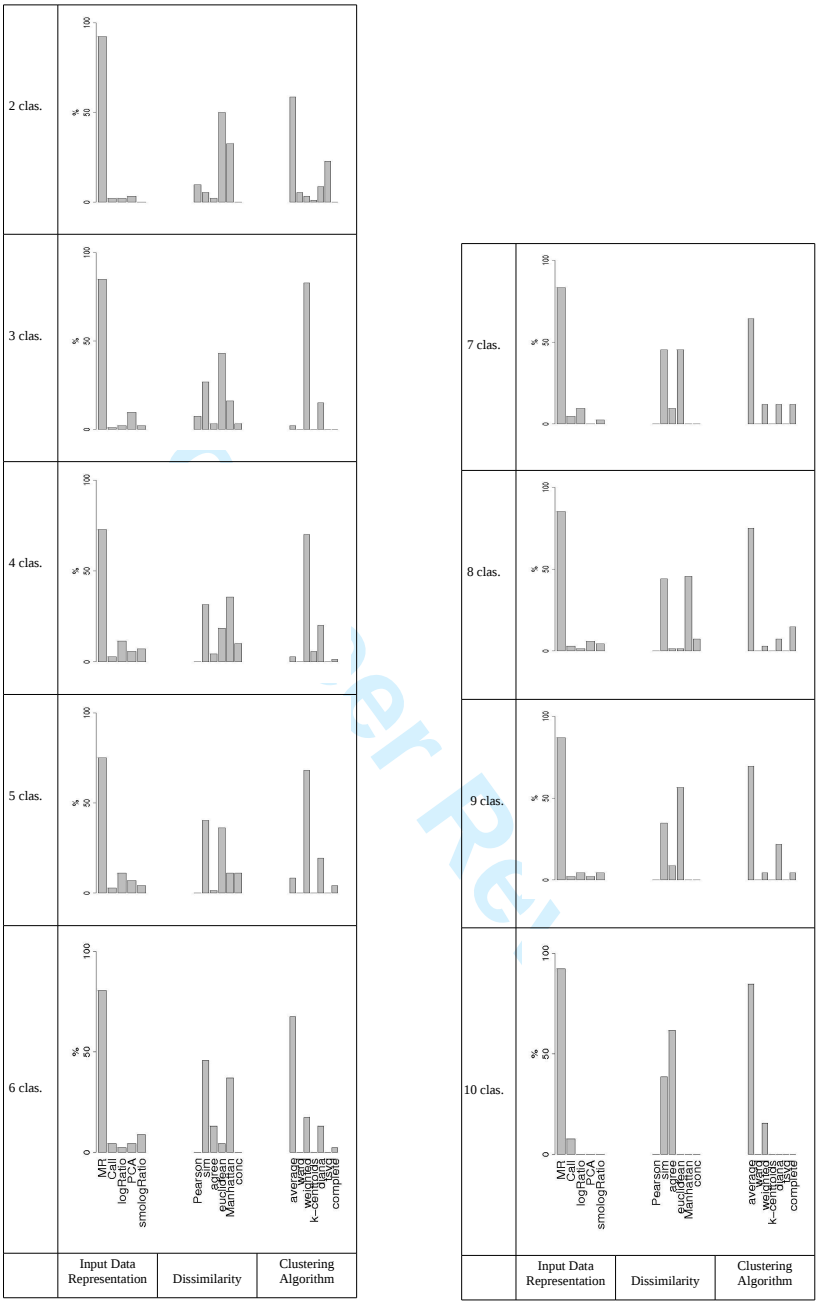


Fig. 2. Frequency of input data representation, dissimilarity measure and clustering algorithm among the class discovery methods declared stable for each partition from 2 to 10 clusters.

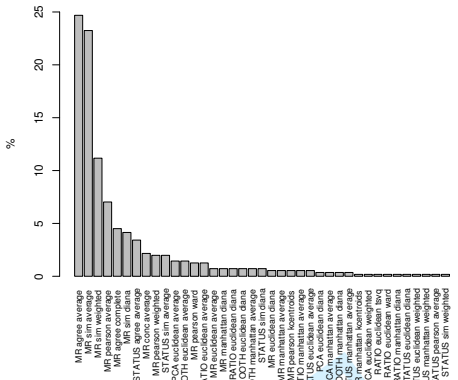


Fig. 3. Frequency of class discovery methods declared stable

REFERENCES

Albertson, D., Collins, C., McCormick, F., and Gray, J. (2003). Chromosome aberrations in solid tumors. *Nat. Genet.*, **34**, 369–376.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pac Symp Biocomput.*, **7**, 6–17.

Bertoni, A. and Valentini, G. (2007). Model order selection for bio-molecular data clustering. *BMC Bioinformatics*, **8**, S4.

Blaveri, E., Brewer, J., Roydasgupta, R., Fridlyand, J., DeVries, S., Koppie, T., Pejavar, S., Mehta, K., Carroll, P., Simko, J., and Waldman, F. (2005). Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clin Cancer Res*, **11**, 7012–7022.

Brock, G., Shaffer, J., Blakesley, R., Lotz, M., and Tseng, G. (2008). Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics*, **10**, 9–12.

Carvalho, B., Ouwerkerk, E., Meijer, G., and Ylstra, B. (2004). High resolution microarray comparative genomic hybridization analysis using spotted oligonucleotides. *J. Clin. Pathol.*, **57**, 644–646.

Douglas, E., Fiegler, H., Rowan, A., Halford, S., Bicknell, D., Bodmer, W., Tomlinson, I., and Carter, N. (2004). Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res*, **64**, 4817–4825.

Gysin, S., Rickert, P., Kastury, K., and McMahon, M. (2005). Analysis of genomic DNA alterations and mRNA expression patterns in a panel of human pancreatic cancer cell lines. *Genes Chromosomes Cancer*, **44**, 37–51.

Handl, J., Knowles, J., and Kell, D. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hupé, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3419–3422.

Hupé, P., La Rosa, P., Liva, S., Lair, S., Servant, N., and Barillot, E. (2007). ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. *Oncogene*, **26**, 6641–6652.

Ishkanian, A., Malloff, C., Watson, S., DeLeeuw, R., Chi, B., Coe, B., Snijders, A., Albertson, D., Pinkel, D., Marra, M., Ling, V., MacAulay, C., and Lam, W. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.*, **36**, 299–303.

Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, **31**, 264–323.

La Rosa, P., Viara, E., Hupé, P., Pierron, G., Liva, S., Neuvial, P., Brito, I., Lair, S., Servant, N., Robine, N., Manié, E., Brennetot, C., Janoueix-Lerosey, I., Raynal, V., Gruel, N., Rouveirol, C., Stransky, N., Stern, M., Delattre, O., Aurias, A., Radvanyi, F., and Barillot, E. (2006). VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics*, **22**, 2066–2073.

Lai, C., Horlings, H., van de Vijver, M., van Beers, E., Nederlof, P., Wessels, L., and Reinders, M. (2007). SIRAC: supervised identification of regions of aberration in aCGH datasets. *BMC Bioinformatics*, **8**, 422.

Lai, W., Johnson, M., Kucherlapati, R., and Park, P. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Legendre, P. and Legendre, L. (1998). *Numerical Ecology*. Elsevier, Amesterdam.

Levine, E. and Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Comput*, **13**, 2573–2593.

Liu, J., Mohammed, J., Carter, J., Ranka, S., Kahveci, T., and Baudis, M. (2006). Distance-based clustering of CGH data. *Bioinformatics*, **22**, 1971–1978.

Liu, J., Ranka, S., and Kahveci, T. (2007). Markers improve clustering of CGH data. *Bioinformatics*, **23**, 450–457.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J., Rostan, S., Nguyen, K., Powers, S., Ye, K., Olshen, A., Venkatraman, E., Norton, L., and Wigler, M. (2003). Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.*, **13**, 2291–2305.

Mattfeldt, T., Wolter, H., Trijic, D., Gottfried, H., and Kestler, H. (2001). Chromosomal regions in prostatic carcinomas studied by comparative genomic hybridization, hierarchical cluster analysis and self-organizing feature maps. *Anal. Cell. Pathol.*, **23**, 29–37.

Mitelman, F. (2005). Cancer cytogenetics update 2005. *Atlas Genet. Cytogenet. Oncol. Haematol.*, **9**, 342–346.

Patil, M., Gutgemann, I., Zhang, J., Ho, C., Cheung, S.-T., Ginzinger, D., Li, R., Dykema, K., So, S., Fan, S.-T., Kakar, S., Furge, K., Buttner, R., and Chen, X. (2005). Array-based comparative genomic hybridization reveals recurrent chromosomal aberrations and Jab1 as a potential target for 8q gain in hepatocellular carcinoma. *Carcinogenesis*, **26**, 2050–2057.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W., Chen, C., Zhai, Y., Dairkee, S., Ljung, B., Gray, J., and Albertson, D. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Pollack, J., Perou, C., Alizadeh, A., Eisen, M., Pergamenschikov, A., Williams, C., Jeffrey, S., Botstein, D., and Brown, P. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.

Quackenbush, J. (2006). Computational approaches to analysis of DNA microarray data. *Methods Inf. Med.*, **45**, 91–103.

Rouveirol, C., Stransky, N., Hupé, P., La Rosa, P., Viara, E., Barillot, E., and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.

Snijders, A., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J., Gray, J., Jain, A., Pinkel, D., and Albertson, D. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, **29**, 263–264.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T., and Lichter, P. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C., Lønning, P., Brown, P., Børresen-Dale, A., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS*, **100**, 8418–8423.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.

van de Wiel, M. and van Wieringen, W. (2007). CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Informatics*, **2**, 55–63.

van Wieringen, W., van de Wiel, M., and Ylstra, B. (2008). Weighted clustering of called array CGH data. *Biostatistics*, **23**, 484–500.

Veltman, J., Fridlyand, J., Pejavar, S., Olshen, A., Korkola, J., DeVries, S., Carroll, P., Kuo, W., Pinkel, D., Albertson, D., Cordon-Cardo, C., Jain, A., and Waldman, F. (2003). Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res*, **63**, 2872–2880.

Vogel, C., Cobleigh, M., and et al., D. T. (2002). Efficacy and safety of trastuzumab as a single agent in first-line treatment of her2-overexpressing metastatic breast cancer. *J Clin Oncol*, **20**, 719–726.

For Peer Review

ARTICLE

High-Resolution Mapping of DNA Breakpoints to Define True Recurrences Among Ipsilateral Breast Cancers

Marc A. Bollet, Nicolas Servant, Pierre Neuvial, Charles Decraene, Ingrid Lebigot, Jean-Philippe Meyniel, Yann De Rycke, Alexia Savignoni, Guillem Rigaill, Philippe Hupé, Alain Fourquet, Brigitte Sigal-Zafrani, Emmanuel Barillot, Jean-Paul Thiery

Background To distinguish new primary breast cancers from true recurrences, pangenomic analyses of DNA copy number alterations (CNAs) using single-nucleotide polymorphism arrays have proven useful.

Methods The pangenomic profiles of 22 pairs of primary breast carcinoma (ductal or lobular) and ipsilateral breast cancers from the same patients were analyzed. Hierarchical clustering was performed using CNAs and DNA breakpoint information. A partial identity score developed using DNA breakpoint information was used to quantify partial identities between two tumors. The nature of ipsilateral breast cancers (true recurrence vs new primary tumor) as defined using the clustering methods and the partial identity score was compared with that based on clinical characteristics. Metastasis-free survival was compared among patients with primary tumors and true recurrences as defined using the partial identity score and by clinical characteristics. All statistical tests were two-sided.

Results All methods agreed on the nature of ipsilateral breast cancers for 14 pairs of samples. For five pairs, the clinical definition disagreed with both clustering methods. For three pairs, the two clustering methods were discordant and the one using DNA breakpoints agreed with the clinical definition. The partial identity score confirmed the nature of ipsilateral breast cancers as defined by clustering of DNA breakpoints in 21 of 22 pairs. The difference in metastasis-free survival of patients with new primary tumors and those with true recurrences was not statistically significant when tumors were defined based on clinical and histologic characteristics (5-year metastasis-free survival: 76%, 95% confidence interval [CI] = 52% to 100% for new primary tumors and 38%, 95% CI = 17% to 83% for true recurrences; $P = .18$; new primary tumor vs true recurrence, hazard ratio = 2.8, 95% CI = 0.6 to 13.7), but the difference was statistically significant when tumors were defined using the partial identity score (5-year metastasis-free survival: 100% for new primary tumors and 29%, 95% CI = 11% to 78% for true recurrences; $P = .01$).

Conclusions DNA breakpoint information more often agreed with the clinical determination than CNAs in this population. The partial identity score, which was calculated based on DNA breakpoints, allows statistical discrimination between new primary tumors and true recurrences that could outperform the clinical determination in terms of prognosis.

J Natl Cancer Inst 2008;100:48–58

Breast-conserving therapy is the preferred treatment for patients with early-stage breast cancer (1). It offers equal local control and overall survival (2) and superior psychosocial outcomes (3,4) compared with modified radical mastectomy. However, an ipsilateral breast cancer recurrence can be traumatizing and can lead to death (2).

When an ipsilateral breast cancer develops, the new tumor can either be a true recurrence—that is, a regrowth of clonogenic cells that were not removed by surgery or killed by radiotherapy—or a new primary tumor that arises from the remaining breast tissue (5). Several definitions have been used to distinguish true recurrences from new primary tumors. Initially, these distinctions were based

Affiliations of authors: Département d'oncologie radiothérapie (MAB, AF), Service de Bio-informatique (NS, PN, GR, PH, EB), Département de Transfert (CD, JPM, JPT), Département de Biologie des tumeurs (IL, BSZ), Service de Biostatistiques (YDR, AS), and Centre National de la Recherche Scientifique, Unité Mixtes de Recherche 144 (CD, PH), Institut Curie, Paris, France; Institute of Molecular and Cell Biology Biopolis A*STAR, Singapore (JPT).

Correspondence to: Marc A. Bollet, MD, Département d'oncologie radiothérapie, Institut Curie, 26, rue d'Ulm, 75248 Paris cedex 05, France (e-mail: marc.bollet@curie.net).

See "Funding" and "Notes" following "References."

DOI: 10.1093/jnci/djm266

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oxfordjournals.org.

on the location of the ipsilateral breast cancer (ie, the farther from the initial primary tumor, the more likely it is to be a new primary tumor) and on shared common histopathologic criteria (eg, type, grade, and hormone receptor status) (6–10). In the quest for additional ways to distinguish new primary breast tumors from true breast cancer recurrences, biologic studies of clonal relationships between the new and original tumor have also been performed. These studies have relied on ploidy (5,11), loss of heterozygosity (12–14), p53 analysis (15), or X chromosome inactivation (16) or have been based on DNA copy number alterations (CNAs) (17–19). CNA data can be obtained by high-resolution techniques, such as array-based comparative genomic hybridization or single-nucleotide polymorphism (SNP) arrays (20). One of the most commonly used ways to look at clonal relatedness using pangenomic data is to perform an unsupervised hierarchical clustering that organizes primary breast tumors and ipsilateral breast cancers on the basis of their overall genomic similarity (18,19). These measures of similarity are summarized in a dendrogram, in which the pattern and length of the branches reflect the relatedness of the samples in terms of DNA CNAs.

Changes in DNA copy numbers occur at chromosomal locations called breakpoints. We hypothesized that the precise locations of these breakpoints could serve as markers for clonal relatedness and that we could distinguish true recurrences from new primary tumors by the number of common breakpoints in the ipsilateral breast cancer and the primary tumor. In this study, we first aimed to test the added value of examining the clustering of breakpoints (over CNAs) when determining the nature of the ipsilateral breast cancer. Second, we aimed to develop a score to quantify the partial identity between two tumors according to their clonal relatedness (determination of the partial identity score). Third, we examined prognosis in terms of metastasis-free survival. In each case, these methods were compared with the clinical determination of the nature of the ipsilateral breast cancer.

CONTEXT AND CAVEATS

Prior knowledge

Detecting changes in DNA copy number using single nucleotide polymorphism arrays has been a useful tool in distinguishing new primary breast tumors from recurrences.

Study design

Comparison of hierarchical clustering of DNA copy number and DNA breakpoints, an identity score based on the DNA breakpoint information, and clinical characteristics to accurately designate ipsilateral breast tumors as new primary tumors or true recurrences in breast tumor pairs from 22 patients.

Contributions

For 14 of the pairs, all methods agreed on the designation of the ipsilateral breast cancer as a new primary tumor or a true recurrence; however, for five pairs and three pairs, both clustering methods and clustering by DNA breakpoints, respectively, agreed with the clinical definition. For 21 pairs, the partial identity score confirmed the designation of the tumor as defined by both clustering methods. Patients with recurrences had poorer metastasis-free survival than patients with new primary tumors, according to the partial identity score, but this difference was not statistically significant using the clinical definition.

Implications

The partial identity score may outperform clinical determination for the prognosis of ipsilateral breast cancers.

Limitations

Freshly frozen tissue samples that contain a large number of cells from both the initial primary tumor and the ipsilateral tumor are needed to perform the DNA breakpoint analyses.

Subjects and Methods

Selection of Patients

Specimens from patients with primary breast cancers and ipsilateral breast cancers were selected from freshly frozen samples of the Institut Curie tissue bank according to the following criteria: the primary tumor was either ductal or lobular invasive breast carcinoma; the patient was 49 years or younger at diagnosis of the initial tumor; all patients were premenopausal; and there was no previous history of cancer, except for one nonmelanoma skin cancer. All patients had been treated at the Institut Curie by breast-conserving surgery, including dissection of the axillary lymph nodes in most patients, followed by radiotherapy to the breast with or without a boost to the tumor bed (external beam radiotherapy or brachytherapy) and/or to the regional lymph node–bearing areas if indicated and, when required, systemic treatment as part of their initial management. For all tumors, histopathologic characteristics were reviewed by one pathologist (B. Sigal-Zafrani).

To ensure that the data would be informative, we restricted genomic analyses to tumors (primary and recurrences) in which at least 50% of cancer cells had been assessed by hematoxylin, eosin, and saffron staining of sections from snap-frozen samples. This study reports a series of 22 patients with assessable pairs of primary breast tumors and ipsilateral breast cancers.

To evaluate the genomic features of a population with similar breast cancers, 44 control patients from the pool of patients with primary tumors who met the above selection criteria were matched to the case patients in accordance with their age at diagnosis and adjuvant treatment. The control patients had not experienced an ipsilateral breast recurrence within the time span of the local recurrence of the index patient.

This research was approved by the institutional review boards of the Institut Curie. No patient refused the use of her tumor specimens for research purposes.

Clinical and Histologic Studies

The histologic/biologic properties of the breast cancers were determined by subjecting tissue sections to immunohistochemical analysis for the estrogen receptor (clone 6F11, 1:200 dilution; Novocastra, Newcastle Upon Tyne, England) and progesterone receptor (clone 1A6, 1:200 dilution; Novocastra) antibodies. Tumors were considered to be positive for these receptors if at least 10% of the invasive tumor cells in a section showed nuclear staining (21).

In accordance with theories of the clonal evolution of tumor cell populations, ipsilateral breast cancers were clinically defined as true recurrences if they had the same histologic subtype (ductal or

lobular) and a similar or increased growth rate, similar or loss of dependence on either estradiol or progesterone, and similar or decreased differentiation as the initial tumor (22).True recurrences also had to share with their primary tumors the same breast quadrant. Thus, new primary tumors were clinically defined as such when the ipsilateral breast cancer had occurred in a different location, had a distinct histologic type, or had less aggressiveness features (lower grade, appearance of hormonal receptors) than the initial tumor.

Genomic Studies

Total genomic DNA was extracted from tissue samples using a variation of the standard phenol : chloroform protocol (23). Genomic DNA was quantified by spectrophotometry using a ND-1000 Spectrophotometer (NanoDrop, Wilmington, DE), and quality was assessed by 0.8% agarose gel electrophoresis.

Genomic DNA from each sample was prepared for microarray hybridization using the GeneChips Mapping 50K Xba Assay Kit (Affymetrix Inc., Santa Clara, CA). Briefly, 250 ng of total genomic DNA was digested with the restriction enzyme XbaI and ligated to an adaptor sequence (XbaI adaptator: 5′-ATTATGAGCACGAC AGACGCCTGATCT-3′ and 5′-CTAGAGATCAGGCGTCTG TCGTGCTCATAA-3′) that recognizes the cohesive four base pair (bp) region (3′-GATC-5′). A generic primer (5′-ATT ATG AGC ACG ACA GAC GCC TGA TCT-3′) that recognizes the adaptor sequence was used to preferentially amplify adaptor-ligated DNA fragments 250–2000 bp in size by the optimized polymerase chain reaction (PCR) conditions, according to the manufacturer's instructions. The amplified DNA was then fragmented by DNase treatment and hybridized to the Affymetrix GeneChips Human Mapping 50K array Xba 240 (Affymetrix), according to the manufacturer's instructions. Washing, staining, and scanning of chips were performed using materials and methods provided by the manufacturer. The pangenomic profiles of the 22 pairs of primary tumors/ipsilateral breast cancers are available on ACTuDB (24) (http://bioinfo.curie.fr/actudb/). Human mapping 50K array Xba 240 annotations and sequence files are available on the Affymetrix website (http://www.affymetrix.com/support/technical/byproduct.affx?product=100k).

Metastasis-Free Survival

Metastasis-free survival was estimated by the Kaplan–Meier method (25) and compared between the groups of patients defined as having been diagnosed with either a true recurrence or a new primary tumor using the log-rank test. The confidence interval (CI) of the hazard ratio was obtained using a semiparametric Cox model (26).

Statistical Methods

Copy Number Alteration Determination. SNP data were gathered from the pangenomic profile and analyzed using the iterative and alternative normalization of copy number SNP array (ITALICS) algorithm with default parameters, which simultaneously normalizes the genomic profile and detects the biologic signal. Briefly, ITALICS alternatively estimates the biologic signal (ie, the DNA copy number at each SNP locus) with the gain and loss analysis of DNA algorithm (27) and normalizes the data to correct the nonrelevant effects (CG content and fragment length of PCR products, oligonucleotide CG content, and SNP effect). These two steps are repeated iteratively to improve the biologic signal estimation until no more improvement is seen. ITALICS outperforms other methods of normalization. The result of this process is a segmented genomic profile that consists of regions of homogeneous DNA and information on their corresponding copy numbers. Each region is given a smoothing value (ie, the median of the SNP copy numbers within the region) and a status (ie, gain, normal, or loss).

We defined a breakpoint as 1) a SNP locus located at a change of status (eg, normal/gain or gain/loss) or as 2) a SNP locus located at a change of smoothing value that occurred within a region of gain or loss, thus defining different levels of gain or loss among these regions. Additional breakpoints were also added at the extremities of the chromosome to take into account their gain or loss whenever applicable. Because some breakpoints could be due to copy number variations that occur in healthy individuals, breakpoints arising in the copy number variable regions in the HapMap collection (28) were excluded. The visualization and further analysis of the data was performed through a graphic user interface, Visualization and analysis of array CGH, transcriptome and other molecular profiles (29).

Hierarchical Clustering. *Similarity between genomic profiles.* We considered two measures of similarity among the genomic profiles of a primary tumor and ipsilateral breast cancer. First, we used the Pearson correlation between their CNA profiles. Second, we used a measure M that is derived from the percent concordance proposed by Waldman et al. (18) and adapted from Dice's formula (30) and corresponds to the number of common breakpoints divided by the mean number of breakpoints in either a primary tumor or an ipsilateral breast cancer. M is computed as follows, for a (i,j) pair.

$$M_{i,j} = \frac{\#(S_i \cap S_j)}{1/2 \times (\#S_i + \#S_j)},$$

in which S_i and S_j are the subsets of breakpoints present in the SNP arrays of the primary tumor, i, and ispilateral breast cancer, j. An example of M is given in Supplementary Fig. 1 (available online).

Two tumors had common breakpoints if the following conditions were fulfilled: 1) the changes in copy number occurred at the exact same locus and 2) the changes in copy number were of the same nature (ie, either an increase or a decrease in numbers) in the two tumors.

Assessing clonal relatedness from a dendrogram. We assumed that clonal unrelatedness was revealed by the clustering apart of the two tumors (primary tumor and ipsilateral breast tumor) from the same patient, reflecting that they were more similar to carcinomas of other patients than to each other. In contrast, the clustering together of two tumors from the same patient indicated clonal relatedness among them. For both measures of similarity (Pearson coefficient and M measure), we used Ward's criteria (31) as an agglomerative method in the hierarchical clustering.

Partial Identity Score. *Score definition.* To distinguish true recurrences from new primary tumors, we developed a partial identity score

Table 1. Patient and tumor characteristics of the 22 patients whose tumors (both PT and IBC) had exploitable SNP arrays*

Pair	Age, y	Family	Prob	BRCA1	BRCA2	pT	pN	Surgical margin, mm	Radiotherapy dose, Gy		No. of cycles of chemotherapy†
									Whole breast	Tumorectomy bed	
P1	23.1	0	20	0	2	1	0	≥4	54	54	4
P2	42.1	1	NA	NA	NA	1	0	≥4	50	50	0
P3	42.6	0	NA	NA	NA	1	0	≥4	54	54	0
P4	48.2	1	44	0	0	1	0	≥4	50	50	0
P5	45.5	0	NA	NA	NA	1	1	≥4	50	60	4
P6	35.7	0	8	0	0	2	0	≥4	51	66	4
P10	46.2	0	NA	NA	NA	2	0	0–3	50	70	0
P11	49.0	1	95	0	1	2	0	≥4	50	64	0
P12	48.9	1	NA	NA	NA	1	0	≥4	52	52	0
P13	45.0	0	NA	NA	NA	2	0	≥4	51	67	6
P14	43.6	0	NA	NA	NA	1	0	0–3	50	50	0
P15	46.1	0	NA	NA	NA	1	0	≥4	50	65	0
P16	48.4	0	NA	NA	NA	1	0	≥4	50	66	0
P18	27.9	1	82	0	0	2	0	0–3	50	70	4
P19	49.1	0	NA	NA	NA	2	0	0–3	51	65	4
P20	47.1	0	NA	NA	NA	2	1	0–3	45	65	4
P21	46.3	0	NA	NA	NA	1	0	DCIS	50	70	0‡
P22	35.0	0	NA	NA	NA	2	2	≥4	50	75	6‡
P23	30.8	0	NA	NA	NA	2	0	≥4	50	66	4
P24	47.7	0	NA	NA	NA	1	1	≥4	50	60	6
P25	43.0	0	NA	NA	NA	1	0	0–3	45	60	0‡
P26	30.5	0	NA	NA	NA	NA	1	≥4	52	70	4‡

* PT = primary tumor; IBC = ipsilateral breast cancer; SNP = single nucleotide polymorphism; Family = family history of breast cancer in the first two degrees (0 = no, 1 = yes); Prob = age-specific risk estimates of breast cancer according to the Claus Model (32); BRCA1 and BRCA2 = mutation found in BRCA1 and BRCA2 (0 = not found, 1 = deleterious, 2 = possibly deleterious, NA = not available); pT = histologic tumor classification according to Union Internationale Contre le Cancer (UICC) (33); pN = histologic lymph node classification according to UICC; DCIS = ductal carcinoma in situ.

† Chemotherapy consisted of 5-fluorouracil, anthracyclines, and cyclophosphamide.

‡ Patients were treated with tamoxifen for 5 years.

that is based on the *M* measure of similarity described above. The score reflects the number of common breakpoints among the ipsilateral breast cancer and the primary tumor. In addition, because very frequent breakpoints may be less informative than frequent ones in estimating the clonal relatedness between two tumors, the added value of each breakpoint was weighted according to its frequency among the samples of 44 control patients. The partial identity score (PS) was thus

$$\mathrm{PS}_{i,j} = \frac{\sum_{k\in(S_i\cap S_j)} (1-F_k)^2}{1/2\times[\sum_{k\in S_i}(1-F_k)+\sum_{k\in S_j}(1-F_k)]},$$

in which F_k represents the frequency of appearance of the breakpoint k calculated in the series of the 44 control breast cancers. An example of a partial identity score is given in Supplementary Fig. 1 (available online).

Statistical testing for partial identity. The partial identity score was calculated for all 462 possible "artificial pairs" (462 = 22 × 21, because each of the 22 primary tumors could be artificially paired with the ipsilateral breast cancer of the 21 other patients, *see* Table 3 notes). The distribution under the null hypothesis, H0, of no partial identity between the two tumors was estimated using all 462 possible artificial pairs. We rejected H0 with a type I error fixed at 5%, that is, we considered that a local recurrence shared partial identity with a primary tumor when the score was higher than the upper 5th percentile in the distribution of artificial pairs. The score was then calculated for the "natural pairs," that is, a primary tumor and its ipsilateral breast cancer occurring in the same patients (*see* Table 3 notes). Ipsilateral breast cancers from pairs with scores higher than this cutoff, that is, with shared partial identity, were considered to be true recurrences.

Robustness of the score. The robustness of the partial identity score was assessed by randomly selecting two subgroups of 15 and 7 patients from the population of 22 breast cancer patients. The first subgroup of 15 patients was used to compute the scores of the artificial pairs and to record the cutoff score corresponding to the 95th percentile. This score was then used to determine the status of each of the natural pairs in the seven patients of the other subgroup. To make the comparison statistically sound, each process was repeated 1000 times. The variation of the cutoff scores was assessed by box plot representation. The consistency of the ipsilateral breast cancer status was calculated as the percentage of extractions when the status of this pair was respectively a true recurrence or a new primary tumor.

All statistical tests were two-sided. *P* values less than .05 were considered to be statistically significant.

Results

Clinical and Histologic Features

The clinical and tumor characteristics of 22 patients whose tumors had exploitable SNP arrays were analyzed (Tables 1 and 2). According to clinical and histologic criteria (Table 2), nine of the 22 ipsilateral breast cancers were new primary tumors and the other

Table 2. Histologic characteristics of the primary tumors and their ipsilateral breast cancers: distinctions between new primary tumors and true recurrences according to clinical criteria or clustering methods*

	Primary tumors					Ipsilateral breast cancers					New primary tumors or true recurrences				
Pair	Type	Grade	ER	PR	Time, y	Location	Type	Grade	ER	PR	CNA	BKP	Clinical	Divergence	Score
P1	D	3	0	40	6.5	1	D	2	90	15	TR	NP‡	NP	CNA	0.020
P2	D	2	90	40	5.3	1	L	1	90	70	TR	NP‡	NP	CNA	0.000
P3	D	3	30	80	3.1	1	D	3	60	90	TR	TR‡	TR	No	0.465
P4	L	1	90	80	3.5	1	L	2	90	80	TR	TR‡	TR	No	0.278
P5	D	2	90	40	2.0	1	D	2	80	90	TR	TR‡	TR	No	0.555
P6	L	1	90	100	3.1	1	L	2	70	70	NP	NP‡	TR	Clinical	0.104
P10	L	3	80	95	5.0	0	D	2	70	40	NP	NP‡	NP	No	0.059
P11	L	3	0	0	6.3	1	D	3	0	0	NP	NP‡	NP	No	0.029
P12	L	2	90	50	2.9	0	L	2	90	0	TR	TR†	NP	Clinical	0.116
P13	D	2	20	85	4.6	1	D	2	95	20	TR	TR‡	TR	No	0.240
P14	L	2	90	60	2.5	1	L	2	0	100	TR	TR‡	TR	No	0.310
P15	D	2	100	80	3.3	1	D	2	70	100	NP	TR‡	TR	CNA	0.127
P16	D	2	80	30	3.8	1	D	1	20	70	TR	TR‡	NP	Clinical	0.317
P18	D	3	0	0	2.2	1	D	2	80	50	NP	NP‡	NP	No	0.004
P19§	D	3	0	0	3.0	1	D	3	0	0	TR	TR‡	TR	No	0.325
P20	D	3	0	0	1.4	0	D	3	0	0	TR	TR‡	NP	Clinical	0.139
P21	D	2	80	0	4.2	1	D	2	70		TR	TR‡	TR	No	0.360
P22§	D	2	20	50	3.5	1	M	3	15	0	TR	TR‡	NP	Clinical	0.394
P23	D	3	0	0	0.8	1	D	3	0	0	TR	TR‡	TR	No	0.341
P24§	D	3	0	0	1.0	1	D	3	0	0	TR	TR‡	TR	No	0.311
P25§	D	3	75	70	2.2	1	D	3	70	15	TR	TR‡	TR	No	0.375
P26	D	3	0	0	1.8	1	D	3	0	0	TR	TR‡	TR	No	0.519

* Type = histologic type (D = ductal, L = lobular, M = micropapillary); Grade = histologic grade; ER = estrogen receptor; PR = progesterone receptor; Location (1 = IBC at the index quadrant, 0 = IBC at a different quadrant); CNA = cluster according to copy number alterations; BKP = cluster according to breakpoints; Clinical = definition according to clinical criteria; NP = new primary tumor; TR = true recurrence.

† NP according to the partial identity score.

‡ Agreement with the definition by the partial identity score.

§ The ipsilateral breast cancers of these pairs received chemotherapy before surgery.

13 were true recurrences. Ipsilateral breast cancers occurred at a median time of 3.1 years after the initial breast cancer diagnosis (range = 0.8–6.5 years). In three of 22 (14%) patients, ipsilateral breast cancers occurred in a different quadrant than the initial tumor; all of these were defined clinically as new primary tumors.

Genomic Studies

The pangenomic profiles of a primary tumor and its ipsilateral breast cancer revealed common breakpoints, with a precision within a SNP that can be used as markers of their clonal relatedness. Pair 5 is given as an illustration (Fig. 1).

The median number of breakpoints per array was statistically significantly higher for ipsilateral breast cancers (median = 71, range = 21–433) than for primary tumors (median = 52, range = 4–646) ($P = .001$) (Table 3). The mean number of common breakpoints per pair was also statistically significantly higher for natural pairs (mean = 18.8, SD = 18.8) than for artificial pairs (mean = 4.1, SD = 3.1) ($P = 0.5 \times 10^{-6}$).

Clustering by Copy Number Alterations or Breakpoints

According to hierarchical clustering by DNA CNAs (Fig. 2) and by breakpoints (Fig. 3), five and six ipsilateral breast cancers, respectively, were new primary tumors. The two clustering methods and the clinical definition agreed for 14 pairs (Table 2). However, for five pairs (P6, P12, P16, P20, P22), the clinical definition disagreed with both clustering methods and, for three others (P1, P2, P15), the clustering by breakpoints disagreed with that by CNAs but agreed with the clinical definition. The recurrences in pairs 1 and 2 were identified as true recurrences by the CNA clustering but as new primary tumors by the clinical definitions because of the reappearance of estrogen receptors in the pair 1 ipsilateral breast cancer and different histologic type (ductal instead of lobular carcinoma) in pair 2. In pair 15, CNA clustering did not find a true recurrence, whereas the clinical definition did. No statistically significant differences in clinical and histologic characteristics between the patients diagnosed with new primary tumors or true recurrences were observed by breakpoint information, apart from a suggestion for patients with new primary tumors to be younger and to have a more frequent family history of breast cancer (Supplementary Table 1, available online).

Partial Identity Score

According to the partial identity score reported for each pair in Table 2, 15 ipsilateral breast cancers were true recurrences and seven were new primary tumors (Fig. 4). With a type I error set at 5%, the partial identity score disagreed with clustering by breakpoints in pair 12 only; the clinical definition was new primary tumor because of a change in tumor location. When the score was determined according to Waldman's percent of concordance without either weighing the influence of the coexistence of breakpoints according to their frequency in a similar population or excluding

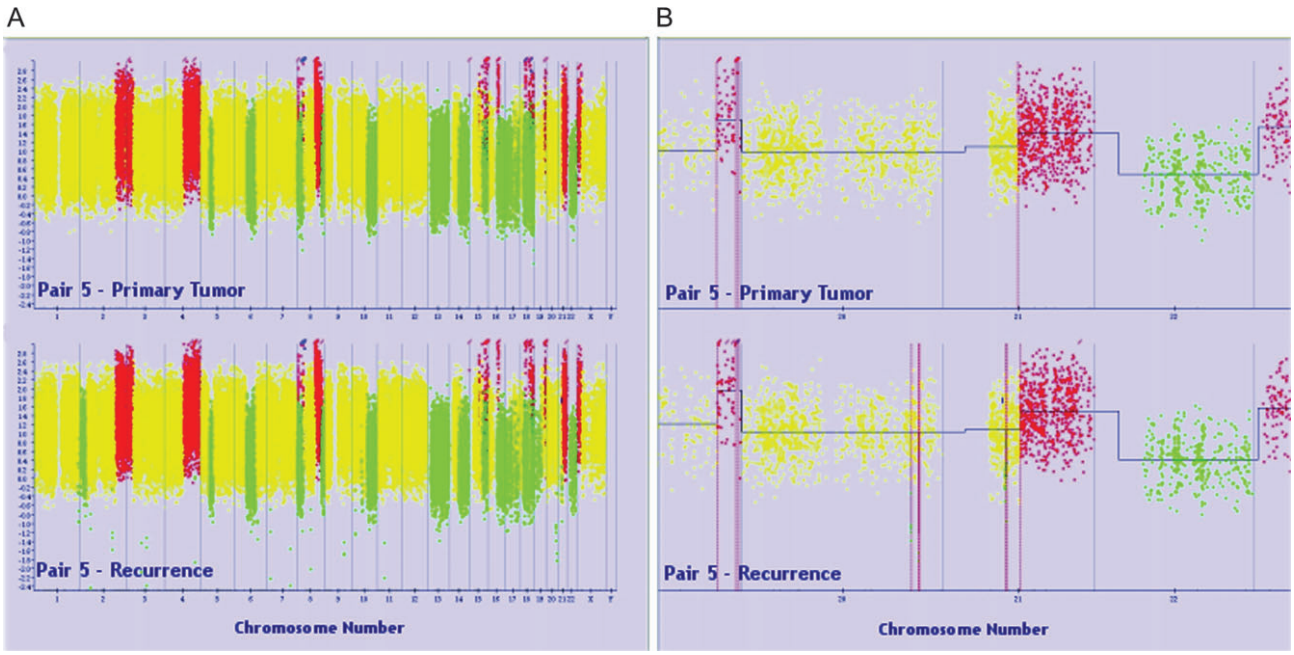


Fig. 1. Genomic profiles of tumors of pair 5 to illustrate the finding of common breakpoints within a single nucleotide polymorphism (SNP). A genomic profile represents the ordered values of the DNA copy numbers obtained as described in "Subjects and Methods". Each **dot** represents the number of DNA copies at each SNP position. Regions with gains are in **red**, with losses in **green**, with no DNA copy number alterations in **yellow**. **A**) Pangenomic profiles. **B**) Profiles of chromosomes 20, 21, and 22. Top primary tumor of pair 5; bottom, ipsilateral breast cancer of pair 5. The **blue horizontal line** represents the smoothing line and the **dotted vertical line** the breakpoint position.

the breakpoints that occur in the copy number variable regions in the HapMap collection, the attribution of the status of three pairs (20 changed from a true recurrence to a new primary, whereas 6 and 12 became true recurrences) and two pairs (10 and 12 changed from new primaries to true recurrences) changed, respectively.

The status of all pairs was confirmed by the 1000 random extractions (Supplementary Table 2, available online). The mean cutoff value was 0.1203 (SD = 0.0102) (Supplementary Fig. 2, available online). The cutoff used to determine the status of the 22 ipsilateral breast cancers, which was defined using all 462 artificial pairs, was 0.1212.

Prognostic Value of the Determination of the Nature of the Ipsilateral Breast Cancer

Patients who were diagnosed with true recurrences had lower metastasis-free survival than those diagnosed with new primary tumors (Supplementary Fig. 3, available online). The difference in metastasis-free survival in the two groups was not statistically significant when they were defined based on clinical and histologic characteristics (5-year metastasis-free survival: 76%, 95% CI = 52% to 100% for new primary tumors and 38%, 95% CI = 17% to 83% for true recurrences; $P = .18$; primary tumors vs true recurrences, hazard ratio = 2.8, 95% CI = 0.6 to 13.7). However, metastasis-free survival was different when the groups were defined according to the partial identity score (5-year metastasis-free survival: 100% for new primary tumors and 29%, 95% CI = 11% to 78% for true recurrences; $P = .01$).

Discussion

DNA breakpoint information was more often in agreement with the clinical definition than that from CNAs to define true recurrences among ipsilateral breast cancers in this population. We developed a partial identity score that is based on DNA breakpoints, which allowed statistical discrimination between new primary tumors and true recurrences. This score outperformed the clinical prognosis determination in terms of metastasis-free survival.

We chose to base our study on a series of young (<50 years old) premenopausal women not only because young age is recognized as one of the most important independent prognostic factors for ipsilateral breast recurrence (34–40) but also to ensure a very high level of homogeneity. In addition, all patients had undergone breast-conserving surgery followed by whole-breast radiotherapy for their initial breast cancers, which were selected as either ductal or lobular invasive carcinomas, and were treated at the same cancer center.

Our results show that some ipsilateral breast cancers share with their primary tumors many DNA CNA breakpoints at the same locations (precision to within a SNP, as illustrated in Fig. 1). From these observations, we produced a method of determining true recurrences that relies on a number of assumptions. The first and most obvious is that the vast majority of breast cancers are of clonal origin. The second is that a tumor retains a substantial number of genomic alterations throughout its evolution. The third assumption, which is key to the method that we have developed, is that the exact locations of the breakpoints that are on the edge of a given change in DNA copy numbers are better hallmarks of a given tumor than the magnitude or width of the genomic alteration itself. For example, because the deletion that causes the loss of Phosphatase and TENsin homolog (PTEN) alters regulatory pathways that lead to precocious development and neoplasia in the mammary gland (41), it can be found in many breast cancers (42–44); however, the exact location of the breakpoints bordering this deletion can be specific to a given tumor. We provide as an

Table 3. Number of common breakpoints in natural (same patient) and artificial (two different patients) pairs of primary tumors (vertically) and ipsilateral breast cancers (horizontally)

No. of BKPs in		No. of BKPs in PT*																					
		77	11	46	16	94	8	22	4	31	55	12	11	58	646	89	69	127	49	60	57	41	72
IBC*	Pair	P1	P2	P3	P4	P5	P6	P10	P11	P12	P13	P14	P15	P16	P18	P19	P20	P21	P22	P23	P24	P25	P26
433	P1	6†	3	12‡	3	8	5§	5	1	4	5	6	1	1	7§	8	6	7	3	8	8	5	12‡
25	P2	0	0†	1	0	1	0	0	0	3‡	0	1	0	0	0	1	0	2	1	0	1	0	0
43	P3	3	2	23†‡§	5	5	2	10§	2§	4	6	5	4	3	4	11	5	7	6	4	8	4	9
26	P4	5	3	7	9†‡§	5	2	7	0	6§	4	4	3	2	0	9‡	3	4	5	3	6	3	5
128	P5	3	3	11	4	64†‡§	1	7	0	4	4	5	2	2	2	8	4	3	8	3	2	3	10
21	P6	3	3	4‡	3	3	3†	4‡	0	4‡	1	4‡	2	0	0	3	1	2	1	1	2	4‡	2
23	P10	3	2	4	3	3	1	3†	1	2	2	1	1	1	3	5‡	1	1	2	1	1	5‡	3
97	P11	5	2	19‡	6	9	1	9	2†§	6§	9	7	6§	5	7§	14	7	10	9	4	12	4	13
35	P12	6‡	3	4	5	4	2	3	0	6†‡§	2	2	3	2	0	4	3	3	3	1	4	4	4
74	P13	3	2	7	3	6	1	5	1	3	18†‡§	4	3	2	2	7	3	3	4	2	2	5	2
35	P14	1	2	7	3	7	3	5	0	3	5	10†‡§	2	1	3	6	3	4	3	2	3	5	4
49	P15	5	2	5	3	4	2	3	0	6‡§	4	1	5†	4	2	3	2	4	3	1	1	2	2
84	P16	2	2	3	2	3	0	2	0	4	2	0	3	23†‡§	1	1	1	3	2	0	3	3	4
53	P18	2	2	9‡	3	3	1	5	1	3	2	3	2	0	2†	7	5	3	2	3	2	3	5
150	P19	9§	4§	18	5	8	2	10§	2§	3	10	5	5	5	7§	42†‡§	13†	11	6	11	10	6	10
93	P20	4	1	6	1	5	0	3	1	2	4	1	2	1	5	7	12†‡	3	4	6	3	3	6
219	P21	2	1	12	3	6	1	5	2§	2	5	3	4	4	6	8	7	63†‡§	6	7	8	3	5
100	P22	5	2	17	5	8	1	10§	1	5	5	5	4	5	3	13	9	10	31†‡§	6	10	5	9
73	P23	7	1	10	3	6	1	7	2§	3	5	5	2	1	5	12	10	6	6	25†‡§	6	3	10
69	P24	6	2	11	5	3	2	6	1	4	5	3	2	3	5	9	5	5	3	7	23†‡§	1	11
42	P25	4	3	9	5	5	2	7	2§	4	5	5	2	2	2	5	4	4	6	1	2	18†‡§	3
88	P26	5	3	11	7	7	1	9	1	6§	5	3	2	4	3	17	5	2	8	5	9	3	43†‡§

* Number of BKPs per tumor. BKP = breakpoint; PT = primary tumor; IBC = ipsilateral breast cancer.

† Numbers correspond to the 22 natural pairs of PTs and their IBCs arising in the same patient; numbers in the other cells correspond to the 462 (22 × 21) artificial pairs of each PT with all other possible IBCs arising in other patients.

‡ Pairs with the most common BKPs per PT.

§ Pairs with the most common BKPs per IBC.

example (Supplementary Fig. 4, available online) the prototype case of PTEN deletion in which the breakpoints are identical between the primary tumor and ipsilateral breast cancer of pair 5 and yet differ in all the other tumors that also harbor a loss of PTEN.

Because clustering is commonly used to determine whether two tumors are clonally related and because it performs better than previously developed similarity scores (18,19), we addressed the issue of whether there was added value in looking at breakpoints rather than at CNAs by comparing clustering by CNAs and by breakpoints to determine the nature of the ipsilateral breast cancer. We concluded from the comparison of clusterings of CNAs and of breakpoints that breakpoint information is more valid than CNA information because when they were discordant, the definition by breakpoints always agreed with the clinical definition, which is routinely used in clinical practice.

A second issue was whether a method could be found to quantify the partial identity between two tumors. We chose to use a partial identity score rather than the results of clustering for a number of reasons. 1) Clustering methods have been designed for exploratory data analysis, so that using a score is more appropriate for a discrimination purpose. 2) A score induces a natural ordering of the pairs from the most dissimilar to the most similar, which is not the case for clustering. 3) The assessment of clonal relatedness by a score can be statistically motivated through the choice of a threshold, as we have demonstrated in the present work. For clustering, clonal relatedness of two tumors depends only on their being clustered apart on the dendrogram, which leads to inconsistent decisions over time. As illustrated by Fig. 3, if pair 2 had not been included in the study, the ipsilateral breast cancer from pair 6 would have been considered as a true recurrence rather than a new primary tumor. Conversely, the assessment of the partial identity score robustness was satisfactory with a narrow range of the cutoff (Supplementary Fig. 2, available online) and with the consistency of the ipsilateral breast cancer status (Supplementary Table 2, available online). Moreover, a score allows one to choose the cutoff that best distinguishes new primary tumors from true recurrences. In this study, we chose a type I error rate at 5% to favor sensitivity for diagnosing true recurrences over the specificity. Further studies will be needed to verify the biologic validity of this choice (Supplementary Fig. 3, available online).

In addition, we chose to weigh the influence of a common breakpoint between the ipsilateral breast cancer and its primary tumor by a factor that takes into account the frequency of this given breakpoint in a population of similar tumors. This weighting changed the determination of three of 22 pairs.

The clinical definition considered an ipsilateral breast cancer as a new primary tumor when the partial identity score did not in three instances. In the first because of a change in location for pairs 12 and 20, in the second because of a lesser degree of differentiation for pair 16, and in the third because of a change in histology for pair 22. The first example illustrates the possibility that a true recurrence can occur at a distance from the first cancer. The second exemplifies the possibility for a true recurrence to have many but not all of the striking alterations present in the primary tumor.

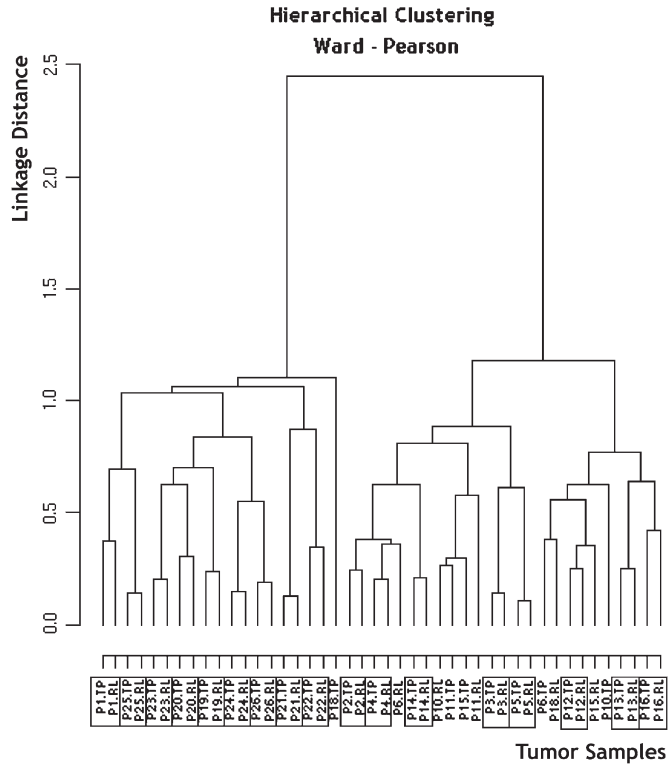


Fig. 2. Dendogram of hierarchical clustering by DNA copy number alterations (Ward–Pearson) of 22 available pairs of primary tumors (TP) and their ipsilateral breast cancer (RL). **Boxes** represent natural pairs with a true recurrence, that is, a pair of tumors from one patient clustered together.

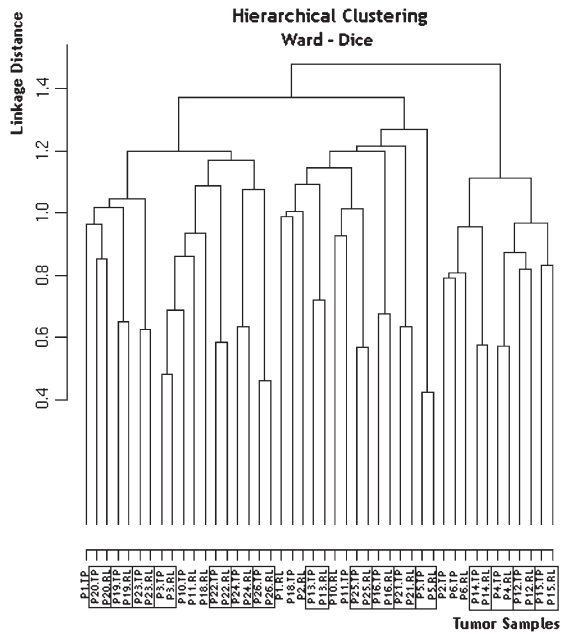


Fig. 3. Dendogram of hierarchical clustering by breakpoints (Ward–Dice) of 22 available pairs of primary tumors (TP) and their ipsilateral breast cancer (RL). **Boxes** represent natural pairs with a true recurrence, that is, a pair of tumors from one patient clustered together.

A criticism that can be made of the clinical definition is that it assumes that a true recurrence is derived from its primary tumor instead of only being related to it. A true recurrence, according to some clinical definitions (5,6,11), cannot be more differentiated than its primary tumor. Usual classifications define differentiation according to histologic grading, DNA ploidy, or the presence of ductal carcinoma in situ. They are based on the assumption that tumors accumulate genetic alterations with time (22,45,46) and that the chronologic order of these alterations reflects the development of a tumor clone. This assumption is, however, challenged by the fact that the ipsilateral breast cancers are neither more aggressive nor more undifferentiated than their primary tumors (47).

The situation with pair 22 illustrates another possible limitation of histologic determination. Here, the clinical status of the ipsilateral breast cancer was of a new primary tumor because its histologic type was a micropapillary carcinoma, whereas the initial tumor was a ductal carcinoma. However, after further histologic analysis, a minor component of micropapillary carcinoma was revealed in the initial carcinoma that otherwise would have been overlooked (Supplementary Fig. 5, available online). This finding implies that, in some instances, the current histologic taxonomy, which is based more on architectural features than on biologic ones, could become obsolete and that some ipsilateral breast cancers could qualify as true recurrences without sharing the same histologic type as their primary tumors.

We observed that patients with true recurrences had lower metastasis-free survival than patients with new primary tumors and that this difference became statistically significant when the partial identity score, instead of clinical definition, was used to define ipsilateral breast cancer types. This observation has been shared by many authors (5,6,10,12). Possible explanations are,

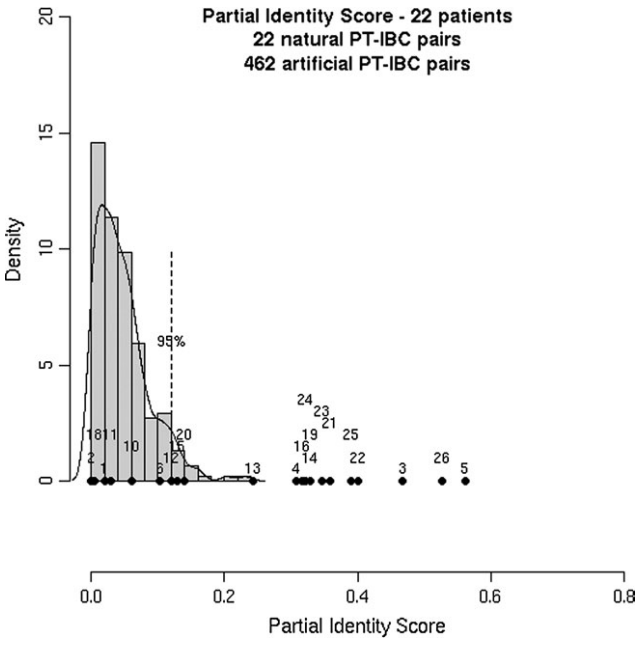


Fig. 4. Partial identity score. Histogram performed on 462 artificial pairs (two different patients) of tumors and representation of the 22 natural (same patient) pairs of primary tumors (PT)/ipsilateral breast cancer (IBC). x-axis: partial identity score (the higher the score, the more likely the IBC is a true recurrence), y-axis: number of artificial pairs in **boxes**. The **vertical dashed bar** represents the upper 5th percentile of the artificial pairs distribution and the threshold above which true recurrences were defined (rejection of the null hypothesis). Each **dot** represents one of the 22 natural pairs (its identifier is written above it).

first, that a true recurrence is the expression of clones that are resistant to adjuvant treatment and therefore could be more difficult to eradicate and, second, that it could be the tip of the iceberg, that is, distant metastases. Conversely, new primary tumors have a prognosis similar to de novo primary cancers but can also reflect a genetic predisposition to develop breast cancer, in the contralateral breast in particular. The clinical implication should therefore be to advocate the use of a systemic treatment in the case of true recurrences and the use of either chemoprevention, such as hormone therapies (48–50) or screening with magnetic resonance imaging (51–53), for patients who are diagnosed with new primary tumors. Here, using breakpoint information led to a better discrimination between new primary tumors and true recurrences in terms of metastasis-free prognosis than the clinical definition.

We also hope that a better distinction among ipsilateral breast cancers of tumors that are genetically related to their primary tumors, that is, true recurrences, will help reveal genetic differences that would provide new information on radioresistance and tumor aggressiveness. To date, little is known about the differential or similarity of the pangenomic expression or the nature of both new primary tumors and ipsilateral breast cancers. Kreike et al. (54) performed a gene expression analysis of 18 000 cDNAs in nine pairs of primary breast cancer with their ipsilateral breast recurrences among women who were younger than 51 years at the time of their initial breast-conserving therapy. Paired data analysis showed no set of genes that had consistently different levels of expression in primary tumors and local recurrences. Another route that has still scarcely been explored is the search for a biologic signature to predict the risk of local recurrence, especially after breast-conserving treatment (54–56). A better distinction between new primary tumors and true recurrences is needed to perform a supervised study based on the occurrence of true recurrences only and not of all ipsilateral breast cancers.

However, our scoring method, which is based on the DNA breakpoint partial identity, has two shortcomings. First, it suffers from the need to conserve unaltered, freshly frozen tissue samples of both the primary tumor and the ipsilateral breast recurrence. This problem should, however, be resolved in time with the possibility of performing the same genomic studies on formalin-fixed paraffin-embedded tissue samples (57–61) or when cryoconservation of either biopsies or fine-needle aspirations (because only 250 ng of DNA is needed, ie, less than 50 000 cells) become standard practice and will make it possible to perform SNP arrays on many more patients. Second, it requires selecting tumors with a cancer cellularity of more than 50%, discarding in the process a number of potentially analyzable tumors. This loss should be diminished in time with both a better selection of frozen tissue material due to the increased experience of the pathologist and the possibility of performing laser capture microdissection.

References

1. Temple WJ, Russell ML, Parsons LL, et al. Conservation surgery for breast cancer as the preferred choice: a prospective analysis. *J Clin Oncol.* 2006;24(21):3367–3373.
2. Clarke M, Collins R, Darby S, et al. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *Lancet.* 2005; 366(9503):2087–2106.
3. Engel J, Kerr J, Schlesinger-Raab A, Sauer H, Holzel D. Quality of life following breast-conserving therapy or mastectomy: results of a 5-year prospective study. *Breast J.* 2004;10(3):223–231.
4. Moyer A. Psychosocial outcomes of breast-conserving surgery versus mastectomy: a meta-analytic review. *Health Psychol.* 1997;16(3): 284–298.
5. Haffty BG, Carter D, Flynn SD, et al. Local recurrence versus new primary: clinical analysis of 82 breast relapses and potential applications for genetic fingerprinting. *Int J Radiat Oncol Biol Phys.* 1993;27(3): 575–583.
6. Huang E, Buchholz TA, Meric F, et al. Classifying local disease recurrences after breast conservation therapy based on location and histology: new primary tumors have more favorable outcomes than true local disease recurrences. *Cancer.* 2002;95(10):2059–2067.
7. Gage I, Recht A, Gelman R, et al. Long-term outcome following breast-conserving surgery and radiation therapy. *Int J Radiat Oncol Biol Phys.* 1995;33(2):245–251.
8. Touboul E, Buffat L, Belkacemi Y, et al. Local recurrences and distant metastases after breast-conserving surgery and radiation therapy for early breast cancer. *Int J Radiat Oncol Biol Phys.* 1999;43(1):25–38.
9. Recht A, Silen W, Schnitt SJ, et al. Time-course of local recurrence following conservative surgery and radiotherapy for early stage breast cancer. *Int J Radiat Oncol Biol Phys.* 1988;15(2):255–261.
10. Komoike Y, Akiyama F, Iino Y, et al. Analysis of ipsilateral breast tumor recurrences after breast-conserving treatment based on the classification of true recurrences and new primary tumors. *Breast Cancer.* 2005;12(2): 104–111.
11. Smith TE, Lee D, Turner BC, Carter D, Haffty BG. True recurrence vs. new primary ipsilateral breast tumor relapse: an analysis of clinical and pathologic differences and their implications in natural history, prognoses, and therapeutic management. *Int J Radiat Oncol Biol Phys.* 2000;48(5): 1281–1289.
12. Schlechter BL, Yang Q, Larson PS, et al. Quantitative DNA fingerprinting may distinguish new primary breast cancer from disease recurrence. *J Clin Oncol.* 2004;22(10):1830–1838.
13. Wang ZC, Buraimoh A, Iglehart JD, Richardson AL. Genome-wide analysis for loss of heterozygosity in primary and recurrent phyllodes tumor and fibroadenoma of breast using single nucleotide polymorphism arrays. *Breast Cancer Res Treat.* 2006;97(3):301–309.
14. Vicini FA, Antonucci JV, Goldstein N, et al. The use of molecular assays to establish definitively the clonality of ipsilateral breast tumor recurrences and patterns of in-breast failure in patients with early-stage breast cancer treated with breast-conserving therapy. *Cancer.* 2007;109(7): 1264–1272.
15. van der Sijp JR, van Meerbeeck JP, Maat AP, et al. Determination of the molecular relationship between multiple tumors within one patient is of clinical importance. *J Clin Oncol.* 2002;20(4):1105–1114.
16. Shibata A, Tsai YC, Press MF, Henderson BE, Jones PA, Ross RK. Clonal analysis of bilateral breast cancer. *Clin Cancer Res.* 1996;2(4): 743–748.
17. Kuukasjarvi T, Karhu R, Tanner M, et al. Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Res.* 1997;57(8):1597–1604.
18. Waldman FM, DeVries S, Chew KL, Moore DH 2nd, Kerlikowske K, Ljung BM. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. *J Natl Cancer Inst.* 2000;92(4):313–320.
19. Teixeira MR, Ribeiro FR, Torres L, et al. Assessment of clonal relationships in ipsilateral and bilateral multiple breast carcinomas by comparative genomic hybridisation and hierarchical clustering analysis. *Br J Cancer.* 2004;91(4):775–782.
20. Zhao X, Li C, Paez JG, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* 2004;64(9):3060–3071.
21. Balaton AL, Coindre JM, Collin F, et al. Recommendations for the immunohistochemical evaluation of hormone receptors on paraffin sections of breast cancer. Study Group on Hormone Receptors using

Immunohistochemistry FNCLCC/AFAQAP. National Federation of Centres to Combat Cancer/French Association for Quality Assurance in Pathology. *Ann Pathol.* 1996;16:144–148.
22. Nowell PC. The clonal evolution of tumor cell populations. *Science.* 1976;194(4260):23–28.
23. Sambrook J, Fritsch EF, Maniatis T. *Molecular Cloning. A Laboratory Manual.* 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.
24. Hupe P, La Rosa P, Liva S, Lair S, Servant N, Barillot E. ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. *Oncogene.* 2007;26:6641–6652.
25. Kaplan EL, Meier P. Nonparametric estimation from incomplete observation. *J Am Stat Assoc.* 1958;53:457–481.
26. Cox DR, Oakes D. *Analysis of Survival Data*. London: Chapman & Hall; 1984.
27. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics.* 2004;20(18):3413–3422.
28. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature.* 2006;444(7118):444–454.
29. La Rosa P, Viara E, Hupe P, et al. VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics.* 2006;22(17):2066–2073.
30. Dice LR. Measures of the amount of ecologic association between species. *Ecology.* 1945;26:297–302.
31. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* 1963;58:236–244.
32. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer.* 1994; 73(3):643–651.
33. Sobin LH, Wittekind C. TNM classification of malignant tumours. New York: Wiley-Liss; 2002.
34. Vrieling C, Collette L, Fourquet A, et al. Can patient-, treatment- and pathology-related characteristics explain the high local recurrence rate following breast-conserving therapy in young patients?. *Eur J Cancer.* 2003;39(7):932–944.
35. Fourquet A, Campana F, Zafrani B, et al. Prognostic factors of breast recurrence in the conservative management of early breast cancer: a 25-year follow-up. *Int J Radiat Oncol Biol Phys.* 1989;17(4):719–725.
36. Borger J, Kemperman H, Hart A, Peterse H, van Dongen J, Bartelink H. Risk factors in breast-conservation therapy. *J Clin Oncol.* 1994;12(4): 653–660.
37. Elkhuizen PH, van de Vijver MJ, Hermans J, Zonderland HM, van de Velde CJ, Leer JW. Local recurrence after breast-conserving therapy for invasive breast cancer: high incidence in young patients and association with poor survival. *Int J Radiat Oncol Biol Phys.* 1998;40(4): 859–867.
38. Elkhuizen PH, Voogd AC, van den Broek LC, et al. Risk factors for local recurrence after breast-conserving therapy for invasive carcinomas: a case-control study of histological factors and alterations in oncogene expression. *Int J Radiat Oncol Biol Phys.* 1999;45(1):73–83.
39. Oh JL, Bonnen M, Outlaw ED, et al. The impact of young age on locoregional recurrence after doxorubicin-based breast conservation therapy in patients 40 years old or younger: how young is "young"?. *Int J Radiat Oncol Biol Phys.* 2006;65(5):1345–1352.
40. Bollet MA, Sigal-Zafrani B, Mazeau V, et al. Age remains the first prognostic factor for loco-regional breast cancer recurrence in young (<40 years) women treated with breast conserving surgery first. *Radiother Oncol.* 2007;82(3):272–280.
41. Li G, Robinson GW, Lesche R, et al. Conditional loss of PTEN leads to precocious development and neoplasia in the mammary gland. *Development.* 2002;129(17):4159–4170.
42. Sapolsky RJ, Hsie L, Berno A, Ghandour G, Mittmann M, Fan JB. High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays. *Genet Anal.* 1999;14(5–6):187–192.
43. Jonsson G, Staaf J, Olsson E, et al. High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization. *Genes Chromosomes Cancer.* 2007;46(6):543–558.
44. Perez-Tenorio G, Alkhori L, Olsson B, et al. PIK3CA mutations and PTEN loss correlate with similar prognostic factors and are not mutually exclusive in breast cancer. *Clin Cancer Res.* 2007;13(12): 3577–3584.
45. Chen LC, Kurisu W, Ljung BM, Goldman ES, Moore D 2nd, Smith HS. Heterogeneity for allelic loss in human breast cancer. *J Natl Cancer Inst.* 1992;84(7):506–510.
46. Lininger RA, Fujii H, Man YG, Gabrielson E, Tavassoli FA. Comparison of loss heterozygosity in primary and recurrent ductal carcinoma in situ of the breast. *Mod Pathol.* 1998;11(12):1151–1159.
47. Sigal-Zafrani B, Bollet MA, Antoni G, et al. Are ipsilateral breast tumour invasive recurrences in young (40 years) women more aggressive than their primary tumours?. *Br J Cancer.* 2007;97(8):1046–1052.
48. Powles TJ, Ashley S, Tidy A, Smith IE, Dowsett M. Twenty-year follow-up of the Royal Marsden randomized, double-blinded tamoxifen breast cancer prevention trial. *J Natl Cancer Inst.* 2007;99(4):283–290.
49. Cuzick J, Forbes JF, Sestak I, et al. Long-term results of tamoxifen prophylaxis for breast cancer—96-month follow-up of the randomized IBIS-I trial. *J Natl Cancer Inst.* 2007;99(4):272–282.
50. Veronesi U, Maisonneuve P, Rotmensz N, et al. Tamoxifen for the prevention of breast cancer: late results of the Italian Randomized Tamoxifen Prevention Trial among women with hysterectomy. *J Natl Cancer Inst.* 2007;99(9):727–737.
51. Kriege M, Brekelmans CT, Boetes C, et al. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. *N Engl J Med.* 2004;351(5):427–437.
52. Kuhl CK, Schrading S, Leutner CC, et al. Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. *J Clin Oncol.* 2005;23(33): 8469–8476.
53. Lehman CD, Gatsonis C, Kuhl CK, et al. MRI evaluation of the contralateral breast in women with recently diagnosed breast cancer. *N Engl J Med.* 2007;356(13):1295–1303.
54. Kreike B, Halfwerk H, Kristel P, et al. Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. *Clin Cancer Res.* 2006;12(19): 5705–5712.
55. Nuyten DS, Kreike B, Hart AA, et al. Predicting a local recurrence after breast-conserving therapy by gene expression profiling. *Breast Cancer Res.* 2006;8(5):R62.
56. Niméus E, Krogh M, Malmström P, Strand C, Fredriksson I, Karlsson P, et al. Gene expression profiling in primary breast cancer distinguishes patients developing local recurrence despite postoperative radiotherapy after breast conserving surgery. In: 29th Annual *San Antonio Breast Cancer Symposium* 2006. San Antonio, TX: 2007;103(1):115–124.
57. Isola J, DeVries S, Chu L, Ghazvini S, Waldman F. Analysis of changes in DNA sequence copy number by comparative genomic hybridization in archival paraffin-embedded tumor samples. *Am J Pathol.* 1994;145(6): 1301–1308.
58. Devries S, Nyante S, Korkola J, et al. Array-based comparative genomic hybridization from formalin-fixed, paraffin-embedded breast tumors. *J Mol Diagn.* 2005;7(1):65–71.
59. Johnson NA, Hamoudi RA, Ichimura K, et al. Application of array CGH on archival formalin-fixed paraffin-embedded tissues including small numbers of microdissected cells. *Lab Invest.* 2006;86(9): 968–978.
60. Oosting J, Lips EH, van Eijk R, et al. High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays. *Genome Res.* 2007;17(3):368–376.
61. Schubert EL, Hsu L, Cousens LA, et al. Single nucleotide polymorphism array analysis of flow-sorted epithelial cells from frozen versus fixed tissues for whole genome analysis of allelic loss in breast cancer. *Am J Pathol.* 2002;160(1):73–79.

Funding

Institut Curie, the "Courir pour la vie, Courir pour Curie" association, the "Odyssea" association and the PHRC 2006 (AOM 06 149).

Notes

M. A. Bollet and N. Servant contributed equally to this work. The authors thank the members of the departments of Tumor Biology (Martial Caly, Blandine Massemin, Michèle Galut), Biostatistics (Eléonore Gravier, Chantal Gautier), Translational Research (David Gentien, Cécile Reyes, Audrey Rapinat, Benoît Albaud, Vincent Lepetit), and Bioinformatics (Philippe La Rosa, Séverine Lair) who participated in this study. The authors are also indebted to Anne Vincent-Salomon, Patricia de Crémoux, Dominique Stoppa-Lyonnet, and particularly Olivier Delattre for their very valuable comments on this work. Finally, they thank all the members of the Institut Curie Breast Cancer Group.

The sponsors had no role in the study design, data collection, interpretation of the results, preparation of the manuscript, or the decision to submit the manuscript for publication.

Manuscript received June 4, 2007; revised October 16, 2007; accepted November 13, 2007.

nature immunology

A critical function for transforming growth factor-β, interleukin 23 and proinflammatory cytokines in driving and modulating human T_H-17 responses

Elisabetta Volpe[1,2], Nicolas Servant[3–5], Raphaël Zollinger[1,2], Sofia I Bogiatzi[1,2], Philippe Hupé[3–6], Emmanuel Barillot[3–5] & Vassili Soumelis[1,2]

Interleukin 17 (IL-17)–producing T helper 17 cells (T_H-17 cells) have been described as a T helper cell subset distinct from T helper type 1 (T_H1) and T_H2 cells, with specific functions in antimicrobial defense and autoimmunity. The factors driving human T_H-17 differentiation remain controversial. Using a systematic approach combining experimental and computational methods, we show here that transforming growth factor-β, interleukin 23 (IL-23) and proinflammatory cytokines (IL-1β and IL-6) were all essential for human T_H-17 differentiation. However, individual T_H-17 cell–derived cytokines, such as IL-17, IL-21, IL-22 and IL-6, as well as the global T_H-17 cytokine profile, were differentially modulated by T_H-17-promoting cytokines. Transforming growth factor-β was critical, and its absence induced a shift from a T_H-17 profile to a T_H1-like profile. Our results shed new light on the regulation of human T_H-17 differentiation and provide a framework for the global analysis of T helper responses.

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

Since the initial description of T helper type 1 (T_H1) and T_H2 cells[1], cytokines have seemed increasingly important for the induction, regulation and function of distinct T helper subsets. In the T_H1-T_H2 paradigm, single cytokines such as interleukin 12 (IL-12) or IL-4 induce the differentiation of T_H1 or T_H2 cells, respectively[2]. Other parameters, such as the dose of antigen or type of costimulation, are able to modulate T_H1 or T_H2 responses[2].

Another subset of T helper cells that produce IL-17 (T_H-17 cells) has been identified as being distinct from T_H1 and T_H2 cells[3,4]. T_H-17 cells have specific functions in antimicrobial immunity[5,6] and autoimmune inflammation[7–9]. In mice, many cytokines are required and act in a coordinated way to induce T_H-17 differentiation, with a critical function for transforming growth factor-β (TGF-β; A002271) and IL-6 (refs. 8,10,11). IL-6 induces IL-21 production, which subsequently favors T_H-17 differentiation in an autocrine way[12–14]. Mouse T_H-17 cells produce not only IL-17 but also IL-21 (ref. 15), IL-22 (ref. 16) and, in some cases, IL-10 (ref. 17). It is unclear at present whether T_H-17 cells can produce additional T helper cytokines and to what extent the requirements for induction of these T_H-17-associated cytokines are similar. Understanding the regulation of the global T_H-17 cytokine profile is essential, as each T helper cytokine has specific functions.

Characterizing the factors driving human T_H-17 differentiation is of particular interest because of the importance of T_H-17 cells in health and disease. Five independent reports have addressed this issue with unexpectedly contradictory results. Three studies showed IL-1β[18], IL-23 (ref. 19) or polyclonal stimulation with antibody to CD3 (anti-CD3) and anti-CD28 (ref. 20) to be sufficient for the generation of human T_H-17 cells, in contrast to the many factors required in mice. Two other studies were not able to differentiate naive CD4 T cells into T_H-17 cells, even with conditions shown to be efficient in mouse or human systems[21,22]. Finally, TGF-β, which has been shown to be essential for mouse T_H-17 differentiation, has been reported as a negative regulator in humans[18,19]. Thus, the requirements for human T_H-17 differentiation remain controversial[23]. Here we show that TGF-β, IL-23 and proinflammatory cytokines (IL-1β and IL-6) were essential components of human T_H-17 differentiation and expression of IL-17A, IL-17F, the IL-23 receptor (IL-23R) and the transcription factor RORγt. However, experimental and computational methods showed that each T_H-17-promoting cytokine had a specific function in the regulation of the global T_H-17 cytokine profile.

RESULTS

Driving IL-17 production

To define the cytokine requirements for the induction of human T_H-17 differentiation, we did a standard naive CD4 T cell differentiation assay in the presence of polyclonal stimulation with anti-CD3 and anti-CD28. We systematically tested all cytokines shown to be involved

[1]Institut National de la Santé et de la Recherche Médicale U653, Paris F-75248, France. [2]Institut Curie, Laboratoire d'Immunologie Clinique, Paris F-75248, France. [3]Institut Curie, Bioinformatique, Paris F-75248, France. [4]Institut National de la Santé et de la Recherche Médicale, U900, Paris F-75248, France. [5]Ecole des Mines de Paris, ParisTech, Fontainebleau F-77300, France. [6]Centre National de la Recherche Scientifique, UMR144, Paris F-75248, France. Correspondence should be addressed to V.S. (vassili.soumelis@curie.net).

Received 20 February; accepted 9 April; published online 4 May 2008; doi:10.1038/ni.1613

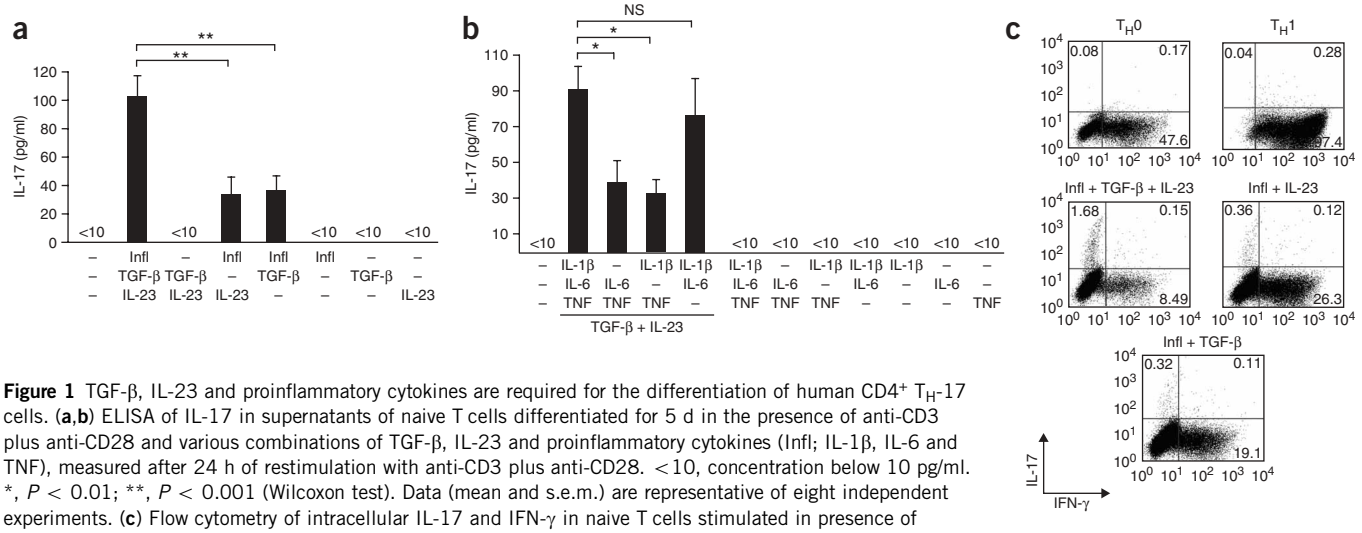


Figure 1 TGF-β, IL-23 and proinflammatory cytokines are required for the differentiation of human CD4$^+$ T_H-17 cells. (**a**,**b**) ELISA of IL-17 in supernatants of naive T cells differentiated for 5 d in the presence of anti-CD3 plus anti-CD28 and various combinations of TGF-β, IL-23 and proinflammatory cytokines (Infl; IL-1β, IL-6 and TNF), measured after 24 h of restimulation with anti-CD3 plus anti-CD28. <10, concentration below 10 pg/ml. *, $P < 0.01$; **, $P < 0.001$ (Wilcoxon test). Data (mean and s.e.m.) are representative of eight independent experiments. (**c**) Flow cytometry of intracellular IL-17 and IFN-γ in naive T cells stimulated in presence of anti-CD3 plus anti-CD28 and various cytokines (above plots), assessed after 6 h of restimulation with anti-CD3 plus anti-CD28. T_H0, no polarizing cytokines; T_H1, IL-12; Infl, IL-1β and IL-6 and TNF. Numbers in quadrants indicate percent IL-17$^+$IFN-γ$^-$ cells (top left), IL-17$^+$IFN-γ$^+$ cells (top right) or IL-17$^-$IFN-γ$^+$ cells (bottom right). Data are from one experiment representative of five independent experiments.

in the polarization of IL-17-producing cells in mouse systems. In the first set of experiments, we independently considered TGF-β, IL-23 and proinflammatory cytokines (IL-1β, IL-6 and tumor necrosis factor (TNF); **Fig. 1a**). None of these components individually was sufficient to induce detectable IL-17 (**Fig. 1a**). A combination of these three components, however, induced high production of IL-17 (**Fig. 1a**). IL-17 dropped to undetectable amounts in the absence of proinflammatory cytokines and decreased by 70% in the absence of TGF-β or IL-23 (**Fig. 1a**), which indicated that the three components were required for optimal IL-17 production. We obtained similar results with CD3$^+$CD4$^+$CD45RA$^+$ peripheral blood naive T cells obtained by negative or positive selection and with total CD4$^+$ T cells from cord blood (data not shown).

In a second set of experiments, we addressed the function of individual proinflammatory cytokines. In the presence of TGF-β and IL-23, the removal of TNF had only a small effect on IL-17 production (**Fig. 1a**). We obtained similar results in the presence or absence of IL-1β and/or IL-6, which confirmed that TNF does not have a substantial effect on IL-17 production (**Supplementary Fig. 1** online), contrary to what has been reported for mice[11]. The removal of IL-1β or IL-6 induced a comparably substantial decrease (over 50%; **Fig. 1b**). Any combination of one, two or three of the proinflammatory cytokines was not sufficient to induce detectable IL-17 production in the absence of TGF-β and IL-23 (**Fig. 1b**). T cell population expansion on day 5 of culture was similar in all cytokine combinations, which indicated that differences in cytokine production could not be attributed to insufficient expansion (**Supplementary Fig. 2a**,**b** online).

Intracellular cytokine staining confirmed that TGF-β, IL-23 and proinflammatory cytokines induced a well defined IL-17-producing cell population (**Fig. 1c**), which dropped by 70% in the absence of TGF-β or IL-23. That paralleled the data obtained by enzyme-linked immunosorbent assay (ELISA). The decrease in IL-17 was accompanied by an increased cell population producing interferon-γ (IFN-γ), up to threefold in the absence of TGF-β (**Fig. 1c**). IL-17-producing cells did not make IFN-γ; this distinguished them from T_H1 cells, which are generated in the presence of IL-12. We detected no IL-17-producing cells among unpolarized T cells (T_H0 cells) or in T_H1 conditions (**Fig. 1c**), which again confirmed the ELISA data (**Fig. 1a**). Using the frequency of the IL-17-producing population, we calculated an average production of 0.006 pg IL-17 per cell. That is similar to the amount of IL-17 produced by *in vitro*–differentiated or *ex vivo* memory T_H-17 cells in other human studies (range, 0.001–0.016 pg/cell)[19,22].

Many endogenous factors present in the T cell cultures could have possibly altered the cytokine requirements for IL-17 production, and it was important to clarify their function in our system. First, we addressed the effect of serum TGF-β, as we used medium containing 10% fetal calf serum for our experiments. When we added a TGF-β-blocking monoclonal antibody to the complete T_H-17 combination, IL-17 decreased considerably, consistent with the inhibition of exogenous TGF-β (**Supplementary Fig. 3a** online). In the absence of exogenous TGF-β, we noted a residual small amount of IL-17, which was not significantly affected by monoclonal antibody to TGF-β ($P = 0.5$; **Supplementary Fig. 3a**). This indicated that endogenous (serum) TGF-β had only a marginal function in our system. We also did TGF-β 'titration', which confirmed that TGF-β acted positively to regulate IL-17 production in a dose-dependent way (**Supplementary Fig. 3b**). Second, we addressed the function of serum itself, which may affect T_H-17 differentiation independently of TGF-β. Although absolute IL-17 production was higher in the absence of serum than in the presence of serum, we found that the cytokine requirements and regulation of IL-17 production were similar in these two types of culture media (**Supplementary Fig. 4** online); TGF-β, IL-23 and proinflammatory cytokines were all required for IL-17 production. We also addressed the function of endogenous IL-4 and IFN-γ, two cytokines described in the mouse to inhibit T_H-17 differentiation[4]. We repeated the same type of experiment in the presence of monoclonal antibodies blocking IL-4 and IFN-γ and found that the cytokine requirements to induce IL-17 production were not affected (**Supplementary Fig. 4**). In summary, none of endogenous factors tested modified the cytokine requirements for the induction of IL-17 production, and TGF-β invariably had a critical function independently of the experimental system.

IL-17 production is associated with typical T_H-17 features

To address whether IL-17 production was associated with the acquisition of typical features of T_H-17 cells, we first measured mRNA encoding various IL-17 family members. Optimal T_H-17 conditions

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

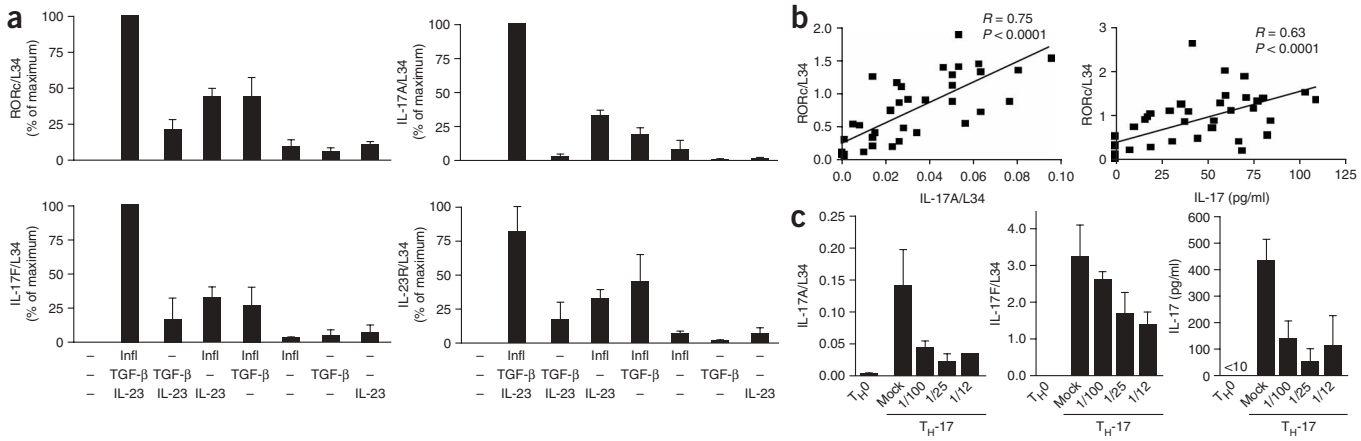


Figure 2 TGF-β, IL-23 and proinflammatory cytokines induce typical T_H-17 features. (**a**) RT-PCR analysis of the expression of RORc, IL-17A, IL-17F and IL-23R mRNA in naive $CD4^+$ T cells differentiated *in vitro* for 5 d in presence of anti-CD3 plus anti-CD28 and various cytokine combinations (below graphs), followed by 24 h of restimulation with anti-CD3 plus anti-CD28. Cycling threshold values are normalized to those of mRNA encoding ribosomal protein L34, and data are normalized to the maximum value obtained for each donor. Data are the mean and s.e.m. of three independent experiments. (**b**) IL-17A transcript and IL-17 protein from cells cultured in the presence of T_H-17-inducing cytokines (IL-1β, IL-6, TNF, TGF-β and IL-23) or in the absence of individual components of that group, correlated to RORc transcripts with the Pearson correlation (*R*, correlation coefficient). Data are representative of six experiments. (**c**) RT-PCR of IL-17A and IL-17F mRNA and ELISA of IL-17 protein in naive T cells cultured in serum-free medium and infected with various dilutions (below graphs) of supernatants containing shRNA-expressing lentiviral vector specific for RORc (shRORc) or empty vector (Mock; negative control) during the first day of T_H-17 differentiation, then washed extensively and cultured for additional 5 d in T_H0 or T_H-17 conditions and analyzed after 24 h of restimulation with anti-CD3 plus anti-CD28. T_H0, no polarizing cytokines; T_H-17, IL-1β, IL-6, TNF, TGF-β and IL-23. Data are the mean and s.e.m. of three independent experiments.

(TGF-β, IL-23 and proinflammatory cytokines) induced the most mRNA encoding IL-17A and IL-17F (**Fig. 2a**), two cytokines associated with T_H-17 differentiation[24]. However, we did not detect measurable amounts of other IL-17 family members, such as IL-17E (IL-25), which is related more to T_H2 responses[25] (data not shown). TGF-β and proinflammatory cytokines were also required for the induction of IL-23R mRNA (**Fig. 2a**), another important characteristic of T_H-17 cells[10,26].

The transcription factor RORγt has been shown to be critical for mouse T_H-17 differentiation[26]. To assess the relationship between human RORγt expression and IL-17 production, we measured the transcription of *RORC*, which encodes the human ortholog of mouse RORγt. We quantified RORc mRNA in many optimal and suboptimal T_H-17-polarizing conditions, similar to those used above (**Fig. 1a**,**b**). We noted a strong correlation between the amount of RORc transcript and IL-17 transcript or protein (**Fig. 2b**), which indicated that RORc expression was both a sensitive and specific marker of human T_H-17 cells and suggested that RORγt could be involved in regulating the production of human IL-17. IL-17F expression was less associated with RORc than was IL-17A expression (**Supplementary Fig. 5** online), which confirmed a published result obtained with mice[27]. To directly address the function of RORγt in controlling IL-17 production, we used short-hairpin RNA (shRNA) to 'knock down' RORc expression. RORc-specific shRNA but not control shRNA induced a decrease of about 50% in RORc mRNA expression (**Supplementary Fig. 6** online); this decrease was sufficient to inhibit IL-17A mRNA and protein at all concentrations of shRNA tested (**Fig. 2c**). The effect on IL-17F expression was weaker but was dose dependent (**Fig. 2c**). Expression of the 'housekeeping' genes *RPL34* (encoding ribosomal protein L34) and *HPRT* (encoding hypoxanthine guanine phosphoribosyl transferase) was not affected by RORc-specific or control shRNA (data not shown); expression of other genes not reported before to depend on RORc, such as those encoding IFN-γ, TNF and IL-23R, was also not affected by shRNA treatment, which indicated that the inhibition of IL-17 was specific (**Supplementary Fig. 6**). Overall, our data show that a combination of TGF-β, IL-23 and proinflammatory cytokines was both necessary and sufficient to induce typical features of T_H-17 differentiation.

Cytokine profiles of T_H1, T_H2 and T_H-17 cells

We then defined the detailed T_H-17 cytokine profile relative to that of standard T_H0, T_H1 and T_H2 subsets. T_H1 differentiation was driven by IL-12, T_H2 differentiation was driven by IL-4, and T_H0 differentiation was driven by polyclonal stimulation (anti-CD3 plus anti-CD28) in the absence of any polarizing cytokines (**Fig. 3**). We did not use blocking monoclonal antibodies in any of these conditions to avoid interfering with potential autocrine loops. We also assessed the expression of transcription factors associated with each of the T helper cell subsets. As expected, T_H1 and T_H2 conditions induced the highest expression of the transcription factors T-bet and GATA-3, respectively, and RORc was highly specific for T_H-17 cells (**Fig. 3a**), which confirmed data obtained with mice[26]. We measured the expression of ten T helper cell–associated cytokines for the T_H0, T_H1, T_H2 and T_H-17 subsets (**Fig. 3b**). IL-17 and IL-6 were the most specific for T_H-17 and were either absent or produced in very small amounts in T_H1 and T_H2 conditions; a second set of cytokines, IL-21, IL-22, TNF and IFN-γ, could be detected in T_H-17 and T_H1 conditions and, notably, IL-21 and IL-22 had similar or higher expression in T_H1 conditions versus T_H-17 conditions; and a third set of cytokines, IL-10, IL-4, IL-5 and IL-13, was produced mainly in T_H2 conditions. Thus, T_H-17 conditions induced the production of many cytokines in addition to IL-17.

Differential regulation of individual T_H-17 cell–derived cytokines

How IL-17-promoting cytokines regulate the production of diverse T_H-17-associated cytokines is not known. To address that issue, we measured T helper cell cytokines in the presence or absence of individual T_H-17-promoting cytokines. As shown before (**Fig. 1a**,**b**),

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

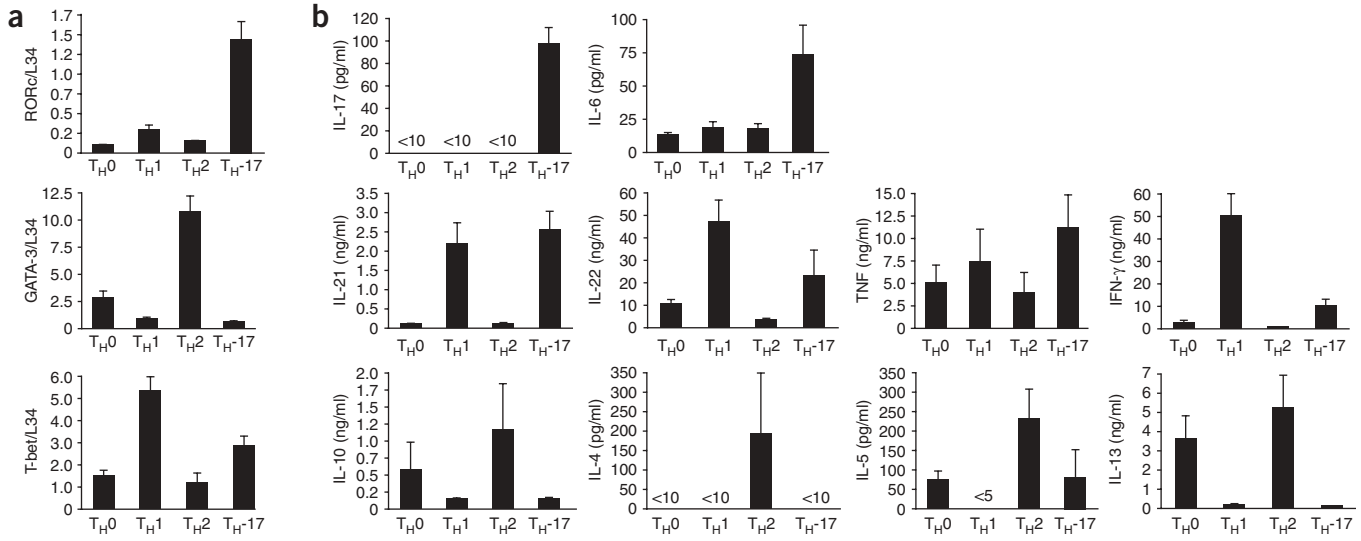


Figure 3 The T_H-17 cytokine profile has specific features but also features that overlap with those of other T helper cell–polarizing conditions. (**a**) RT-PCR analysis of the expression of T-bet, RORc and GATA-3 mRNA in naive T cells differentiated with anti-CD3 plus anti-CD28 in T_H0, T_H1, T_H2 or T_H-17 conditions and then restimulated for 24 h with anti-CD3 plus anti-CD28. Cycling threshold values are normalized to those of L34. Data are the mean and s.e.m. of three independent experiments. (**b**) Cytometric bead assay or ELISA of IL-17, IL-21, IL-22, IL-4, IL-5, IL-6, IL-10, IL-13, IFN-γ and TNF in culture supernatants of naive T cells differentiated in T_H0, T_H1, T_H2 or T_H-17 conditions and then restimulated for 24 h with anti-CD3 plus anti-CD28. T_H0, no polarizing cytokines; T_H1, IL-12; T_H2, IL-4; T_H-17, IL-1β, IL-6, TNF, TGF-β and IL-23. Data are the mean and s.e.m. of eight independent experiments.

removal of any of the five T_H-17-promoting cytokines, except for TNF, decreased the production of IL-17 by over 50% (**Fig. 4**). Notably, each T_H-17-associated cytokine was regulated in a specific way: IL-22 was generally stable, even in the absence of critical T_H-17-inducing cytokines such as IL-1β, IL-6 and IL-23, and the removal of TGF-β induced a significant increase in IL-22 production, which indicated that IL-17 and IL-22 were differentially regulated by TGF-β (**Fig. 4**). IL-21, shown before to be an important autocrine factor in the induction of mouse T_H-17 differentiation[12,13], was specifically dependent on IL-23 (**Fig. 4** and **Supplementary Fig. 7** online). Notably, although T_H-17 conditions induced only low production of IL-10, this production was higher in the absence of IL-1β (**Fig. 4**). We obtained an opposite result with IFN-γ, which indicated that IL-1β differentially regulated IL-10 and IFN-γ. Finally, IL-6 production was mostly dependent on TGF-β and IL-23, a regulation that is more closely related to that of IL-17. Overall, each T_H-17-associated cytokine was regulated in a specific way. This suggested that individual T_H-17-promoting cytokines might not only control the amount of IL-17 produced but also modulate quantitatively and qualitatively the global T helper cytokine profile, potentially inducing a shift in the type of T cell response.

T_H-17 profile modulated by cytokines

To test the hypothesis that T_H-17-promoting cytokines might drive or modulate the global T helper cytokine profile, we measured all ten T helper cytokines in control T helper cell conditions (T_H0, T_H1 and T_H2), in 'optimal' T_H-17 conditions and after the removal of individual T_H-17-promoting cytokines (**Fig. 5a**). We obtained complete data sets (ten cytokines in nine polarizing conditions) from six independent experiments. To allow for comparison among profiles, we normalized values obtained for each cytokine

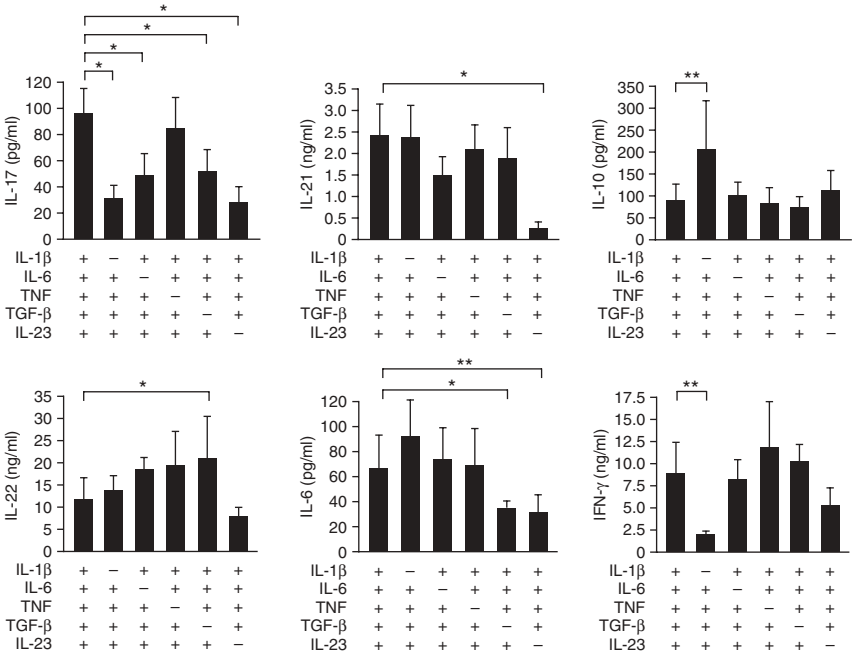


Figure 4 T_H-17 cell–derived cytokines are differentially regulated by T_H-17-promoting cytokines. ELISA and/or cytometric bead assay of IL-17, IL-6, IL-21, IL-22, IL-10 and IFN-γ in supernatants of naive T cells differentiated with anti-CD3 plus anti-CD28 in the presence of IL-1β, IL-6, TNF, TGF-β and IL-23 (T_H-17 conditions) or in the absence of individual components of that group (below graphs) and then restimulated for 24 h with anti-CD3 plus anti-CD28. *, $P < 0.05$; **, $P < 0.01$ (Wilcoxon test). Data are the mean and s.e.m. of seven independent experiments.

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

to the maximum value obtained for that cytokine across the whole data set for each donor (**Fig. 5a**). Three cytokine sets characterized each profile: IL-10, IL-4, IL-5 and IL-13 were highest in T_H2 conditions; IFN-γ and IL-22 were highest in T_H1 conditions; and IL-17, IL-6 and TNF were highest in T_H-17 conditions. IL-21 was similarly high in T_H1 and T_H-17 conditions (**Fig. 5a**). Overall, the T_H-17 profile was distinct from those of T_H1 and T_H2. The analysis of cytokine profiles generated in the absence of each T_H-17-promoting factor identified a notable diversity. It became apparent that each of these cytokines not only controlled IL-17 production but induced substantial changes in the global T_H-17 cytokine profile. For example, removing IL-1β decreased IL-17 and IFN-γ and increased IL-10; removing TGF-β induced a decrease in the entire T_H-17 'cluster' and an increase in the T_H1 'cluster'; and removing IL-23 decreased the T_H-17 sets without any substantial change in the other cytokines (**Fig. 5a**).

We sought to determine whether the profiles generated in suboptimal T_H-17 priming conditions (through the removal of individual T_H-17-promoting cytokines) represented new T helper cell profiles or whether they were related to any of the standard T_H1, T_H2 or T_H-17 profiles. Computational methods were necessary because of the complexity of the data. We used cluster analysis as an exploratory tool to identify similarities among all the profiles (**Fig. 5b**). The agglomerative coefficient of 0.85 obtained indicated that the clustering allowed the separation of samples into clusters of conditions. The resampling similarity index of 0.093 reflected highly robust clusters and suggested high statistical significance. The standard T helper conditions showed that the T_H1, T_H2 and T_H-17 conditions segregated into different clusters, which confirmed their distinct cytokine profiles. The T_H0 condition segregated with the T_H2 condition, probably because of the baseline production of small amounts of T_H2 cytokines in the absence of T_H1 and T_H-17 cytokines (**Fig. 3b**). Among the suboptimal T_H-17 conditions, the removal of TNF induced only a minor change in the profile, which clustered with the optimal T_H-17 condition. Notably, the removal of TGF-β induced a shift in the profile, which clustered with the T_H1 condition; the removal of IL-1β, IL-6 or IL-23 induced profiles with no distinct similarity to the T_H1, T_H2 or T_H-17 conditions (**Fig. 5b**).

We used principal component analysis to complement the cluster analysis and applied this to the average cytokine values obtained for the six donors for each T helper condition (**Fig. 5c**). This analysis can be viewed as a simplification of the data projected along two axes (the two principal components) that best represented the entire data set and preserved maximum dispersion of the data[28]. We projected the T helper cell profiles generated in each polarizing condition onto this two-dimensional space (**Fig. 5c**). Each polarizing condition presented in this way was thus a simplified representation of the cytokine profiles shown for the same polarizing conditions described above (**Fig. 5a**). The two principal components enabled good discrimination among the T_H0, T_H1, T_H2 and T_H-17 profiles, which confirmed that they represented distinct entities (**Fig. 5c**). The removal of TNF induced the smallest deviation from the T_H-17 profile; in the absence of TGF-β, the remaining IL-23 and proinflammatory cytokines induced a profile that was more closely related to that of T_H1, which was 'converted' to a T_H-17 profile in the presence of TGF-β.

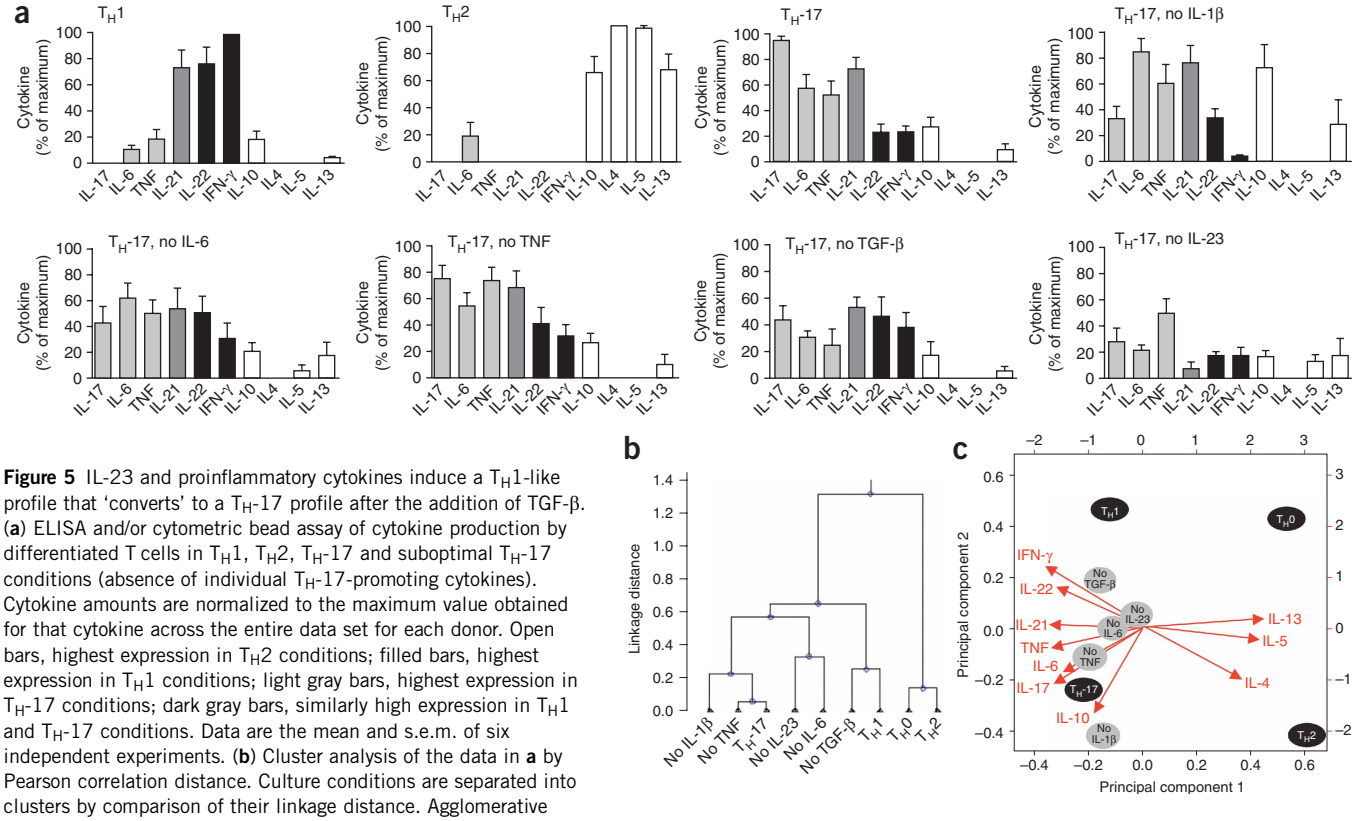


Figure 5 IL-23 and proinflammatory cytokines induce a T_H1-like profile that 'converts' to a T_H-17 profile after the addition of TGF-β. (**a**) ELISA and/or cytometric bead assay of cytokine production by differentiated T cells in T_H1, T_H2, T_H-17 and suboptimal T_H-17 conditions (absence of individual T_H-17-promoting cytokines). Cytokine amounts are normalized to the maximum value obtained for that cytokine across the entire data set for each donor. Open bars, highest expression in T_H2 conditions; filled bars, highest expression in T_H1 conditions; light gray bars, highest expression in T_H-17 conditions; dark gray bars, similarly high expression in T_H1 and T_H-17 conditions. Data are the mean and s.e.m. of six independent experiments. (**b**) Cluster analysis of the data in **a** by Pearson correlation distance. Culture conditions are separated into clusters by comparison of their linkage distance. Agglomerative coefficient, 0.85 (reflects data structure; values near 1 indicate well separated clusters); resampling similarity index, 0.093 (values near 0 indicate a robust cluster). (**c**) Principal component analysis of the data in **a**. T helper conditions (ovals) are positioned in a space defined by the principal components 1 and 2, which are the two 'best' axes representing the entire data set. Black ovals, T_H0, T_H1, T_H2 and T_H-17 profiles; gray ovals, removal of cytokines from the T_H-17 conditions. Red arrows indicate T helper cell–derived cytokines that contribute to the differences among culture conditions. The direction and length of such vectors indicate the importance of each T helper cytokine in discriminating the T helper profiles (**Supplementary Methods** online).

The cytokine vectors indicated the importance of the individual T helper cell–derived cytokines in the discrimination of the different T helper profiles according to the length and direction of the vector (**Fig. 5c**). For example, the IL-4 and IFN-γ vectors pointed in opposite directions, which indicated that they were inversely correlated. Accordingly, the T_H2 profile was determined not only by IL-4 but also by the lack of IFN-γ (opposite vector). The T_H1-like profile induced in the absence of TGF-β 'segregated away' from T_H-17 along the second principal component. According to the cytokine vectors, the presence of IFN-γ and IL-22, along with the decreased IL-17 and IL-6, explained the separation of these two profiles (**Fig. 5c**). Similarly, IL-10, along with smaller amounts of IL-17 and IL-6, explained the separation between the T_H-17 profile and the T_H-17 profile without IL-1β. Our computational analysis of the global T helper profiles thus demonstrated that the removal of individual T_H-17-promoting cytokines generated a diversity of distinct T helper cytokine profiles. Although TNF had a small effect on the global T_H-17 profile, the removal of TGF-β induced substantial changes and a shift toward a T_H1-like profile.

DISCUSSION

Here we have shown that TGF-β, IL-23 and proinflammatory cytokines were essential in driving and regulating four key aspects of human T_H-17 differentiation: IL-17 production; the acquisition of T_H-17-specific features; individual T_H-17-derived cytokines; and the global T_H-17 cytokine profile. The T_H-17 pathway has been linked to the pathogenesis of several autoimmune diseases, including psoriasis[29], experimental allergic encephalomyelitis[9], arthritis[7] and colitis[30]. It is also crucial in immunity to mycobacteria[31] and *Candida albicans*[6]. In mice, several studies have shown that TGF-β and IL-6 are essential in driving T_H-17 differentiation[8,10,11], with regulatory T cells as a potential source of TGF-β[11]. This indicates that proinflammatory cytokines in the absence of TGF-β are not sufficient to induce a T_H-17 response, which gives a central function to TGF-β in the generation of both regulatory T cells and T_H-17 cells. Two studies in human systems have questioned the importance of TGF-β, showing that IL-1β[18] or IL-23 (ref. 19) is sufficient to induce T_H-17 differentiation and that TGF-β negatively regulates this response[18,19]. Those studies[18,19] used much longer T cell assays than the mouse studies[8,10,11] and detected substantial IL-17 production in control medium[18] or IL-2 alone[19], which suggests that T cells might have less stringent requirements for IL-17 production. In our study, although we obtained a lower yield in IL-17-producing cells, we did not find measurable IL-17 in the absence of polarizing cytokines (T_H0) either by ELISA or by intracellular flow cytometry. This suggests that the standard 5-day T cell assay is less sensitive but more specific than systems with longer culture duration. In these conditions, we found that TGF-β was required for optimal human T_H-17 differentiation and the acquisition of typical T_H-17-associated features, such as expression of IL-17A, IL-17F, IL-23R and RORc.

Other factors that might explain the discrepancies among human T_H-17 studies[18–22] include the following: the cytokine combinations used, which do not always overlap[18,19,22]; the use of exogenous IL-2, which produces population expansion of differentiated T_H-17 cells[18,32]; and the use of monoclonal antibody blocking IL-4 and/or IFN-γ[18,33], two cytokines that inhibit mouse T_H-17 differentiation[4]. The differences might also be due to the serum added to the culture medium, which usually contains TGF-β and may also affect T_H-17 differentiation in a TGF-β-independent way. Because of the many experimental parameters that could potentially affect T_H-17 differentiation, we confirmed our basic findings in six different systems: $CD3^+CD4^+CD45RA^+$ peripheral blood naive T cells purified by positive or negative selection; total cord blood CD4 T cells; serum-containing medium; serum-containing medium and monoclonal antibody blocking TGF-β; serum-free medium; and serum-free medium and monoclonal antibody blocking IL-4 and IFN-γ. Although the absolute amount of IL-17 varied, the cytokine requirements to induce optimal IL-17 production in each of these experimental systems were similar.

Another important issue that remains controversial is the function of IL-23, a cytokine of the IL-12 family associated with T_H-17 responses[34]. In mice, IL-23 acts on IL-23R-expressing differentiated T_H-17 cells to induce their population expansion *in vitro*[11] and *in vivo*[35] but does not influence T_H-17 differentiation[10,11]. In humans, IL-23 is sufficient for T_H-17 differentiation[19]. Our data have confirmed an important function for IL-23 in human T_H-17 differentiation but only in synergy with TGF-β and proinflammatory cytokines, which were essential for inducing IL-23R expression.

Studies have shown that human and mouse T_H-17 cells can produce IL-22 (refs. 16,18). However, IL-22 is also produced by polyclonally stimulated naive $CD4^+$ cells[36] and T_H1 cells[37] and is inhibited by T_H-17-inducing conditions in memory $CD4^+$ T cells[20]. Our study has provided further evidence that IL-22 is not specific for T_H-17 cells and could have even higher production by T_H1 cells. This could explain the different functions of IL-17 and IL-22 in inflammatory responses[38] and autoimmune diseases[39].

In mice, IL-21 is dependent on IL-6 (refs. 12,23) and is produced in T_H-17 conditions but not in T_H1 or T_H2 conditions[13,15]. In contrast, we have shown that human IL-21 was produced in T_H1-polarizing conditions as well as T_H-17-polarizing conditions and that its production depended on IL-23. Given those results and our data on IL-23R expression, we can infer a sequence wherein TGF-β and proinflammatory cytokines induce IL-23R, which enables IL-23 to induce IL-21 production in a second step. The production of IL-21 during T_H-17 responses might enhance B cell immunity[40], which is involved in the physiopathology of autoimmune diseases such as lupus erythematosus and multiple sclerosis.

IL-10 is produced by mouse T_H-17 cells driven by TGF-β and IL-6 and confers regulatory functions on them[17]. Our results suggest that in humans, the presence of IL-1β in a T_H-17 environment could inhibit IL-10 production. During the resolution of inflammation, a decrease in or lack of IL-1β may simultaneously decrease the production of IL-17 and enhance the production of IL-10, which would further favor immune contraction through its anti-inflammatory properties[41].

Studies of T_H1 and T_H2 cells have shown that T helper subsets can produce a broader array of cytokines than initially described and that some T cell cytokines have limited specificity for a given T helper subset but are potentially associated with important functional properties, such as proinflammatory, for TNF[42], or regulatory, for IL-10 (refs. 43–45). Thus, focusing on a single T helper cell–derived cytokine, such as IFN-γ for T_H1 cells or IL-4 for T_H2 cells, gives only a partial view of a complex T helper cell response. Here we have shown that optimal T_H-17-polarizing conditions also drove the production of an array of cytokines, including IL-21, IL-22, IL-6, TNF and IFN-γ. Although each of these cytokines has different functions, they could collectively affect the global outcome of a T_H-17 response. Most notably, we have also shown that T_H-17 cell–derived cytokines were regulated in a specific way. We speculate that *in vivo*, priming of naive T cells might occur in optimal but also suboptimal polarizing conditions, depending on spatiotemporal factors. Each polarizing environment might induce a different T helper cytokine profile, contributing to the diversity and regulation of an immune response. In our *in vitro*

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

model, proinflammatory cytokines and IL-23 polarized CD4 T cells toward a T_H1-like profile, so they produced mainly IFN-γ, IL-21 and IL-22. The addition of TGF-β in such an environment, which could mimic the onset of TGF-β-producing regulatory T cells, induced a switch toward a typical T_H-17 profile. It has been reported that TGF-β 'antagonizes' T_H1 responses by inhibiting expression of IFN-γ as well as of T-bet[46,47]. We have shown that TGF-β might act in a more global way and was able to convert a T_H1-like profile into a T_H-17 response.

To our knowledge, this is the first study to analyze global T helper cell cytokine profiles with computational approaches. This could open new perspectives for the pharmacological modulation of T helper responses and could help elucidate and/or allow the prediction of the outcome of a specific therapeutic intervention. Applied to the T_H-17 cytokine profile, our data could help identify and target pathogenic components while preserving or enhancing protective aspects in the same T helper response.

METHODS

Purification of naive CD4$^+$ T lymphocytes from adult blood. Peripheral blood mononuclear cells were separated by Ficoll-Hypaque centrifugation (Amersham Biosciences) from buffy coats obtained from samples from healthy blood donors (Saint Antoine-Crozatier Blood Bank, Paris). CD4$^+$ T Lymphocytes were purified by immunomagnetic depletion with the human CD4$^+$ T Cell Isolation Kit II (Miltenyi Biotec), with the addition of biotinylated anti-CD45RO (C2400-67; USBiological). Naive CD4$^+$ T cells (CD3$^+$CD4$^+$CD45RA$^+$CD45RO$^-$) had a purity of over 96%, as shown by flow cytometry (**Supplementary Fig. 8** online). For some experiments, peripheral blood naive CD4$^+$ T cells were isolated with the CD4$^+$ T Cell Isolation Kit II (Miltenyi Biotec), followed by staining with allophycocyanin–anti-CD4 (VIT4; Miltenyi Biotec) and phycoerythrin–anti–CD45RA (PNIM1834; Immunotech) and cell sorting of double-positive cells (purity, over 99%) with a FACSAria (BD Bioscience). Human cord blood was obtained by an ethically approved convention (Necker Hospital, Paris), and total CD4$^+$ T cells were purified with the CD4$^+$ T Cell Isolation Kit II (Miltenyi Biotec).

T helper cell differentiation assay. Naive CD4$^+$ T cells were cultured in 48-well plates (Falcon) at a density of 8×10^4 cells per well in Yssel's medium (a gift from H. Yssel) containing 10% (vol/vol) FCS (Hyclone) or X-VIVO 15 serum-free medium (Lonza) in presence of Dynabeads CD3/CD28 T Cell Expander (one bead per cell; Invitrogen) and the following cytokines: IL-1β (10 ng/ml), IL-6 (20 ng/ml), TNF (10 ng/ml), TGF-β (1 ng/ml), IL-23 (100 ng/ml), IL-4 (25 ng/ml) and/or IL-12 (10 ng/ml; R&D Systems). For some experiments, anti-TGF-β (human LAP; 27235; R&D Systems), anti-IFN-γ (B27; BD Biosciences) and/or anti-IL-4 (34019; R&D Systems) were added to the cultures at a concentration of 10 μg/ml. For interference with RORc function, in some experiments lentiviral vectors were used that contained a plasmid encoding shRNA selected for its ability to suppress RORc mRNA expression (33658; Open Biosystems) or empty pLKO.1 vector (Open Biosystems), generated as described[48]. After 5–6 d, cells were collected and washed extensively and their viability was determined by trypan blue exclusion. Cells (1×10^6 cells/ml) were restimulated for 6 h (for flow cytometry intracellular staining) or for 24 h (for ELISA and RT-PCR) with Dynabeads CD3/CD28 T Cell Expander (one bead per cell). For shRNA experiments, naive T cells were cultured in serum-free medium at a density of 1×10^5 cells per well in 96-well round-bottomed plates and were infected with various concentrations of lentiviral vector expressing shRNA or with empty vector (negative control) for the first day of T_H-17 differentiation. Cells were washed extensively and were cultured for an additional 5 d in T_H0 or T_H-17 conditions. IL-17A and IL-17F transcripts and IL-17 protein were analyzed after 24 h of restimulation with anti-CD3 and anti-CD28.

Analysis of cytokine production. Cytokines in culture supernatants were measured by IL-17 ELISA (R&D Systems), IL-21 ELISA (eBioscience) or IL-22 ELISA (Antigenix) or with IL-4, IL-5, IL-6, IL-10, IL-13, IFN-γ or TNF cytometric bead assay Flex Sets (BD Bioscience) according to the manufacturer's instructions. Cells producing IFN-γ and IL-17 were analyzed by intracellular cytokine staining after the addition of brefeldin (10 μg/ml) during the final 3 h of restimulation. Cells were made permeable with Cytofix/Cytoperm reagents (BD Biosciences). Cells were stained with fluorescein isothiocyanate–conjugated anti-IFN-γ (4S.B3; BD Pharmingen) and phycoerythrin-conjugated anti-IL-17 (eBio 64DEC17; eBioscience) and washed and then were analyzed by flow cytometry (FACScan; Becton Dickinson).

Real-time quantitative RT-PCR. Total RNA was extracted with an RNeasy Micro kit (Qiagen). A mixture containing random hexamers, oligo(dT)$_{15}$ (Promega) and SuperScript II Reverse Transcriptase (Invitrogen) was used for cDNA synthesis. Transcripts were quantified by real-time quantitative PCR on an ABI PRISM 7900 sequence detector (Applied Biosystems) with Applied Biosystems predesigned TaqMan Gene Expression Assays and Absolute QPCR ROX mix (Thermo Fisher Scientific). The following probes were used (Applied Biosystems assay identification numbers in parentheses): IL-17A (Hs00174383_m1), IL-17F (Hs00369400_m1), RORc (Hs01076112_m1), IL-23R (Hs00332759_m1), T-bet (Hs00203436_m1), GATA-3 (Hs00231122), TNF (Hs 00174128_m1) and IFN-γ (Hs00174143_m1). For each sample, mRNA abundance was normalized to the amount of ribosomal protein L34 (Hs00241560_m1).

Statistical analysis. A nonparametric two-tailed Wilcoxon test was used for pairwise comparisons of cytokines. *P* values of 0.05 or less were considered statistically significant. The Pearson correlation coefficient was used to assess the significance of correlation among IL-17A, IL-17F mRNA or IL-17 protein and RORc. Data for the clustering and principal component analysis (**Supplementary Methods** online) were corrected for the 'donor effect' through the application of a linear model. For information summaries, replicates were aggregated in each condition to their barycentric value for each cytokine and the principal component analysis was computed with these variables. The Pearson correlation distance and the Ward's criteria as an agglomerative method were used for hierarchical clustering analysis.

Accession code. UCSD-Nature Signaling Gateway (http://www.signaling-gateway.org): A002271.

Note: Supplementary information is available on the Nature Immunology website.

ACKNOWLEDGMENTS

We thank O. Lantz, S. Denépoux, F. Barrat, C. Théry, P. Benaroch, I. Fernandez, Z. Maciorowsky and H. Kitamura for suggestions and critical reading of the manuscript; and Z. Maciorowsky, C. Guérin and A. Viguier for cell sorting. Yssel's medium was a gift from H. Yssel (Institut National de la Santé et de la Recherche Médicale). Supported by the European Community Sixth Framework Programme (Marie Curie Excellence Grant 014162).

AUTHOR CONTRIBUTIONS

E.V. did experiments and drafted the manuscript; N.S. did computational and statistical analysis; R.Z. did quantitative RT-PCR analysis and helped with the computational data analysis; S.I.B. did some experiments; P.H. did computational and statistical analysis; E.B. supervised the computational and statistical analysis; and V.S. designed and supervised the study and wrote the manuscript.

Published online at http://www.nature.com/natureimmunology
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Mosmann, T.R. & Coffman, R.L. TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties. *Annu. Rev. Immunol.* **7**, 145–173 (1989).
2. O'Garra, A. Cytokines induce the development of functionally heterogeneous T helper cell subsets. *Immunity* **8**, 275–283 (1998).
3. Park, H. *et al.* A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nat. Immunol.* **6**, 1133–1141 (2005).
4. Harrington, L.E. *et al.* Interleukin 17–producing CD4$^+$ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nat. Immunol.* **6**, 1123–1132 (2005).
5. Aujla, S.J., Dubin, P.J. & Kolls, J.K. Th17 cells and mucosal host defense. *Semin. Immunol.* **19**, 377–382 (2007).
6. LeibundGut-Landmann, S. *et al.* Syk- and CARD9-dependent coupling of innate immunity to the induction of T helper cells that produce interleukin 17. *Nat. Immunol.* **8**, 630–638 (2007).

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

7. Murphy, C.A. *et al.* Divergent pro- and antiinflammatory roles for IL-23 and IL-12 in joint autoimmune inflammation. *J. Exp. Med.* **198**, 1951–1957 (2003).
8. Bettelli, E. *et al.* Reciprocal developmental pathways for the generation of pathogenic effector T_H17 and regulatory T cells. *Nature* **441**, 235–238 (2006).
9. Langrish, C.L. *et al.* IL-23 drives a pathogenic T cell population that induces autoimmune inflammation. *J. Exp. Med.* **201**, 233–240 (2005).
10. Mangan, P.R. *et al.* Transforming growth factor-β induces development of the T_H17 lineage. *Nature* **441**, 231–234 (2006).
11. Veldhoen, M., Hocking, R.J., Atkins, C.J., Locksley, R.M. & Stockinger, B. TGFβ in the context of an inflammatory cytokine milieu supports de novo differentiation of IL-17-producing T cells. *Immunity* **24**, 179–189 (2006).
12. Zhou, L. *et al.* IL-6 programs T_H-17 cell differentiation by promoting sequential engagement of the IL-21 and IL-23 pathways. *Nat. Immunol.* **8**, 967–974 (2007).
13. Nurieva, R. *et al.* Essential autocrine regulation by IL-21 in the generation of inflammatory T cells. *Nature* **448**, 480–483 (2007).
14. Korn, T. *et al.* IL-21 initiates an alternative pathway to induce proinflammatory T_H17 cells. *Nature* **448**, 484–487 (2007).
15. Wei, L., Laurence, A., Elias, K.M. & O'Shea, J.J. IL-21 is produced by Th17 cells and drives IL-17 production in a STAT3-dependent manner. *J. Biol. Chem.* **282**, 34605–34610 (2007).
16. Liang, S.C. *et al.* Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides. *J. Exp. Med.* **203**, 2271–2279 (2006).
17. McGeachy, M.J. *et al.* TGF-β and IL-6 drive the production of IL-17 and IL-10 by T cells and restrain T_H-17 cell–mediated pathology. *Nat. Immunol.* **8**, 1390–1397 (2007).
18. Acosta-Rodriguez, E.V., Napolitani, G., Lanzavecchia, A. & Sallusto, F. Interleukins 1β and 6 but not transforming growth factor-β are essential for the differentiation of interleukin 17–producing human T helper cells. *Nat. Immunol.* **8**, 942–949 (2007).
19. Wilson, N.J. *et al.* Development, cytokine profile and function of human interleukin 17–producing helper T cells. *Nat. Immunol.* **8**, 950–957 (2007).
20. Chen, Z., Tato, C.M., Muul, L., Laurence, A. & O'Shea, J.J. Distinct regulation of interleukin-17 in human T helper lymphocytes. *Arthritis Rheum.* **56**, 2936–2946 (2007).
21. Evans, H.G., Suddason, T., Jackson, I., Taams, L.S. & Lord, G.M. Optimal induction of T helper 17 cells in humans requires T cell receptor ligation in the context of Toll-like receptor-activated monocytes. *Proc. Natl. Acad. Sci. USA* **104**, 17034–17039 (2007).
22. van Beelen, A.J. *et al.* Stimulation of the intracellular bacterial sensor NOD2 programs dendritic cells to promote interleukin-17 production in human memory T cells. *Immunity* **27**, 660–669 (2007).
23. Laurence, A. & O'Shea, J.J. T_H-17 differentiation: of mice and men. *Nat. Immunol.* **8**, 903–905 (2007).
24. Weaver, C.T., Hatton, R.D., Mangan, P.R. & Harrington, L.E. IL-17 family cytokines and the expanding diversity of effector T cell lineages. *Annu. Rev. Immunol.* **25**, 821–852 (2007).
25. Wang, Y.H. *et al.* IL-25 augments type 2 immune responses by enhancing the expansion and functions of TSLP-DC-activated Th2 memory cells. *J. Exp. Med.* **204**, 1837–1847 (2007).
26. Ivanov, I.I. *et al.* The orphan nuclear receptor RORγt directs the differentiation program of proinflammatory IL-17$^+$ T helper cells. *Cell* **126**, 1121–1133 (2006).
27. Ivanov, I.I., Zhou, L. & Littman, D.R. Transcriptional regulation of Th17 cell differentiation. *Semin. Immunol.* **19**, 409–417 (2007).
28. Ringner, M. What is principal component analysis? *Nat. Biotechnol.* **26**, 303–304 (2008).
29. Zaba, L.C. *et al.* Amelioration of epidermal hyperplasia by TNF inhibition is associated with reduced Th17 responses. *J. Exp. Med.* **204**, 3183–3194 (2007).
30. Yen, D. *et al.* IL-23 is essential for T cell-mediated colitis and promotes inflammation via IL-17 and IL-6. *J. Clin. Invest.* **116**, 1310–1316 (2006).
31. Khader, S.A. *et al.* IL-23 and IL-17 in the establishment of protective pulmonary CD4$^+$ T cell responses after vaccination and during *Mycobacterium tuberculosis* challenge. *Nat. Immunol.* **8**, 369–377 (2007).
32. Amadi-Obi, A. *et al.* T_H17 cells contribute to uveitis and scleritis and are expanded by IL-2 and inhibited by IL-27/STAT1. *Nat. Med.* **13**, 711–718 (2007).
33. Chen, Z. & O'Shea, J.J. Th17 cells: a new fate for differentiating helper T cells. *Immunol. Res.* published online 3 January 2008 (doi:10.1007/s12026-007-8014-9).
34. Kastelein, R.A., Hunter, C.A. & Cua, D.J. Discovery and biology of IL-23 and IL-27: related but functionally distinct regulators of inflammation. *Annu. Rev. Immunol.* **25**, 221–242 (2007).
35. Veldhoen, M., Hocking, R.J., Flavell, R.A. & Stockinger, B. Signals mediated by transforming growth factor-β initiate autoimmune encephalomyelitis, but chronic inflammation is needed to sustain disease. *Nat. Immunol.* **7**, 1151–1156 (2006).
36. Xie, M.H. *et al.* Interleukin (IL)-22, a novel human cytokine that signals through the interferon receptor-related proteins CRF2–4 and IL-22R. *J. Biol. Chem.* **275**, 31335–31339 (2000).
37. Gurney, A.L. IL-22, a Th1 cytokine that targets the pancreas and select other peripheral tissues. *Int. Immunopharmacol.* **4**, 669–677 (2004).
38. Zenewicz, L.A. *et al.* Interleukin-22 but not interleukin-17 provides protection to hepatocytes during acute liver inflammation. *Immunity* **27**, 647–659 (2007).
39. Kreymborg, K. *et al.* IL-22 is expressed by Th17 cells in an IL-23-dependent fashion, but not required for the development of autoimmune encephalomyelitis. *J. Immunol.* **179**, 8098–8104 (2007).
40. Leonard, W.J. & Spolski, R. Interleukin-21: a modulator of lymphoid proliferation, apoptosis and differentiation. *Nat. Rev. Immunol.* **5**, 688–698 (2005).
41. Moore, K.W., de Waal Malefyt, R., Coffman, R.L. & O'Garra, A. Interleukin-10 and the interleukin-10 receptor. *Annu. Rev. Immunol.* **19**, 683–765 (2001).
42. Ito, T. *et al.* TSLP-activated dendritic cells induce an inflammatory T helper type 2 cell response through OX40 ligand. *J. Exp. Med.* **202**, 1213–1223 (2005).
43. Trinchieri, G. Interleukin-10 production by effector T cells: Th1 cells show self control. *J. Exp. Med.* **204**, 239–243 (2007).
44. Jankovic, D. & Trinchieri, G. IL-10 or not IL-10: that is the question. *Nat. Immunol.* **8**, 1281–1283 (2007).
45. O'Garra, A. & Vieira, P. T_H1 cells control themselves by producing interleukin-10. *Nat. Rev. Immunol.* **7**, 425–428 (2007).
46. Lin, J.T., Martin, S.L., Xia, L. & Gorham, J.D. TGF-β1 uses distinct mechanisms to inhibit IFN-γ expression in CD4$^+$ T cells at priming and at recall: differential involvement of Stat4 and T-bet. *J. Immunol.* **174**, 5950–5958 (2005).
47. Gorelik, L., Constant, S. & Flavell, R.A. Mechanism of transforming growth factor β-induced inhibition of T helper type 1 differentiation. *J. Exp. Med.* **195**, 1499–1505 (2002).
48. Manel, N., Unutmaz, D. & Littman, D.R. Human T_H-17 differentiation requires transforming growth factor-β and induction of the nuclear receptor RORγt. *Nat. Immunol.* (in the press).

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

Oncogene (2007) 26, 6641–6652
© 2007 Nature Publishing Group All rights reserved 0950-9232/07 $30.00
www.nature.com/onc

ORIGINAL ARTICLE

ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors

P Hupé[1,2,3], P La Rosa[1,3], S Liva[1], S Lair[1], N Servant[1] and E Barillot[1]

[1]*Institut Curie, Service Bioinformatique, Paris, France and* [2]*Institut Curie, CNRS UMR 144, Paris, France*

In recent years, an increasing number of projects have investigated tumor genome structure, using microarray-based techniques like array comparative genomic hybridization (array-CGH) or single nucleotide polymorphism (SNP) arrays. The forthcoming studies have to integrate these former results and compare their findings to the existing sets of copy number data for validation. These sets also form the basis from which many comparative retrospective analyses can be carried out. Nevertheless, exploitation of this mass of data relies on a homogeneous preparation of copy number data, which will make it possible to compare them together, and their integration into a unified bioinformatics environment with ad hoc analysis tools and interfaces. To our knowledge, no such data integration has been proposed yet. Therefore the biologists and clinicians involved in cancer research urgently need such an integrative tool, which motivated us to undertake the construction of a database for array-CGH and other DNA copy number data for tumors (ACTuDB). When available, the associated clinical, transcriptome and loss of heterozygosity data were also integrated into ACTuDB. ACTuDB contains currently about 1500 genomic profiles for tumors and cell lines for the bladder, brain, breast, colon, liver, lymphoma, neuroblastoma, mouth and pancreas, together with data for replication timing experiments. The CGH array data were processed, using ad hoc algorithms (probe mapping, breakpoint detection, gain or loss status assignment and visualization) developed at Institut Curie. The database is available from http://bioinfo.curie.fr/actudb/ and can be browsed with a user-friendly interface. This database will be a useful resource for the genomic profiling of tumors, a field of highly active research. We invite research groups involved in tumor genome profiling to submit their data to ACTuDB.
Oncogene (2007) **26,** 6641–6652; doi:10.1038/sj.onc.1210488; published online 14 May 2007

Keywords: DNA copy number; database; tumors; bioinformatics platform; molecular profiles; clinical data

Correspondence: P Hupé, Service Bioinformatique, Institut Curie, 26 rue d'Ulm, Paris 75248 cedex 05, France.
E-mail: Philippe.Hupe@curie.fr or actudb@curie.fr
[3]These authors contributed equally to this work.
Received 7 December 2006; revised 13 March 2007; accepted 14 March 2007; published online 14 May 2007

Introduction

Genome alterations are a hallmark of cancer (Albertson *et al*., 2003; Pinkel and Albertson, 2005). Several microarray-based techniques can be used to identify copy number changes in the genome at an unprecedented high resolution, from the megabase range down to a few tens of kilobases. These techniques include array comparative genomic hybridization (array-CGH), the use of cDNA arrays or oligonucleotide arrays and single nucleotide polymorphism (SNP) arrays (see Ylstra *et al*., 2006 for a review of the various platforms). The identification of genome alterations is useful in several ways. First, the characterization of these copy number changes can provide insight into tumor progression mechanisms. Second, this method can be used to identify genes involved in tumor progression: tumor suppressor genes and oncogenes are thought to be located in regions of loss and gain, respectively. Finally, variation in genomic alterations could be used to classify tumors molecularly, facilitating the diagnosis of new patients and assessments of their prognosis.

Many studies have been carried out on bladder cancer (Veltman *et al*., 2003; Blaveri *et al*., 2005; Stransky *et al*., 2006), brain cancer (Bredel *et al*., 2005; Kotliarov *et al*., 2006), breast cancer (Pollack *et al*., 2002; Fridlyand *et al*., 2006), colon cancer (Douglas *et al*., 2004; Nakao *et al*., 2004), liver cancer (Patil *et al*., 2005), lymphoma (de Leeuw *et al*., 2004), neuroblastoma (Janoueix-Lerosey *et al*., 2005; Mosse *et al*., 2005), mouth cancer (Snijders *et al*., 2005), pancreas cancer (Gysin *et al*., 2005) and replication timing (Woodfine *et al*., 2004; Janoueix-Lerosey *et al*., 2005). Comparisons of the results of experiments from different laboratories, on different types of cancer, are required to validate results or hypotheses and to improve our understanding of the recurrent alterations involved in cancer. To our knowledge, no database allows today comparison in a rigorous way, that is no database have solved the question yet of how to make heterogeneous array copy number data comparable. We have, therefore, defined and carried out a protocol for homogeneous pretreatment of DNA copy number data and integration into ACTuDB (Array CGH Tumor DataBase) – a database compiling published array-CGH datasets that can be used for the browsing, visualization and analysis of tumor profiles via a user-friendly interface. In the Results section, the user interface, several analysis

scenarios and meta-analyses are described. We then describe in the Materials and Methods section the various datasets available in ACTuDB and present how these data have been analysed with our algorithms to allow direct comparisons.

Results

Access and data analysis

Data browsing in ACTuDB is based on VAMP software (La Rosa *et al*., 2006). VAMP is a graphical user interface for the visualization and analysis of array-CGH, transcriptome and other molecular profiles. We describe below the way in which data are organized, visualized and queried.

Data organization. Datasets are stored as 'projects' under the name of the first author and the year of publication. Within a project, the user can access the data either in 'Chromosome' mode (only the data for one particular chromosome are loaded) or in 'Genomic' mode (the profiles for all the concatenated chromosomes are loaded). The database loads and compares data from different projects, possibly corresponding to different array technologies or with different array designs.

Data visualization. VAMP software offers several possibilities for visualization, for example, the classical CGH karyotype view (see Figure 1) and genome-wide multitumor comparison views (see Figures 2–5) are available, facilitating the comparison of different arrays. Additional information for each clone or DNA region can be retrieved interactively from various public databases, through external links (NCBI, UCSC, Ensembl). We advise the reader to refer to the VAMP documentation available at the following URL: http://bioinfo.curie.fr/vamp/doc.

Transverse analysis of array-CGH experiments

Users can easily carry out transverse analyses for a set of tumors from one or several projects. We present here three scenarios of potential interest to biologists.

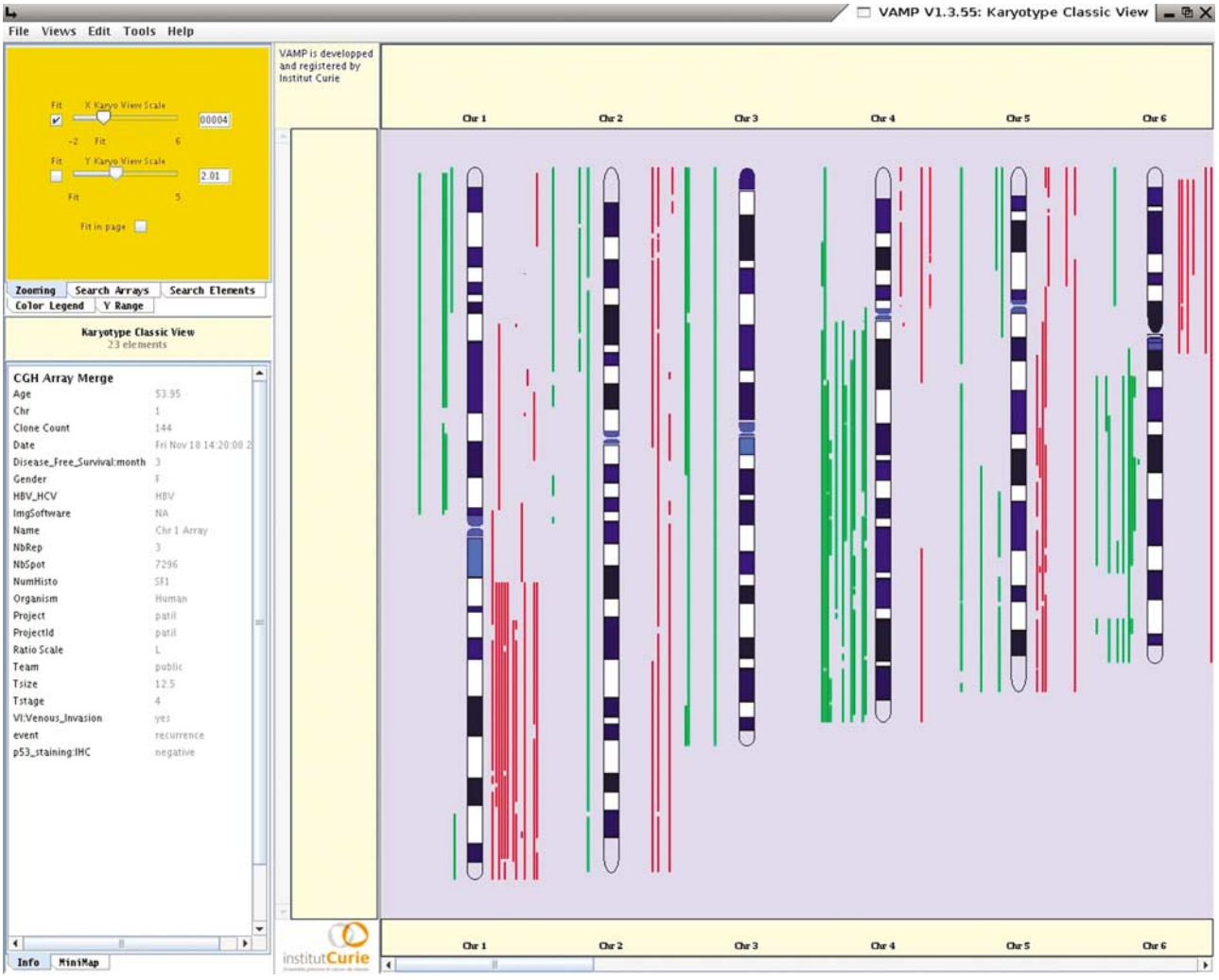


Figure 1 The classical karyotype view (dataset from Patil *et al*., 2005).

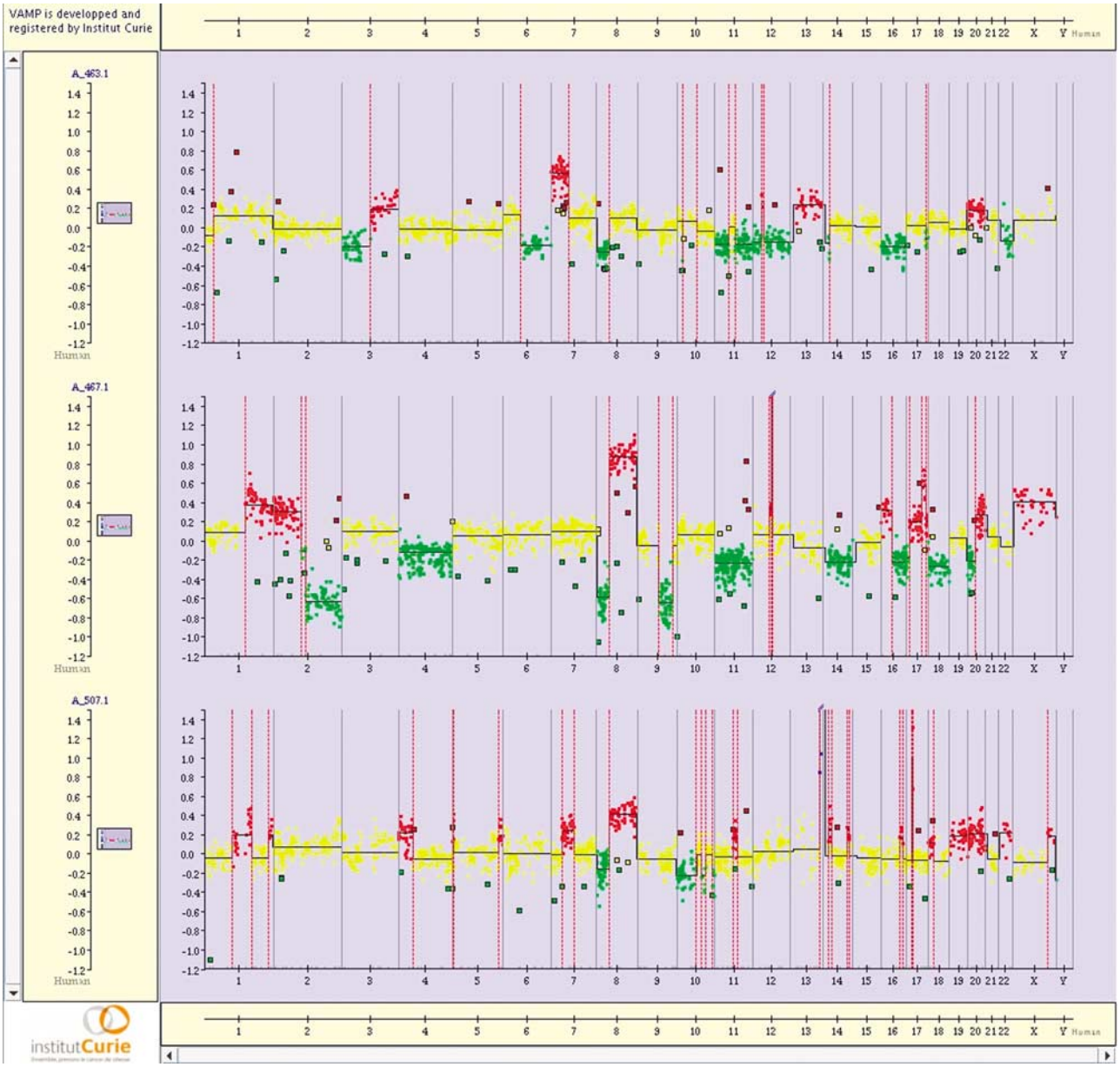


Figure 2 Genome-wide multitumor comparison views. The results of the GLAD algorithm can also be displayed: the black line corresponds to the smoothing value, breakpoints are shown as vertical red dashed lines and outliers are circled in black (dataset from Blaveri *et al*., 2005).

Finding informative genomic regions. The identification of systemic alterations within a set of tumors is central to the analysis of CGH array data for two reasons. First, it can pinpoint new candidate genes, as tumor suppressor genes and oncogenes are thought to be present in regions of loss and gain, respectively. Second, some alterations may be significantly correlated with clinical phenotype and may therefore be useful for diagnosis and prognosis. The simplest way to identify informative regions is to work at the probe level. For each probe, the fraction of tumors with gains and losses over the dataset is displayed in the FrAGL view (see Figure 6). Instead of looking for individual probes carrying genome alterations, it is often fruitful to consider the geography of the genome and to look for whole regions. Rouveirol *et al*. (2006) have described algorithms for this analysis (see Figure 4), and defined two categories of regions, included in ACTuDB:

- *Minimal regions* of gain (or loss) correspond to the intersection for all tumors of the gained (or lost) regions. These regions are minimal, in that no breakpoint evidence is available to narrow the region further.
- *Recurrent regions* of gain (or loss) are defined as regions gained (or lost), with the same extremities (breakpoint positions), in a sufficient number of tumors.

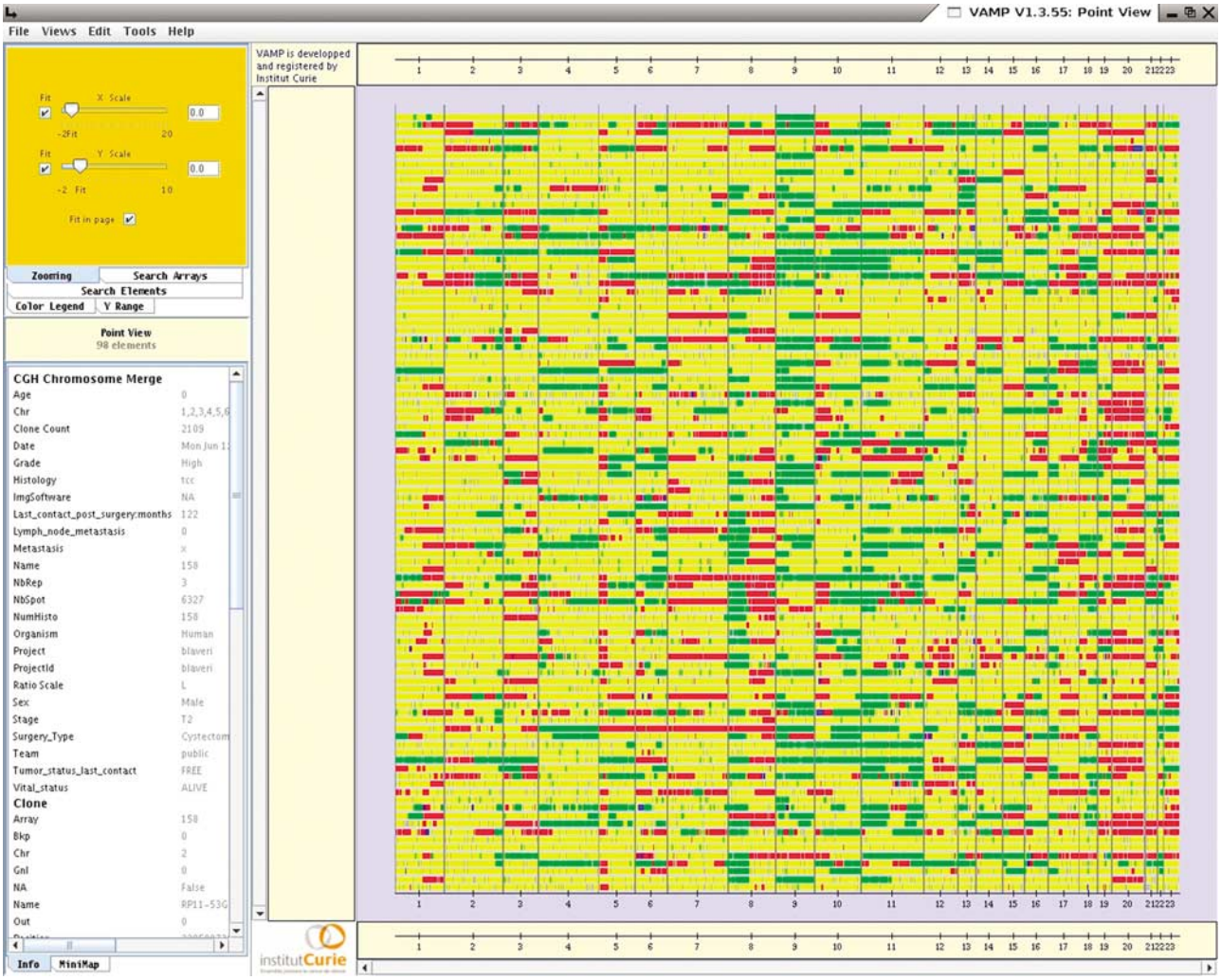


Figure 3 Dotplot view (dataset from Blaveri *et al*., 2005).

Identification of new subgroups of tumors. Class discovery is a central objective of genomic studies. ACTuDB facilitates such analyses, through hierarchical clustering: different object and cluster distances (Euclidean, Pearson correlation, Manhattan; Ward, Single linkage, Complete linkage, Group Average) for log-ratios, status or smoothing values (the smoothing value of a region is the signal statistically inferred from the signals for all its probes by the GLAD algorithm; the status of a region is loss, normal, gain or amplicon, and is also assessed by GLAD). It is also possible to cluster the data based on the minimal/recurrent regions of alteration described above or on user-defined regions. The use of region status rather than information for all probes has two advantages. It eliminates redundancy between contiguous probes and gives the same weighting to each region, which is important because very small regions (such as amplicons) may be highly relevant for diagnosis/prognosis. This approach also makes it easy to provide a biological interpretation for clustering, as it is generally based on less than a hundred regions rather than thousands of probes. It should be noted that if users wish to cluster data from different projects with different designs, the only way to do this is to make use of regions.

Comparison of array-CGH and transcriptome data. Such comparisons are possible when the available transcriptome profiles have been integrated into the database. This is the case for the dataset of Pollack *et al.* (2002), Patil *et al.* (2005) and Stransky *et al.* (2006). The user can investigate whether gene expression displays particular alterations. Typically, the biologist will search for genes overexpressed in amplified regions or underexpressed in lost regions (see Figure 7).

Comparison of array-CGH and loss of heterozygosity data. In the dataset of Kotliarov *et al.* (2006), both the copy number and loss of heterozygosity (LOH) have been studied with the Affymetrix GeneChip® Human Mapping 100K SetS. In Figure 8, displayed are the DNA copy number profiles for chromosomes 10 and 13, and corresponding LOH profiles for sample HF0505

Figure 4 Example of informative genomic regions for 37 colorectal cancers from Douglas *et al*. (2004). Genome alterations already reported by Douglas *et al*. (2004) were identified by our software (alterations are represented by vertical bandings ranging from dark to light pink for gain regions, dark to light green for loss regions and blue for amplified regions – amplicons are arbitrarily defined here as regions with $\log_2$-ratio > 2). These alterations include gain of chromosomes 20, 13, 8q and loss of 8p, 18q or 17q11.2-q12 amplification as reported by Douglas *et al*. (2004), gain of chromosome 7p and loss of chromosomes 4, 14q, 15q, 17p, 21q and 22q. The left panel shows the alterations identified on chromosome 8 with respect to cytogenetic banding.

Xba. For chromosome 10 the DNA copy number is diploid, whereas the LOH profile gives a strong evidence of loss of heterozigosity; this leads to the conclusion that chromosome 10 is isodisomic. For chromosome 13, LOH profile confirms the loss region identified on the DNA copy number profile.

Meta-analysis examples

We present here two meta-analyses examples on colon and bladder cancer. The detailed guidelines to perform these analyses are available as Supplementary Information.

Comparison of the frequency of alterations in two colon datasets. For each probe, the fraction of tumors with gains and losses over the Nakao *et al*. (2004) dataset and the Douglas *et al*. (2004) dataset has been computed separately and displayed in a FrAGL view (see Supplementary Information for details – colon-meta-analysis.pdf file). The results show similar pattern between the two datasets: similar frequency are observed for gains of chromosomes 20 (>65%), 8q (~40%), losses of chromosomes 8p (>40%) and 18q (>60%). The chromosome 13 gain tends to be more frequent in Douglas *et al*. (2004) (~60%) than in Nakao *et al*. (2004) (~35%).

Amplicon in bladder cancer. The minimal amplified regions have been identified for the three bladder cancer datasets (Veltman *et al*., 2003; Blaveri *et al*., 2005; Stransky *et al*., 2006) (see Supplementary Information for details – bladder-meta-analysis.pdf file). Amplicons are located at 6p22, 8q22-q23, 11q13 and are present in the three datasets. The gene list within the regions is

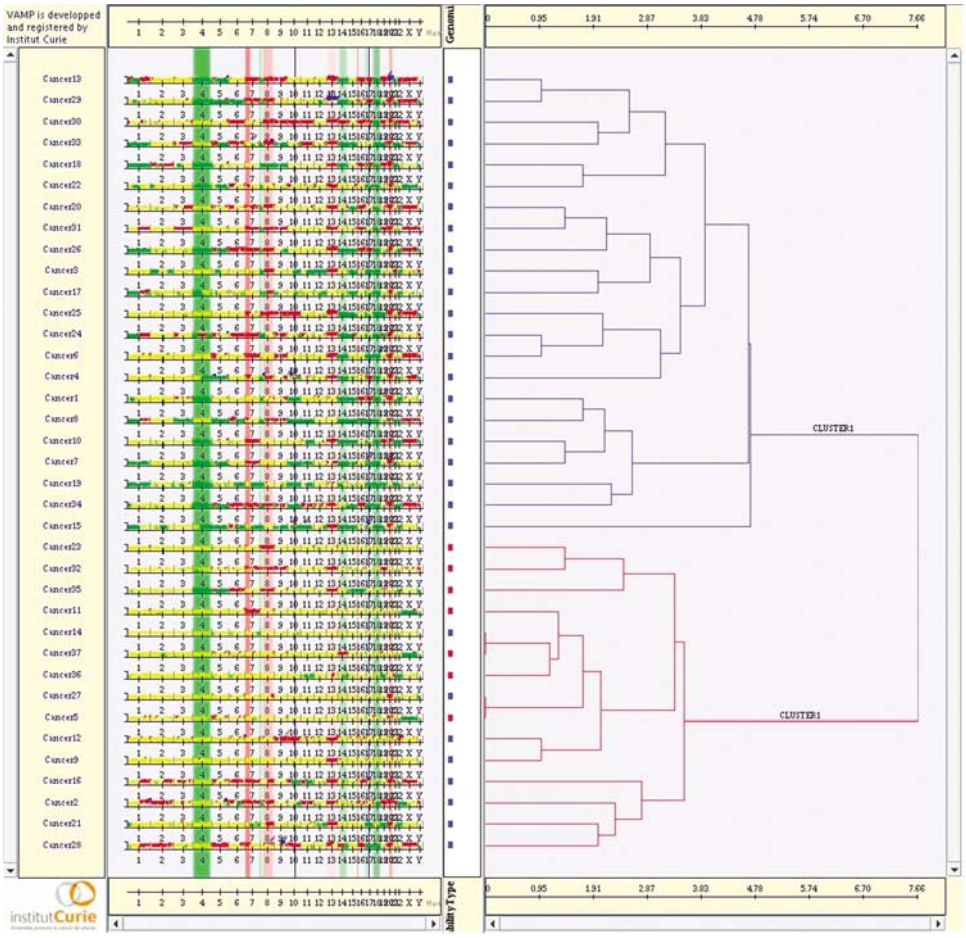


Figure 5 Clustering results based on recurrent alterations (as identified in Figure 4) and comparison with clinical data (CIN + in blue, MSI + in red) for the dataset from Douglas *et al*. (2004).

given as Supplementary Information (gene-list.pdf file); among them we can mention some of them involved in human tumors: E2F3, CDKAL1, SOX4 on chromosome 6, COX6C on chromosome 8 and CCND1 on chromosome 11. We then focus on the chromosome 6 analysis using only the Stransky *et al*. (2006) dataset to compare the genome alterations with respect to gene expression level. E2F3 gene expression was found to be highly correlated with DNA copy number level (correlation = 0.88). This correlation was already reported by Hurst *et al*. (2004).

Discussion

ACTuDB is a database that compiles array-CGH profiles and clinical data for tumors and can be browsed using VAMP software. All data in ACTuDB have been homogeneously pretreated as microarray probe mapping, breakpoint detection and gain/loss assignment. When available, expression data can also be compared with the genomic profiles. Transverse analysis can be carried out, by searching for minimal/recurrent alterations and compare their frequency between different projects or identifying subgroups of tumors with clustering techniques. Informative regions can be compared with genome annotations, such as gene position, to pinpoint candidate genes or with replication timing experiments. ACTuDB provides a user-friendly interface with more advanced analysis algorithms than existing databases (Baudis, 2006). It also offers the possibility of adding other types of genome annotations potentially relevant for biologists. The database currently contains BAC, cDNA and oligonucleotide array data but can accept any kind of profile related to DNA copy number obtained with different platforms. In addition, any kind of molecular profile can be incorporated into ACTuDB, provided that data are available for genome position and there is a quantitative value to plot. Typical molecular profiles that can be integrated into the database include DNA copy number data, transcriptome data, ChIP-on-chip (Chromatine Immuno Precipitation) data and LOH data. We encourage our colleagues to submit their datasets to ACTuDB upon publication. They should contact us at actudb@curie.fr and follow the guidelines provided at http://bioinfo.curie.fr/actudb/.

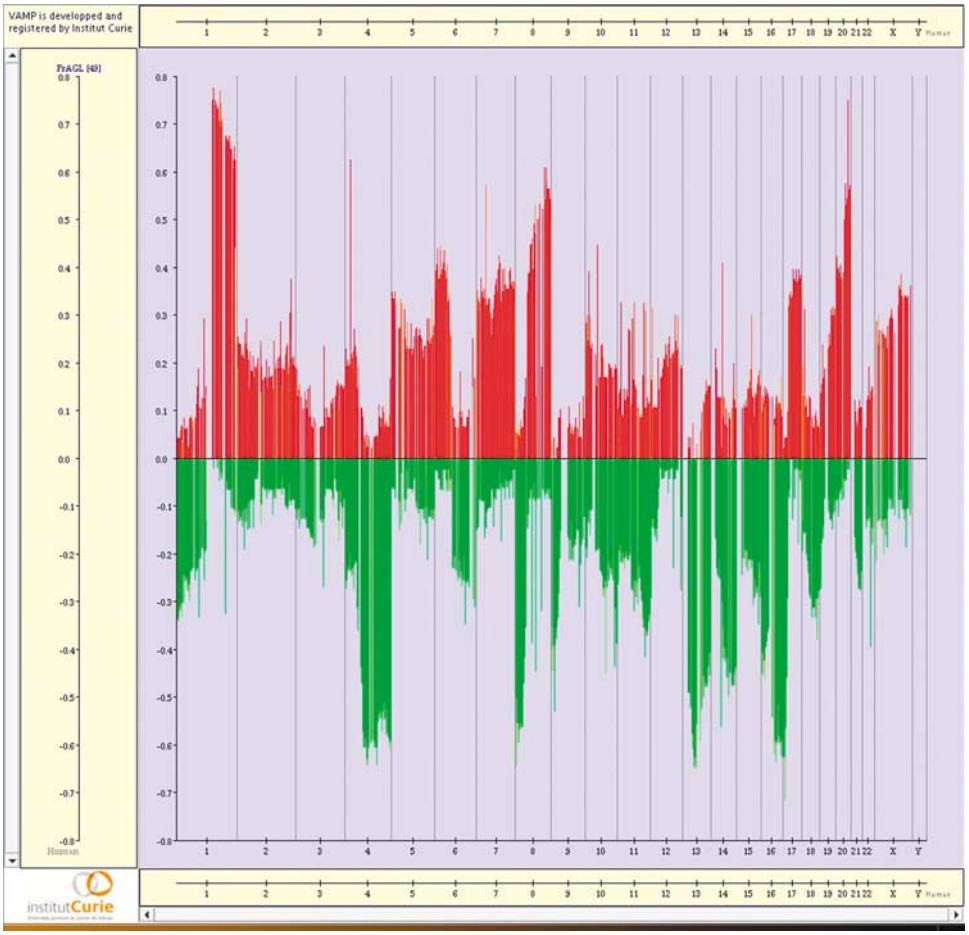


Figure 6 FrAGL (Frequency of Amplicon, Gain and Loss) view. The values correspond to the percentage of gained and lost clones identified with GLAD over the whole dataset from Patil *et al*. (2005). Recurrent genome alterations, such as 1q, 6p 8q, 20q gains and 4q, 8p, 13q 16q, 17p losses, can be clearly identified on this plot, as reported by the authors of the original publication.

Materials and methods

Data content

The data integrated into ACTuDB have been collected from available published data for the analysis of tumor genomic profiles by BAC, cDNA array-CGH or SNP arrays. The datasets are summarized in Table 1 and are presented more in detail in what follows.

Array-CGH datasets for cancers. The datasets in ACTuDB consist mostly of DNA copy number profiles for tumors of various origins as listed below (bladder, brain, breast, colon, liver, lymphoma, mouth, neuroblastoma and pancreas). Associated clinical and gene expression data have also been included in the database, when available.

- *Bladder cancer*: copy number changes were studied in bladder tumors at different stages (Veltman *et al*., 2003), and small regions with high levels of amplification or homozygous deletion were identified. In a study of 98 tumors, Blaveri *et al*. (2005) confirmed the alterations reported by Veltman *et al*. (2003) and identified additional alterations. Various statistical analyses demonstrated that copy number variation between pairs of known oncogenes and tumor suppressor genes was associated with changes in pathways known to be involved in bladder cancer (retinoblastoma and p53-MDM2 pathways). Stransky *et al*. (2006) proposed an approach for identifying candidate regions controlled by epigenetic mechanisms in cancer and for the characterization of one of these regions combining DNA copy number data and transcriptome data.
- *Brain cancer*: Bredel *et al*. (2005) carried out copy number profiling for 54 gliomas differing in histogenesis and tumor grade, using a 42 000-clone cDNA array. The analysis led to the delimitation of the precise (gene-specific) boundaries of known and new chromosomal alterations. Genes involved in gliomagenesis were identified. A subset of these genes was shown to be associated with the genetic subgroups of glial brain tumors (astrocytic or oligodendrocytic phenotype). Kotliarov *et al*. (2006) identified in 178 gliomas novel regions of copy number alteration and LOH using the Affymetrix GeneChip Human Mapping 100K SetS.
- *Breast cancer*: Pollack *et al*. (2002) profiled DNA copy number alterations of 44 primary breast tumors and 10 cell lines, using a cDNA array also used for transcriptome profiling of the same samples. This study was original in the direct comparison of DNA copy number changes and gene expression for a subset of four breast cancer cell lines and 37 tumors. Fridlyand *et al*. (2006) identified three subtypes of tumor from CGH array data for 67 breast cancer samples.
- *Colon cancer*: Nakao *et al*. (2004) identified many small, previously uncharacterized genomic regions in 125 primary

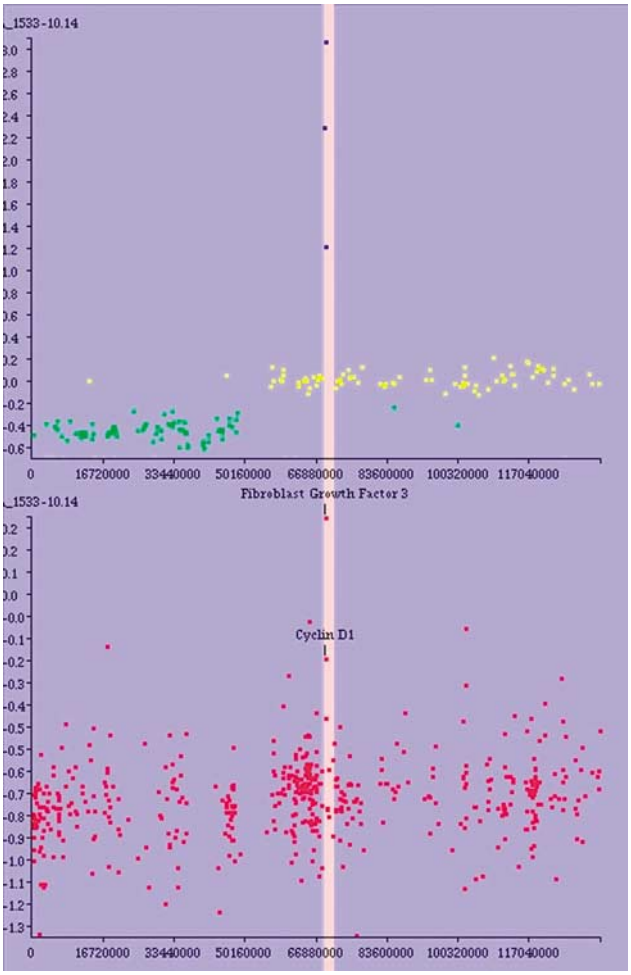


Figure 7 Array-CGH (top profile) and expression ratio (second profile in descending order) of the same human tumors. The expression ratio was computed from Affymetrix U95 array of a bladder tumor sample and of a reference sample with no alteration on chromosome 11 (Stransky *et al*., 2006). This confrontation pinpoints the probable implication of the gene cyclin D1 and fibroblast growth factor 3 in this tumor. The second profile is the ratio of the tumoral transcriptome array to the reference transcriptome profile.

tumors. The frequency of alteration was compared with clinical phenotypes: tumor stage and location, and the patient's age and sex were found to have little effect, whereas microsatellite instability had a significant effect. Douglas *et al*. (2004) reported genomic alterations in 48 cell lines and 37 tumors and that samples displaying chromosomal instability (CIN) presented a larger number of alterations than samples displaying microsatellite instability (MSI).

- *Liver cancer*: the study by Patil *et al*. (2005) aimed to identify chromosomal aberrations in 49 hepatocellular carcinomas. Like Pollack *et al*. (2002), Patil *et al*. (2005) compared genome alterations with gene expression quantified in a previous study (Chen *et al*., 2002). They found that high levels of Jab1 expression were significantly correlated with DNA copy number gain at 8q.
- *Lymphoma*: only one of the papers contributing data to ACTuDB dealt with non-solid tumors. Cell lines from patients with mantle cell lymphoma (MCL), an aggressive non-Hodgkin's lymphoma, were studied by de Leeuw *et al*. (2004), with the aim of identifying secondary genomic alterations concomitant with the translocation t(11;14) and determining whether the eight cell lines tested, all of which are widely used as models, adequately represented MCL.
- *Mouth cancer*: Snijders *et al*. (2005) analysed 89 oral squamous cell carcinomas with the aim of defining minimum common amplified regions. They then used expression analysis to identify candidate driver genes in amplicons and to deduce the genetic pathways involved in the disease.
- *Neuroblastoma*: Mosse *et al*. (2005) characterized 42 cell lines. Janoueix-Lerosey *et al*. (2005) identified the genome alterations in 28 neuroblastoma cell lines and compared their location with replication timing profiles. They found an association between breakpoint position and early replication regions.
- *Pancreas cancer*: Gysin *et al*. (2005) studied 25 cell lines from patients with pancreatic cancer, investigating copy number abnormalities and trying to understand precisely how these genetic alterations interact to generate the aberrant pathophysiology of the cancer. By combining these results with those from expression arrays, the authors identified candidate genes contributing to cancer cell invasion and metastasis.

Genome annotation data. Comparisons of genomic profiles with genomic annotation of any kind can help to elucidate the mechanisms involved in tumor progression. In ACTuDB, the user can visualize annotations, such as human gene structure, microRNA genes and genomic variants. Moreover, the results of two replication timing studies are provided. It may be interesting to compare breakpoint locations with replication timing pattern, as it has been suggested that chromosome breakpoints occur preferentially within early replicating regions (evidence to support this hypothesis has been obtained for neuroblastoma (Janoueix-Lerosey *et al*., 2005)). The annotation data are detailed below:

- *Gene structure*: the structural profiles of genes are updated from the UCSC Genome Browser (http://genome.ucsc.edu; Karolchik *et al*., 2003) (release hg18). All the information available for the gene, such as its name, position, intron–exon structure and alternative splicing variants, are provided. Coding exons are shown as red blocks and introns, as horizontal red lines connecting the exons. The 5′ and 3′ untranslated regions (UTR) are displayed in dark green and a lighter green, respectively.
- *Genomic variants*: The Database of Genomic Variants (http://projects.tcag.ca/variation), which was first described by Iafrate *et al*. (2004), includes genomic variants – defined as DNA regions larger than 1 kb and presenting copy number variation among a panel of unrelated individuals. When available, a list of diseases previously shown to be associated with this region is provided.
- *MicroRNA*, which was first described by Lee *et al*. (1993), is now considered to be a major factor in cell regulation. MicroRNA expression can affect the cell cycle and survival mechanisms. MicroRNA loss or amplification has been reported in several types of cancer (Calin and Croce, 2006), and microRNA expression profiles can be used to classify human cancers (Lu *et al*., 2005). We have collected the microRNA data from the miRBase Sequence Database (http://microrna.sanger.ac.uk/sequences) (Release 9.0) of the Sanger Institute (Griffiths-Jones, 2006) in ACTuDB.

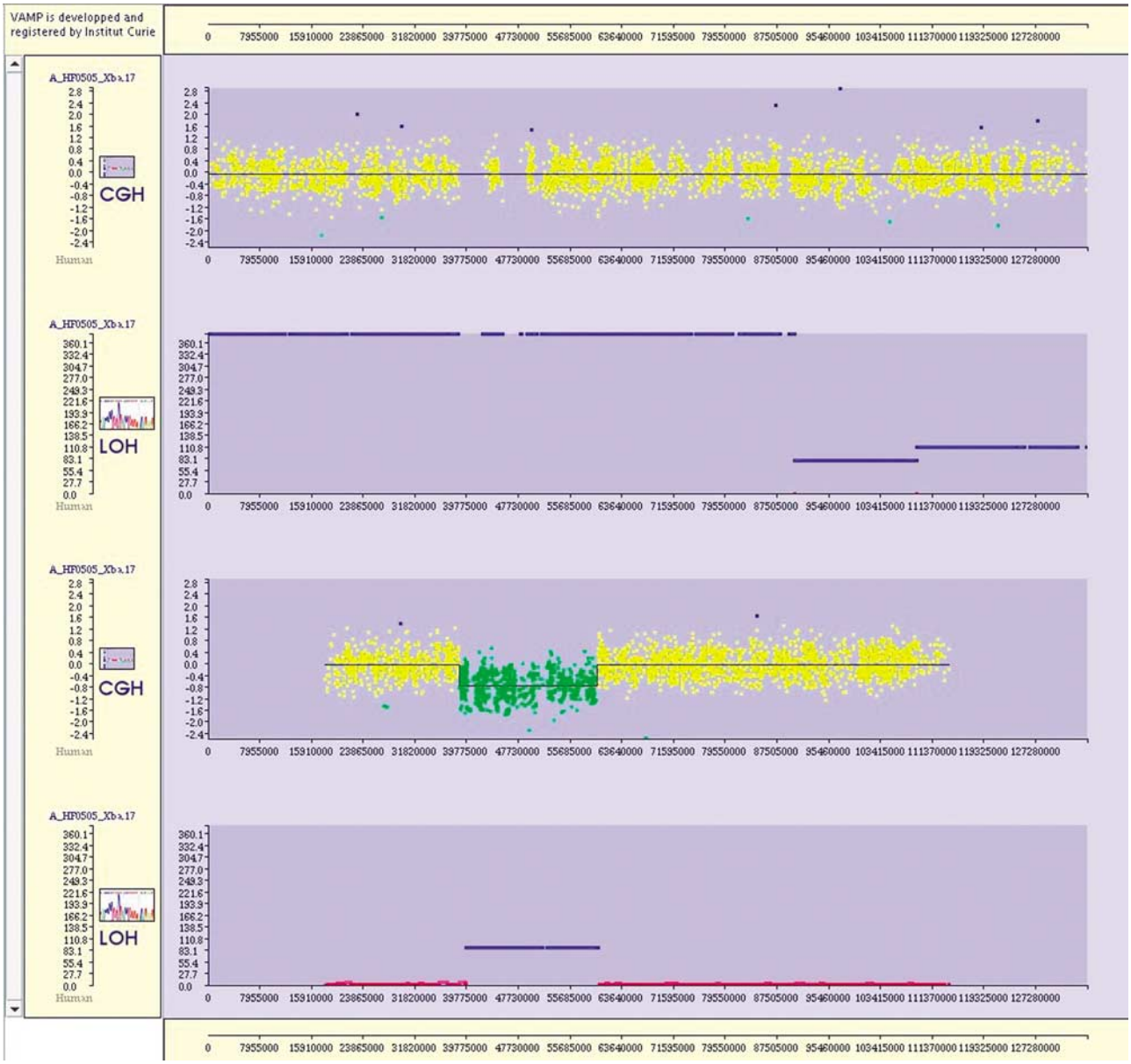


Figure 8 DNA copy number profile of chromosome 10 (top profile) and corresponding LOH profile (second profile from top), and DNA copy number profile of chromosome 13 (third profile from top) and corresponding LOH profile (bottom profile) for sample HF0505 Xba from the Kotliarov *et al.* (2006) dataset. For DNA copy number, the results of the GLAD algorithm are displayed (yellow correspond to normal region and green to loss region). For LOH profile $\log_{10}(P\text{-value})$ from Affymetrix software is displayed (this $\log_{10}(P\text{-value})$ is small for heterozygous regions and large for LOH regions): blue regions correspond to LOH and red regions do not show LOH.

- *Replication timing*: Woodfine *et al.* (2004) suggested that array-CGH technology could be used to assess the replication timing of sequences during the S phase of the cell cycle. They used human lymphoblastoid cells and found a positive correlation between replication timing and various genome parameters, including GC content, gene density and transcriptional activity. Janoueix-Lerosey *et al.* (2005) also assessed the replication timing pattern of seven neuroblastoma cell lines and obtained similar results. They also found that the breakpoint frequency in 28 neuroblastoma cell lines was higher in early replicated regions than elsewhere.

Protocol for array pretreatment before integration into ACTuDB

Probe mapping. A pipeline has been developed for mapping the probes onto a common human genome sequence reference: for each publication, the genome position is based on the last Working Draft version (current version is 36.1), with updating for each new release. The microarray data collected in ACTuDB were collected with three different types of probe, each involving a specific mapping process, as described below:

- *BacEnd clones* were mapped, using information from public databases and an internal database from Institut Curie. The

Table 1 List of publications from which microarray data have been integrated into ACTuDB

Author	*Cancer*	*Sample number*	*Molecular profile*	*Platform*	*Clinical data*
Fridlyand *et al*. (2006)	Breast	67 tumors	DNA copy number	BAC array HumArray1.14 and 2.0, UCSF	ER, TP53, tumor and vital statuses, radiation site, recurrence, CGH subtype, stage, grade, follow-up, treatment, …
Kotliarov *et al*. (2006)	Brain	178 tumors	DNA copy number	GeneChip® Human Mapping 100K SetS	
	Brain	178 tumors	LOH	GeneChip Human Mapping 100K SetS	
Stransky *et al*. (2006)	Bladder	57 tumors	DNA copy number	BAC array HumArray2.0, UCSF	Sex, grade, TNM, primary tumor (yes/no)
	Bladder	57 tumors	Expression	Affymetrix U95A/Av2	
Blaveri *et al*. (2005)	Bladder	98 tumors	DNA copy number	BAC array HumArray2.0, UCSF	Age, sex, stage, grade and lymph node status
Bredel *et al*. (2005)	Brain	54 tumors	DNA copy number	cDNA array	
Gysin *et al*. (2005)	Pancreas	25 cell lines	DNA copy number	BAC array HumArray2.0, UCSF	
Janoueix-Lerosey *et al*. (2005)	Neuroblastoma	28 cell lines	DNA copy number	BAC array	
	Replication timing	7 cell lines			
Mosse *et al*. (2005)	Neuroblastoma	42 cell lines	DNA copy number	BAC array	
Patil *et al*. (2005)	Liver	44 tumors	DNA copy number	BAC array HumArray1.14, UCSF	Age, sex, stage, tumor size, disease-free survival, HBV and HCV status, venous invasion, encapsulation
		5 cell lines			
Snijders *et al*. (2005)	Mouth	89 tumors	DNA copy number	BAC array HumArray2.0, UCSF	Age, sex, TP53 status, location and differentiation
de Leeuw *et al*. (2004)	Lymphoma	8 cell lines	DNA copy number	BAC array SMRT array	
Douglas *et al*. (2004)	Colon	37 tumors	DNA copy number	BAC/PAC array	Genome instability
		48 cell lines			
Nakao *et al*. (2004)	Colon	125 tumors	DNA copy number	BAC array HumArray1.14, UCSF	Age, sex, stage, location microsatellite instability (BAT26 marker)
Woodfine *et al*. (2004)	Fibroblast	1 cell line	DNA copy number	BAC array	
Veltman *et al*. (2003)	Bladder	41 tumors	DNA copy number	BAC array HumArray1.11, UCSF	Stage
		8 normal DNA			
Chen *et al*. (2002)	Liver	207 tumors	Expression	cDNA array	The patients are the same as in Patil *et al*. (2005)
Pollack *et al*. (2002)	Breast	44 tumors	DNA copy number	cDNA array	
		10 cell lines			
	Breast	5 chromosome X			
		37 tumors	Expression		
		4 cell lines			
Snijders *et al*. (2001)	Fibroblast	15 cell lines	DNA copy number	BAC array HumArray1.14, UCSF	
Summary					
DNA copy number		834 tumors 186 cell lines			
LOH Expression		8 normal DNA 178 tumors			
Replication timing		301 tumors 4 cell lines 8 cell lines			
Total		1519			

Abbreviations: ACTuDB, Array CGH Tumor DataBase; CGH, comparative genomic hybridization; LOH, loss of heterozygosity.

mapping process consists of four-ordered steps, with each step corresponding to a database query. We used the information from the first database in which the probe appeared, ignoring all other databases. The databases were searched in the following order: (i) the UCSC Genome Browser annotation database (release hg18) (Karolchik

Table 2 Percentage of probes mapped for each publication

Author	*Percentage*
Fridlyand *et al*. (2006)	93.2
Kotliarov *et al*. (2006)	100.0
Stransky *et al*. (2006) (DNA copy number)	92.9
Stransky *et al*. (2006) (expression)	94.8
Blaveri *et al*. (2005)	92.9
Bredel *et al*. (2005)	91.7
Gysin *et al*. (2005)	92.6
Janoueix-Lerosey *et al*. (2005)	97.4
Mosse *et al*. (2005)	61.7
Patil *et al*. (2005)	92.8
Snijders *et al*. (2005)	93.2
de Leeuw *et al*. (2004)	71.5
Douglas *et al*. (2004)	99.3
Nakao *et al*. (2004)	93.0
Woodfine *et al*. (2004)	99.8
Veltman *et al*. (2003)	74.2
Chen *et al*. (2002)	83.8
Pollack *et al*. (2002)	88.5
Snijders *et al*. (2001)	89.7
BacEnd clones	77.1
IMAGE clones	84.3
Affymetrix probes sets	94.8
Affymetrix probes sets (expression array)	94.8
Affymetrix probes sets (SNP array)	100.0
Total	89.8

Abbreviation: SNP, single nucleotide polymorphism.

et al., 2003) for the BacEnd; (ii) the SANGER/DECIPHER database (http://www.sanger.ac.uk/PostGenomics/decipher) using the 1 Mb clone, the 32 K clone set and the international clone sets information; (iii) the Institut Curie database in which each BacEnd sequence is mapped onto the NCBI build 36.1, using the BLAT algorithm from the UCSC Genome Browser (Kent, 2002); (iv) the UCSC Genome Browser annotation database (hg18) for STS. The percentage of clones mapped at each stage in the process was (i) 38.3, (ii) 35.0, (iii) 2.6 and (iv) 2.0. Thus, overall, a mean of 77.1% of the BacEnd clones were mapped.

- *IMAGE clones* were mapped from the UCSC Genome Browser annotation database (hg18). An average of 84.3% of IMAGE clones were mapped.
- *Affymetrix probes sets* were mapped from the UCSC Genome Browser annotation database (hg18) for expression arrays and from the SANGER/DECIPHER database for SNP arrays: the average of mapped probes sets were 94.8 and 100%, respectively.

The percentage of mapped probes for each publication is given in Table 2.

Breakpoint detection. Genomic profiles were analysed using the GLAD algorithm (Hupé *et al*., 2004). GLAD identifies chromosomal regions with same DNA copy number, delimited by breakpoints. A label (Gain, Normal, Loss or Amplicon) is assigned to each region, based on its median DNA copy number. Amplicons are defined as probes with a signal ratio greater than 2. Outliers are also detected and correspond to probes with a signal value significantly different from the region in which they lie. The GLAD algorithm was also used to analyse the replication timing experiments, but with other parameters, as described by Janoueix-Lerosey *et al*. (2005).

Hardware requirements and implementation

ACTuDB is based on a client–server architecture. On the client side, a Java-enabled HTML 4.0-compliant browser (Firefox, Safari, Internet Explorer) is required. Users must also configure their Java Virtual Machine according to the instructions given at http://bioinfo.curie.fr/actudb/. The user's computer must have at least 512 Mb of memory. On the server side, ACTuDB combines an XML repository with a dynamic web interface written in HTML and a Java applet. The website is powered by an Apache server.

Acknowledgements

We thank Olivier Delattre and Alain Aurias (Institut Curie, INSERM U509), who made information from the Institut Curie clone database available within ACTuDB. We thank our colleagues at Institut Curie for their help in setting up ACTuDB: Stéphane Tsacas, Jean-Gabriel Dick and Fraŋcois-David Collin (Institut Curie) for system, network and database administration. This work was supported partly by the EC contract ESBIC-D (LSHG-CT-2005-518192).

References

Albertson DG, Collins C, McCormick F, Gray JW. (2003). Chromosome aberrations in solid tumors. *Nat Genet* **34**: 369–376.

Baudis M. (2006). Online database and bioinformatics toolbox to support data mining in cancer cytogenetics. *BioTechniques* **40**: 269–271.

Blaveri E, Brewer JL, Roydasgupta R, Fridlyand J, DeVries S, Koppie T *et al*. (2005). Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clin Cancer Res* **11**: 7012–7022.

Bredel M, Bredel C, Juric D, Harsh GR, Vogel H, Recht LD *et al*. (2005). High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res* **65**: 4088–4096.

Calin GA, Croce CM. (2006). MicroRNA-cancer connection: the beginning of a new tale. *Cancer Res* **66**: 7390–7394.

Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J *et al*. (2002). Gene expression patterns in human liver cancers. *Mol Biol Cell* **13**: 1929–1939.

de Leeuw RJ, Davies JJ, Rosenwald A, Bebb G, Gascoyne RD, Dyer MJS *et al*. (2004). Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum Mol Genet* **13**: 1827–1837.

Douglas EJ, Fiegler H, Rowan A, Halford S, Bicknell DC, Bodmer W *et al*. (2004). Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res* **64**: 4817–4825.

Fridlyand J, Snijders AM, Ylstra B, Li H, Olshen A, Segraves R *et al*. (2006). Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer* **6**: 96.

Griffiths-Jones S. (2006). miRBase: the microRNA sequence database. *Methods Mol Biol* **342**: 129–138.

Gysin S, Rickert P, Kastury K, McMahon M. (2005). Analysis of genomic DNA alterations and mRNA expression patterns in a panel of human pancreatic cancer cell lines. *Genes Chromosomes Cancer* **44**: 37–51.

Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**: 3413–3422.

Hurst CD, Fiegler H, Carr P, Williams S, Carter NP, Knowles MA. (2004). High-resolution analysis of genomic copy number alterations in bladder cancer by microarray-based comparative genomic hybridization. *Oncogene* **23**: 2250–2263.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y *et al*. (2004). Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949–951.

Janoueix-Lerosey I, Hupé P, Maciorowski Z, La Rosa P, Schleiermacher G *et al*. (2005). Preferential occurrence of chromosome breakpoints within early replicating regions in neuroblastoma. *Cell Cycle* **4**: 1842–1846.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT *et al*. (2003). The UCSC genome browser database. *Nucleic Acids Res* **31**: 51–54.

Kent WJ. (2002). BLAT – The BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Kotliarov Y, Steed ME, Christopher N, Walling J, Su Q, Center A *et al*. (2006). High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. *Cancer Res* **66**: 9428–9436.

La Rosa P, Viara E, Hupé P, Pierron G, Liva S, Neuvial P *et al*. (2006). VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics* **22**: 2066–2073.

Lee RC, Feinbaum RL, Ambros V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**: 843–854.

Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D *et al*. (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**: 834–838.

Mosse YP, Greshock J, Margolin A, Naylor T, Cole K, Khazi D *et al*. (2005). High-resolution detection and mapping of genomic DNA alterations in neuroblastoma. *Genes Chromosomes Cancer* **43**: 390–403.

Nakao K, Mehta KR, Fridlyand J, Moore DH, Jain AN, Lafuente A *et al*. (2004). High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis* **25**: 1345–1357.

Patil MA, Gutgemann I, Zhang J, Ho C, Cheung S-T, Ginzinger D *et al*. (2005). Array-based comparative genomic hybridization reveals recurrent chromosomal aberrations and Jab1 as a potential target for 8q gain in hepatocellular carcinoma. *Carcinogenesis* **26**: 2050–2057.

Pinkel D, Albertson DG. (2005). Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **37**(Suppl): 11–17.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE *et al*. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA* **99**: 12963–12968.

Rouveirol C, Stransky N, Hupé P, La Rosa P, Viara E, Barillot E *et al*. (2006). Computation of recurrent minimal genomic alterations from CGH data. *Bioinformatics* **22**: 849–856.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J *et al*. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* **29**: 263–264.

Snijders AM, Schmidt BL, Fridlyand J, Dekker N, Pinkel D, Jordan RCK *et al*. (2005). Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene* **24**: 4232–4242.

Stransky N, Vallot C, Reyal F, Bernard-Pierrot I, de Medina SGD, Segraves R *et al*. (2006). Regional copy number-independent deregulation of transcription in cancer. *Nat Genet* **38**: 1386–1396.

Veltman JA, Fridlyand J, Pejavar S, Olshen AB, Korkola JE, DeVries S *et al*. (2003). Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res* **63**: 2872–2880.

Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD *et al*. (2004). Replication timing of the human genome. *Hum Mol Genet* **13**: 191–202.

Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA. (2006). BAC to the future! Or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res* **34**: 445–450.

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc).

Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W477–W481***
doi:10.1093/nar/gkl215

CAPweb: a bioinformatics CGH array Analysis Platform

Stéphane Liva[1,*], Philippe Hupé[1,2], Pierre Neuvial[1], Isabel Brito[1], Eric Viara[1], Philippe La Rosa[1] and Emmanuel Barillot[1]

[1]Institut Curie, Service Bioinformatique and [2]Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 Paris Cedex 05, France

Received February 14, 2006; Revised and Accepted March 24, 2006

ABSTRACT

Assessing variations in DNA copy number is crucial for understanding constitutional or somatic diseases, particularly cancers. The recently developed array-CGH (comparative genomic hybridization) technology allows this to be investigated at the genomic level. We report the availability of a web tool for analysing array-CGH data. CAPweb (CGH array Analysis Platform on the Web) is intended as a user-friendly tool enabling biologists to completely analyse CGH arrays from the raw data to the visualization and biological interpretation. The user typically performs the following bioinformatics steps of a CGH array project within CAPweb: the secure upload of the results of CGH array image analysis and of the array annotation (genomic position of the probes); first level analysis of each array, including automatic normalization of the data (for correcting experimental biases), breakpoint detection and status assignment (gain, loss or normal); validation or deletion of the analysis based on a summary report and quality criteria; visualization and biological analysis of the genomic profiles and results through a user-friendly interface. CAPweb is accessible at http://bioinfo.curie.fr/CAPweb.

INTRODUCTION

In recent years, array-CGH (comparative genomic hybridization) has become the technology of choice for large scale investigations of DNA copy number changes between two genomes. Today, CGH arrays allow the ratio of DNA copy number between a test and a reference sample to be simultaneously assessed in 2000 to 30 000 positions in the genome, giving a resolution of between 1.5 Mb to 100 kb (1,2). Its main applications are the study of diseases in which the DNA copy number varies in certain locations of the genomes, due to either constitutional mutations (hereditary or *de novo*), such as human genetic diseases (3) or somatic changes, such as in cancers (4). The identification of regions of altered DNA gives valuable information about the genes involved in the disease, and many projects have been launched worldwide to determine the genome structure of tumour cells (4). Array-CGH is also an important source of information for studying genome evolution, for example in bacteria (5) or mammals (6). We have developed a Web tool, called CAPweb (CAP: CGH array Analysis Platform), for bioinformatics analysis of CGH arrays. This tool combines the following tasks: (i) data management, (ii) array normalization, (iii) automatic breakpoint detection and assessment of gain and loss regions, (iv) quality control and (v) a graphical user interface for browsing and analysing the genomic profiles.

Several tools have recently been developed for analysing CGH array data, such as CGH-Explorer (7), ArrayCyGHt (8), CGHPRO (9), WebArray (10) or ArrayCGHbase (11), although the only web-accessible servers are ArrayCyGHt, WebArray and CAPweb. Among these three, only CAPweb allows project management and the upload of raw data files without pre-processing. It also offers unique features for the analysis and visualization of array-CGH data. CAPweb accepts raw data from the main microarray image analysis software. As far as we are aware, CAPweb is the only platform dedicated to biologists that allows the complete analysis of raw CGH arrays from the raw data to visualization and biological interpretation.

DESCRIPTION

The CAPweb server allows the user to store, analyse and manage his or her data. We will now describe its operation (Figure 1). A tutorial is accessible at http://bioinfo.curie.fr/tutorial/CAPweb/capweb_tutorial.html.

*To whom correspondence should be addressed. Tel: +33 0 1 4234 65 31; Fax: +33 0 1 42 34 65 28; Email: capweb@curie.fr

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

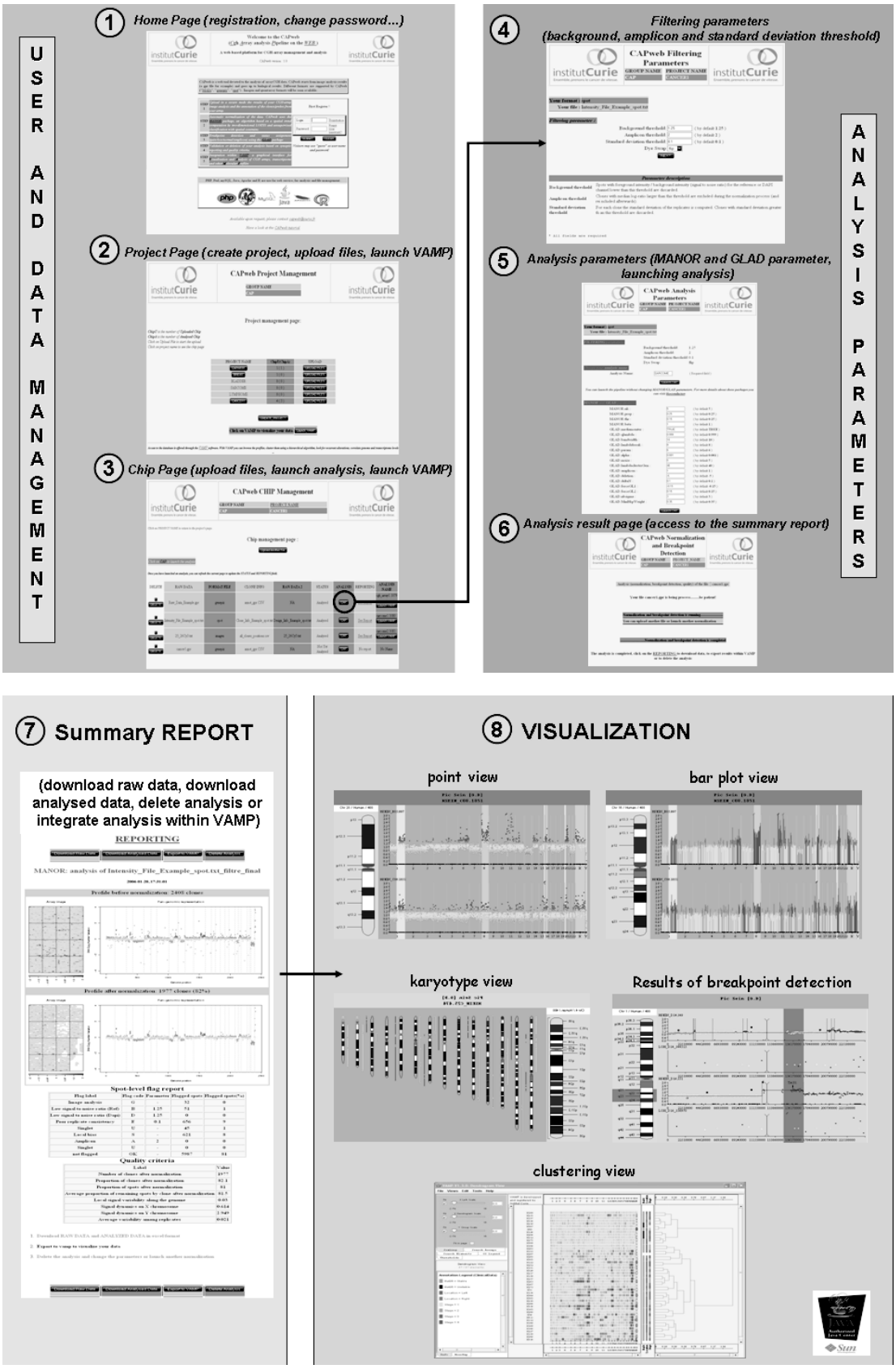


Figure 1. Different views of CAPweb Interface showing how the CGH array analysis proceeds, see text for details.

User registration, data upload and management

The first step of the analysis is user registration [Figure 1(1)], which ensures the confidentiality of the submitted data. The user is sent a login/password by email and can then create one or more projects to upload data files [Figure 1(2)]. Several input formats from microarray image analysis software are currently supported: Genepix (http://www.moleculardevices.com/pages/instruments/gn_genepix4000.html), Imagene (http://www.biodiscovery.com/index/imagene), Spot (12) and MAIA (13). CAPweb requires only two types of file: (i) a raw intensity file (one file for Genepix and MAIA, two files for Imagene and Spot) and (ii) a genomic position file mapping each spot to a name and its position on the genome under CSV (semi colon separator) format.

For each project, the 'Array Management' page [Figure 1(3)] lists all the arrays, their analysis status and the summary report file, and allows new analyses to be launched.

The array files are permanently stored on the server: the user can only browse the arrays of his or her projects, and only the user is allowed to delete them.

CGH array analysis

From the 'Array management' page, the user can launch the array analyses. The analyses are run in the background, allowing the user to use CAPweb for other analyses.

Data Normalization (MANOR). As in all microarray analyses, CGH array data must be normalized to correct for experimental artefacts while preserving the true biological signal. For this goal, CAPweb uses the Bioconductor package MANOR, which includes spot and clone filtering steps that discards spots having too low a signal-to-noise ratio or clones with a poor replicate consistency, and, most importantly, it includes a spatial normalization step. This step aims to correct for spatial effects on the arrays. We identified these as the predominant experimental artefact in the array-CGH data we have studied. The corresponding algorithm is based on a spatial trend estimation and a signal segmentation method with a spatial constraint, as described in P. Neuvial *et al.* (manuscript submitted).

Breakpoint detection and assessment of gain and loss region (GLAD). This step aims to identify chromosomal regions having an identical DNA copy number, which are delimited by breakpoints. CAPweb uses the Bioconductor package GLAD, which implements an algorithm described in (14). This method first uses the spatial structure of array-CGH data to adaptively calculate a smoothed signal value for each clone. These smoothed signal values are then used to detect breakpoints and outliers, and then genomic regions having the same underlying copy number are clustered together.

Quality control. Various statistical criteria can help the user assess the quality of the array. These include intra-replicate variability, genomic neighbour variability, the percentage of spots filtered out after image analysis and the amplitude of signal gap between regions having a different DNA copy number. These quality criteria are reported in an HTML summary report file, which also displays key features of the normalization process: array image and genomic profile before and after normalization, and a summary of the normalization. This file [Figure 1(7)] allows the user to compare the quality of the data before and after analysis. Based on this information, the user may choose to keep or discard the analysis.

This data analysis step can be run without an extensive knowledge of the underlying statistical algorithms by using default parameters. Default parameters have been calibrated by comparing quality criteria for various parameter value in two datasets: one from UCSF (218 arrays, Spot format, as a collaboration with Dan Pinkel), and one from Institut Curie/INSERM U509 (181 arrays, Genepix format). This part is described in detail elsewhere (P. Neuvial *et al.* manuscript submitted). However, CAPweb allows the user to choose the value of several parameters for filtering, spatial normalization and breakpoint detection. The summary report also helps in comparing the results of analyses carried out with different parameter values [Figure 1 (4–6)].

Visualization (VAMP) and biological analysis

Once the first level of array analysis has finished, the user can visualize and further analyse the data through a graphical user interface: VAMP—visualization and analysis of array-CGH, transcriptome and other molecular profiles (P. La Rosa *et al.* manuscript submitted) [Figure 1 (8)]. Several visualization types are proposed, such as the classical CGH karyotype view or the genome-wide multi-tumour comparison view. These allow the user to easily compare different arrays. Additional information concerning each clone or DNA region can be interactively retrieved from different public databases through external links. Other functions for analysing CGH data are provided within the interface, such as looking for minimal or recurrent regions of alterations (15), clustering, etc.

VAMP allows the user to display genomic profiles at various resolutions [from the whole genome to small regions (clone level)]. All the analyses results (breakpoint detection, assignment of gain/lost region, quality criteria, etc.) can also be displayed within VAMP. VAMP has many other functions for navigation, querying and analysis that we have not explained here; we refer the reader to the documentation and demo for further details (http://bioinfo.curie.fr/vamp/doc).

Note that the user can analyse at least 200 arrays with 1GB of memory.

IMPLEMENTATION

The CAPweb server is based on freely available components (Figure 2). The database for user management and array management was built on mySQL. PHP scripts ensure registration and project management. Perl scripts control the launching of statistical analyses written in R. A Java applet and XML files are used for the visualization. CAPweb integrates the MANOR and GLAD R packages and the VAMP software, all of which were developed at the Institut Curie.

The security in CAPweb is based on mysql authentication and cookie session. Uploaded data are considered strictly confidential. The CAPweb server is also available upon request for local installation on Unix/Linux/MacOS X operating systems.

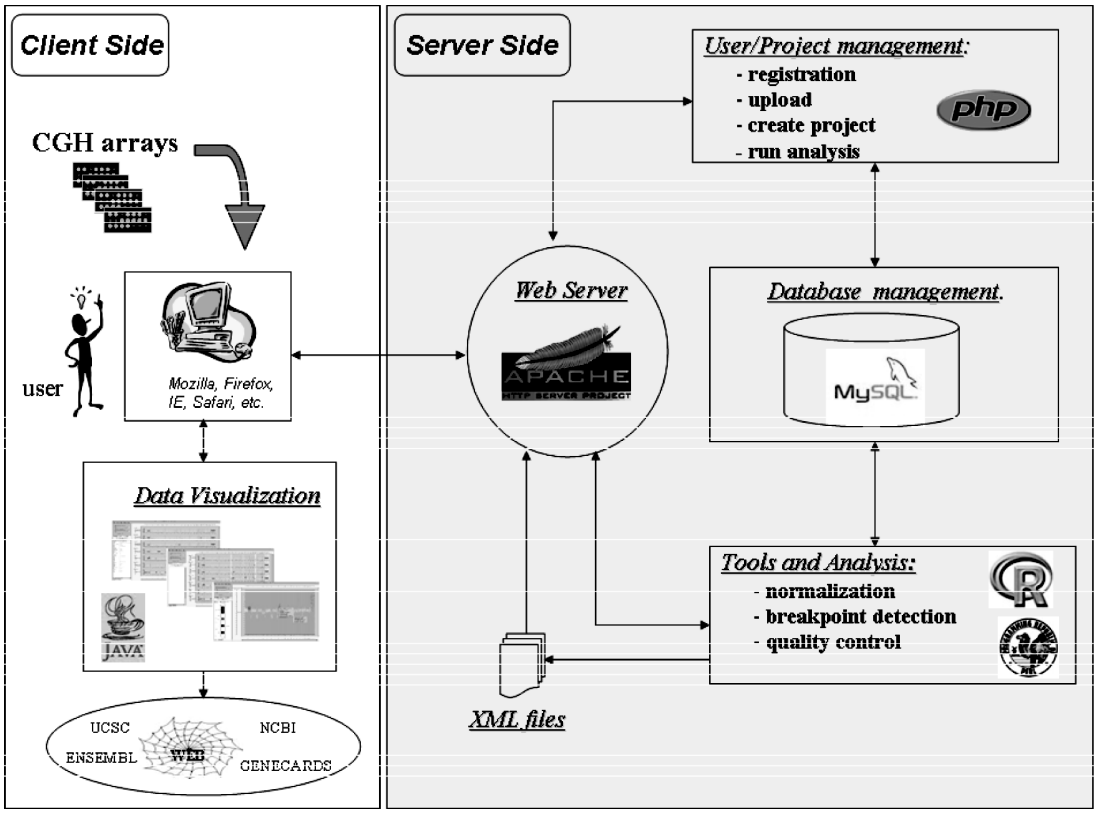


Figure 2. CAPweb software architecture, see text for details.

CONCLUSION

Array-CGH is a popular technology that is now used in many projects ranging from the characterization of tumours to the study of genome evolution. As with any large scale technology, its exploitation relies heavily on the availability of bioinformatics tools for managing and analysing the data. Many bioinformatics algorithms and interfaces have been developed but biologists have lacked a web-based platform for integrating these tools in a user-friendly manner. CAPweb offers this service and combines array normalization, quality control, breakpoint detection and the biological interpretation of the results. It also helps with data management. Currently, the public CAPweb server at the Institut Curie contains 800 arrays.

In this paper we have presented CAPweb 1.0 version. A new version is currently being developed, which will allow the user to analyse high density oligonucleotide arrays, such as Affymetrix GeneChip® Arrays or Nimblegen™ Arrays, to integrate any clinical information, and to add gene expression profiles so that copy number profiles can be compared and correlated to them.

ACKNOWLEDGEMENTS

This work was supported by the Institut Curie, the Centre National de la Recherche Scientifique, the Cancéropole Ile-de-France, the Région Ile-de-France and the association 'Courir pour la vie, Courir pour Curie'. The authors thank all our colleagues who have tested CAPweb and suggested improvements: G. Pierron, C. Brennetot, A. Idbaih, E. Manié (Institut Curie) and S. Law (UCSF). Funding to pay the Open Access publication charges for this article was provided by Institut Curie.

Conflict of interest statement. None declared.

REFERENCES

1. Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. *et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet.*, **29**, 263–264.
2. Ishkanian,A.S., Malloff,C.A., Watson,S.K., DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A. *et al.* (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genet.*, **36**, 299–303.
3. Lockwood,W.W., Chari,R., Chi,B. and Lam,W.L. (2006) Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur. J. Hum. Genet.*, **14**, 139–148.
4. Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nature Genet.*, **37**, 11–17.
5. Fukiya,S., Mizoguchi,H., Tobe,T. and Mori,H. (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative hybridization microarray. *J. Bacteriol.*, **186**, 3911–3921.
6. Wilson,G.M., Flibotte,S., Missirlis,P.I., Marra,M.A., Jones,S., Thornton,K., Clark,A.G. and Holt,R.A. (2006) Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. *Genome Res*, **16**, 173–181.
7. Lingjaerde,O.C., Baumbush,L.O., Liestol,K., Glad,I.K. and Borresen-Dale,A.L. (2005) CGH-explorer, a program for analysis of array-CGH data. *Bioinformatics*, **6**, 821–822.

8. Kim,S.Y., Nam,S.W., Lee,S.H., Park,W.S., Yoo,N.J., Lee,J.Y. and Chung,Y.J. (2005) ArrayCyGHt, a web application for analysis and visualization of array-CGH data. *Bioinformatics*, **21**, 2554–2555.
9. Chen,W., Erdogan,F., Ropers,H., Lenzner,S. and Ullmann,R. (2005) CGHPRO, a comprehensive data analysis tool for array CGH. *BMC Bioinformatics*, **6**, 85.
10. Xia,X., McClelland,M. and Wang,Y. (2005) WebArray, an online platform for microarray data analysis. *BMC Bionformatics*, **6**, 306.
11. Menten,B., Pattyn,F., De Preter,K., Robbrecht,P., Michels,E., Buysse,K., Mortier,G., De Paepe,A., van Vooren,S., Vermeesh,J. *et al*. (2005) ArrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. *BMC Bioinformatics*, **6**, 124.
12. Jain,A.N., Tokuyasu,T.A., Snidjers,A.M., Segraves,R., Albertson,D.G. and Pinkel,D. (2002) Fully automatic quantification of microarray image data. *Genome Res*., **12**, 325–332.
13. Novikov,E. and Barillot,E. (2005) A robust algorithm for ratio estimation in two-color microarray experiments. *J. Bioinform. Comput. Biol*., **6**, 1411–1428.
14. Hupé,P., Stransky,N., Thiery,J.P., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss DNA regions. *Bioinformatics*, **20**, 3413–3422.
15. Rouveirol,C., Stransky,N., Hupé,P., La Rosa,P., Viara,E., Barillot,E. and Radvanyi,F. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 17 2006, pages 2066–2073
doi:10.1093/bioinformatics/btl359

Genome analysis

VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles

Philippe La Rosa[1,*], Eric Viara[1], Philippe Hupé[1,2], Gaëlle Pierron[3], Stéphane Liva[1], Pierre Neuvial[1], Isabel Brito[1], Séverine Lair[1], Nicolas Servant[1], Nicolas Robine[1,4], Elodie Manié[3], Caroline Brennetot[3], Isabelle Janoueix-Lerosey[3], Virginie Raynal[3], Nadège Gruel[3], Céline Rouveirol[2], Nicolas Stransky[2], Marc-Henri Stern[3], Olivier Delattre[3], Alain Aurias[3], François Radvanyi[2] and Emmanuel Barillot[1]

[1]Institut Curie, Service Bioinformatique, 26 rue d'Ulm, Paris, 75248 cedex 05, France, [2]Institut Curie, CNRS UMR 144, 26 rue d'Ulm, Paris, 75248 cedex 05, France, [3]Institut Curie, INSERM U509, 26 rue d'Ulm, Paris, 75248 cedex 05, France and [4]Institut Curie, CNRS, Université Pierre et Marie Curie UMR 7147, 26 rue d'Ulm, Paris, 75248 cedex 05, France

Received on January 11, 2006; revised on May 31, 2006; accepted on June 25, 2006
Advance Access publication July 4, 2006
Associate Editor: Nikolaus Rajewsky

ABSTRACT

Motivation: Microarray-based CGH (Comparative Genomic Hybridization), transcriptome arrays and other large-scale genomic technologies are now routinely used to generate a vast amount of genomic profiles. Exploratory analysis of this data is crucial in helping to understand the data and to help form biological hypotheses. This step requires visualization of the data in a meaningful way to visualize the results and to perform first level analyses.

Results: We have developed a graphical user interface for visualization and first level analysis of molecular profiles. It is currently in use at the Institut Curie for cancer research projects involving CGH arrays, transcriptome arrays, SNP (single nucleotide polymorphism) arrays, loss of heterozygosity results (LOH), and Chromatin ImmunoPrecipitation arrays (ChIP chips). The interface offers the possibility of studying these different types of information in a consistent way. Several views are proposed, such as the classical CGH karyotype view or genome-wide multi-tumor comparison. Many functionalities for analyzing CGH data are provided by the interface, including looking for recurrent regions of alterations, confrontation to transcriptome data or clinical information, and clustering. Our tool consists of PHP scripts and of an applet written in Java. It can be run on public datasets at http://bioinfo.curie.fr/vamp

Availability: The VAMP software (Visualization and Analysis of array-CGH,transcriptome and other Molecular Profiles) is available upon request. It can be tested on public datasets at http://bioinfo.curie.fr/vamp. The documentation is available at http://bioinfo.curie.fr/vamp/doc

Contact: vamp@curie.fr

1 INTRODUCTION

Array Comparative Genome Hybridization (array-CGH) is a recently developed technology based on DNA microarrays (Pinkel *et al*., 1998; Snijders *et al*., 2001; Solinas-Toldo *et al*., 1997; Ishkanian *et al*., 2004) that can be used to investigate DNA copy number differences between two samples. A CGH array-generally consists of spotted clones of genomic sequences (e.g. bacterial artificial chromosomes) that cover part or all of the genome. Both DNA samples are labeled with distinct fluorescent dyes and undergo competitive hybridization onto the CGH array. The array is then scanned with a scanner or a CCD camera, and the acquired image is analyzed (gridding, spot addressing, spot segmentation, spot quantification, outlier detection), normalized (to remove as much as possible any systematic spatial or intensity biases, e.g. Neuvial *et al*., (2005), duplicate statistical analysis is then carried out (each clone is generally spotted in several copies), and adequate statistical algorithms detect any loss or gain regions (Hupé *et al*., 2004; Olshen *et al*., 2004; Fridlyand *et al*., 2004; Jong *et al*., 2003; Picard *et al*., 2005; Eilers and de Menezes, 2005; Bilke *et al*., 2005). CGH arrays are often used in cancer research because chromosome aberrations are thought to be causal in tumor progression (Albertson *et al*., 2003; Pinkel and Albertson, 2005). Here, normal DNA is used as reference and the test sample would be tumoral biopsy DNA. The normal sample has two copies of each genomic region, whereas tumor DNA may show losses or gains in certain DNA regions. Measurement of the signal intensities of the reference and tumor samples for each clone makes it possible to determine the lost or gained regions in the tumor sample. Further analyses can include the determination of recurrent loss or gain of DNA regions, clustering of samples and determination of candidate oncogenes and candidate tumor suppressor genes within the altered regions (based on their annotations or on their transcription level). It is also possible to link array-CGH results to the clinical phenotype or to biological parameters through, for example, supervised classification or correlation analysis. The visualization of the data is a crucial step in the analysis procedure and is essential for hypothesis formulation and model-free reasoning. We have developed, in the framework of large-scale array-CGH projects, a graphical user interface that allows several visualization modes of the CGH profiles and offers several data analysis tools. The software also displays a large variety of genomic profiles, such as transcriptome,

*To whom correspondence should be addressed.

 © The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

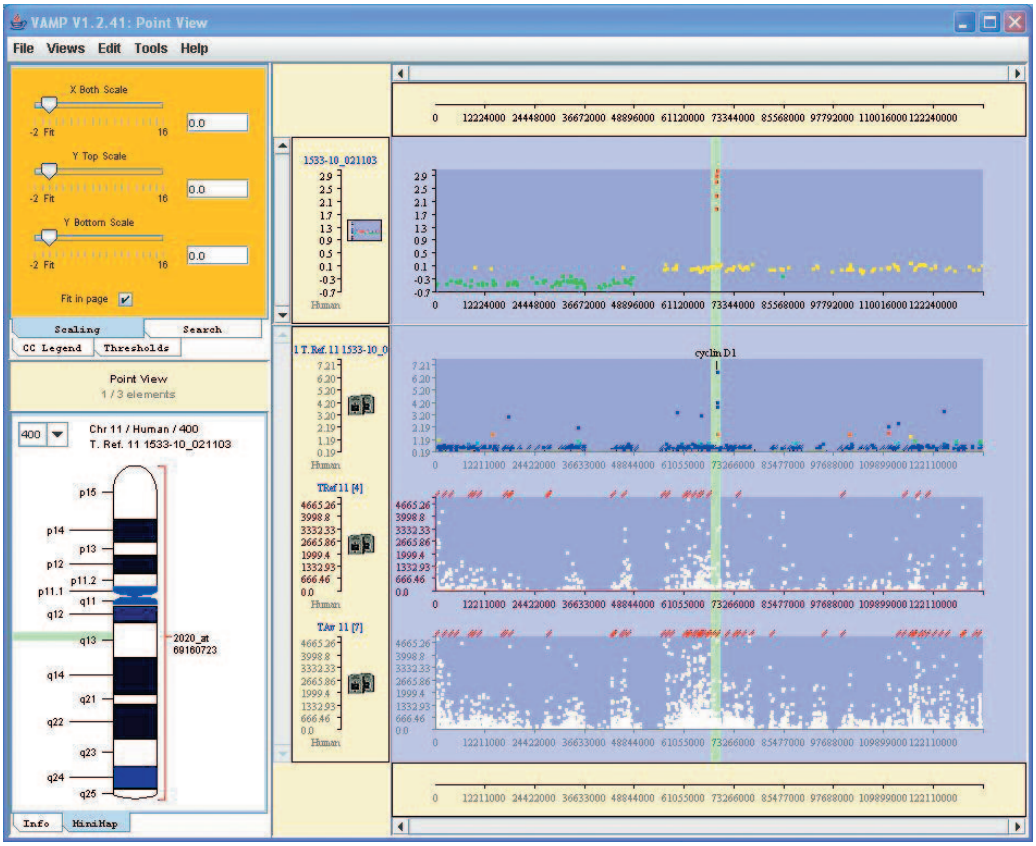


Fig. 1. Array-CGH (top profile) versus transcriptome ratio (second profile in descending order), computed for Affymetrix U95 array of a bladder tumor sample and of a reference sample. This confrontation pinpoints the probable implication of the oncogene cyclin D1 in this tumor. The third and fourth profiles in descending order correspond to a reference profile (average normal bladder tissue profile) and the profile of the tumor under study, respectively. The second profile is the ratio of the fourth to the reference profile.

Loss Of Heterozygosity (LOH), Vogelstein *et al*. (1989), Single nucleotide polymorphism (SNP) arrays (Bignell *et al*., 2004; Huang *et al*., 2004) and ChIP chip [Chromatin ImmunoPrecipitation coupled with microarrays, Buck and Lieb (2004)] profiles and allows addition of new tools for data treatment or analysis. We have called the software VAMP for 'Visualization and Analysis of Molecular Profiles'. In this article we first detail how data are visually presented in VAMP, and then we explain how the user interacts with the software and which functionalities are offered for data analysis. Finally, we describe the software architecture of VAMP.

2 RESULTS

2.1 Data representation

VAMP was designed to graphically represent any genomic profile along the genome axis. We started the development of VAMP for array-CGH data, but we have extended it to accept, on the same window, any kind of profile. We currently use the software for expression arrays, SNP arrays, LOH results and ChIP chip profiling, in addition to array-CGH. VAMP is currently used for three species (human, mouse and yeast) but the addition of a new species is straightforward. It is possible to visualize simultaneously, on the same window, different types of profiles for a given species, e.g. array-CGH and mRNA expression profiles of a tumor (Fig. 1). All profiles in a window are drawn on the *x*-axis with the same scale (the genome sequence), which allows an easy comparison of profiles.

A typical VAMP window is divided into three areas (Fig. 2): the main frame consists of the graphical display of the profiles; the top left frame controls zoom, search and drawing options; the bottom left frame offers the choice between textual information (Fig. 3) on the object under the mouse pointer, or context information, called MiniMap (Fig. 2).

2.1.1 Main frame VAMP currently offers several types of visualization that can be displayed in the main frame: (1) List View, (2) Profile View (Fig. 2) (3) Karyotype View (Fig. 3), (4) Dot Plot View (Fig. 4). These views all allow simultaneous visualization of several profiles (the only limitation is the memory size of the computer running VAMP, or more precisely, the memory allocated to the Java virtual machine: for example with an 800 Mb Java virtual machine memory, 700 microarrays (each with 3500 probes) can be loaded simultaneously).

- List View: the List View lists the names of all the arrays currently loaded and can be used for selecting or keeping track of the data under study.

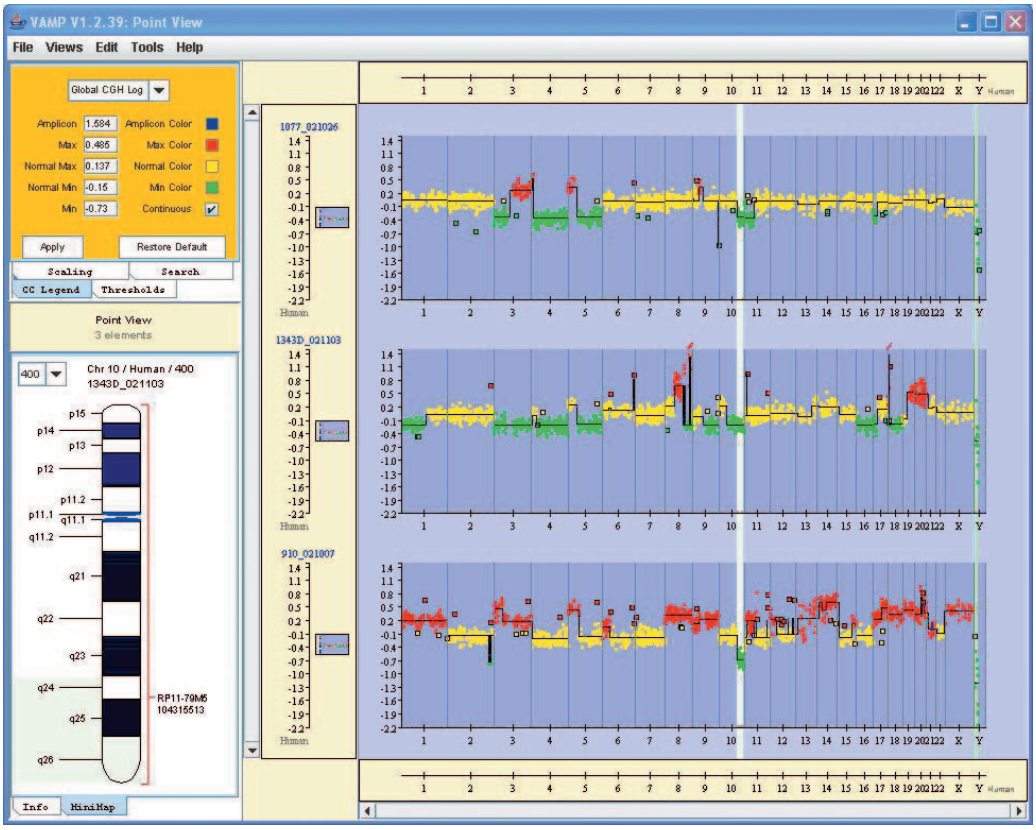


Fig. 2. Genomic View, main frame: profiles along all the concatenated chromosomes; top left: zoom control, search and drawing options; bottom left: textual information on the object under the mouse pointer or (in this figure) chromosome context information (MiniMap). The regions spanning the three tumors highlighted in green are those that are lost in all tumors (short arm of chromosome 10, and Y chromosome); these are called minimal regions.

- Profile View: the Profile View (Fig. 2) can display the profiles as points, barplots or curves. It can be split into two frames, as in Figure 1. The upper frame can, for example, contain a profile for reference when browsing a collection of profiles in the lower frame. The two frames have separate control of Y-scale and Y-scrolling, but have the same X-scale and X-scrolling. The Profile View can also display symbols for chromosome telomeres and centromeres, and can show the results of CGH ratio statistical analysis (e.g. breakpoints, or smoothed signal values, see Fig. 2).
- Karyotype View: the Karyotype View (Fig. 3) displays profiles having the well-known classical CGH rendering: vertical representations of chromosomes with cytogenetic banding and contiguous representation of sample profiles.
- Dot Plot View: the Dot Plot View does not consider the microarray probe positions on the genome, but only their ranks. It displays a collection of samples as a heat map based on the level of signal for each probe (Fig. 4).

By default, points or barplots are colored according to the signal intensity (generally using ratios of the two channels or log-ratios) using a continuous scale from red to yellow to green. All the previously mentioned views for the CGH data can be colored as a function of the array-CGH data analysis. Typically, gained DNA regions are displayed in red, lost regions in green, amplicons in blue and normal in yellow.

Whatever view is chosen, the profiles can be represented in Genomic mode or Chromosome mode. The Genomic mode simply depicts the profiles along all the concatenated chromosomes. It is the most usual representation, and allows comparison of profiles from different samples or comparison of different types of profiles from a given sample. The Chromosome mode is similar to the Genomic mode except that it only displays one particular chromosome. It is also possible to merge several chromosomes and to represent those chromosomes useful for the study.

- New Views: our object-oriented architecture easily allows us to add new types of views that can be associated with particular actions or data processing. For example, the Minimal Region functionality is associated with a particular type of view. Therefore, when profiles are pasted in the window, the Minimal Region View automatically displays the array-CGH profiles with the DNA regions recurrently lost or gained in the samples (Fig. 2).

2.1.2 Top left frame This frame controls zoom, search and drawing options. Zooming is independent on *X* and *Y* axes, and all profiles in the same window have the same zoom control, except

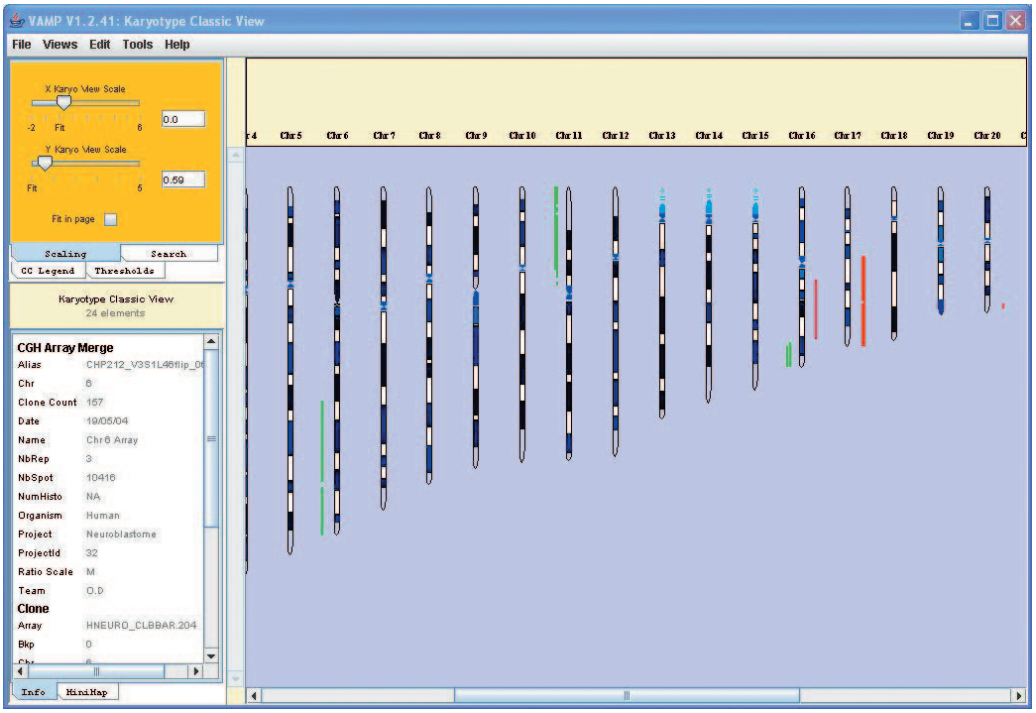


Fig. 3. Karyotype View, classic rendering of CGH data, loss regions in green, gain in red.

for *Y* zooming of the reference profile. The search can be carried out on any property attached to the arrays or the clones/probes held in an XML (eXtended Markup Language) data file or in the database (see Fig. 6 and the Software architecture presentation below). For XML data files, the list of properties is not limited,but is established at run time, leading to a very flexible search option. Drawing options include color-coding for signal values, and the threshold values to be applied; they can be either global to the application or restricted to one profile (local). User preferences can be saved on your computer in a XML configuration file.

2.1.3 Bottom left frame (Object information and context frame) The bottom left frame can either display textual information on the object under the mouse pointer (Fig. 3) or context information, called MiniMap (Fig. 2). The textual information consists of mandatory fields (object genomic position, signal value, project name, organism and data type) and any other type of complementary information stored in the XML data file. For example, in array-CGH profiles we currently display general information about the clone under the mouse pointer (name, chromosome, number of valid replicates, rank and position on the sequence, signal ratio and standard deviation, size of the clone, CGH status—gain/lost/normal) as well as information about the array (name, number of spots, number of clones, number of replicates, chromosomes covered, ratios or log-ratios) and information about the sample (sample id, project name, date). MiniMap is a special view type that gives some context on what the user is examining in the main frame: (1) a cytogenetic representation of the chromosome under the mouse pointing, with (2) a rule delimiting the region of the chromosome displayed on the main frame and (3) the name and position of the object (array-CGH clone, transcriptome microarray probe, etc.) under the mouse pointer. In this view, the display can be automatically updated when the user moves the mouse.

2.2 User interaction

All user actions are accessible either through a Menu on the menubar, or through pointing to or clicking objects. When using VAMP, the session can be saved in local XML files. Reloading the file later on allows the continuation of the analysis within the context of the previous work, or allows the exchange of results and data with colleagues. All user preferences can also be stored in local XML files. Drag and drop capability is offered for any profile, from one window to any other window, the rendering being automatically adapted (e.g. from a dot plot view to a karyotype view). An advanced printing function is offered, either in visible mode (only the profiles that are visible on the screen are printed), or in global mode (all profiles in the view are printed). A template is offered for defining the output of the printing (this can, for example, include several frames in an arbitrary composition, to which text or images can be added). It can be used for defining and printing standardized outputs. The user can also interactively monitor the print preferences.

2.3 Data analyses

VAMP allows addition of any new piece of software for data analysis and visualization of the results. Several functionalities have already been implemented either as plug-ins or within the VAMP Java source code. VAMP was initially developed for the analysis of CGH-arrays of tumoral samples. As VAMP is actually an interface, it is assumed that the microarray data have already been normalized, and also, for CGH data, that breakpoints have been established and

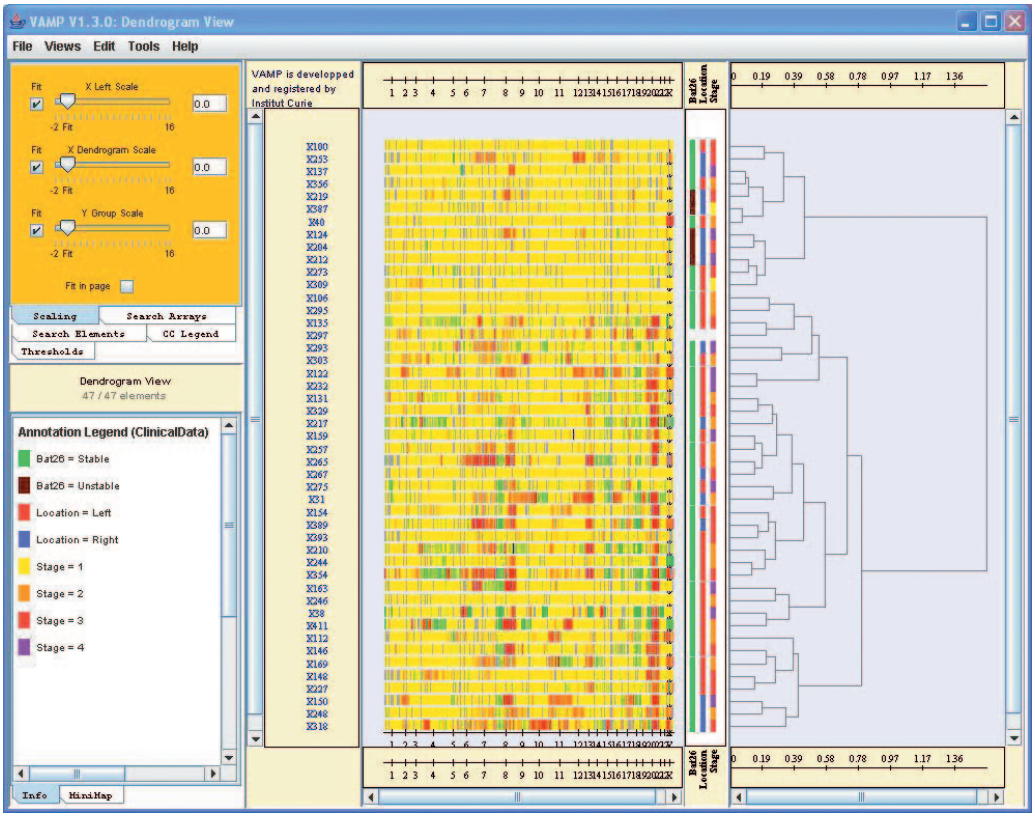


Fig. 4. VAMP interface, dotplot view of array-CGH profiles (middle panel), and dendrogram resulting from a hierarchical clustering (right panel). In between, color-coded clinical information about the samples, with a legend (bottom left). Data from Nakao *et al.* (2004).

regions of DNA loss or gain inferred. VAMP can then display in the profile frame (Fig. 2) the breakpoint positions, the status of each region (by default, green for loss, yellow for normal, red for gain, blue for amplicons), and the estimation of the signal value in each region, which is computed, for example, using smoothing techniques (Hupé *et al*., 2004). VAMP also allows the defining of the gain and loss regions by simply applying a threshold to the signal ratios. Examples of data analyses available within VAMP are given below and are described in more detail in the software documentation (http://bioinfo.curie.fr/vamp/doc).

Finding common alterations among a collection of CGH- array profiles. CGH array analysis principally consists in finding common regions of alterations, i.e. regions that are lost in many tumors. It is essential in these studies to distinguish between recurrent and random alterations. Recurrent alterations pinpoint regions involved in tumoral progression, whereas random alterations are simply the consequence of the general instability that affects the genome of a tumor. Among the recurrent alterations we distinguish the minimal regions and the recurrent regions. Minimal regions are extracted by intersecting the profiles of many tumors and looking for a sufficient number of alterations in the tumors (this parameter is set by the user) over the smallest possible region of the profile (Fig. 2). Tumoral progression obeys a selection principle, and it would be expected that the genes that need to be altered for a cell to become tumoral must be located in the smallest possible intersection of all alterations of a region. Recurrent regions are defined differently: in a given tumor, an alteration is bounded by two extremities, which can be a breakpoint or a chromosome end; when a sufficient number of tumors have the same extremities, these extremities define a recurrent region. We have implemented a linear algorithm that detects such minimal and recurrent regions, which is described in (Rouveirol *et al*., 2006). Gained regions appear in red in the main frame, and lost regions appear in green (Fig. 2). Amplicons (defined as gained regions with signal-ratio above a threshold typically equal to two) are colored in blue. The tumors that support a region of alteration may be optionally shadowed in the region, and for each region the user can sort these tumors.

Clustering profiles. Clustering is a general technique for unsupervised data classification widely used in microarray data analysis. A VAMP function offers the possibility to perform a hierarchical clustering (Kaufman and Rousseuw, 1990) on the profiles in the dot plot view. This can cluster genes and tumors from transcriptome arrays, or tumors from a CGH profile. In a CGH profile, the clustering uses the smoothed values of the CGH profile as variables and the Euclidean distance and Ward method for group distance computation. VAMP displays the results as a cluster view including a heat map and the trees resulting from the clustering algorithm (Fig. 4).

Comparing profiles. The Menu proposes several different data manipulation procedures for the profiles such as loading any type of

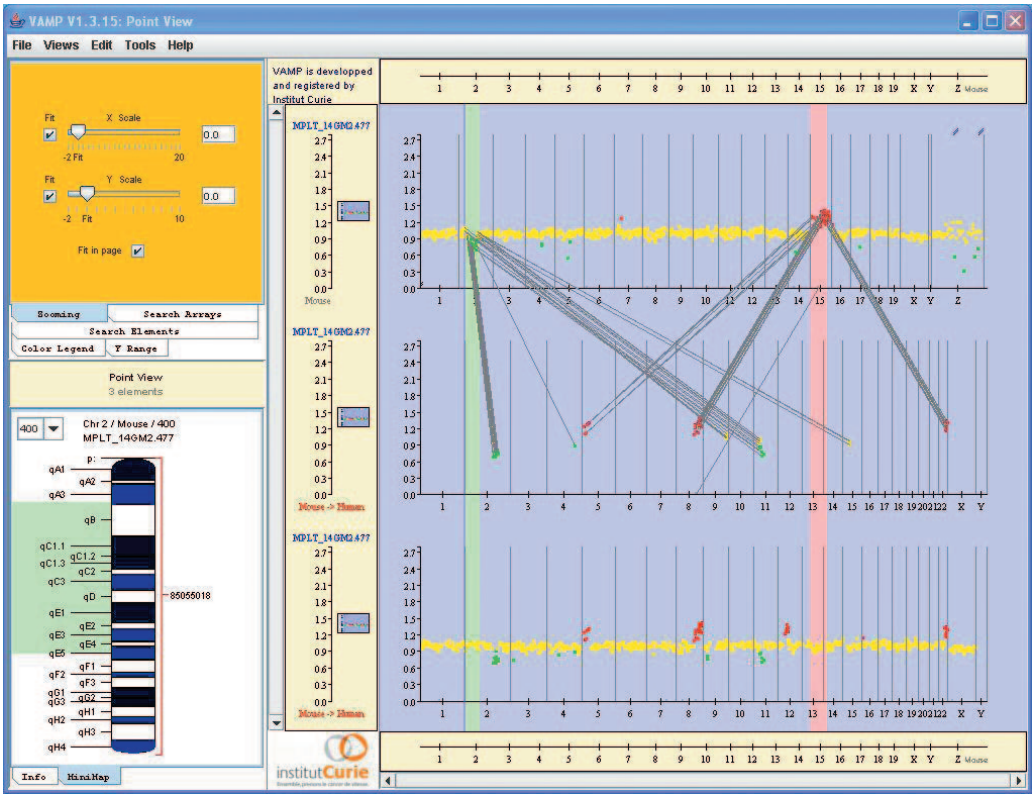


Fig. 5. Array-CGH profile for a mouse tumor (top) and its syntenic projection, i.e. a humanized array-CGH profile after mapping each mouse clone onto the human genome (bottom) and projection for two regions (middle profile) with resulting synteny relationships. Mapping is done from each clone of the mouse profile onto the location of the most similar sequence of the human genome. Mouse clones with ambiguous syntenic locations have not been mapped onto the human genome.

profile (CGH, expression, LOH, ChIP chip—an icon at the left of each profile shows the type of loaded profile) for a given sample (e.g. a typical application of VAMP is the simultaneous visualization of the DNA alterations and gene under- and over-expression in a region, Fig. 1); defining a profile as a reference and calculating the ratio of a profile to the reference (useful for one-color microarrays such as Affymetrix); averaging profiles; drawing marks (vertical bars) or regions (such as the green regions in Fig. 2) across all profiles (and simultaneously on the MiniMap); and many others.

Confrontation with sample annotation. Clinical data, or any other sample annotations, present in the XML files can be used for filtering tumors or for sorting them. This data can be visualized as color-coded bars in an annotation frame on the left of the profiles, and can be easily compared with a clustering result (Fig. 4).

Synteny analysis. VAMP can display the syntenic projection of a profile onto the genome of another species, in which that genome serves as a reference; a typical application is the projection of a mouse array-CGH profile onto the human genome (Fig. 5). In our case if an unambiguous syntenic locus was found, the mapping was done from each clone of the mouse profile onto the location of the most similar sequence of the human genome. The synteny relationships can be shown, for a selection of regions of the genome, as links from each clone of the profile to the location of the most similar sequence of the reference genome.

Other functions. The right mouse button brings up a menu with several actions associated to the clone/probe currently under the mouse pointer. These include: centering the profile around the current position; drawing of a vertical bar through all the profiles (to define a locus or a region); and linking to external web pages from NCBI clone or MapViewer (http://www.ncbi.nlm.nih.gov/mapview and Wheeler *et al*., 2005), UCSC Genome Browser (http://genome.ucsc.edu and Kent *et al*., 2002), Ensembl Contig View or CytoView (http://www.ensembl.org and Hubbard *et al*., 2005), *Saccharomyces* Genome Database (http://www.yeastgenome.org). New links are defined in a XML configuration file and adding them is straightforward. Most data and results (profiles, minimal regions, etc.) can be exported and saved in full text, csv (comma separated values) or HTML format. We refer the reader to the user manual for a description of the other functions.

2.4 Software architecture and requirements

The software architecture is shown in Figure 6. The core of the interface consists of a Java applet, and was developed using the Swing library. It runs on any operating system supporting Java 1.4.2

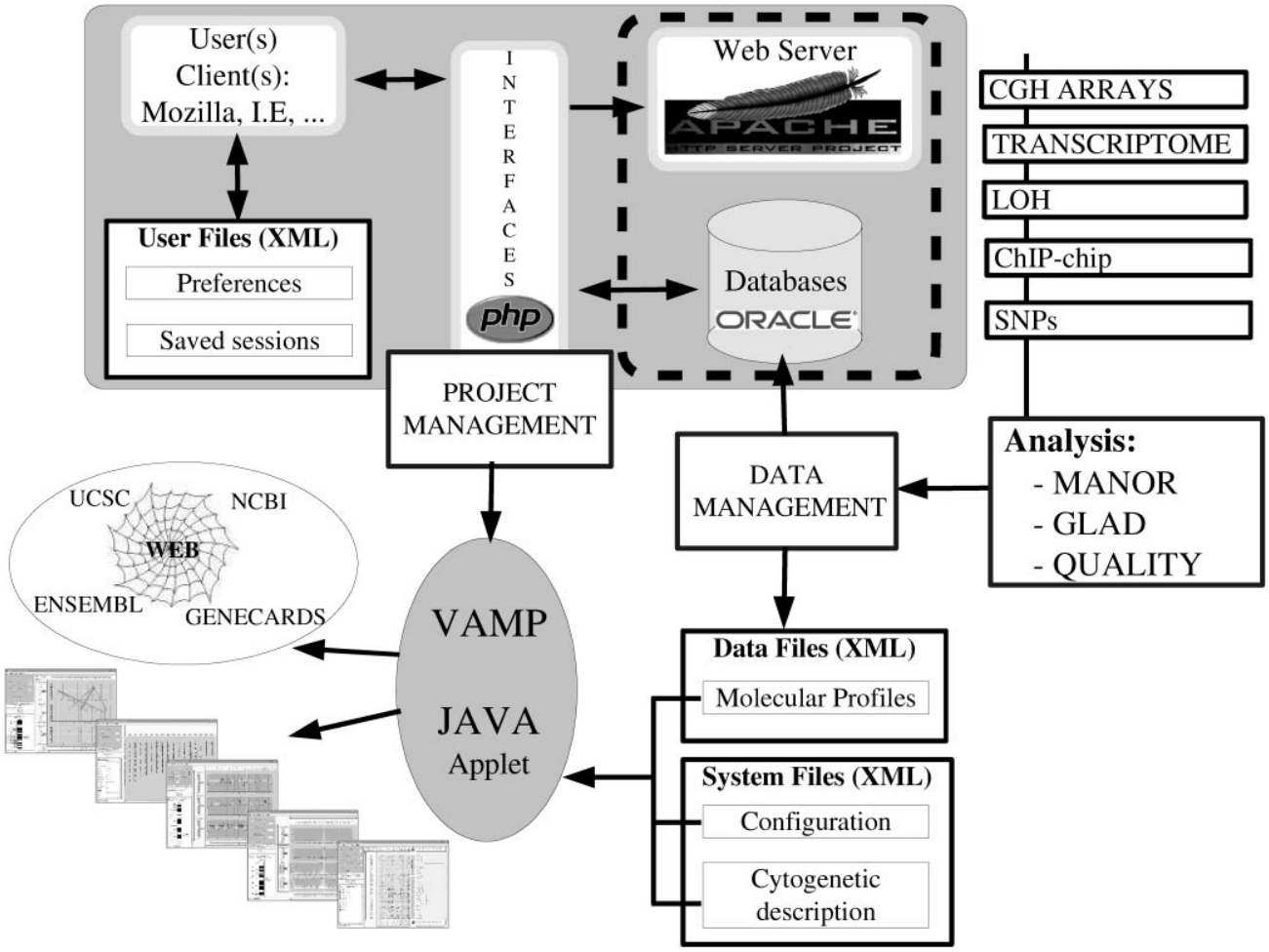


Fig. 6. Software architecture of a microarray environment based on VAMP. VAMP can also be used as a local application.

(we recommend computers with a minimum of 1 Gb memory, although 256 Mb is enough for small projects). The data used by the program are of several types:

- The genome profile information, which are retrieved either from a relational database management server (currently Oracle™) or from XML data files. These include the signal value for each clone/probe and its genomic location.
- The system files (also in XML), which includes the cytogenetic description of the genome under study and the configuration parameters (environment variables for file and URL management). Cytogenetic banding files for human ISCN 400, 550 and 850 descriptions, as well as mouse and yeast genome descriptions are also available. The user files, which consist of the user visualization preferences and saved sessions.

VAMP can be used either as a local application, with all data and configuration files directly accessible to the client, or as an applet, with all data and configuration files installed on a server. In this mode, only the user configuration file is stored locally on the client machine.

VAMP can be easily installed on any platform running Java 1.4.2. All that is needed is to convert the microarray data into XML files, with a specific syntax described in a DTD (XML Document Type Definition). The use of a database management server is not mandatory, although it is recommended for large-scale projects. Arbitrary complementary profile information can be added to the XML files, and this information can be displayed by the interface.

3 DISCUSSION

We have developed a graphical user interface for the visualization and analysis of any type of genomic profile, with an emphasis on array-CGH. VAMP is currently used in cancer genomic projects on human and mouse samples and in studying the proteins involved in the reparation, recombination and replication of DNA in yeast. It is used in Institut Curie and many labs in Europe and the United States. Several publications describing data analysis with VAMP are coming soon. Janoueix-Lerosey *et al*., (2005) describe the use of VAMP for replication timing data analysis (http://microarrays.curie.fr/publications/U509/reptiming). In Institut Curie, ∼3600 microarray profiles have been interfaced with VAMP to date.

VAMP aids greatly in finding genes of clinical and biological importance from CGH, transcriptome, LOH, ChIP chip profiles and SNP arrays. VAMP improves upon existing solutions such as SeeGH (Chi *et al*., 2004), CGHPRO (Chen *et al*., 2005), CGH-Analyzer (Margolin *et al*., 2005) or general purpose spreadsheet software, because it offers many different modes of visualization, allows the display of several samples and of several types of profiles simultaneously, and offers many data analysis functions. VAMP can be compared with other general-purpose genomic browsers such MapView (NCBI), Genome Browser of UCSC or Ensembl. VAMP is well suited to handle sample profiles and to analyse this type of data, which the other genomic browsers are not designed to do. Therefore, in cancer research it addresses a real need and is a useful tool for biologists and clinicians. Our software is fully portable and only requires a computer running Java 1.4.2 and data in XML format.

VAMP can be run on public datasets at http://bioinfo.curie.fr/vamp. The array-CGH data from Snijiders *et al*. (2001, 2005), Pollack *et al*. (2002), Veltman *et al*. (2003), Nakao *et al*. (2004), Douglas *et al*. (2004), de Leeuw *et al*. (2004), Gysin *et al*. (2005), Patil *et al*. (2005) and Bredel *et al*. (2005) are currently browsable. Expression profiles are also available for the samples from Pollack *et al*. (2002).

ACKNOWLEDGEMENTS

This work was supported by the Institut Curie, the Centre National de la Recherche Scientifique, the Institut National de la Santé et de la Recherche Médicale, the CNRG and the Ligue contre le Cancer.

Conflict of Interest: none declared.

REFERENCES

Albertson,D.G. *et al*. (2003) Chromosome aberrations in solid tumors. *Nat. Genet*., **34**, 369–76.

Bignell,G.R. *et al*. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res*., **14**, 287–295.

Bilke,S. *et al*. (2005) Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays. *Bioinformatics*, **21**, 1138–1145.

Bredel,M. *et al*. (2005) High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res*., **65**, 4088–4096.

Buck,M.J. and Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.

Chen,W. *et al*. (2005) CGHPRO—a comprehensive data analysis tool for array CGH. *BMC Bioinformatics*, **6**, 85.

Chi,B. *et al*. (2004) SeeGH—a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, **5**, 13.

de Leeuw,R.J. *et al*. (2004) Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum. Mol. Genet*., **13**, 1827–1837.

Douglas,E.J. *et al*. (2004) Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res*., **64**, 4817–4825.

Eilers,P.H.C. and de Menezes,R.X. (2005) Quantile smoothing of array CGH data. *Bioinformatics*, **21**, 1146–1153.

Fridlyand,J. *et al*. (2004) Application of hidden markov models to the analysis of the array CGH data. *J. Multivari. Anal*. (Special Issue on Multivariate Methods in Genomic Data Analysis), **90**, 132–153.

Gysin,S. *et al*. (2005) Analysis of genomic DNA alterations and mRNA expression patterns in a panel of human pancreatic cancer cell lines. *Genes Chromosomes Cancer*, **44**, 37–51.

Huang,J. *et al*. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics*, **1**, 287–299.

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, 447–453.

Hupé,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Ishkanian,A.S. *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet*., **36**, 299–303.

Janoueix-Lerosey,I. *et al*. (2005) Preferential occurrence of chromosome breakpoints within early replicating regions in neuroblastoma. *Cell Cycle*, **4**, 1842–1846.

Jong,K. *et al*. (2003) Chromosomal breakpoint detection in human cancer. In Raidl,G.R., Cagnoni,S., Cardalda,J.J.R., Corne,D.W., Gottlieb,J., Guillot,A., Hart,E., Johnson,C.G., Marchiori,E., Meyer,J.-A. and Middendorf,M. (eds), *Applications of Evolutionary Computing, EvoWorkshops2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, EvoSTIM*, vol. 2611 of *LNCS*. Springer-Verlag, University of Essex, England, UK.

Kaufman,L. and Rousseuw,P. (1990) *Finding Groups in Data—An Introduction to Cluster Analysis*, Wiley Series in Probability and Mathematical Sciences. John Wiley & Sons.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.

Margolin,A. *et al*. (2005) CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics*, **21**, 3308–3311.

Nakao,K. *et al*. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, **25**, 1345–1357.

Neuvial,P., Hupé,P., Brito,I., Liva,S., Manié,E., Brennetot,C., Radvanyi,F., Aurias,A. and Barillot,E. (2005) Spatial normalization of array-CGH data. *BMC Bioinformatics*, **7**, 264.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Patil,M.A. *et al*. (2005) Array-based comparative genomic hybridization reveals recurrent chromosomal aberrations and Jab1 as a potential target for 8q gain in hepatocellular carcinoma. *Carcinogenesis*, **26**, 2050–2057.

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., **37** (Suppl.1), 11–17.

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet*., **20**, 207–211.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Rouveirol,C. *et al*. (2005) Computation of recurrent minimal genomic alterations from CGH data. *Bioinformatics*, **22**, 849–856.

Snijders,A.M. *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet*., **29**, 263–4.

Snijders,A.M. *et al*. (2005) Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, **24**, 4232–4242.

Solinas-Toldo,S. *et al*. (1997) Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Veltman,J.A. *et al*. (2003) Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res*., **63**, 2872–2880.

Vogelstein,B. *et al*. (1989) Allelotype of colorectal carcinomas. *Science*, **244**, 207–11.

Wheeler,D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, 39–45.

ELECTRONIC JOURNAL OF ONCOLOGY
© John Libbey Eurotext
Bull Cancer 2006 ; 93 (8) : E81-9

Challenges in the stratification of breast tumors for tailored therapies

J.-P. Thiery
X. Sastre-Garau
B. Vincent-Salomon
X. Sigal-Zafrani
J.Y. Pierga
C. Decraene
J.P. Meyniel
E. Gravier
B. Asselain
Y. De Rycke
P. Hupe
E. Barillot
S. Ajaz
M. Faraldo
M.A. Deugnier
M. Glukhova
D. Medina,
and the Breast Cancer Group*

Institut Curie, 26, rue d'Ulm,
75248 Paris Cedex 05
<jean-paul.thierry@curie.fr>

Abstract. Studying the molecular stratification of breast carcinoma is a real challenge considering the extreme heterogeneity of these tumors. Many patients are now treated following recommendation established at several NIH and St Gallen consensus conferences. However a significant fraction of these breast cancer patients do not need adjuvant chemotherapies while other patients receive inefficacious therapies. High density gene expression arrays have been designed to attempt to establish expression profiles that could be used as prognostic indicators or as predictive markers for response to treatment. This review is intended to discuss the potential value of these new indicators, but also the current weaknesses of these new genomic and bioinformatic approaches. The combined analysis of transcriptomic and genomic alteration data from relatively large numbers of well annotated tumor specimens may offer an opportunity to overcome the current difficulties in validating recently published non overlapping gene lists as prognostic or therapeutic indicators. There is also hope for identifying and deciphering signal transduction pathways driving tumor progression with newly developed algorithms and semi quantitative parameters obtained in simplified *in vitro* or *in vivo* models for specific transduction pathways. ▲

Keywords : breast cancer, molecular stratification

The number of cancers is increasing steadily in Western countries in relation with aging and with changes in societal behaviors. The conventional therapies, including surgery and radiotherapy, have benefited from major technical advances. Chemotherapy has improved with the discovery of more efficacious molecules, adjusted schedules of administration and better drug combinations. Targeted therapies are already contributing to longer relapse-free survival. These targeted therapies include hormonotherapy, kinase inhibitors and monoclonal antibodies to specific receptors. The therapeutic protocols are based on more stringent clinico-histopathological criteria. However, it is well known that only a fraction of patients will respond to these therapies. The quality of response can be evaluated more readily in the neo-adjuvant setting than in adjuvant therapies following surgery. It is now clear that current stratification methods are still relatively inadequate to define precise prognosis and to predict response to treatment. Current strategies are therefore aimed at establishing more accurate methods for the stratification of cancers and for tailored therapies. However, this goal is a major challenge still facing numerous difficulties. Advanced oncogenomic approaches offer new tools to improve stratification and to discover predictors of response to treatment. Over the last few years, an increasing number of publications have shown that small groups of identifier genes could be used to determine the molecular status of tumours. In this review we shall discuss some of the potential value of these new findings in breast cancers, albeit stressing some crucial issues to be solved. We shall also discuss whether the development of preclinical models based on transgenic mice can provide new insights into the molecular complexities of cancer focusing on breast cancers.

Classification of breast cancers

Breast cancers comprise very heterogeneous diseases, which are imperfectly described by histopathological and clinical parameters. Although ductal invasive carcinomas is the most frequently encountered histological type, other entities such as ductal in situ carcinomas are becoming increasingly frequent in westernised countries due, in part, to earlier detection of the disease. Ductal carcinomas *in situ* encompasses different histological types including comedo, cribriform and papillary. They can also be stratified as low or high nuclear grade carcinoma. Almost 80% of high-grade in situ carcinoma are characterized by overexpression of HER2. Invasive lobular carcinomas is also a well-defined histological entity; the lack of expression of E-cadherin is a hallmark of lobular cancers as a result of mutations in a large proportion of these

* The Breast cancer group: B. Asselain, A. Aurias, E. Barillot, F. Campana, P. de Crémoux, V. Diéras, O. Delattre, A. Fourquet, M.-F. Poupon, F. Radvanyi, J.-Y. Pierga, L. Mignot, R. Salmon, A. Salomon, B. Sigal-Zafrani, D. Stoppa Lyonnet, A. Tardivon, F. Thibault, J.-P. Thiery, P. This.

tumours. Invasive carcinomas also include histological variants such as mucinous, tubular, medullary and papillary carcinoma. Most mucinous carcinomas have good outcome. Medullary carcinomas are poorly differentiated ductal carcinoma, which present frequently an unexpected favorable outcome, probably related to a high sensitivity of the tumour to chemo- or radio-therapy. These tumours are characterized by a high frequency of p53 mutations and by the lack of oestrogen and progesterone receptors. Interestingly, they exhibit a luminal-basal like mixed phenotype. Micropapillary carcinomas have an unfavorable prognosis often giving rise to microembolisms and lymph node metastasis. Mixed phenotypes are often encountered; for instance, a large number of ductal invasive carcinomas contain foci of in situ carcinomas and in some cases lobular carcinomas. There is no definitive data permitting one to establish tumor progression scenarios for these different entities. Atypical hyperplasia can give rise to ductal *in situ* carcinoma and ductal in situ carcinoma can develop into ductal invasive carcinoma. The classical distinction between ductal and lobular carcinoma does not provide faithful indication for the cell types at the origin of the carcinoma. Tumour heterogeneity is a hallmark of all tumours, particularly of breast carcinomas. In addition to the cellular heterogeneity of the carcinoma cells, tumours contain a variable proportion of stromal cells including endothelial cells, myofibroblasts, lymphocytes and macrophages. Staging and grading is a critical step to evaluate the wide range in extension and dedifferentiation of breast carcinoma diseases.

Staging and grading

Staging

The stage describes the extension of the tumour locally or at a distance from the primary site. According to the International Union against Cancer (UICC), staging should be based on clinical characteristics at the time of diagnosis, including the size of the tumour (T), the status of loco-regional lymph nodes (N) and the presence of metastasis (M). Pathological characteristics of the tumour, determined following surgery, provide additional and more precise definition of the T, called pT, and status of lymph nodes, pN. Each tumour site displays particular characteristics depending on the pattern of local invasion and metastasis. A revised classification was proposed by the American Joint Committee on staging system for breast cancer [1].

– stage I corresponds to T1N0M0 tumours (invasive tumour measuring less or equal to 2 cm) with no lymph node involvement and no metastasis ;

– stage II includes from T1T2 N1 (ipsilateral lymph node involvement) M0 tumors to T3 (over 5 cm) N0 M0 tumors ;

– stage III is even more heterogenous, comprising T0T1N2M0 and T4N0 or T4N1 or T4N2M0, and any T N3M0. T4 is a tumour with extension to chest wall or skin; N2 corresponds to involvement of ipsilateral axillary lymph nodes fixed or mated or of internal mammary nodes and N3 includes ipsilateral infraclavicular lymph node involvement in addition to axillary lymph node metastasis ;stage IV tumors include any T any N and M1 (for distant metastasis).

Grading

The grade reflects the morphology and proliferative capacity of the primary tumour. The microscopic analysis of tumour samples has two major objectives: first, to establish a diagnosis of cancer and determine the type of cancer; second, to produce a histoprognostic index, which is used to determine the type of treatment. For breast cancer, the histoprognostic index is based on analysis of three criteria [2]. First, architecture, measuring the degree of differentiation, is ranked 1 to 3; the score 3 describes tumours with less than 10% glandular structure; second, anisokaryosis (variation in size of the nucleus) is ranked 1 to 3; and third, proliferation is assessed by the mitotic index (number of mitoses per 10 microscopic fields). Tumours are defined as grade 1 when the combined score is between 3 and 6, grade 2 for 6-7 and grade 3 for 8 or 9.

Classical diagnostic and prognostic markers

There are few classical markers routinely used in breast carcinoma. The presence of oestrogen and progesterone receptors is determined mostly by immunohistochemical methods. Tumours are classified as positive if at least 1-10% of the carcinoma cells are labelled [3]. Approximately 80% of breast carcinomas are oestrogen receptor (ER) positive. Of these, about 75% demonstrate hormone responsiveness. Level of expression of HER2 is also critical since these patients can potentially benefit from trastuzumab-based immunotherapy, in conjunction with a cytotoxic drug. Between 15 an 25% of breast carcinomas are HER positive, namely, ovexpressing HER2 often as a result of gene amplification at the HER2 locus.

Micrometastasis

Carcinoma cells can disseminate through the lymph and blood vessels associated with the tumour bed. Lymph node involvement is an important aspect of staging. However, the routine examination of lymph node does not involve the search for micrometastatic invasion. There is an increasing interest to determine as early as possible minimal distant dissemination of carcinoma cells. The search for micrometastatic tumour cells in lymph node is now benefiting from the sentinel lymph node technique [4]. This technical approach is applied principally for T1 tumours as an alternative procedure to initial axillary lymph node dissection. Vital dye and/or radioactive colloid are injected in the peritumoral space prior to surgery. The first lymph node(s) draining the tumours (sentinel lymph node) is easily detected and excised. The identification of tumour-free sentinel lymph node avoids routine axillary dissection, reducing thus morbidity and cost. The prognostic value of a few carcinoma cells detected only by immunohistochemistry is still a matter of debate. Cancer cells have also been detected in the blood of patients. Blood-born tumour cells are more easily detected in patients carrying relatively large primary tumour masses. The detection of these cells may be particularly important as a surrogate marker to predict the response to systemic treatments. A recent study has shown that the search for circulating carcinoma cells in blood can be facilitated by an automated detection of cytokeratin positive cells following immunomagnetic enrichment using an anti Ep-CAM antibody. Patients with more than 5 circulating carcinoma cells per 7.5 ml of blood have shorter median and overall survival. Beneficial treatment is suggested for a fraction of patients who showed less circulating tumour cells following initiation of chemotherapy [5]. Although the detection of carcinoma cells in blood is

much less demanding and better accepted by patients, it is clearly less sensitive than detection in the bone marrow [6]. Rare tumour cells have been detected in the bone marrow medulla, the only other accessible site for most carcinoma. Many studies have confirmed that immunocytochemical techniques involving anticytokeratin antibodies provide relatively reproducible results [7]. The data obtained to date clearly show that 5-25% of patients carrying T1 tumours (≤ 2 cm) already have disseminated cancer cells in the bone marrow (frequency of 1 per 106 mononucleated cells). The presence of micrometastatic tumour cells provides a novel independent prognostic indicator for recurrence and survival [8]. A European consortium has recently analysed data pooled from 4703 patients and confirmed the crucial importance of bone marrow micrometastasis for prognosis in multivariate analysis [9]. Bone marrow culture may in some instance reveal the presence of micrometastatic tumors cells not detected by direct immunocytochemistry on bone marrow samples [10]. Methods for enrichment of micrometastatic cells are urgently needed to improve this diagnostic since these methods are susceptible to artefacts resulting from the capture of non-carcinomatous Ep-CAM positive cells [11]. Very interestingly, CGH of chromosomes from micrometastatic cells isolated in M0 patients revealed many less alterations than those obtained from M1 patients, indicating the bone marrow micrometastasis can occur at early stages of tumour progression challenging the dogma that micrometastatic cells are derived from the most advanced primary tumour foci [12].

High density molecular profiling

Genomic alterations

The genome of breast carcinoma is remarkably unstable, possibly as a result of early dysfunction of DNA replication, repair and recombination machineries. Numerous chromosomal aberrations have already been extensively described by classical cytogenetic approaches. The comparative genomic hybridisation technique, and more recently the high density arrays, have revealed an extraordinary complexity of genomic alterations. A provisional list established in 2003 has described the frequency of loss and gain on each chromosome [13]. Remarkably, low level gains on chromosomes are more frequent than losses and amplification of loci. There is a high proportion of loci that can be affected in both directions, losses or gain. LOH and CGH studies are not providing overlapping results, suggesting that LOH events are not describing the behaviour of individual genes but rather variably large regions. Poor outcome correlates with distinct patterns of alteration as seen by LOH and CGH studies. Amplification of specific loci already allows one to define subgroups of breast carcinomas. Aside from the well described gene amplification at the HER2 locus, comprising 7 genes, other loci including CCND1, MDM2, MYC and EGFR have been characterised. A recent study using fluorescent *in situ* hybridisation on tissue arrays of more than 2000 breast carcinoma specimens showed that co-amplifications are more prevalent than previously described [14]. For instance, almost 30% of CCND1-amplified tumours harbour other amplicons. More strikingly CCND1 amplification was observed in 43% of HER2 amplified tumours and in 56% of MDM2 amplified tumours. A CGH array with selected BAC encoding major regions of interest in breast cancer was used to screen a limited number of advanced breast carcinoma. This study identified relatively frequent amplicons coding for 112 candidate genes; out of these, 44 were validated [15]. Recently, a basal-like phenotype was found with a subset of ductal invasive breast by high-density arrays for molecular profiling of transcripts. This phenotype had already been identified by immunohistochemical characterisation of cytokeratins almost 20 years ago. The CGH analysis of microdissected tumour cells from grade 3 basal CK14-positive and negative tumours revealed that the majority of basal-like tumours have significantly less genomic alterations than the CK14-negative grade 3 tumours. Hierarchical clustering identified a subgroup which contains 40% of basal-like tumours which had a worse prognosis than the other basal-like tumours [16]. This study exemplifies the difficulty in stratifying breast carcinoma even in the case of a relatively well defined molecular entity by using only cytokeratins immunocytochemistry. Refined genomic and transcriptomic approaches can detect heterogeneity in an otherwise fairly homogenous ER, PR and HER2 negative group.

With the advent of new high density arrays which can scan the genome at much higher definition, such as BAC arrays with more than 30,000 clones, long oligonucleotide and SNP arrays, one can expect to see many more alterations. These new data will require new software for signal analysis and precise determination of affected loci. In this respect, an algorithm was developed to detect breakpoints and outliers, and to assign a status to each loci from array CGH data [17]. This software has also been adapted to carry out the same analysis on SNP data.

Promising data will emerge from studies aimed at defining tumour evolution in breast tumours. One crucial issue is to determine to what extent ductal in situ gives rise to ductal invasive carcinoma and lobular *in situ* carcinoma leads to lobular invasive carcinoma. A similar issue concerns local regional relapses; to what extent are they clonally derived from the primary tumour? One pilot study based on a 2400 BAC clone CGH array showed that a majority of synchronous lobular *in situ* and lobular invasive carcinoma are clonally related [18].

Point mutations

Recently, a major effort has been deployed to sequence gene candidates from breast tumour lines and breast carcinoma specimens. The data are compiled and published regularly by the Sanger centre (cosmic database; The Sanger Institute: catalogue of somatic mutations). The p53 protein is mutated in 20-40% of breast cancers (see http:oewww-p53.iarc.fr/index.html). Recent studies reveal that PI3K is mutated in more than 25% of breast cancers. The mutations are frequently found in the catalytic site. Pioneer studies with limited number of samples could not show correlation with anatomoclinical data [19-22]. Studies with larger number of patients show correlation with the oestrogen receptor, lymph node and HER2 status [23, 24]. In addition, mutation in PI3K and the loss of PTEN are mutually exclusive [24]. Two activating mutations in PI3K have been shown to transform normal mammary epithelial cells suggesting that such mutations could contribute to tumour progression [25]. Mutations are also relatively frequent in CDKN2A. An extensive screen has been performed recently to search for mutations in the kinase gene superfamily. This study, carried out in a limited number of breast carcinoma, shows that only a few tumours

accumulated mutations in a large number of kinases while most other tumours do not carry any mutations [26].

Transcriptomics

High density RNA profiling became possible with the advent of new technological developments, including array spotters, radio-labelled or fluorescent nucleotides, and phosphoro-imagers or sensitive laser-based scanners. The first studies showed the great utility of high density molecular profiling of tumours. The first series of breast carcinomas analysed by the Stanford group who pioneered cDNA arrays showed that distinct patterns could be established for individual tumours and that tumours analysed before and after chemotherapy resembled each other more than tumours coming from other patients. Lymph node metastasis profiles were also more closely related to their primary tumour profiles than to those of other tumours [27]. Subsequent studies using the same technology revealed a new molecular taxonomy for breast cancers. One major ER negative cluster contains HER2 positive, basal-like and normal breast-like tumours. The ER positive cluster can be subdivided into three distinct luminal A, B, C subtypes [28]. Most remarkably, the newly identified basal phenotype is associated with shorter survival times similar to the amplified HER2 group. The ER positive luminal B and C subtypes also showed poorer prognosis than the luminal A subtype. These findings were confirmed in another study showing that the luminal A and B, the normal-like, basal-like and HER2 phenotypes were found in two independent sets of data with similar prognostic values to the previous study. Interestingly, a large fraction of the BRCA1 tumours exhibit a basal-like phenotype [29]. Immunohistochemical approaches can be applied with a limited number of markers to identify about 75% of the basal-like tumours. This study showed that a subset of basal tumours exhibited an HER1 overexpression as compared to other tumour types. This simple approach stratified ER and HER2 negative tumours using only cytokeratins 5/6 and 17, HER1 and c-Kit. The relative frequency of the different subtypes in a large group of specimens was 15% for basal-like, 23% for HER2 positive and 40% for ER-positive tumours, 22% of the tumours could not be classified [30].

Surprisingly, RNA profiling studies of premalignant *in situ* and invasive carcinoma revealed similar profiles, suggesting that global gene alteration patterns are already acquired in atypical ductal hyperplasia. Differences were, however, found between different stages and subtle differences were found between in situ and invasive forms [31].

RNA profiling can also be used to search for differential gene expression in well defined histological entities such as lobular and ductal invasive carcinoma. It can also provide the information for the construction of class predictors, in the so-called supervised classification analysis. Supervised classification based on gene expression identified a limited list of genes that can classify accurately lobular and ductal invasive carcinoma. Some of the genes may indicate distinct molecular pathways for local invasion [32]. RNA profiling was also used to define poor prognosis gene signatures. A pioneering study identified a list of 70 genes that can predict relapse within 5 years of diagnosis in patients with node negative T1T2 tumours less than 55 years old [33]. In a second study, lymph node negative and positive tumours were analysed to evaluate the predictive power of the 70-gene signature. This gene signature was found to be more powerful than prognosis based on anatomical/clinical conventional criteria adopted in consensus conferences in St-Gallen or at NIH [34]. A prognostic score also could be given by a wound-response gene expression signature, since wound response is a biological hallmark of tumour progression. The integration of the 70 gene signature with the wound signature in a decision tree improved significantly the stratification of patients at high risk of metastasis [35].

A 17-gene pan-metastatic signature was found to be shared by different types of adenocarcinomas, possibly suggesting that the metastatic potential is encoded in the primary tumour and not by a small subset of carcinoma cells undergoing a Darwinian type selection throughout progression [36]. A similar conclusion was reached by comparing RNA profiling of a limited number of primary and matched metastatic breast cancer tumours [37]. An extensive study was recently carried out on a large collection containing mostly T1T2N0 tumours from patients who had not received adjuvant chemotherapy. A 76-gene signature was identified with good sensitivity but moderate specificity on a validation set. A 5.5 hazard ratio was obtained in multivariate analysis as compared to 2.6 for stage II and III versus stage I, demonstrating the potential value of this new signature. This signature shared only three genes in common with the van't Veer signature [38].

Most of the studies so far have used different algorithmic and biostatistical approaches to find a group of genes whose expression profile predicts disease progression. Some studies are based on the use of metagenes, *i.e.* a group of genes behaving similarly are first identified to construct a decision tree in a Bayesian approach. These studies, combining clinical and genomic data, allow to establish probability predictions of lymph node status and recurrences with a predictive accuracy of 90% [39, 40].

Considering the formidable heterogeneity of breast tumours, it is not surprising that multiple gene expression prognostic signatures have been found so far. There is an advantage to establishing breast cancer gene signatures in clinically more homogenous cohorts such as ER and age status, two well established prognostic parameters. The van't Veer cohort was analysed using these criteria and new algorithms modifying the training set to eliminate those patients that were not correctly classified during a cross-validation procedure led to the definition of a new 50-gene signature [41]. This gene signature was more clearly focusing on one pathway than previous signatures. In this set of selected genes, overexpression of the cell cycle associated genes were clearly identifying the poor prognosis group. It is indeed a valuable approach to determine signatures associated with a potentially dominating pathway.

A signature related to p53 status was recently published and outperformed the stratification established on p53 sequence data. The 32-gene signature was able to identify two major groups of patients defined as p53 wild type and p53 mutated. The two groups contain a small proportion of patient whose p53 status did not fit with their group status. However, these misclassified tumours were most likely correctly assigned for their p53 functionality. For instance, tumours with low wild type p53 expression may behave like mutated p53 tumours [42]. Organ-specific metastasis is a long debated issue since the pioneering work of Stephen Paget. The molecular profiling of the MDA MB 231 pleural effusion metastatic cell line, selected for its ability to uniquely metastasise to bone, showed that a small set of genes was associated with organ

specificity [43]. This set of genes differs from those conferring a general poor prognosis included in the original 70-gene signature [33]. This list of genes has been tested on a cohort of breast carcinoma showing the possibility of identifying the tumours which will metastasise to bone. A similar study has been reported to define metastasis to lung [44]. These signatures need to be validated on a much larger cohort in order to determine whether these organ-specific signatures remain valid for metastases occurring at multiple sites.

Weaknesses in the transcriptomic approach

The rapidly increasing number of non-overlapping lists of genes selected for prognostic purposes and for prediction of response to treatment by different teams has already prompted several studies to identify the origin of these discrepancies. These issues have been discussed in a recent review [45]. The data from seven studies comprising lung, breast, hepatocellular carcinoma, medulloblastoma, non-Hodgkin's lymphoma and acute lymphocytic leukaemia were reanalysed by creating multiple random training sets to study the stability of the molecular signatures and the proportion of misclassification. The genes selected for prognosis are crucially dependent on the choice of patients included in the training set. Clearly, the proportion of misclassified patients in the validating step decreases when the number of patients was increased in the training set. Most of these studies could not prove that they perform better than random [46]. Another study [47] showed that 50 patients is clearly a minimum for a training set to achieve some significance, but a few hundreds are required to build a clinically useful predictor. The 70-gene list for prognosis of breast cancer metastasis [33] was also analysed independently to evaluate its robustness. One important finding is that many genes are correlated with survival but the differences in their correlation coefficients are small and the correlation fluctuates strongly when the set of patients is even partially modified, probably because of the high heterogeneity of the disease [48]. The conclusion from these studies is that gene signatures derived from high density microarrays are not unique and not necessarily easily reproducible from one platform to another platform [49]. However it is very likely that the main cause of this lack of robustness is linked to tumour heterogeneity and relatively poor quality of RNA preparation in a fraction of the samples, due in part to inadequate collection and preservation procedures. To circumvent this major difficulty, a very large number of high-quality samples, selected on histoprognostic and immunohistochemical criteria, are required to diminish heterogeneity. Laser microdissection has been utilised by several teams for such studies; however, this approach also suffers from a number of drawbacks including the preparation of a reasonable quantity of high quality RNA to avoid two amplification steps. Better methods need to be developed for RNA preparation from formalin-fixed paraffin-embedded specimens. Multiplex PCR may overcome these difficulties, especially for the new major clinical trials aimed at defining the best multiparametric histological and molecular markers for prognosis or response to treatment [50].

Prediction of response to treatment

Gene classifiers

Molecular profiling is now thought to provide indicators which will replace or complement the standard markers such as stage, grade and HER2 and ER status. The surrogate markers used in the neo-adjuvant setting for pathologic complete response, in comparison to partial response, stable disease and tumour progression, have proven useful to establish a limited list of gene predictors. Pathological complete response is certainly correlated with lower risk of relapse and death, but it is in no way a perfect surrogate for cure. In reality, the response is rather a continuum than a very discrete entity, which renders difficult supervised analyses [51]. In one study, a 74-gene predictor was shown to identify non-responders with an overall accuracy of 78%, but recognized only three out of seven complete responders. However this gene set was established using a limited number of tumours in the training cohort, possibly not including other genes that could identify complete responders in the validating set [52].

The quality of response to paclitaxel followed by 5-fluorouracil, doxorubicin and cyclophosphamide chemotherapy was evaluated using the molecular stratification described above. Very interestingly, the basal-like and HER groups responded much better than the luminal subtypes; the normal-like type had almost no response. Noticeably, the gene predictors for the basal subgroup were not overlapping with those predicting response in the HER2 group strongly suggesting different mechanisms mediating response in the two ER negative tumor types [53].

The response to docetaxel was evaluated in a limited number of core biopsy samples from breast cancer patients undergoing neo-adjuvant therapy. A 92-gene predictor list was able to classify, with 90% specificity and 85% sensitivity, in a leave-one-out validation procedure [54, 55]. A complementary study showed that residual tumour profiling was very similar in each case and resembled that of the initially fully resistant tumours. These results show that some specific transduction pathways could confer sensitivity to docetaxel such as stress-related DNA damage and apoptosis, while cell cycle arrest and survival confer resistance. However, in another study with a small number of patients, no specific gene expression profile was identified for response to doxorubicin-cyclophosphamide or doxorubicin-doxetaxel, which advocated for larger cohorts [56].

The search for predictors of response to treatment is currently being studied in different laboratories. A collection of 60 ER-positive tumours was analysed to identify differentially expressed genes between responders and non responders to tamoxifen as a monotherapy following primary surgery. HOXB13 and IL17BR mRNA levels, determined by semi-quantitative PCR, are sufficient to predict outcome in an independent set of samples. Interestingly, increased HOXB3 was observed in non-responding tumours. In vitro constitutive expression of this gene confers motile and invasive properties to the MCF10A mammary cell line. HOXB3 interferes in the control of ER signalling by an unknown mechanism, as is the case for EGFR and HER2 signalling, which are known to alter the response to tamoxifen [57]. This important finding, however, was not validated on an independent collection of tumours [58] stressing the crucial importance of analysing very large and more homogenous cohorts of tumours.

A 64-gene signature distinguishing good and poor prognosis was established on a training set comprising node-negative and node-positive patients who did or did not receive adjuvant therapy. This set of genes was complemented by a risk factor score. The training set showed that the patients could

be subdivided in three clusters; the first cluster contained mostly patients who did well without treatment and the third cluster corresponded to patients who did poorly with treatment but may benefit from other protocols. However, the second cluster was not informative. It was not identifying a group of patients who did poorly without treatment and who, therefore, could have benefited from treatment. This study was potentially aimed at determining which patients could escape systemic chemotherapy and which patients could be treated with an alternative therapeutic protocol to overcome failure from the conventional treatment [59].

Resistance to trastuzumab is encountered relatively frequently; however, the mechanism by which this resistance is acquired remains unknown. One recent study has addressed this issue by establishing a carcinoma cell line from a patient resistant to trastuzumab. This cell line shares many characteristics of the primary tumour; although it has an amplified HER2 locus, this cell line has a mixed basal and luminal phenotype. This tumour type is, therefore, atypical since the HER2 cluster is mostly of the luminal phenotype. The lack of inhibition of AKT phosphorylation by trastuzumab is so far the only detected alteration in signalling. However, PI3K inhibitors have not been used in this study to determine whether resistance to trastuzumab can be overcome [60]. Resistance may also be acquired through a steric hindrance mechanism mediated by MUC4 at the cell surface. Diminished expression of MUC by RNA interference resulted in increased binding to trastuzumab, potentially abrogating resistance to this therapy [61].

Defining resistance

Multidrug resistance is a well know phenomenon applied to most tumour tissues. Numerous studies have addressed mechanisms driving this resistance. The RNA profiling of the 48 ABC transporters, established by PCR in the NCI cell line collection, compared the ability to respond to a panel of 1429 drugs [62] in a much better correlated manner than a previous study based on expression profile of 9000 transcripts [63]. A surprising result was that MDR1 (ABCB1) overexpression potentiated the cytotoxic activity of some drugs rather than resistance. This study opens a new strategy to overcome drug resistance in a more rational way.

Murine models

Numerous transgenic murine models of breast carcinoma have now been developed through the targeting of oncogenes, mostly using the MMTV or WAP promoters. The two promoters are specifically expressed in the luminal epithelium, but the MMTV promoter is also expressed in some other epithelia and is expressed at an early stage in mammary gland development, prior to the terminal differentiation into secretory cells. A major effort has been devoted to classify precisely the proliferative lesions [64]. Most tumours forming in genetically engineered mice are morphologically distinct from spontaneous MMTV or chemically induced tumours. Many of these tumours in genetically engineered models are not closely related to human breast tumours as they exhibit squamous metaplasia. However, they have been extremely useful to assess the role of known oncogenes. Tumours induced by each oncogene have a specific morphological and molecular signature as revealed by a recent study of KRAS2 expression signature in mouse and human lung cancers [65]. For instance, HER2 tumours are composed of solid sheets of carcinoma cells without glandular differentiation. The c-MYC expressing tumours have large cells with pleiomorphic nuclei with a coarse chromatin and prominent nucleoli. RAS tumours form papillary-like tumours resembling transitional cell carcinoma of the bladder. The Ret 1 tumours form small crowded glands with large pleiomorphic nuclei. HER2 and SV40Tag transgenes can produce ductal carcinoma *in situ* of the comedo-type resembling human tumours. Papillary carcinoma can be obtained with the cyclin D1 transgene.

The phenotype of multigenic transgene derived tumours is often determined by the dominating oncogene such as c-MYC. Much care should be paid to the genetic background of the mouse and different phenotypes are obtained with MMTV or WAP promoters. Human and mouse tumours differ notably with respect to their relative sensitivity to hormones, their stroma, their capacity to metastasise and their pattern of metastasis. Using terminal differentiation markers, luminal myoepithelial and mesenchymal phenotypes have been identified in a large variety of mouse tumors. Three types of neoplasms have been described; simple carcinoma, complex carcinoma possibly originating from a stem cell, and carcinoma undergoing an epithelial-mesenchymal transition (EMT). Remarkably, an EMT phenotype [66, 67] has been described in c-MYC, RAS and SV40 Tag driven tumours [68]. Mammary epithelial cells expressing Met and Myc can develop into tumours mixed luminal and myoepithelial, when transplanted in the mammary fat pad suggesting that these tumours arose from a bipotent progenitor [69].

Recently, the analysis of an epithelial cell line derived from the mouse mammary gland taken at the mid gestation stage showed remarkable epithelial cell plasticity. When deprived from EGF, these mammary epithelial cells acquired a fibroblastic phenotype and expressed characteristic markers of the basal phenotype such as K5/14 and P-cadherin [70]. Their injection in vivo in the cleared mammary fat pad clearly showed their capacity to produce luminal cells. These findings indicating that these basal cells display progenitor properties together with the demonstration that the Wnt/b-catenin pathway is playing a crucial role in the maintenance of progenitor cells in different epithelia prompted experiments to target a truncated β-catenin (resulting in constitutive activation of the pathway) into the basal myoepithelial layer. The myoepithelial layer has not prompted as many studies as the luminal layer and many less transgenic mice have been targeted to this basal layer. As the myoepithelial layer is in direct contact with the extracellular environment and interacts also directly with the luminal layer, any alteration of these cells could direct consequences for the luminal layer. Its disappearance in carcinomas suggests a direct role in controlling invasive behaviour of DCIS [71]. The truncated b-catenin transgene induced excessive lateral branching and precocious lobulo-alveolar development of the mammary gland at mid-gestation. Most interestingly, hyperplastic foci were observed in the basal layer. These cells expressed basal cytokeratins, but not smooth muscle a-actin, indicating their undifferentiated state. Multiparous mice also exhibited squamous carcinoma and most importantly invasive carcinoma with a strong basal phenotype [72]. These transgenic mice potentially represent a useful model to study breast carcinoma of the basal phenotypes. Moreover the formation of undifferentiated basal tumors can be interpreted as the amplification of a

population of basal-type mammary progenitors. The mammary gland is hypothesised to contain one epithelial stem/progenitor cell in every 2000 cells [73]. Several studies have described attempts to isolate the stem/progenitors cells in the mouse using various approaches [74]. These cells may be evidenced *in vivo* as a subpopulation of BrdU long-term label retaining epithelial cells in the mouse and human mammary tissues. Mammary epithelial cells belonging to the so-called long-term label retaining cells are found to divide asymmetrically and to retain their template strand. These cells also self renew, thus they may represent the mammary stem cells [75]. Other studies using surface markers have shown that Sca-1 positive epithelial cells from the mouse mammary gland had a much higher regenerative potential in vivo than the Sca-1 negative cells [76]. Very recently, two studies reported the isolation of a cell population from the mouse mammary epithelium that is able, at the clonal level, to give rise to the entire mammary gland if transplanted *in vivo*. Interestingly, these progenitor cells were characterized by high surface levels of integrins and cytoskeletal markers of basal epithelial cells [77, 78].

Cancer stem cells

The presence of cancer stem cells has long been hypothesized in solid tumours. Strong evidence was already provided in the seventies through the analysis of teratocarcinomas [79]. The enrichment of metastatic breast carcinoma cells derived from pleural effusions using CD44 and CD24 as sorting criteria showed that as low as 100 cells could form a malignant tumour [80]. This pioneering study opened the road for new investigations in cancer stem cells to further enrich and characterise the phenotype and response to treatment. Breast cancer stem cell research is at very early stages and many issues remain unsolved. The isolation of stem/progenitor cells capable of self-renewing was successfully achieved with a few primary high grade, ER positive tumours specimens. As low as 100 tumour cells could form a tumour in SCID mice. The phenotype of these cells was CD44-positive, CD24-negative, Oct4-positive and connexin 43-negative [81]. An important issue is whether these cells are enriched in the so-called side population and whether this is related to the level of expression of the ABCG2 transporter [82]. The phenotype of the stem/precursor cells needs to be more accurately defined. In addition, there may be several distinct types of progenitor cells as has been already established for normal mammary gland [83, 84]. The fact that self-renewing progenitors were not identified from aggressive ER negative tumours suggests that culture conditions may not have been suitable. Alternatively, stem cells from different types of breast cancers may express different phenotypic markers. Clearly, conventional drugs are unlikely to efficiently eradicate quiescent stem cells. In the same manner these cells may well be resistant to radiotherapy and would then be responsible for local or distant relapses. It would be intriguing to characterise such cells in bone marrow micrometastasis and correlate their presence with tumour progression in these patients.

Concluding remarks

Breast carcinomas comprise a very large set of remarkably heterogeneous tumours. The conventional treatment of breast cancers has made substantial progress over the last 20 years. The new targeted therapies, although permitting longer-term survival, have so far failed to cure metastatic diseases. At best, metastatic cancer patients could benefit from protocols treating a chronic disease. However, as it is already known from chronic myeloid leukaemia patients treated with Gleevec™, resistance can be acquired through specific mutations of the Abelson tyrosine kinase, leading eventually to progression to an acute phase. The current dogma is that one must treat cancer stem/progenitor cells in addition to the actively proliferating cells. Defining such stem/progenitor cells, which may be quite heterogeneous themselves, requires more basic studies on normal stem/progenitor cells in order to understand their phenotypes. Well designed transgenic models may help refine our understanding of cancer stem/progenitor cells. Another major effort is to define better molecular markers, which, in conjunction with the well established histoprognostic markers, will permit tailored individual therapy. The major challenge is indeed in the remarkable heterogeneity of breast carcinoma. Progress has been made in unravelling this issue with high-density arrays, landscaping the genome and the transcriptome of breast tumours. Other high-density screening epigenetic modifications, such as the methylome and posttranslational modifications such as the phosphokinome, have considerable potential to further define the molecular status of each tumour. These combined studies may bring more robustness to the transcriptomic data. They may also offer new potential targets for therapies. The more conventional surrogate markers now routinely used, such as bone micrometatases, must also be considered for prognosis and for evaluation of response to therapy. This formidable task which the research community now faces will provide major benefits to breast cancer patients in the near future. ▼

References

1. Singletary SE, Allred C, Ashley P, Bassett LW, Berry D, Bland KI, *et al.* Revision of the American Joint Committee on Cancer staging system for breast cancer. *J Clin Oncol* 2002 ; 20 : 3628-36.

2. Rampaul RS, Pinder SE, Elston CW, Ellis IO. Prognostic and predictive factors in primary breast cancer and their role in patient management : The Nottingham Breast Team. *Eur J Surg Oncol* 2001 ; 27 : 229-38.

3. Harvey JM, Clark GM, Osborne CK, Allred DC. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *J Clin Oncol* 1999 ; 17 : 1474-81.

4. Kim T, Giuliano AE, Lyman GH. Lymphatic mapping and sentinel lymph node biopsy in early-stage breast carcinoma. *Cancer* 2006 ; 106 : 4-16.

5. Cristofanilli M, Budd GT, Ellis MJ, Stopeck A, Matera J, Miller MC, *et al.* Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N Engl J Med* 2004 ; 351 : 781.

6. Pierga JY, Bonneton C, Vincent-Salomon A, de Crémoux P, Nos C, Blin N, *et al.* Clinical significance of immunocytochemical detection of tumor cells using digital microscopy in peripheral blood and bone marrow of breast cancer patients. *Clin Cancer Res* 2004 ; 10 : 1392-400.

7. Pantel K, Cote RJ, Fodstad O. Detection and clinical importance of micrometastatic disease. *J Natl Cancer Inst* 1999 ; 91 : 1113-24.

8. Braun S, Pantel K, Muller P, Janni W, Hepp F, Kentenich CR, *et al.* Cytokeratin-positive cells in the bone marrow and survival of patients with stage I, II, or III breast cancer. *N Engl J Med* 2000 ; 342 : 525-33.

9. Braun S, Vogl FD, Naume B, Janni W, Osborne MP, Coombes RC, *et al.* A pooled analysis of bone marrow micrometastasis in breast cancer. *N Engl J Med* 2005 ; 353 : 793-802.

10. Pierga JY, Bonneton C, Magdelenat H, Vincent-Salomon A, Nos C, Pouillart P, *et al.* Clinical significance of proliferative potential of occult metastatic cells in bone marrow of patients with breast cancer. *Br J Cancer* 2003 ; 89 : 539-45.

11. Choesmel V, Anract P, Hoifodt H, Thiery JP, Blin N. A relevant immunomagnetic assay to detect and characterize epithelial cell adhesion molecule-positive cells in bone marrow from patients with breast carcinoma : immunomagnetic purification of micrometastases. *Cancer* 2004 ; 101 : 693-703.

12. Schmidt-Kittler O, Ragg T, Daskalakis A, Granzow M, Ahr A, Blankenstein TJ, *et al.* From latent disseminated cells to overt metastasis : genetic analysis of systemic breast cancer progression. *Proc Natl Acad Sci USA* 2003 ; 100 : 7737-42.

13. O'Connell P. Genetic and cytogenetic analyses of breast cancer yield different perspectives of a complex disease. *Breast Cancer Res Treat* 2003 ; 78 : 347-57.

14. Al-Kuraya K, Schraml P, Torhorst J, Tapia C, Zaharieva B, Novotny H, *et al.* Prognostic relevance of gene amplifications and coamplifications in breast cancer. *Cancer Res* 2004 ; 64 : 8534-40.

15. Nessling M, Richter K, Schwaenen C, Roerig P, Wrobel G, Wessendorf S, *et al.* Candidate genes in breast cancer revealed by microarray-based comparative genomic hybridization of archived tissue. *Cancer Res* 2005 ; 65 : 439-47.

16. Jones C, Ford E, Gillett C, Ryder K, Merrett S, Reis-Filho JS, *et al.* Molecular cytogenetic identification of subgroups of grade III invasive ductal breast carcinomas with different clinical outcomes. *Clin Cancer Res* 2004 ; 10 : 5988-97.

17. Hupe P, Stransky N, Thiéry JP, Radvanyi F, Barillot E. Analysis of array CGH data : from signal ratio to gain and loss of DNA regions. *Bioinformatics* 2004 ; 20 : 3413-22.

18. Hwang ES, DeVries S, Chew KL, Moore DH, Kerlikowske K, Thor A, *et al.* Patterns of chromosomal alterations in breast ductal carcinoma in situ. *Clin Cancer Res* 2004 ; 10 : 5160-7.

19. Samuels Y, Velculescu VE. Oncogenic mutations of PIK3CA in human cancers. *Cell Cycle* 2004 ; 3 : 1221-4.

20. Bachman KE, Argani P, Samuels Y, Silliman N, Ptak J, Szabo S, *et al.* The PIK3CA gene is mutated with high frequency in human breast cancers. *Cancer Biol Ther* 2004 ; 3 : 772-5.

21. Campbell IG, Russell SE, Choong DY, Montgomery KG, Ciavarella ML, Hooi CS, *et al.* Mutation of the PIK3CA gene in ovarian and breast cancer. *Cancer Res* 2004 ; 64 : 7678-81.

22. Lee JW, Soung YH, Kim SY, Lee HW, Park WS, Nam SW, *et al.* PIK3CA gene is frequently mutated in breast carcinomas and hepatocellular carcinomas. *Oncogene* 2005 ; 24 : 1477-80.

23. Li SY, Rong M, Grieu F, Iacopetta B. PIK3CA mutations in breast cancer are associated with poor outcome. *Breast Cancer Res Treat* 2006 ; 96 : 91-5.

24. Saal LH, Holm K, Maurer M, Memeo L, Su T, Wang X, *et al.* PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. *Cancer Res* 2005 ; 65 : 2554-9.

25. Isakoff SJ, Engelman JA, Irie HY, Luo J, Brachmann SM, Pearline RV, *et al.* Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells. *Cancer Res* 2005 ; 65 : 10992-1000.

26. Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet* 2005 ; 37 : 590-2.

27. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, *et al.* Molecular portraits of human breast tumours. *Nature* 2000 ; 406 : 747-52.

28. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001 ; 98 : 10869-74.

29. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003 ; 100 : 8418-23.

30. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, *et al.* Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res* 2004 ; 10 : 5367-74.

31. Ma XJ, Salunga R, Tuggle JT, Gaudet J, Enright E, McQuary P, *et al.* Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci USA* 2003 ; 100 : 5974-9.

32. Korkola JE, DeVries S, Fridlyand J, Hwang ES, Estep AL, Chen YY, *et al.* Differentiation of lobular versus ductal breast carcinomas by expression microarray analysis. *Cancer Res* 2003 ; 63 : 7167-75.

33. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002 ; 415 : 530-6.

34. Van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002 ; 347 : 1999-2009.

35. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci USA* 2005 ; 102 : 3738-43.

36. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003 ; 33 : 49-54.

37. Weigelt B, Glas AM, Wessels LF, Witteveen AT, Peterse JL. van't Veer LJ. Gene expression profiles of primary breast tumors maintained in distant metastases. *Proc Natl Acad Sci USA* 2003 ; 100 : 15901-5.

38. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005 ; 365 : 671-9.

39. Nevins JR, Huang ES, Dressman H, Pittman J, Huang AT, West M. Towards integrated clinico-genomic models for personalized medicine : combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hum Mol Genet* 2003 ; (12 Spec Iss 2 : R153-7).

40. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, *et al.* Gene expression predictors of breast cancer outcomes. *Lancet* 2003 ; 361 : 1590-6.

41. Dai H, van't Veer L, Lamb J, He YD, Mao M, Fine BM, *et al.* A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Res* 2005 ; 65 : 4059-66.

42. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, *et al.* An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA* 2005 ; 102 : 13550-5.

43. Minn AJ, Kang Y, Serganova I, Gupta GP, Giri DD, Doubrovin M, *et al.* Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. *J Clin Invest* 2005 ; 115 : 44-55.

44. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, Giri DD, *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* 2005 ; 436 : 518-24.

45. Brenton JD, Carey LA, Ahmed AA, Caldas C. Molecular classification and molecular forecasting of breast cancer : ready for clinical application? *J Clin Oncol* 2005 ; 23 : 7350-60.

46. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays : a multiple random validation strategy. *Lancet* 2005 ; 365 : 488-92.

47. Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, *et al.* Estimating data size requirements for classifying DNA micrarray data. *J Comp Biol* 2003 ; 10 : 119-42.

48. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer : is there a unique set? *Bioinformatics* 2005 ; 21 : 171-8.

49. Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, *et al.* Multi-platform, multi-site, microarray-based human tumor classification. *Am J Pathol* 2004 ; 164 : 9-16.

50. Van't Veer LJ, Paik S, Hayes DF. Gene expression profiling of breast cancer : a new tumor marker. *J Clin Oncol* 2005 ; 23 : 1631-5.

51. Ellis M, Ballman K. Trawling for genes that predict response to breast cancer adjuvant therapy. *J Clin Oncol* 2004 ; 22 : 2267-9.

52. Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, *et al.* Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *J Clin Oncol* 2004 ; 22 : 2284-93.

53. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, *et al.* Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res* 2005 ; 11 : 5678-85.

54. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, *et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* 2003 ; 362 : 362-9.

55. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Tham YL, *et al.* Patterns of resistance and incomplete response to docetaxel by gene expression profiling in breast cancer patients. *J Clin Oncol* 2005 ; 23 : 1169-77.

56. Hannemann J, Oosterkamp HM, Bosch CA, Velds A, Wessels LF, Loo C, *et al.* Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. *J Clin Oncol* 2005 ; 23 : 3331-42.

57. Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, *et al.* A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 2004 ; 5 : 607-16.

58. Reid JF, Lusa L, De Cecco L, Coradini D, Veneroni S, Daidone MG, *et al.* Limits of predictive models using microarray data for breast cancer clinical treatment outcome. *J Natl Cancer Inst* 2005 ; 97 : 927-30.

59. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy : derived and validated in two population-based cohorts. *Breast Cancer Res* 2005 ; 7 : R953-R964.

60. Tanner M, Kapanen AI, Junttila T, Raheem O, Grenman S, Elo J, *et al.* Characterization of a novel cell line established from a patient with Herceptin-resistant breast cancer. *Mol Cancer Ther* 2004 ; 3 : 1585-92.

61. Nagy P, Friedlander E, Tanner M, Kapanen AI, Carraway KL, Isola J, *et al.* Decreased accessibility and lack of activation of ErbB2 in JIMT-1, a herceptin-resistant, MUC4-expressing breast cancer cell line. *Cancer Res* 2005 ; 65 : 473-82.

62. Szakacs G, Annereau JP, Lababidi S, Shankavaram U, Arciello A, Bussey KJ, *et al.* Predicting drug sensitivity and resistance : profiling ABC transporter genes in cancer cells. *Cancer Cell* 2004 ; 6 : 129-37.

63. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, *et al.* A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000 ; 24 : 236-44.

64. Cardiff RD, Anver MR, Gusterson BA, Hennighausen L, Jensen RA, Merino MJ, *et al.* The mammary pathology of genetically engineered mice : the consensus report and recommendations from the Annapolis meeting. *Oncogene* 2000 ; 19 : 968-88.

65. Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, *et al.* An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat Genet* 2005 ; 37 : 48-55.

66. Thiéry JP. Epithelial-mesenchymal transitions in tumour progression. *Nat Rev Cancer* 2002 ; 2 : 442-54.

67. Vincent-Salomon A, Thiery JP. Host microenvironment in breast cancer development : epithelial-mesenchymal transition in breast cancer development. *Breast Cancer Res* 2003 ; 5 : 101-6.

68. Mikaelian I, Blades N, Churchill GA, Fancher K, Knowles BB, Eppig JT, *et al.* Proteotypic classification of spontaneous and transgenic mammary neoplasms. *Breast Cancer Res* 2004 ; 6 : R668-R679.

69. Welm AL, Kim S, Welm BE. Bishop JM. MET and MYC cooperate in mammary tumorigenesis. *Proc Natl Acad Sci USA* 2005 ; 102 : 4324-9.

70. Deugnier MA, Faraldo MM, Janji B, Rousselle P, Thiéry JP, Glukhova MA. EGF controls the in vivo developmental potential of a mammary epithelial cell line possessing progenitor properties. *J Cell Biol* 2002 ; 159 : 453-63.

71. Faraldo MM, Teulière J, Deugnier MA, Taddei de la Hosseraye I, Thiéry JP, Glukhova M. Myopepithelial cells in the control of mammary development and tumorigenesis : data from genetically modified mice. *J Mammary Gland Biol Neoplasia* 2005 ; 10 : 211-9.

72. Teulière J, Faraldo MM, Deugnier MA, Shtutman M, Ben-Ze'ev A, Thiéry JP, *et al.* Targeted activation of beta-catenin signaling in basal mammary epithelial cells affects mammary development and leads to hyperplasia. *Development* 2005 ; 132 : 267-77.

73. Smith GH, Medina D. A morphologically distinct candidate for an epithelial stem cell in mouse mammary gland. *J Cell Sci* 1988 ; 90 : 173-83.

74. Woodward WA, Chen MS, Behbod F, Rosen JM. On mammary stem cells. *J Cell Sci* 2005 ; 118 : 3585-94.

75. Smith GH. Label-retaining epithelial cells in mouse mammary gland divide asymmetrically and retain their template DNA strands. *Development* 2005 ; 132 : 681-7.

76. Welm BE, Tepera SB, Venezia T, Graubert TA, Rosen JM, Goodell MA. Sca-1(pos) cells in the mouse mammary gland represent an enriched progenitor cell population. *Dev Biol* 2002 ; 245 : 42-56.

77. Stingl J, Eirew P, Ricketson I, Shackleton M, Vaillant F, Choi D, *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* 2006 ; 439 : 993-7.

78. Shackleton M, Vaillant F, Simpson KJ, Stingl J, Smyth GK, Asselin-Labat ML, *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* 2006 ; 439 : 84-8.

79. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature* 2001 ; 414 : 105-11.

80. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison S, Clarke MF. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci USA* 2003 ; 100 : 3983-8.

81. Ponti D, Costa A, Zaffaroni N, Pratesi G, Petrangolini G, Coradini D, *et al.* Isolation and in vitro propagation of tumorigenic breast cancer cells with stem/progenitor cell properties. *Cancer Res* 2005 ; 65 : 5506-11.

82. Patrawala L, Calhoun T, Schneider-Broussard R, Zhou J, Claypool K, Tang DG. Side population is enriched in tumorigenic, stem-like cancer cells, whereas ABCG2+ and ABCG2- cancer cells are similarly tumorigenic. *Cancer Res* 2005 ; 65 : 6207-19.

83. Clarke RB. Isolation and characterization of human mammary stem cells. *Cell Prolif* 2005 ; 38 : 375-86.

84. Stingl J, Raouf A, Emerman JT, Eaves CJ. Epithelial progenitors in the normal human mammary gland. *J Mammary Gland Biol Neoplasia* 2005 ; 10 : 49-59.

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 7 2006, pages 849–856
doi:10.1093/bioinformatics/btl004

Data and text mining

Computation of recurrent minimal genomic alterations from array-CGH data

C. Rouveirol[1,*], N. Stransky[2], Ph. Hupé[2,3], Ph. La Rosa[3], E. Viara[3], E. Barillot[3] and F. Radvanyi[2]

[1]LRI, UMR CNRS 8623, Université, Paris Sud, bât 490 91405 Orsay cedex, France, [2]UMR CNRS, 144 and [3]Service de Bioinformatique, Institut Curie, 26 rue d'Ulm 75248 Paris cedex 05, France

Received on June 16, 2005; revised on December 28, 2005; accepted on January 13, 2006
Advance Access publication January 24, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: The identification of recurrent genomic alterations can provide insight into the initiation and progression of genetic diseases, such as cancer. Array-CGH can identify chromosomal regions that have been gained or lost, with a resolution of ~1 mb, for the cutting-edge techniques. The extraction of discrete profiles from raw array-CGH data has been studied extensively, but subsequent steps in the analysis require flexible, efficient algorithms, particularly if the number of available profiles exceeds a few tens or the number of array probes exceeds a few thousands.

Results: We propose two algorithms for computing minimal and minimal constrained regions of gain and loss from discretized CGH profiles. The second of these algorithms can handle additional constraints describing relevant regions of copy number change. We have validated these algorithms on two public array-CGH datasets.

Availability: From the authors, upon request.

Contact: celine@lri.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Cancer is a genetic disease. Tumour initiation and progression result from the activation of oncogenes and the inactivation of tumour suppressor genes (Fearon and Vogelstein, 1990; Vogelstein and Kinzler, 2004). Genomic instability is a hallmark of cancer and most cancers display various genomic alterations, such as losses, gains and amplifications of chromosome regions. Cancer-associated gains and amplifications are thought to be responsible for oncogene activation and chromosomal deletions are thought to result in the inactivation of at least one copy of a tumour suppressor gene, the other copy being inactivated by a point mutation or other genetic or epigenetic event (Pinkel and Albertson, 2005).

Large-scale analysis of genomic alterations is now possible with array-CGH (comparative genomic hybridization) (Solinas-Toldo *et al*., 1997; Pinkel *et al*., 1998). Fragments of genomic DNA are spotted as probes on a glass slide and hybridized with a mixture of tumour and normal DNA, labelled with two different fluorophores. An alternative approach to the spotting of genomic DNA fragments is the use of cDNA arrays (Pollack *et al*., 2002) or oligonucleotide arrays (Lucito *et al*., 2003; Herr *et al*., 2005). The data obtained with array-CGH techniques should provide a rapid, precise identification of the chromosomal regions altered in tumours. An increasing number of tools [see, in particular, CGHAnalyzer (Margolin *et al*., 2005), ChARMView (Myers *et al*., 2005)] have recently become available for managing, discretizing, visualizing sets of CGH profiles. CGHAnalyzer also provides statistical tools for supervised or unsupervised analysis of sets of genes, based on copy number status, with or without discretization. However, transverse analyses of array-CGH profiles to define genomic regions frequently subject to copy-number change are frequently performed manually, as a preliminary step before further analysis [see, among others Veltman *et al*. (2005), Schraders *et al*. (2005), de Leeuw *et al*. (2004)]. Attempts have recently been made to construct common alteration regions automatically (Aguirre *et al*., 2004; Tonon *et al*., 2005), but this crucial task is still mostly carried out on a manual, *ad hoc* basis. No general, reusable formalization or tool for finding common or recurrent alteration regions in a CGH-array dataset is currently available. We define a recurrent region as a sequence of altered probes common to a set of CGH profiles and a minimal recurrent region as a recurrent region that contains no smaller recurrent region. In many cases, the accurate determination of minimal regions of chromosomal alterations is the first, crucial step towards the identification of new oncogenes and tumour suppressor genes. If the number of array-CGH profiles to be analysed approaches a few tens, or if there are more than a few thousand array probes, it is very difficult to develop a global view of all the genomic alterations in the dataset, and therefore to identify recurrent regions of gain and loss. Characterization of the minimal regions of alteration can also improve our understanding of tumour progression, even before identification of the genes involved. Minimal regions can be used as new variables for the analysis of array-CGH profiles, e.g. to explore the patterns of copy number alterations in groups of tumours. As these minimal regions are thought to convey concise, biologically meaningful information, their use should improve both supervised and unsupervised classification analyses.

We propose a formalization and two algorithms for computing minimal copy number alteration regions. The first step is identification—starting from normalized array-CGH data—of the chromosomal regions altered in a tumour. We recently described a method, the Gain and Loss Analysis of DNA (GLAD) algorithm

*To whom correspondence should be addressed.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

(Hupé *et al*., 2004), for the automatic detection of breakpoints, using array-CGH profiles. GLAD assigns a status (Gain, Loss or Normal) to each chromosomal region. We now describe two algorithms that compute recurrent alteration regions from such sets of discretised profiles generated by GLAD. The first one, MAR, efficiently computes all minimal recurrent alteration regions from a set of discretized profiles. MAR may identify too many minimal regions (see Section 4), the manual validation of which may be time-consuming for biologists. In such situations, the tumour biologist may have expert knowledge concerning what makes an alteration region relevant. It may be possible to express this knowledge as a set of simple constraints, such as a minimum frequency of a given alteration region in a dataset, or the number of observations defining the border of the alteration region. We therefore propose a second algorithm, CMAR, that computes minimal constrained regions of chromosomal gain and loss from the discretized CGH profiles generated by GLAD. This algorithm can easily be extended to handle additional constraints. Although CMAR is less computationally efficient than MAR (quadratic rather than linear in terms of the number of probes describing the profiles), it should generate fewer, potentially more relevant alteration regions. The parameters of the current constraints implemented in CMAR can easily be adapted to any given dataset.

In Section 2, we introduce the terminology and notations required for the two algorithms and present the first algorithm. Section 3 introduces a number of constraints and their properties, and presents the extension of the first algorithm to the computation of minimal constrained regions. Section 4 provides an experimental validation of the approach, using public CGH data for various types of cancer and, finally, Section 5 sums up the advantages and current limitations of the method and indicates promising directions for further research.

2 MINIMAL REGIONS

The notion of a minimal common alteration region has not been formalized as such in the bioinformatics community. This concept is, however, used by biologists searching for candidate genes involved in tumour initiation and progression, using data describing genomic alterations across sets of genomic profiles. We provide here a formalization based on Formal Concept Analysis theory (Ganter and Wille, 1999).

2.1 Formalization

We assume that we have, as input data, a three-value discrete matrix describing each observation in terms of gain, loss and normal probes. It is straightforward to transform such a discrete matrix into two boolean contexts, M_g and M_l, describing gain and loss events in array-CGH profiles, respectively, with no loss of information.

DEFINITION 1. *A context is a triplet* (O, P, M) *where* O *is a finite set of observations of size* N_O, P *is a finite set of probe attributes of size* N_P *and* M *is a binary relationship between* O *and* P, $(M \subseteq O \times P)$. *For simplicity, we will also refer to* M *as a context when* O *and* P *are clearly known.*

The contexts M_g and M_l are computed such that $M_g(o,p) = 1$ if probe p is gained in o, $M_g(o,p) = 0$ if p is not gained in o, $M_l(o,p) = 1$ if probe p is lost in o and $M_l(o, p) = 0$ if p is not lost in o.

Table 1. A boolean representation of an array-CGH dataset

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
o_1	0	0	1	0	1	0	1	1	0	0	1
o_2	0	0	1	0	1	1	1	1	0	1	0
o_3	0	0	1	1	1	1	0	0	0	1	0
o_4	1	0	1	1	1	1	0	0	0	0	0
o_5	1	1	1	1	1	1	1	1	0	1	0

The computations of minimal gain and loss regions can therefore be handled as identical, independent problems. We will now illustrate our algorithms in the context described in Table 1.

With this representation of the array-CGH data, recurrent gained and lost genomic regions can first be seen as rectangles of ones in two 1–0 matrices. This problem has been extensively studied in Data Mining, in the area of frequent itemset mining [see the seminal paper Agrawal and Srikant (1994)]. The problem of computing closed (Pasquier *et al*., 1999) and constrained (Ng *et al*., 1998) patterns has recently received much attention, particularly for large and dense extraction contexts, such as those for most DNA array data (Pang *et al*., 2003; Besson *et al*., 2005).

The problem dealt with here is more specific: the set of probes P is totally ordered by the relationship $\preceq_P$, where $\preceq_P$ is the ordering of probes in the genome. Consequently, the patterns of interest are not subsets of P, but are instead sequences of probes.

We introduce the following notation for the representation and handling of our specific sequences. Given two probe attributes p_i and p_j, $p_i \preceq_P p_j$ if and only if $i \leq j$. A sequence of contiguous probes is denoted by $[p_i..p_j]$. For simplicity, a single probe sequence $[p_i..p_i]$ is denoted by p_i. Strict inclusion between sets or sequences is denoted by $\subset$, whereas loose inclusion is denoted by $\subseteq$.

DEFINITION 2. *A lattice is a partially ordered set* $(L, \preceq)$ *such that any two nodes* n_1 *and* $n_2 \in L$ *have a single least upper bound* (*lub*) *and a single greatest lower bound* (*glb*). *The lub* $s \in L$ *of two nodes* n_1 *and* $n_2 \in L$ *is such that* $n_1 \preceq s$ *and* $n_2 \preceq s$ *and there is no other node* $s' \in L$ *such that* $n_1 \preceq s'$, $n_2 \preceq s'$ and $s' \prec s$. *Symmetrically, given any two nodes,* n_1 *and* $n_2 \in L$, *the glb* $g \in L$ *of* n_1 *and* n_2 *is such that* $g \preceq n_1$, $g \preceq n_2$ *and there is no other node* $g' \in L$ *such that* $g' \preceq n_1$ *and* $g' \preceq n_2$ *and* $g \prec g'$.

The set of all probe sequences of P is denoted by $S(P)$. $(S(P),\subseteq)$ is a lattice, isomorphic to the lattice of intervals of $[1..N_P]$. Therefore, for simplicity, a sequence of $S(P)$,$[p_i..p_j]$ with $1 \leq i \leq j \leq N_P$, will be denoted $[i..j]$.

DEFINITION 3. *Let* s *be a sequence of probes of* $S(P)$. *The extension of* s, *given the context* M, *denoted ext*(s, M) [*or ext*(s), *when there is no ambiguity concerning the reference context*], *is the set of all observations* o_i *of* M *such that all probe attributes of* s *are set to* 1 *in* o_i, *denoted in short as* $s \subseteq o_i$. *The frequency of* s *is the size of its extension.*

Some sequences of $S(P)$, the closed ones, are remarkable in the context M: they are the largest sequences occurring in a given set of profiles. Closed sets and sequences are useful for Data Mining because they provide a complete and compact representation of all possible solutions to a mining problem.

Formally, a closure operation on $S(P)$ can be defined as follows.

DEFINITION 4. *Let s a sequence of P, the closure of s given a context M, denoted $closure_P(s, M)$ be the largest supersequence of s that has the same extension as s. A sequence s is closed if $closure_P(s, M) = s$. Hereafter, a closed sequence of $S(P)$ will be referred to as a region.*

The closure of a sequence s can be computed iteratively by intersecting the largest supersequences of s in each observation of s extension. As a consequence, a closed sequence of P, $s = [p_i..p_j] \in S(P)$ of extension e, is one for which $e \cap \text{ext}(p_{i-1}) \neq e$ and $e \cap \text{ext}(p_{i+1}) \neq e$.

EXAMPLE 1. In the context of Table 1, $[p_4..p_5]$ is a sequence of $S(P)$ of extension $\{o_3, o_4, o_5\}$, but is not closed. Its closure is $[p_3..p_6]$.

From the above definitions, we can derive the following properties. The proofs of these properties are provided in Appendix 1 of the Supplementary Material.

PROPOSITION 1. *Let us denote by $R(P)$ the set of closed sequences of P. $(R(P), \subseteq)$ forms a lattice.*

EXAMPLE 2. The set of all closed sequences of the context defined in Table 1 is $[p_1..p_1]$, $[p_3..p_3]$, $[p_5..p_5]$, $[p_7..p_8]$, $[p_{10}..p_{10}]$, $[p_{11}..p_{11}]$, $[p_5..p_6]$, $[p_5..p_8]$, $[p_3..p_6]$, $[p_1..p_8]$. The lattice of all such sequences is given in figure 1 of Appendix 1 of the Supplementary Material.

DEFINITION 5. *A region $r \in R(P)$ is minimal if there is no other region $r' \in R(P)$ such that $r' \subset r$.*

EXAMPLE 3. In the context of Table 1, $[p_3..p_6]$ is a region, but not a minimal one, because $[p_5..p_6]$ is a closed subsequence of $[p_3..p_6]$. Note that $ext([p_3..p_6]) = \{o_3,o_4,o_5\} \subset \text{ext}([p_5..p_6]) = \{o_2,o_3,o_4,o_5\}$.

Note that the minimal regions are the smallest elements of this lattice. The sequential organization of probes in the genome could be used to design an efficient algorithm for detecting minimal regions.

2.2 Computing minimal regions

This algorithm, MAR is based on a transformation of the context, provided that the computation of minimal zones does not require access to the extension of probes and requires only knowledge concerning changes of extension in the genome: breakpoints. This approach is conceptually similar to the traditional definition of minimally altered regions based on multiple alignments of alterations. In this case the region is the intersection of the aligned alterations and is therefore delimited by the breakpoints that narrow down the intersection the most.

DEFINITION 6. Given a context (O, P, M), *we define a* breakpoint *as an index i, $1 < i < N_P$ such that there is at least one observation o for which $M(o, p_i) = 0$ and $M(o, p_{i+1}) = 1$ or vice versa. Additionally, 1 is a breakpoint if $M(o, p_1) = 1$ and N_P is a breakpoint if $M(o, pN_e) = 1$. If $M(o, p_i) = 0$ and $M(o, p_{i+1}) = 1$, i is an in-breakpoint; if $M(o, p_p) = 1$ and $M(o, p_{i+1}) = 0$, i is an out-breakpoint. Note that 1 can only be an in-breakpoint and N_P an out-breakpoint. If b is a breakpoint, we will denote shift_in(b) as the set of all observations $o_i \in O$ such that $M(o_i, b) = 0$ and $M(o_i, b + 1) = 1$ and shift_out(b) as the set of observations o_i such that $M(o_i, b) = 1$ and $M(o_i, b + 1) = 0$.*

```
Compute_Minimal_Regions(M);
Lin :=(ϕ); Lout := (ϕ)
For every observation O in M
    if M(O,1) = 1 then Lin := Lin ∪ 0
    if M(O,NP) = 1 then Lout := Lout ∪ NP
    for every other probe P of M
        if M(O,P) = 0 and M(O,P+1) = 1
           then Lin := Lin ∪ P
        else
           if M(O,P) = 1 and M(O,P+1) = 0
              then Lout := Lout ∪ P
MR = {(I,O) with I ∈ Lin  and O ∈ Lout |
      there is no  J ∈ Lin ∪ Lout s.t. I<J<O}
Return(MR)
```

Fig. 1. MAR: algorithm for computing all minimal alteration regions

Note that for a given index i, $1 \leq i \leq N_P$ can be an in- and an out-breakpoint simultaneously. Figures 2 and 3, in Appendix 1 of the Supplementary Material, illustrate the notions introduced in the definition. In the MAR algorithm, we will make use of the following:

THEOREM 1. *A region $r = [in..out]$ is a minimal region if and only if* (1) *in is an in-breakpoint and out is an out-breakpoint and* (2) *there is no breakpoint b such that $in < b < out$.*

The proof of this Theorem 1 is provided in Appendix 1 of the Supplementary material. From this theorem, we can readily derive a linear algorithm for finding all minimal regions (Fig. 1). This algorithm clearly has complexity in $O(N_o * N_p)$ with N_O the number of observations and N_P the number of probes.

EXAMPLE 4. $L_{\text{in}} = \{p_1, p_3, p_5, p_7, p_{10}, p_{11}\}$ and $L_{\text{out}} = \{p_1, p_3, p_5, p_6, p_8, p_{10}, p_{11}\}$. The minimal regions of the context of Table 1 are $[p_1..p_1]$, $[p_3..p_3]$, $[p_5..p_5]$, $[p_7..p_8]$, $[p_{10}..p_{10}]$, $[p_{11}..p_{11}]$.

Note that if the genome studied consists of several chromosomes and if we set the constraint that a region does not overlap two chromosomes, the above algorithm will be iteratively applied to all chromosomes in the genome.

3 CONSTRAINED MINIMAL REGIONS

The definition of gain/loss regions above may yield a large number of minimal regions (see Section 4), the manual validation of which may be time-consuming for biologists. Biologists may have expert knowledge about what constitutes a relevant alteration region that is much more useful than a frequency test. We therefore introduce the notion of a minimal constrained region, which extends the Definition 5 to regions that satisfy a particular combination of properties $C = C_1, \ldots, C_n$.

DEFINITION 7. *A region r is minimal for the conjunction of constraints $C = C_1, \ldots, C_n$ if and only if r satisfies each C_i, $1 \leq i \leq n$ and there is no region r', $r' \subset r$ such that r' satisfies C.*

3.1 Constraints—properties for use in the search for minimal regions

We have identified the following constraints as relevant for finding recurrent chromosomal regions of gain/loss. These constraints

concern either the sequence or the extension of the region:

- Minimum/maximum frequency of the region in M.
- Minimum/maximum size of the region in number of probes.
- The region's extension contains/does not contain a given observation.
- The region is well bounded (see Definition 9).

The first three of these constraints are intuitive and have been extensively studied in the domain of data mining [see, among others De Raedt and Kramer (2001)]. Some of these constraints are anti-monotone with respect to set inclusion ($\subseteq$) and can be used to search the lattice efficiently for subsets of P (and of sequences of P) satisfying these constraints.

DEFINITION 8. *A constraint C_{am} is anti-monotone with respect to $\subseteq$ if, for all sets s and g such that $g \subset s$, if s satisfies C_{am}, then g satisfies it also.*

EXAMPLE 5. Setting a minimum frequency or a maximum size for a region, or imposing that a particular observation belongs to the extension of a region are anti-monotone constraints.

These constraints can be used to search efficiently for constrained closed sequences, avoiding the exploration of parts of the search space that cannot contain solutions, based on current information collected during the search. For instance, if a set or sequence of probes does not satisfy an anti-monotone constraint C, there is no need to explore and evaluate its supersequences, because they will not satisfy C. In particular, if a sequence s is infrequent in a given context M, all supersequences of s are infrequent in M, and need not be evaluated.

Other properties of constraints may be useful for improving the efficiency of pattern search [e.g. monotone or convertible constraints (Ng *et al*., 1998)], but are not dealt with in this paper (see Supplementary Material for a discussion). For instance, our experience with CGH data analysis led us to use the following constraint, which is neither anti-monotone nor monotone, but is nonetheless essential for selecting relevant regions.

DEFINITION 9. *Given a context (O, P, M), a region $r = [in..out]$ is well bounded on the left given a fixed parameter b if and only if there are at least b observations o_i in ext(r, M)* such that $M(o_i, in - 1) = 0$ *and $M(o_i, in) = 1$. In other words, in is an in-breakpoint for at least b observations of the extension of r. Symmetrically, r is well bounded on the right if out is an out-breakpoint for at least b observations of ext(r, M). A region is well bounded if and only if it is well bounded on both the left and right.*

EXAMPLE 6. Let us consider well bounded regions with $b = 2$. In the context of Table 1, $[p_5..p_5]$ and $[p_7..p_8]$ are minimal regions according to the Definition 5, but $[p_5..p_5]$ is not well bounded on the right, whereas $[p_7..p_8]$ is not well bounded on the left. $[p_3..p_6]$ is a well bounded supersequence of $[p_5..p_5]$, and there is no well bounded supersequence of $[p_7..p_8]$.

Well boundedness is not anti-monotone, as demonstrated in the above example.

The above definition can be relaxed to the following definition, which is more suitable for our biological context.

DEFINITION 10. *Given a context (O, P, M), a region $r = [in..out]$ of extension e is well fuzzy bounded on the left given parameters b and m (m is referred to hereafter as the margin parameter) if and only if there are at least b observations of the extension of r that switch from 0 to 1 in the interval $[(in - m)..in]$. Formally, for all in-breakpoints i, $in - m \leq i \leq in$, such that shift_in(i) $\cap\, e \neq \emptyset$, $|\cup_i(shift\_in(i) \cap e)| \geq b$. The definition is symmetric for regions well fuzzy bounded on the right. A region r is well fuzzy bounded for parameters b and m if and only if it is both well fuzzy bounded on both the left and right for b and m.*

This definition is illustrated in figure 3 of Appendix 1.

EXAMPLE 7. Given the bound $b = 2$, the smallest m for which $[p_5..p_5]$ is well fuzzy bounded is $m = 1$. $[p_7..p_8]$ is well fuzzy bounded for $m = 2$.

Computing constrained minimal regions (Fig. 2) is a more complex problem than computing minimal regions, as demonstrated by the Example 6. If the problem is defined exclusively in terms of anti-monotonic constraints, the MAR algorithm can be used to find all minimal regions, and those minimal regions satisfying the anti-monotonic constraints can then be selected. However, if non-anti-monotonic constraints are involved, a level-wise exploration (Mannila and Toivonen, 1997) of the $R(P)$ lattice should be carried out, and this exploration should be as efficient as possible. In the following, we assume, without loss of generality, that the set of constraints on the solution regions can be split into AC, a conjunction of anti-monotone constraints with respect to $\subseteq$ and OC, a conjunction of non-anti-monotone constraints for the problem.

CMAR searches $R(P)$, the lattice of closed probe sequences breadth first. The first sequences it considers are minimal regions, as computed with the algorithm in Figure 1, because no smaller sequence of $S(P)$ can be closed, according to Theorem 1. If a candidate region r satisfies all constraints of the problem (i.e. both AC and OC), then r is a solution. Regions that do not satisfy AC

```
Compute_Minimal_Constrained_Regions(M,AC,OC);
Li, Lo:= Lists of in and out breakpoints
Cand_Reg(1) := Compute_Minimal_Regions(M) (see MAR algorithm)
CMR := ∅; L:= 1
While Cand_Reg(L)  ≠ ∅
  For every region R ∈ Cand_Reg(L) do
      Cand_Reg(L) := Cand_Reg(L) - R
      if R satisfies AC then
         if R satisfies  OC then CMR := CMR ∪ R
         else FailedOC(L) := FailedOC(L) ∪ R
      else  FailedAC := FailedAC ∪ R
  EndFor
  Cand_Reg(L+1) := Next_Cand(
                   FailedOC(L),FailedAC,CMR)
   L := L+1
Return(CMR)
```

Fig. 2. CMAR: algorithm for computing all constrained minimal alteration regions.

```
Next_Cands (FailedOC,FailedAC,M, CMR)
 NL_Cands :=  ∅
 For each R = [in_i..out_j] ∈ FailedOC  do
 (1)   ClosR := { closure_P([(in_i − 1)..out_j],M),
                  closure_P([in_i..(out_j + 1)],M) }
 (2)   NL_Cands := Prune(Prune_Min (ClosR),
                         FailedAC,CMR))
 Return (NL_Cands)
```

Fig. 3. Candidate generation.

are stored to prune the search space (Fig. 3). If a region r satisfies AC but does not satisfy OC, it will be used to generate candidate regions for the next level.

When generating candidate minimal zones of level $L+1$ (Fig. 3), the algorithm first generates all the smallest closed supersequences of regions that failed against OC at level l (Step 1 of Algorithm 3). It then checks that none of the resulting regions is either a superset of a smaller region already in CMR (minimality constraint) or of a region that failed against AC (anti-monotonicity of AC).

EXAMPLE 8. Given the context of Table 1 and its minimal regions for Example 4, we will search for all minimal regions that have both a minimal support of two (AC) and are delimited in at least two observations (OC). At level 1, $[p_3..p_3]$ and $[p_{10}..p_{10}]$ succeed against AC and OC, $[p_1..p_1]$, $[p_5..p_5]$ and $[p_7..p_8]$ fail against OC and $[p_{11}..p_{11}]$ fails against AC. $[p_1..p_1]$ cannot be left extended, its right extension, $[p_1..p_8]$, is a superset of $[p_3..p_3]$, and should therefore be disregarded. The left extension of $[p_5..p_5]$ is $[p_3..p_6]$ and its right extension is $[p_5..p_6]$. The smallest of these two regions, $[p_5..p_6]$ is not well delimited on the left. Finally, $[p_5..p_6]$ cannot be left-extended any further without covering a minimal region. The result is thus $\{[p_3..p_3], [p_{10}..p_{10}]\}$. If OC is modified to be 'the region should be fuzzy delimited in at least two observations with margin $m = 1$', the above solution can be extended with $[p_5..p_5]$.

THEOREM 2. *The above algorithm is complete—it generates all minimal closed sequences of P that satisfy the constraints of the problem.*

The proof of this theorem is detailed in Appendix 1 of the Supplementary Material.

CMAR differs from algorithms that compute closed constrained itemsets in the context of biological constraints (Pang *et al*., 2003; Besson *et al*., 2005), because it handles sequential data. CMAR therefore searches the lattice of intervals of $[1..N_P]$, the size of which is $N_P(N_P-1)/2$, i.e. much smaller than the search space for itemsets, of size 2^{N_P}. It therefore does not need to rely on the Galois connexion used in the other approaches, to search the power-set of observations, $2^{|O|}$, which in most applications[1] is larger than N_P^2. The main difference between CMAR and state-of-the art sequence mining algorithms (Pei *et al*., 2002; Yan *et al*., 2003) is the type of sequences handled. The sequences CMAR handles are totally pre-aligned on a fixed set of probes spread throughout a given genome (here, the human genome), explaining why the algorithm has a quadratic worst-time complexity, whereas the other methods handle unaligned data streams. As a consequence, CMAR requires a simpler and more efficient partial ordering and fully exploits the characteristics of the handled sequences to generate candidate closed sequences efficiently (see Algorithm in Fig. 3). Finally, and unusually in the context of pattern mining (Mannila and Toivonen, 1997), this algorithm computes the most general (rather than the most specific) patterns satisfying the set of constraints.

Our approach can also be seen, from a different viewpoint, as a kind of biclustering algorithm (Madeira and Oliveira, 2004) for discrete data, with the user explicitly setting constraints concerning the shape of the 1-containing rectangles that he or she wishes to extract from the 0–1 context matrix (the height of the rectangle sets the frequency threshold, the closeness constraint ensures that this rectangle has maximal width for a given set of observations, etc.). The ordering of probes in the genome optimizes the efficiency of search for rectangles of 1s satisfying the constraints (Gionis *et al*., 2004).

[1]If the number of observations to handle exceeds 20 or 30.

3.2 Complexity

The algorithm enumerates closed sequences of $S(P)$, starting from the smallest ones. The number of probe sequences, and therefore of regions, is finite [in the worst case, $O(N_P^2)$], so the algorithm terminates. It stops when no candidate regions can be generated for a given level (i.e. all closed sequences of level l are either supersequences of a region of CMR or FailedAC). In other words, the complexity of the algorithm is in $O(N_P^2)$, but it is efficient as AC prunes small sequences and small sequences satisfy all the constraints of the problem. A more detailed discussion of complexity issues with CMAR can be found in Appendix 2 of the Supplementary Material.

4 VALIDATION

In this section, we validate the proposed algorithms by applying them to two different public datasets, containing CGH-array data for two kinds of tumour: colorectal tumours studied with BAC arrays (Nakao *et al*., 2004) and breast tumours studied with cDNA arrays (Pollack *et al*., 2002). These datasets have been handled as uniformly as possible: each dataset was first discretized, pre-processed and provided as input to the algorithms, which then computed the minimal (constrained) recurrent regions. Finally, genes were extracted from the obtained regions; the regions were visualised with VAMP software (http://bioinfo.curie.fr/projects/VAMP) and analysed manually.

Discretization was performed using the GLAD algorithm, as previously described (Hupe *et al*., 2004). The default parameters of the *R* function *glad.R* were used. The status (i.e. Gain, Normal, Loss) given by the Label assignment step is used as the input for the computation of minimal recurrent regions.

Missing values, which are frequent in microarray experiments (some spots and/or clones are discarded owing to poor quality), need not be preprocessed. During the minimal region computation step, unmeasured probes take the value of their neighbouring probes, as assigned by GLAD. They are therefore, by default, included in neighbouring regions.

GLAD automatically detects outliers, which are difficult to handle in the minimal region computation step: outliers may correspond to noise (e.g. mislocated probes, polymorphisms, etc.), or to highly valuable information (i.e. very narrow alteration regions). The taking into account of outliers during the minimal region computation step may yield very short (i.e. one-probe-long) regions, the statistical relevance of which may be difficult to evaluate. For this reason, many approaches simply ignore outliers and one-probe-long regions (Aguirre *et al*., 2004). We have implemented an outlier selection procedure (see details in Appendix 3 of the Supplementary Material) that makes use of the distributions of gain and loss outlier $\log_2$ ratios to select gain and loss outliers with significantly large (for gain outliers) or small (for loss outliers) $\log_2$ ratios. A similar strategy has been implemented in the CLAC approach (Wang *et al*., 2005). Outliers which are not selected are set as unmeasured.

Table 2. Excerpt of minimal regions for colon cancer data (Nakao *et al*., 2004)

Min. region	G/L	Freq	Gene	Cytoband(s)	No. BACs	No. of genes	Length (Mb)
CTD-2141B2..RP11-58G19	*l*	0.32	APC	5q22.2..5q22.3	3	16	3.0
RP11-9M11..RP11-19K9	*l*	0.19	PTEN	10q22.3..10q24.2	9	115	14.7
RP11-13J23..RP11-207J8	*l*	0.23	SMAD3	15q22.2..15q23	10	74	8.9
GS-185O2*	*l*	0.64	DCC	18q21.2	1	11	2.9
RP11-43K24	*l*	0.62	SMAD4	18q21.1	1	18	2.7
GS1-259E18..RP11-188A12	*g*	0.35	BRAF	7q32.1..7q35	23	113	14.4
RP11-265K5..RP11-73M19	*g*	0.23	'FGFR1 amplicon'	8p12..8p11.1	11	51	9.7
RP11-128G18..RP11-237F24	*g*	0.49	MYC	8q24.21	4	8	3.6
RP11-10D18..RP11-94A182*	*g*	0.66	STK6	20q13.31	17	93	11.9
GS-385N22*	*l*	0.53	—	18q21.33	1	18	2.3
RP11-29H19..RP11-169A62*	*g*	0.66	—	20q13.12	4	64	2.56

Each region is described with its bounding BACs (a single BAC if the region is one BAC long), its status (*g* for gain, *l* for loss), its frequency in the dataset, the gene associated with colorectal cancer tumour progression that it contains, its bounding cytobands (a single cytoband, if applicable), its length in BAC number, in gene number and in Mb. Regions are sorted according to their status and their location in the genome. The top part of the table lists regions that contain genes related to cancer, the lower part of the table gives the most frequently lost and gained regions not present in the top part of the table (denoted by *).

All datasets were treated with both MAR and CMAR algorithms. MAR does not handle constraints and has no parameters that must be adjusted. The current version of CMAR has three such parameters: minimal frequency threshold, and the bound and margin parameters (see Definition 10). These parameters can be adjusted according to the characteristics of the dataset and we describe briefly here and more precisely in the appendices, the adjustment of these parameters for the datasets studied. First, a region *r* should have a minimum frequency of 10% in the dataset; second, *r* should be bounded on the left and right in at least two profiles. Note that these left and right delimiting profiles are not necessarily the same. These constraints are very permissive: the minimum frequency is low (i.e. much lower than the frequency used in most current approaches), making it possible to detect relevant regions with a low recurrence rate. Setting *b* to 2 ensures that a region is not delimited because of noise in a single profile, thereby increasing the biological relevance of the regions obtained.

Finally, as a means of setting the value for the last parameter, the margin *m* (see Definition 10), we have studied the distribution of distances between two consecutive breakpoints on the same chromosome, for both gain and loss regions, and both in and out breakpoints (see Appendices 4 and 5 of the Supplementary Material). Intuitively, the distance between two related in- or out-breakpoints (related in the sense that they correspond to the same region) should be smaller than the distance between two unrelated in or out breakpoints (see figure 3 in Appendix 1 of Supplementary Material). The left and right margins for computing gain and loss regions can be set to the *n*-th percentile of such distributions. Basically, increasing the margin for a given bound has the effect of both increasing the number of regions and decreasing the mean size of regions. We will discuss here the results for *m* equals the first quartile of breakpoint distance distributions, for both gain and loss regions. This seems to provide a good compromise between the size of the minimal regions obtained and the number of regions obtained. This value of *m* also gives very good results in terms of the number of known oncogenes and tumour suppressor genes occurring in the constrained minimal regions.

For both datasets and for the parameter setting described below, full lists of minimal regions, and associated genes for regions containing 20 or fewer genes, are provided in Appendices 7a and b of the Supplementary Material for the Nakao *et al*. (2004) dataset and Appendices 8a and b of the Supplementary Material for the Pollack *et al*. (2002) dataset.

4.1 Colon cancer dataset

The (Nakao *et al*., 2004) dataset describes 125 CGH profiles, generated with a resolution of 1.5 Mb, on a human array. Each sample is described in terms of 2120 clones, 2081 of which were selected after pre-processing. A summary of the computed minimal regions can be found in Table 2.

MAR computed 142 minimal gain regions from this dataset and 173 minimal loss regions. Based on predefined constraints, CMAR computed 121 minimal constrained regions, 55 gain regions and 66 loss regions. We found that 17% of the total number of human genes considered, as defined in Appendix 3 of the Supplementary Material, belonged to gained regions whereas 16% of these genes belonged to lost regions. All the regions identified by Nakao *et al*., 2004 were identified by this algorithm, including the regions on chromosomes 8p and 20q. The mean length of gain regions was 7 BACs and 61 genes. Loss regions were slightly smaller: 5 BACs and 46 genes. The size of the regions in BAC clones ranged from 1 to 61, with 85% of the regions containing no more than 10 BACs. Most of the oncogenes and tumour suppressor genes known to be involved in colorectal cancer are found in the minimal regions of alterations (Table 2). Serpin genes, which have been identified as potential tumour suppressor genes, are located in the frequent minimal region of loss GS-385N22.

4.2 Breast cancer dataset

We used the dataset described by Pollack *et al*., (2002), for which both mRNA and DNA copy numbers had been determined with cDNA arrays. This dataset describes 41 profiles, 4 cell lines and 37 tumours, originally described in terms of 6095 cDNA probes, including 5758 retained after pre-processing. The cDNA technology is less sensitive for the detection of losses (Bilke *et al*., 2005), and this dataset seems much more noisy than the colon cancer dataset described in section 4.1: before pre-processing, about 1500 minimal regions were identified in the dataset, >90% of which were one

probe long. This high level of noise and the tendency of breast cancer tumours to display a high level of genetic rearrangement made it much more difficult to set the threshold for selecting outliers. We observed the distribution of outliers' $\log_2$ ratios for both tumoural and normal additional profiles (denoted X0, XX, XXX, XXXX and XXXXX), which the authors initially used to assess the sensitivity of the cDNA technique (the details can be found in Appendix 5 of the Supplementary Material).

MAR computed 350 minimal gain regions and 302 minimal loss regions, whereas CMAR computed 71 altered regions, including 36 loss and 35 gain regions. These regions contained 6.4 and 1.6 % of all the genes considered, respectively. The mean lengths of the regions of gain and loss were 1.6 and 9.8 cDNA probes, respectively, or 8.7 and 34 genes, respectively. Most of the gain regions identified by Pollack *et al*., (2002) or known to be involved in breast cancer are found in this list: I:773724 contains CCND1, I:825577..I:783729 contains ERBB2, a close but different region, I:236059 contains GRB7. The algorithm also identified regions containing RPS6KB1, NCOA3, ABC1, TP53. Clusterin, which has been identified both as a potential oncogene and as a tumour suppressor gene, is located in the frequently lost region, I:810358. PDCD4, a putative tumour suppressor gene, is located in the frequently lost region, I:328567.. I:268258. Some of the regions frequently lost and gained seem to be fragmented, lying very close to one another (see in particular, the various regions on 8q24 and 17q, in Appendix 8a of the Supplementary Material). Some of these neighbouring alteration regions probably correspond to a single minimal region as these two regions are separated by a single or a small number of cDNAs. This would be consistent with the findings of most studies that the minimal regions of amplification on 17q12–17q21 always contain both ErbB2 and GRB7.

5 DISCUSSION

As datasets describing copy-number genomic alterations in sets of samples obtained from large-scale analyses become increasingly common, the need for adequate formalization and tools for analysing such discrete datasets is also increasing.

We propose here two algorithms dedicated to the computation of minimal recurrent alteration regions. The first computes all minimal regions observed in a set of discretized alteration profiles. We then introduced a set of constraints to increase the efficiency of selection of biologically relevant regions, generating a second algorithm designed to compute all the minimal constrained alteration regions. The identification of minimal regions is extremely important in the search for genes involved in tumour progression. If the minimal regions are small enough (i.e. do not contain too many genes), the genes located in these regions can be studied in more detail. The genes located in a region of loss can be screened for inactivating mutations in the remaining allele. The tumour biologist involved in this study established, by reviewing the literature, a list of the most common oncogenes and tumour suppressor genes (putative or proven) involved in breast and colon tumours, and most of these genes were found to be located in the minimal regions identified. Moreover, as expected, the status of the regions (gained or lost) was consistent with the supposed function of the genes involved: 'gained' for the oncogenes and 'lost' for the tumour suppressor genes.

Although many biological studies have handled minimal regions for cancer-related studies [e.g. (Veltman *et al*., 2003; Tonon *et al*., 2005; Schraders *et al*., 2005; de Leeuw *et al*., 2004; Veltman *et al*., 2005)], very few studies (Aguirre *et al*., 2004; Tonon *et al*., 2005; Diskin *et al*., 2005) have tried to formalize the notion of relevant recurrent minimal regions of alterations or the process for automatically computing such regions across a set of observations. Aguirre *et al*., 2004 and Tonon *et al*., 2005 have made the most sophisticated attempt to date to formalize the process for computing common alteration regions in sets of CGH profiles. They introduce a method that selects relevant alteration regions based on both smoothed log2-ratios and frequency in the data. However, this method seems to focus more on high-amplitude deviations for the definition of potentially interesting regions. An empirical comparison of this method with CMAR will become possible once Aguirre *et al*. (2004) make their code available.

In this paper, we have dealt with datasets obtained with a BAC arrays of ∼2100 probes and a cDNA array of ∼6000 probes. Comprehensive segmental copy number arrays covering the whole genome (Ishkanian *et al*., 2004) and oligonucleotide arrays (Lucito *et al*., 2003; Herr *et al*., 2005) have recently been developed. We checked the generality of our approach by applying CMAR to a dataset describing eight mantle cell lymphoma (MCL) cell line profiles, obtained with tiling BAC technology (de Leeuw *et al*., 2004). Each profile in this dataset is described in terms of 32 433 probes (each spotted in triplicate), making it possible to evaluate the scaling-up capabilities of CMAR. With the same parameters as in the publication, with the margin parameter set as for the two previous datasets (see Appendix 6 of the Supplementary Material), CMAR obtained the regions listed in Appendix 9a of the Supplementary Material. This informal comparison showed a good overlap with the regions obtained by de Leeuw *et al*. (2004).

The identification of minimal regions should make it possible to decrease considerably the number of variables associated with a given tumour. Rather than having to know the status of all the probes used in the array, copy number alterations can be coded as the status of the minimal regions only, reducing the complexity from 2000 to 6000 variables (in the examples we have studied) to a few tens or hundreds of variables. Machine learning or statistics techniques, which could not be applied efficiently to the initial CGH data, could be applied to the simplified dataset. We are currently extracting association rules relating combinations of alteration regions to biological (specific gene mutations) or anatomical/clinical attributes, such as the stage of tumours (Rouveirol and Radvanyi, 2005).

This work could be developed in many different directions. First, CMAR performed well with data that had a low signal-to-noise ratio. Performance may be poorer in the presence of high levels of noise or considerable sample rearrangement, as we observed that some minimal regions computed from the Pollack dataset seemed to be fragmented, possibly owing to noise. However, most of the important cancer-related genes were still found in the minimal region computed for this dataset. An additional parameter could be added to CMAR to merge these regions, as proposed by Aguirre *et al*., 2004. This would involve minor changes to the minimal regions obtained in this case, as 10% of these regions were contiguous. This also suggests that another type of constraint may be more suitable for coping with noisy data. One such constraint might involve the computation of chromosomal regions with a high density of alterations rather than fully altered in a set of observations.

Genomic alterations have recently been studied with arrays composed of 100 000 SNPs (Matsuzaki *et al*., 2004) or oligonucleotide arrays (Lucito *et al*., 2003; Herr *et al*., 2005). These arrays differ from the arrays considered in this study in providing datasets with a much larger number of attributes. The CMAR algorithm, as demonstrated by the first experiment conducted on a tiling array dataset (with ∼30 000 probes), should be easy to adapt to the determination of minimal regions of alteration in genomic data obtained with much denser arrays.

ACKNOWLEDGEMENTS

The authors are particularly grateful to Ch. Froidevaux for her constant support and H. Radvanyi for fruitful discussions and A.V. Salle for her support during experimentations. They thank the anonymous referees for their pertinent suggestions. The authors also thank Alex Edelman & Associates for careful reading of the manuscript. This work was initiated, partially supported as part of European project HKIS IST-2001-38153 and mostly carried out while the first author was seconded to the CNRS and working in the Molecular Oncology Group, UMR 144. This work was supported by the CNRS, the Institut Curie and the Ligue Nationale Contre le Cancer, Comité d'Ile de France (Laboratoire Associé).

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. and Srikant,R. (1994) Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*. Morgan Kaufmann, San Francisco, CA, pp. 487–499.

Aguirre,A.J. *et al*. (2004) High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc. Natl Acad. Sci. USA*, **101**, 9067–9072.

Besson,J. *et al*. (2005) Constraint-based concept mining and its application to microarray data analysis. *Intell. Data Anal*., **9**, 59–82.

Bilke,S. (2005) Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays. *Bioinformatics*, **217**, 1138–1145.

de Leeuw,R.J. (2004) Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum. Mol. Genet*., **13**, 1827–1837.

De Raedt,L. and Kramer,S. (2001) The level-wise version space algorithm and its application to molecular fragment finding. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)* Morgan Kaufmann, San Francisco, CA, pp. 853–859.

Diskin,S.J., Eck,T., Greshock,J., Mosse,Y.P., Naylor,T., Stoeckert,C.J., Weber,B.L., Maris,J.M. and Grant,G.R. (2005) Statistical analysis of aCGH (STAC) a novel method for analysing multiple experiments. *AACR 2005*, Anaheim, CA.

Fearon,E.R. and Vogelstein,B. (1990) A genetic model for colorectal tumorigenesis. *Cell*., **61**, 759–767.

Ganter,B. and Wille,R. (1999) Formal Concept Analysis – Mathematical Foundations Springer, Berlin.

Gionis,A., Mannila,H. and Seppänen,J. (2004) Geometric and combinatorial tiles in 0-1 data. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. Springer Verlag, Berlin, pp. 173–184.

Herr,A. *et al*. (2005) High-resolution analysis of chromosomal imbalances using the Affymetrix 10K SNP genotyping chip. *Genomics*, **85**, 392–400.

Hupé,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Ishkanian,A.S. *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet*., **36**, 299–303.

Lucito,R. *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res*., **13**, 2291–2305.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **1**, 24–45.

Mannila,H. and Toivonen,H. (1997) Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Disc*., **1**, 241–258.

Margolin,A.A. *et al*. (2005) CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics*, **21**, 3308–3311.

Matsuzaki,H. *et al*. (2004) Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109–111.

Myers,C.L. *et al*. (2005) Visualization-based discovery and analysis of genomic aberrations in microarray data. *BMC Bioinformatics*, **6**, 146.

Nakao,K. *et al*. (2004) High resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, **25**, 1345–1357.

Ng,R.T., Lakshmanan,V.S., Han,J. and Pang,A. (1998) Exploratory mining and pruning optimizations of constrained associations rules. In *Proceedings of the 1998 ACM SIGMOD International Conference Management of Data*. Seattle, WA, pp. 13–24.

Pang,F., Cong,G., Tung,A., Yang,J. and Zaki,M. (2003) Carpenter : Finding closed patterns in long biological datasets. In *Proceedings of the SIGKDD'03*. ACM, pp. 637–642.

Pasquier,N. *et al*. (1999) Efficient mining of association rules using closed itemset lattices. *Inform. Syst*., **24**, 25–46.

Pei,J., Han,J. and Wang,W. (2002) Mining sequential patterns with constraints in large databases. In *Proceedings of the 2002 ACM CIKM Conference (2002),* Mac Lean, VA, pp. 18–25.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., **37**, S11–17.

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet*., **20**, 207–211.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci*., **99**, 12963–12968.

Rouveirol,C. and Radvanyi,F. (2005) Local pattern discovery in array-CGH data. In K. Morik, J.F. Boulicaut and A. Siebes (eds), *Local Pattern Detection*. Internal Seminar. Dagstuhl Castle, Revised Selected Paper. pp. 135–152, Springer.

Schraders,M. *et al*. (2005) Novel chromosomal imbalances in mantle cell lymphoma detected by genome-wide array-based comparative genomic hybridization. *Blood*, **105**, 1686–1693.

Solinas-Toldo,S. *et al*. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Tonon,G. *et al*. (2005) High-resolution genomic profiles of human lung cancer. *Proc. Natl Acad. Sci*., **102**, 9625–9630.

Veltman,I. *et al*. (2005) Identification of recurrent chromosomal aberrations in germ cell tumors of neonates and infants using genomewide array-based comparative genomic hybridization. *Genes Chromosomes Cancer*, 367–376.

Veltman,J.A. *et al*. (2003) Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res*., **63**, 2872–2880.

Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med*., **10**, 789–799.

Wang,P. *et al*. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.

Yan,X., Han,J. and Afshar,R. (2003) Clospan: Mining closed sequential patterns in large datasets. In *Proceedings of the 2003 SIAM Data Mining Conference (SDM 2003)*. San Francisco, CA, pp. 166–177.

[Cell Cycle 4:12, 1842-1846, December 2005]; ©2005 Landes Bioscience

Report

Preferential Occurrence of Chromosome Breakpoints within Early Replicating Regions in Neuroblastoma

Isabelle Janoueix-Lerosey[1,2]
Philippe Hupé[3,5]
Zofia Maciorowski[4]
Philippe La Rosa[5]
Gudrun Schleiermacher[1,2]
Gaëlle Pierron[6]
Stéphane Liva[5]
Emmanuel Barillot[2]
Olivier Delattre[1,2,*]

[1]Laboratoire de Pathologie Moléculaire des Cancers; [2]INSERM, U509; [3]UMR 144 CNRS; [4]Service de Cytométrie; [5]Service Bioinformatique; [6]Unité de Génétique Somatique, Institut Curie ; Paris, France

*Correspondence to: Olivier Delattre; Institut Curie; INSERM U509; Laboratoire de Pathologie Moléculaire des Cancers; 26 rue d'Ulm; Paris, 75248 France; Tel.: +33.1.42.34.66.79; Fax: +33.1.42.34.66.30; Email: delattre@curie.fr

Received 09/13/05; Accepted 10/14/05

Previously published online as a Cell Cycle E-publication:
http://www.landesbioscience.com/journals/cc/abstract.php?id=2257

KEY WORDS

neuroblastoma, replication timing, translocations breakpoints, genomic microarrays

ABBREVIATIONS

AWS	Adaptive Weights Smoothing
BIR	Break-Induced Replication
CGH	Comparative Genomic Hybridisation
Cy3	Cyanin 3
Cy5	Cyanin 5
DSB	Double Strand Break
NB	Neuroblastoma

ACKNOWLEDGEMENTS

We would like to thank Kathryn Woodfine and Nigel Carter from the Wellcome Trust Sanger Institute for providing the last release of their replication timing data. This work was supported by grants from the Ligue Nationale Contre le Cancer (Equipe labellisée). The construction of the 3.3k BAC array was supported by grants from the Carte d' Identité des Tumeurs (CIT) Program of the Ligue Nationale Contre le Cancer.

ABSTRACT

Neuroblastoma (NB) is a frequent paediatric extra cranial solid tumor characterized by the occurrence of unbalanced chromosome translocations, frequently, but not exclusively, involving chromosomes 1 and 17. We have used a 1 Mb resolution BAC array to further refine the mapping of breakpoints in NB cell lines. Replication timing profiles were evaluated in 7 NB cell lines, using DNAs from G_1 and S phases flow sorted nuclei hybridised on the same array. Strikingly, these replication timing profiles were highly similar between the different NB cell lines. Furthermore, a significant level of similarity was also observed between NB cell lines and lymphoblastoid cells. A segmentation analysis using the Adaptative Weights Smoothing procedure was performed to determine regions of coordinate replication. More than 50% of the breakpoints mapped to early replicating regions, which account for 23.7% of the total genome. The breakpoints frequency per 10^8 bases was therefore 10.84 for early replicating regions, whereas it was only 2.94 for late replicating regions, these difference being highly significant ($p < 10^{-4}$). This strong association was also observed when chromosomes 1 and 17, the two most frequent translocation partners in NB were excluded from the statistical analysis. These results unambiguously establish a link between unbalanced translocations, whose most likely mechanism of occurrence relies on break-induced replication, and early replication of the genome.

INTRODUCTION

Neuroblastoma (NB) is a frequent neoplasm of childhood that derives from primitive cell of the sympathetic nervous system.[1] The specific genetic alterations of NB tumors and cell lines have been explored through a panel of techniques, including Southern blot, FISH, allelotyping, chromosomal CGH (Comparative Genomic Hybridisation) and 24-color or spectral karyotyping.[2-9] Interestingly, these analyses revealed that most rearrangements were unbalanced. At the difference of balanced translocations, which exchange two chromosomal segments, giving rise to two derivative chromosomes without any gain or loss of genetic material, unbalanced translocations replace the distal segment of one chromosome by material from another chromosome. Unbalanced translocations are therefore characterized by the presence of only one derivative chromosome, and result in loss and/or gain of genetic material. Recently, a combined chromosomal CGH and 24-color karyotyping analysis of 27 NB cell confirmed that unbalanced translocations are 10 times more frequent than balanced translocations.[6] This combined analysis also lead to a low resolution mapping of nonreciprocal translocation breakpoints, revealing that the distribution of chromosome breakpoints was not random but skewed toward certain chromosomes particularly rich in early replicating regions. Recently, a powerful replication timing assay using genomic microarrays that allows quantifying the change in genomic copy number occurring during the S phase of the cell cycle has been described and used to determine the replication profile of a human lymphoblastoid cell line of normal karyotype.[10] In order to more precisely map and compare the positions of breakpoints with respect to replication timing, we now report the use of a genomic array containing 3400 BAC/PAC clones, to explore both the position of the unbalanced translocations breakpoints and the patterns of replication along each chromosome in several NB cell lines. We show that the replication timing profiles are highly similar among different NB cell lines and that a general conservation also exists between NB cell lines and lymphoblastoid cells. We unambiguously establish that unbalanced translocations in NB cell lines occur preferentially in early replicating regions.

©2005 LANDES BIOSCIENCE. DO NOT DISTRIBUTE

MATERIALS AND METHODS

Cell lines culture. NB cell lines (CLB-Ge, GI-M-EN, IMR32, KCNR, SJNB-8, SK-N-AS and SK-N-BE) were grown in DMEM or RPMI-1640 media supplemented with 10% foetal calf serum, as previously described.[6] Cells were harvested 24–48 hours after division to maximize the proportion of cells in S phase. Cells were centrifuged at 1500 rpm for 5 mins and washed twice with 1x PBS. The pellet was resuspended in 1 ml of 1x PBS and cells were fixed by the addition of 3 ml cold ethanol. Fixed cells were stored at 4°C until stained with propidium iodide for cell sorting.

Flow sorting and DNA extraction. Fixed cells were centrifuged 5 mins at 2800 rpm then resuspended in staining buffer containing 0.05% Tween 80, propidium iodide at 25 µg/ml and RNase at 25 µg/µl. This mixture was incubated 15 min at 37°C just prior to sorting. Stained nuclei were separated into different phases of the cell cycle using a FACSVantage diva SE cell sorter (BD Biosciences). For each cell line, two sorts were performed: in a first experiment, nuclei were sorted into total S and G_1 phase fractions whereas, in a second experiment, nuclei were sorted into S1 (first quarter of the S phase) and S2 (second quarter of the S phase) fractions. Typical ranges of cellular DNA content were 1.24, 1.4 and 1.45 for S1, S2 and total S fractions, respectively. The sorted fractions were checked for purity by analysis on a second sorter (FacsSort, BD Biosciences). To extract DNA, an equal volume of 2x lysis buffer (100 mM Tris pH = 7.5, 100 mM EDTA pH = 8, 2% SDS and 500 µg/µl proteinase K) was added to each fraction and the nuclei were incubated overnight at 50°C. DNA extraction was then performed using standard procedure.

Whole genome amplification. A whole genomic amplification was performed to amplify DNA for S1 fractions of 6 NB cell lines. No amplification was performed for the CLB-Ge cell line since, in that case, a sufficient amount of DNA was obtained from the sorted cells. We used the GenomiPhi DNA Amplification kit containing the Phi29 DNA Polymerase (Amersham Biosciences). 50 ng of genomic DNA were used for amplification, according to the manufacturer's instructions, which led to a yield of approximately 10 µg.

Genomic microarray preparation. Our genome wide arrays contain 3400 PAC/BAC clones spaced at approximately 1 Mb intervals, spotted in triplicate on Ultra Gaps slides (Corning). Chromosomes 1, 17 and 22 exhibit a higher coverage, the number of clones per Mb being 1.8, 2 and 1.5, respectively. Among chromosomes 1 and 17, the regions from 1pTel to 30 Mb on 1p arm and from 27.5 to 37.5 Mb on 17q arm are also covered at a higher density. All clones were validated by ends' sequencing. Precise coordinates of the clones were obtained by BLAT analysis on the May 2004 release of the Human Genome Browser Gateway (HGBG) (http://genome.ucsc.edu/).

DNA labelling, hybridisation and array analysis. Random primed labelling with dCTP-Cy3 or dCTP-Cy5 and hybridisations were performed as previously described.[11] Images were acquired using a dual laser scanner (Axon 4000B scanner, Axon Instruments). Spots quantification was done with GenePix Pro 5.1 imaging software (Axon Instruments).

Normalization. Normalization was performed using the MANOR algorithm, which enables to correct spatial effects (Neuvial et al manuscript in preparation). In brief, this algorithm proceeds in four steps: (1) estimation of the spatial trend on the background signal of Cy 3 channel using two-dimensional LOESS, (2) segmentation of the array into spatial areas with similar trend values, using the unsupervised classification algorithm NEM (Neighborhood Expectation Maximization) including spatial constraints, (3) definition of areas affected by spatial bias and (4) subtraction of the spatial trend from the log(Cy5/Cy3) values. After this process, clone replicates are averaged.

Statistical analysis

Breakpoint identification. We used the DAGLAD algorithm (Deletion, Amplification, Gain and Loss Analysis of DNA) to identify the breakpoints and genomic alterations in array-CGH profiles (Hupé et al manuscript in preparation). This algorithm is an improvement of the GLAD algorithm (Gain and Loss Analysis of DNA) which has been previously described.[12] This algorithm is based on the Adaptive Weights Smoothing (AWS) procedure and penalized likelihood. It allows the delineation of regions with similar DNA copy number. A breakpoint corresponds to the transition between two consecutive regions.

Correlation of replication timing data for the various NB samples. For each pair-wise comparison of S/G_1 ratios between two cell lines, a correlation coefficient, r, was calculated. Then, we tested the hypothesis of r being different from 0 using the Student's t-test on the t-statistic = $r.(N - 2)^{1/2} / (1 - r^2)^{1/2}$ with (N - 2) degrees of freedom, N being the sample size.

Determination of replication timing profiles using segmentation analysis. To determine the replication timing profiles, we used the AWS procedure to define piece wises of coordinate replication. For robustness consideration we decided to estimate the piecewise constant function over an averaged profile rather than for each cell line separately. For each profile, the log-ratios were centered on their median. The standard-deviation was estimated based on Inter Quartile Range as previously described[12] and centered profiles were scaled at unit variance then averaged. Finally, the averaged profile was rescaled by the average of the standard-deviation calculated over the set of profiles. The length of regions with coordinate replication is expected to be highly variable. In particular, the segmentation of the genomic profile must enable to delineate very small regions. Choosing a λ in the AWS procedure, as previously described,[12] equal to the 0.65 quantile of the $\chi^2(1)$ distribution led to an optimal segmentation result. Thresholds of smoothing values were chosen such that early replicating regions account for about 20% of the genome. For the total S phase and S1 fraction of NB cell lines, regions with a smoothing value greater or equal to 1.52 and 1.36, respectively, were considered as early replicating regions, whereas for Sanger's data (lymphoblastoid cell line),[10] we used a threshold of 1.68. To evaluate the proportion of early replicating regions common to both NB cell lines and lymphoblastoid cells in total S phase, we computed the intersection of the early replicating regions defined using the above described procedure between both types of samples.

Correlation between breakpoints' position and replication timing. We compared the distribution of breakpoints in regions of early or late replication. It is necessary to take into account the respective length in bases for each type of regions. Let's assume that the number of breakpoints in regions early and late replicated follows two independent Poisson processes N_E and N_L with parameters λ_E and λ_L. Let's denote L_E and L_L the lengths of the early and late replicating regions, respectively, and χ_E and χ_L the number of breakpoints in those regions.

The joint distribution is:

$$P(N_E = x_E, N_L = x_L) = \exp(-\mu_E)\, \mu_E^{x_E} / x_E ! \,.\, \exp(-\mu_L)\, \mu_L^{x_L} / x_L !$$

with: $\mu_E = \lambda_E L_E$ and $\mu_L = \lambda_L L_L$

Then the conditional probability is the following:

$$P(N_E = x_E \mid N_E + N_L = x_E + x_L) = C(x_E, x_E + x_L)\, \theta^{x_L} (1 - \theta)^{x_E}$$

with: $\theta = \mu_E / \mu_E + \mu_L$

The conditional distribution follows a binomial law $B(x_E + x_L, \theta)$ which is used to test the null hypothesis $H_0 : \lambda_E = \lambda_L$ with respect to the hypothesis $H_1 : \lambda_E \neq \lambda_L$. Under the null hypothesis, $\theta = L_E / L_E + L_L$.

RESULTS

Replication timing of the genome of NB cell lines. We used the microarray replication timing assay recently described[10] to determine the replication timing profile of 7 NB cell lines: CLB-Ge, GI-M-EN, IMR32, KCNR, SJNB-8, SK-N-AS and SK-N-BE. Flow sorted S1 (first quarter of the S phase) or total S phase DNAs labelled with Cy5 were cohybridised with flow sorted G_1 DNA labelled with Cy3. Using this approach, the measured ratio of S/G_1 phase DNAs is, for each clone, a direct measure of the average sequence copy number in the S phase fraction. Chromosome gains or losses, which are identical in the S and G_1 fractions, do not influence this ratio. For each clone, this ratio measures the proportion of nuclei in which this particular sequence has been replicated and hence, evaluates the

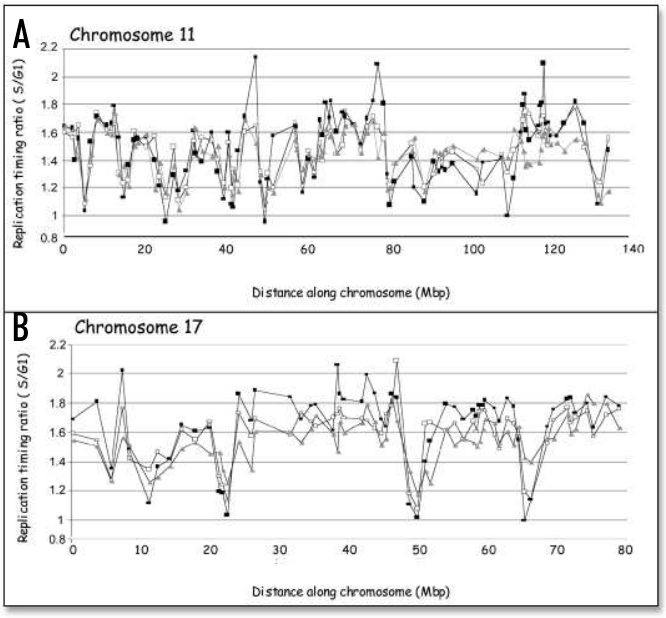


Figure 1. Replication timing profiles of chromosomes 11 and 17 in NB cell lines. DNAs from flow sorted nuclei in G_1 and S phase of the cell cycle were labelled with different fluorochromes and cohybridised on microarrays. Early replicating regions correspond to clones with a S/G_1 ratio close to 2:1. Circles, IMR32; squares, KCNR; triangles, SKNAS.

time at which this sequence replicates. This ratio was further scaled by the value of the median DNA content of the S phase (1.45 for total S, 1.24 for the S1 fraction). Indeed, identical amounts of DNA represent a lower number of S nuclei as compared to G_1 nuclei. Figure 1 shows the replication timing profiles of chromosomes 11 and 17 in three cell lines after sorting and labelling of the total S phase. Complete replication profiles of the seven cell lines are available on the following Web site (http://microarrays.curie.fr / Preferential occurrence of chromosome breakpoints within early replicating regions in neuroblastoma / Section 1). Consistent with previous observations, the pattern of replication was not uniform along chromosomes, early replicating regions being clearly interspersed with late replicating regions. Strikingly, the profiles obtained for the different NB cell lines were highly similar: pair-wised correlations of Cy5/Cy3 ratios across the genome for the analysis of the total S phase DNAs were highly significant (p-values < 10^{-4}, see Materials and Methods). Similar significances were observed for pair-wised comparisons of S1 fractions. Altogether, these results indicated a general similarity of the temporal replication program in these cell lines.

Given these strong similarities we considered the various NB cell lines as replicates and calculated an average replicating ratio for the seven samples, in total S or S1 fractions (see Materials and Methods). Consistent with previous data, early replicating regions appeared to be more abundant in chromosomes 1, 15, 17, 19 and 22 whereas chromosomes 4, 13, 18, 21 were globally late replicating chromosomes (Fig. 2A). The correlation of Cy5/Cy3 ratios across the genome between $S1/G_1$ and total S/G_1 fractions was highly significant ($r = 0.88$ and $p < 10^{-4}$) (see chromosome 17 in Fig. 2B). Strong correlations were observed between experiments with directly labelled DNA (total S phase) or with DNA subjected to whole genome amplification prior to labelling (S1 fraction). This indicated that the amplification process does not impair the measurements of the replication timing ratios.

Definition of regions of coordinated replication. In order to more precisely determine regions of coordinate replication, we used the Adaptive Weights Smoothing (AWS) procedure to perform a segmentation analysis on the data averaged from the 7 NB cell lines in total S phase or S1 (see Materials and Methods). Figure 3 shows the segmentation results obtained for chromosomes 11 and 17 using the total S fraction. Chromosome 17 is composed of early replicating segments of large size interspersed with six main segments of late replication (Fig. 3B). The resolution of the array enables to define precisely the borders and size of these segments, respectively of 1.5 Mb, 5.1 Mb, 2.9 Mb, 2.4 Mb, 2.5 Mb and 2.7 Mb. As expected, the six late replicating segments appear to correspond to regions of poor gene content.

We then sought to compare the replication timing profile of the NB cell lines in total S phase to that obtained for a human lymphoblastoid cell line of normal karyotype after a total S phase sorting.[10] In that respect, we obtained, from Kathryn Woodfine and Nigel Carter, the Cy5/Cy3 ratios corresponding to this lymphoblastoid cell line and performed a segmentation analysis of these ratios using the same AWS procedure as for the NB cell lines. The patterns of replication along each chromosome showed a significant level of similarity between the various NB cell lines that have been studied and the lymphoblastoid cell line: indeed, 60% of the early replicating regions defined in NB cell lines appear to be also early replicating in lymphoblastoid cells (http://microarrays.curie.fr, Section 2).

Replication timing and position of breakpoints. We then investigated the relationship between the breakpoints' position corresponding to unbalanced translocated chromosomes and the replication profiles in NB cells. Array-CGH profiles for the 27 NB cell lines are available on the Website (Section 3). Breakpoint positions were determined using the DAGLAD software (Hupé et al manuscript in preparation), a recently modified version of GLAD.[12] A total of 142 breakpoints associated to color transition by 24 color karyotype and gain or loss of genetic material by chromosomal CGH could be further mapped by array-CGH. We investigated the position of these breakpoints with respect to replication timing. Using the replication profile obtained with total S phase, 53.5% of the breakpoints associated to an unbalanced translocation mapped to early

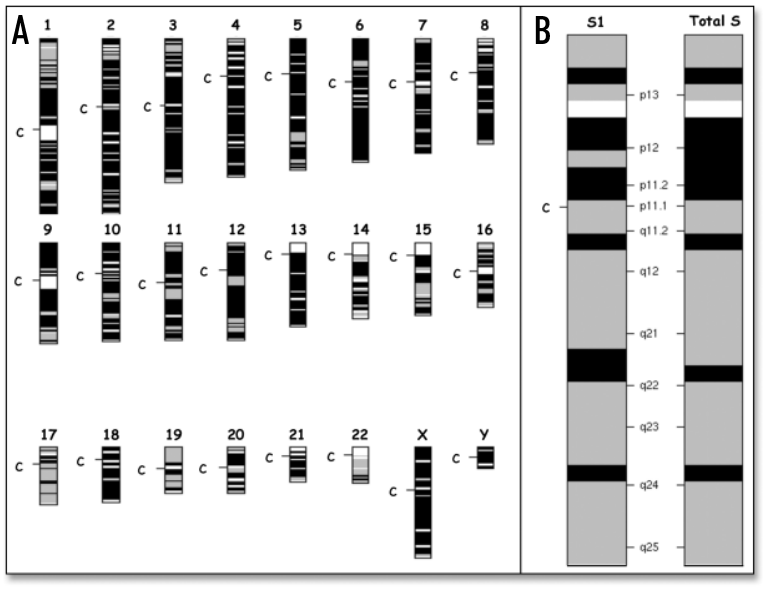


Figure 2. Replication timing pattern of NB cell lines. General view of the mean replication pattern of the whole genome in NB for the early S phase (S1) fraction. For better visualisation, three steps were subsequently applied: (1) for each clone, we calculated a mean replication ratio on the 7 analysed NB cell lines as described in Materials and Methods; (2) the genome was divided into 2.5 Mb intervals and the average ratio for the clones included in each 2.5 Mb interval was calculated, and (3) ratios were colored accordingly: ratio <1.36, black; ratio > = 1.36, grey; white, not determined. Early replicating regions therefore appear in grey. (B) Expanded views of chromosome 17 for S1 and total S fractions. For total S, ratios were colored accordingly: ratio <1.52, black; ratio > = 1.52, grey; white, not determined.

replicating regions, which only account for 23.7% of the total genome. These data may be visualized on the Web site (Section 4). Therefore, the breakpoints frequencies per 10^8 bases were 10.84 and 2.94 for early and late replicating regions, respectively. Similarly, using the profiles obtained with the S1 fractions, these breakpoint frequencies were 10.64 and 3.44 for early and late replicating regions, respectively (Table 1). Using the binomial law these differences in breakpoints rate were very highly significant ($p<10^{-4}$). To rule out a skewed analysis linked to coincidental occurrence of breakpoints and early replicating regions within particular chromosomes, we computed the statistics removing one chromosome at a time. These were still highly significant. Moreover, the p-value computed without the chromosomes 1 and 17 was still strongly significant indicating that the preferential occurrence of breakpoints in early replicating regions is a general characteristic of unbalanced translocations in NB.

DISCUSSION

In this study, we applied the method described by Woodfine et al.[10] to identify early and late replicating regions in NB cells. The power of this assay comes from the use of a genomic microarray and therefore represents a simple genome-wide method to measure the replication profile of a specific cell type. We observed a high similarity between the replication profiles measured for the seven analysed NB cell lines. Moreover, 60% of the early replicating regions defined in NB from the average pattern calculated for the seven samples appear also as regions of early replication in lymphoblastoid cells. These results are in agreement with recently published observations showing that the DNA replication-timing profiles of chromosome 22 is highly similar between different human cells.[13] Our data further indicate that the bulk of the temporal replication program in human cells is conserved throughout the genome, with differences in particular regions potentially associated to tissue-specific gene regulations.

We then examined the distribution of the breakpoints corresponding to unbalanced translocations in NB cell lines with respect to the replication timing status (early or late). Indeed, previous 24-color karyotyping has suggested that NB translocation breakpoints occur predominantly within early replicating regions.[6] However, these preliminary experiments were conducted with low resolution cytogenetic methods and did not rely on a specific analysis of the replication timing of NB cells. The present determination of replication profiles and mapping of breakpoints in NB on the same matrix providing a 1 Mb resolution unambiguously show the preferential occurrence of unbalanced translocations in early replicating regions. Further improvement of the resolution may be achieved with arrays containing overlapping tile path clones; indeed, it has been shown recently that a few regions of sharp change in the replication timing profile of chromosome 6 that were not detected at 1 Mb resolution may appear when using a genomic tile path array representing this chromosome.[14]

According to their genetic alterations, NB tumors may be classified into two main groups.[1] The first one includes tumors presenting a triploid number of chromosomes, without chromosome rearrangements but only loss or gain of entire chromosomes. The second type contains tumors of near diploid or tetraploid karyotype with unbalanced translocations, which are associated to gain and loss of material of the two implicated chromosomes and characterized by breakpoints arising preferentially outside pericentromeric sites, in early replicating regions. These features distinguish NB from carcinomas. Indeed, recent observations in various carcinomas have suggested that polysomy induces chromosome instability with unbalanced translocations mainly occurring in pericentromeric, late replicating regions.[15] In contrast, terminal and reciprocal (balanced) translocations are more frequent in carcinomas with a near diploid karyotype. Different hypotheses may account for this higher prevalence of translocations breakpoints in early replication regions. The concept that tissue specific chromosomal compartmentalization

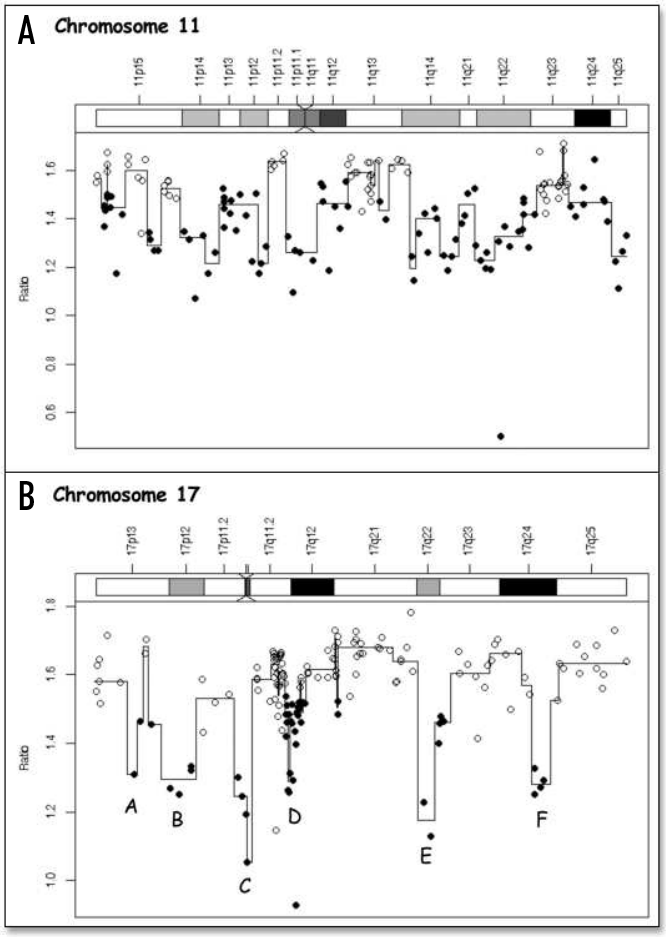


Figure 3. Optimal segmentation for chromosomes 11 and 17 for total S phase. Ratios were colored accordingly: ratio <1.52, filled circles; ratio >=1.52, open circles. See Materials and Methods for details.

Table 1 **Breakpoints associated to unbalanced translocations are preferentially observed in early replicating regions**

	Region length in Mb	Number of BP observed	BP frequency by 10^8 bases
Total S–Early replicating regions	701	76	10.84
Total S–Late replicating regions	2245	66	2.94
S1–Early replicating regions	564	60	10.64
S1–Late replicating regions	2382	82	3.44

A total of 142 breakpoints (BP) characterized by conventional cytogenetic approaches for 27 NB cell lines were further refined using array-CGH. Early and late replicating regions were defined after segmentation analysis of the data obtained for early (S1) or total S phase DNA (see Materials and Methods for details). For each type of region, the BP frequency per 10^8 was calculated according to (Number of BP / Region length in Mb)*100.

within the nucleus may lead to tumor-specific translocations has recently emerged and is supported by data indicating a physical proximity of chromosomes undergoing translocations.[16] Moreover, the chromosomal compartmentalization seems to be a highly regulated mechanism that preferentially localise early replicating regions in the interior nuclear compartment whereas late replicating regions localize predominantly in the peripheral compartment.[17] Preferential occurrence of 1;17 translocations in NB could therefore rely on a spatial proximity of early replicating donor and recipient chromosomes. Interestingly, early replicating regions mainly correspond to transcriptionally active and gene rich regions, the open chromatin structure of which may favor recombination. In that respect, it has recently been shown in yeast that transcription strongly stimulates mitotic recombination.[18]

A model to account for the unbalanced translocations observed in NB is the occurrence of reciprocal translocations between non-homologous chromatides during the S/G_2 phases of the cell cycle, followed by unequal segregation of both derivatives.[2] In such a model, apart from cases where a selective advantage is provided by one of the derivatives, a random segregation of normal and translocated chromosomes during mitosis should result in the transmission to daughter cells of both derivatives in one fourth of the cases. Such a frequency of reciprocal translocations is much greater than what is observed in NB. An alternative mechanism leading to the accumulation of unbalanced translocations in NB could rely on Break Induced Replication (BIR). In this mechanism, well described in the yeast *Saccharomyces cerevisiae*, the 3' end of a double strand break (DSB) invades a non homologous chromosome and replicates to the telomere, hence leading to an unbalanced translocation.[19] Rad51-dependent and Rad-51-independent BIR mechanisms have been described.[20,21] During the S phase, DSBs may arise at stalled replication forks and a deficient signalling or repair of these DSBs may serve as a starting point for BIR. Interestingly, interstitial telomere sequences lying at chromosome breakpoints have recently been observed in NB cells, suggesting that a telomere maintenance pathway may be altered in these cells.[22] The presence of both unbalanced translocations and telomere insertions suggests that a thorough analysis of the S phase checkpoint may help to understand both replication and telomere abnormalities in NB cells. Finally, a number of studies provided support for a role of segmental duplications in certain chromosome rearrangements.[23-25] Sequencing of several breakpoints in NB cell lines may enable this hypothesis to be explored.

Altogether, abnormal signalling of DSBs, nuclear vicinity of heterologous chromosomes, particular chromatin structure of early replicating regions may converge to favor the occurrence of BIR and, hence of unbalanced translocations in NB.

References

1. Brodeur GM. Neuroblastoma: Biological insights into a clinical enigma. Nat Rev Cancer 2003; 3:203-16.
2. Caron H, van Sluis P, van Roy N, de Kraker J, Speleman F, Voute PA, Westerveld A, Slater R, Versteeg R. Recurrent 1;17 translocations in human neuroblastoma reveal nonhomologous mitotic recombination during the S/G_2 phase as a novel mechanism for loss of heterozygosity. Am J Hum Genet 1994; 55:341-7.
3. Savelyeva L, Corvi R, Schwab M. Translocation involving 1p and 17q is a recurrent genetic alteration of human neuroblastoma cells. Am J Hum Genet 1994; 55:334-40.
4. Van Roy N, Laureys G, Van Gele M, Opdenakker G, Miura R, van der Drift P, Chan A, Versteeg R, Speleman F. Analysis of 1;17 translocation breakpoints in neuroblastoma: Implications for mapping of *neuroblastoma* genes. Eur J Cancer 1997; 33:1974-8.
5. Van Roy N, Van Limbergen H, Vandesompele J, Van Gele M, Poppe B, Salwen H, Laureys G, Manoel N, De Paepe A, Speleman F. Combined M-FISH and CGH analysis allows comprehensive description of genetic alterations in neuroblastoma cell lines. Genes Chromosomes Cancer 2001; 32:126-35.
6. Schleiermacher G, Janoueix-Lerosey I, Combaret V, Derre J, Couturier J, Aurias A, Delattre O. Combined 24-color karyotyping and comparative genomic hybridization analysis indicates predominant rearrangements of early replicating chromosome regions in neuroblastoma. Cancer Genet Cytogenet 2003; 141:32-42.
7. Schleiermacher G, Peter M, Michon J, Hugot JP, Vielh P, Zucker JM, Magdelenat H, Thomas G, Delattre O. Two distinct deleted regions on the short arm of chromosome 1 in neuroblastoma. Genes Chromosomes Cancer 1994; 10:275-81.
8. Trakhtenbrot L, Cohen N, Rosner E, Gipsh N, Brok-Simoni F, Mandel M, Amariglio N, Rechavi G. Coexistence of several unbalanced translocations in a case of neuroblastoma: The contribution of multicolor spectral karyotyping. Cancer Genet Cytogenet 1999; 112:119-23.
9. Cohen N, Betts DR, Trakhtenbrot L, Niggli FK, Amariglio N, Brok-Simoni F, Rechavi G, Meitar D. Detection of unidentified chromosome abnormalities in human neuroblastoma by spectral karyotyping (SKY). Genes Chromosomes Cancer 2001; 31:201-8.
10. Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, Debernardi S, Mott R, Dunham I, Carter NP. Replication timing of the human genome. Hum Mol Genet 2004; 13:191-202.
11. Fix A, Peter M, Pierron G, Aurias A, Delattre O, Janoueix-Lerosey I. High-resolution mapping of amplicons of the short arm of chromosome 1 in two neuroblastoma tumors by microarray-based comparative genomic hybridization. Genes Chromosomes Cancer 2004; 40:266-70.
12. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: From signal ratio to gain and loss of DNA regions. Bioinformatics 2004, 20:3413-22.
13. White EJ, Emanuelsson O, Scalzo D, Royce T, Kosak S, Oakeley EJ, Weissman S, Gerstein M, Groudine M, Snyder M, Schubeler D. DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. Proc Natl Acad Sci USA 2004; 101:17771-6.
14. Woodfine K, Beare DM, Ichimura K, Debernardi S, Mungall AJ, Fiegler H, Collins VP, Carter NP, Dunham I. Replication timing of human chromosome 6. Cell Cycle 2005; 4:172-6.
15. Kost-Alimova M, Fedorova L, Yang Y, Klein G, Imreh S. Microcell-mediated chromosome transfer provides evidence that polysomy promotes structural instability in tumor cell chromosomes through asynchronous replication and breakage within late-replicating regions. Genes Chromosomes Cancer 2004; 40:316-24.
16. Parada LA, McQueen PG, Misteli T. Tissue-specific spatial organization of genomes. Genome Biol 2004; 5:R44.
17. Stein GS, Zaidi SK, Braastad CD, Montecino M, van Wijnen AJ, Choi JY, Stein JL, Lian JB, Javed A. Functional architecture of the nucleus: Organizing the regulatory machinery for gene expression, replication and repair. Trends Cell Biol 2003; 13:584-92.
18. Aguilera A. The connection between transcription and genomic instability. Embo J 2002; 21:195-201.
19. Malkova A, Ivanov EL, Haber JE. Double-strand break repair in the absence of RAD51 in yeast: A possible role for break-induced DNA replication. Proc Natl Acad Sci USA 1996; 93:7131-7136.
20. Davis AP, Symington LS. RAD51-dependent break-induced replication in yeast. Mol Cell Biol 2004; 24:2344-51.
21. Ira G, Haber JE. Characterization of RAD51-independent break-induced replication that acts preferentially with short homologous sequences. Mol Cell Biol 2002; 22:6384-6392.
22. Schleiermacher G, Bourdeaut F, Combaret V, Picrron G, Raynal V, Aurias A, Ribeiro A, Janoueix-Lerosey I, Delattre O. Stepwise occurrence of a complex unbalanced translocation in neuroblastoma leading to insertion of a telomere sequence and late chromosome 17q gain. Oncogene 2005; 24:3377-84.
23. Debeer P, Mols R, Huysmans C, Devriendt K, Van de Ven WJ, Fryns JP. Involvement of a palindromic chromosome 22-specific low-copy repeat in a constitutional t(X; 22)(q27;q11). Clin Genet 2002; 62:410-4.
24. Pujana MA, Nadal M, Guitart M, Armengol L, Gratacos M, Estivill X. Human chromosome 15q11-q14 regions of rearrangements contain clusters of LCR15 duplicons. Eur J Hum Genet 2002; 10:26-35.
25. Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, Nair P, Brothman AR, Stallings RL. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. Genes Chromosomes Cancer 2005; 44:305-19.

A.2 Theorems

Theorem 1 (singular value decomposition) *Let $\mathbf{M}_{nm}$ be a matrix of size n, m (with $n > m$), then its SVD is expressed as follows:*

$$\begin{aligned} \mathbf{M}_{nm} &= \mathbf{U}_{nm}\mathbf{S}_{mm}\mathbf{V}'_{mm} \\ \mathbf{U}'\mathbf{U} &= \mathbf{I} \\ \mathbf{V}'\mathbf{V} &= \mathbf{I} \\ \mathbf{S} &= diag\{s_1, \ldots, s_m\} \text{ with } s_1^2, \ldots, s_m^2 \text{ eigenvalue of } \mathbf{MM}' \text{ and } \mathbf{M}'\mathbf{M} \end{aligned}$$

Theorem 2 (McNemar's test) *Let C_A and C_B be two supervised classification algorithms trained on the same training set and tested on the same test set. For each test sample, we record how it was classified by the two algorithms and construct the following contingency table:*

n_{00} = *number of samples misclassified by both* C_A *and* C_B	n_{01} = *number of samples misclassified by* C_A *but not by* C_B
n_{10} = *number of samples misclassified by* C_B *but not by* C_A	n_{11} = *number of samples misclassified by neither* C_A *nor* C_B

Under the null hypothesis that the two algorithms have the same prediction performance the following statistic is distributed as a $\chi^2(1)$:

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Theorem 3 (variance decomposition) *Let $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_p)$ be a (n, p) matrix of p continuous variables. Let $\mathcal{G}_1, \cdots, \mathcal{G}_K$ be K classes with n_k observations ($\sum_k n_k = n$). Let $\mathbf{X}_k$ be the submatrix of $\mathbf{X}$ corresponding to the observations in $\mathcal{G}_k$. Let $\boldsymbol{\mu} = E(\mathbf{X})$, $\boldsymbol{\Sigma} = V(\mathbf{X})$, $\boldsymbol{\mu}_k = E(\mathbf{X}_k)$ and $\boldsymbol{\Sigma}_k = V(\mathbf{X}_k)$. Then $\boldsymbol{\Sigma}$ can be expressed as follows:*

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B$$

with

$$\boldsymbol{\Sigma}_W = \frac{1}{n}\sum_{k=1}^{K} n_k \boldsymbol{\Sigma}_i \text{ and } \boldsymbol{\Sigma}_B = \frac{1}{n}\sum_{k=1}^{K} n_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})'(\boldsymbol{\mu}_k - \boldsymbol{\mu})$$

$\boldsymbol{\Sigma}$ is named the total variance-covariance matrix, $\boldsymbol{\Sigma}_W$ the within-class variance-covariance matrix and $\boldsymbol{\Sigma}_B$ the between-class variance-covariance matrix.

Theorem 4 (Cochran's theorem) *Let $X_1, \cdots, X_n$ be n independent and identically distributed random variables with $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. Let us define the following random variables:*

$$\hat{\mu} = \frac{1}{n}\sum_{i \leq n} X_i \text{ and } \hat{\sigma}^2 = \frac{1}{n-1}\sum (X_i - \hat{\mu})^2$$

Then:

$$\hat{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \tag{i}$$

$$(n-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1} \tag{ii}$$

$$\hat{\mu} \text{ and } \hat{\sigma}^2 \text{are independent} \tag{iii}$$

Theorem 5 (Bienaymé-Chebyshev inequality) *Let X be a random variable. Then, for any positive ϵ:*

$$P(|X - E(X) \geq \epsilon|) \leq \frac{V(X)}{\epsilon^2}$$

Theorem 6 (Gaussian vector) *Let $X_1, \cdots, X_n$ be Gaussian random variables. If they are independent then $a_1X_1 + \cdots a_nX_n$ is gaussian for any choice of real numbers $a_1, \cdots, a_n$. The vector $X = (X_1, \cdots, X_n)$ is said to be Gaussian.*

Asymptotic statistic theorems can be found in Van der Vaart (1998).

Theorem 7 (Slutsky's theorem) *Let (X_n) and (Y_n) be sequences of univariate random variables. If (X_n) converges in distribution to X and (Y_n) converges in probability to a constant c, then:*

$$X_n + Y_n \xrightarrow{d} X + c \tag{i}$$

$$X_n Y_n \xrightarrow{d} cX \tag{ii}$$

$$X_n / Y_n \xrightarrow{d} X/c \text{ if } c \neq 0 \tag{iii}$$

Theorem 8 (continuous mapping) *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $P(X \in C) = 1$.*

$$X_n \xrightarrow{p} X \text{ implies } g(X_n) \xrightarrow{p} g(X) \tag{i}$$

$$X_n \xrightarrow{d} X \text{ implies } g(X_n) \xrightarrow{d} g(X) \tag{ii}$$

Theorem 9 (central limit theorem) *Let $X_1, \cdots, X_n$ be n independent and identically distributed random variables with mean μ and variance σ^2. Then:*

$$n^{-1/2} \sum_{i \leq n} \frac{X_i - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

A.3 Supplementary figures of Chapter 3

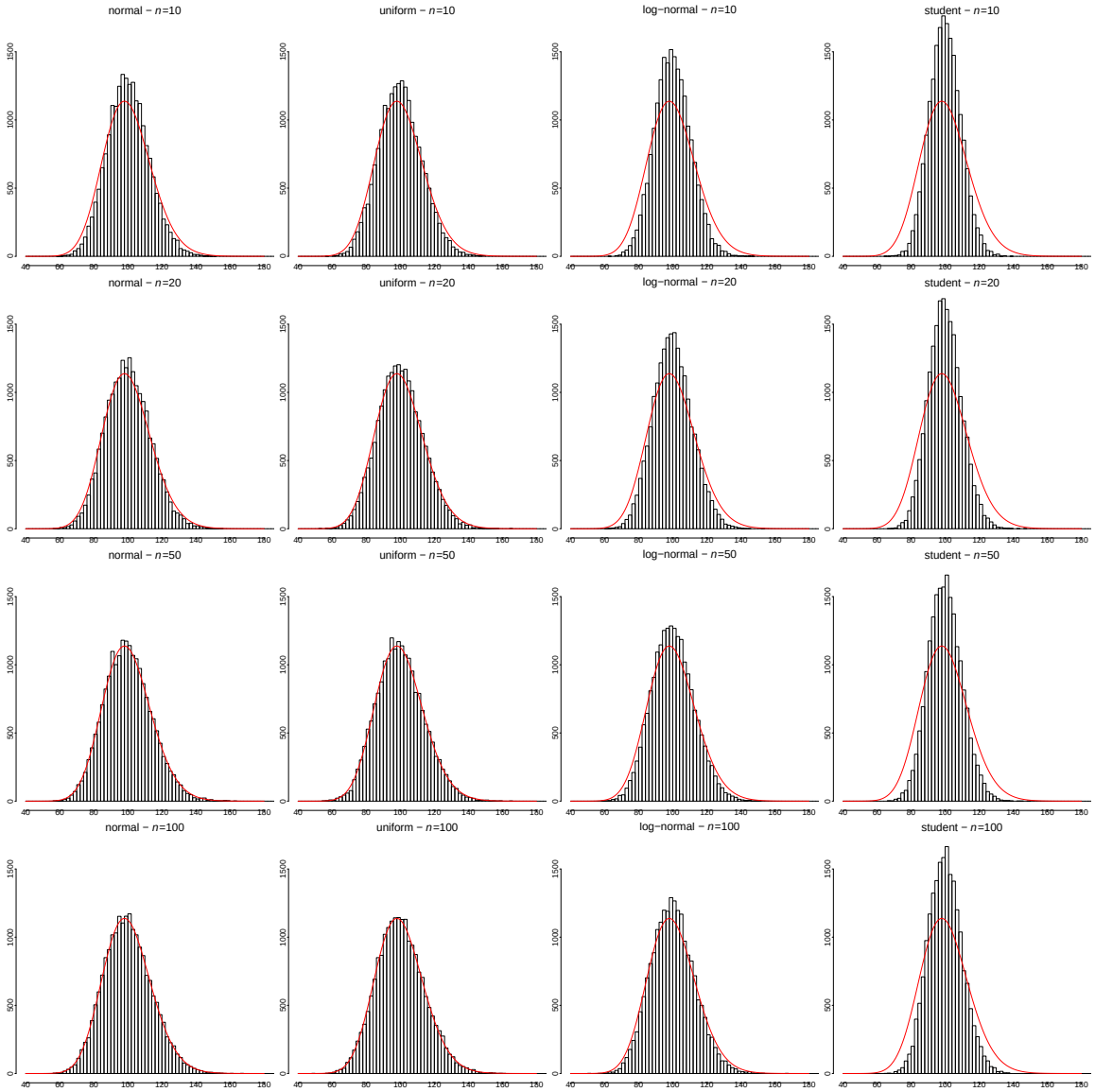


Figure A.1: Histograms for simulations under $\mathbf{H_0}$ without correlation - Histograms of the statistical criterion over 20000 simulations. The data have been simulated using the normal, uniform, log-normal and student distributions. n corresponds to the number of observations. The number of variables has been set at 100. The $\chi^2(100)$ probability density function is displayed as a red line.

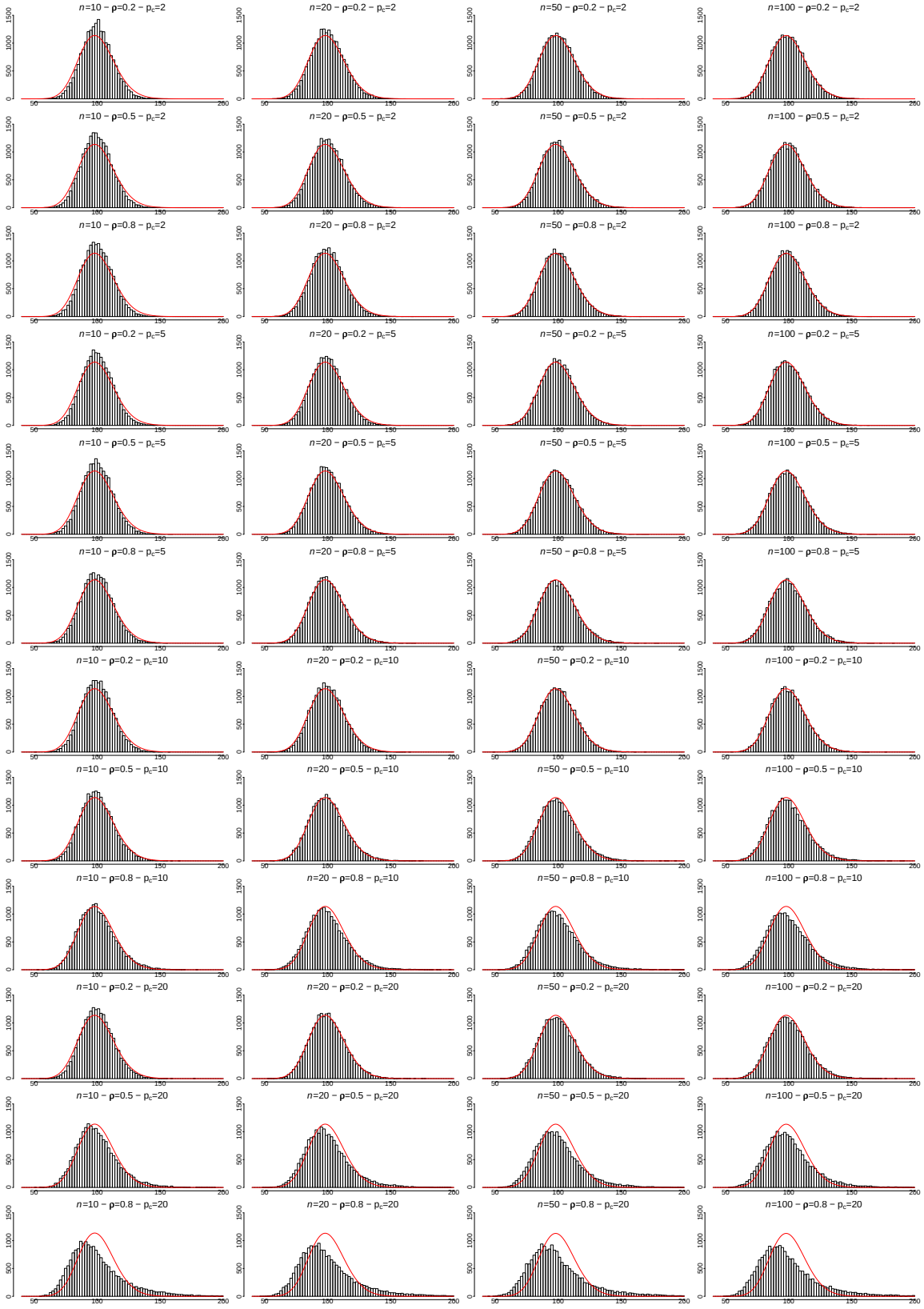


Figure A.2: Histograms for simulations under $\mathbf{H_0}$ with correlation - Histograms of the statistical criterion over 20000 simulations. The data have been simulated using the normal distribution. n corresponds to the number of observations, ρ is the correlation value, and p_c is the number of correlated variables. The number of variables has been set at 100. The $\chi^2(100)$ probability density function is displayed as a red line.

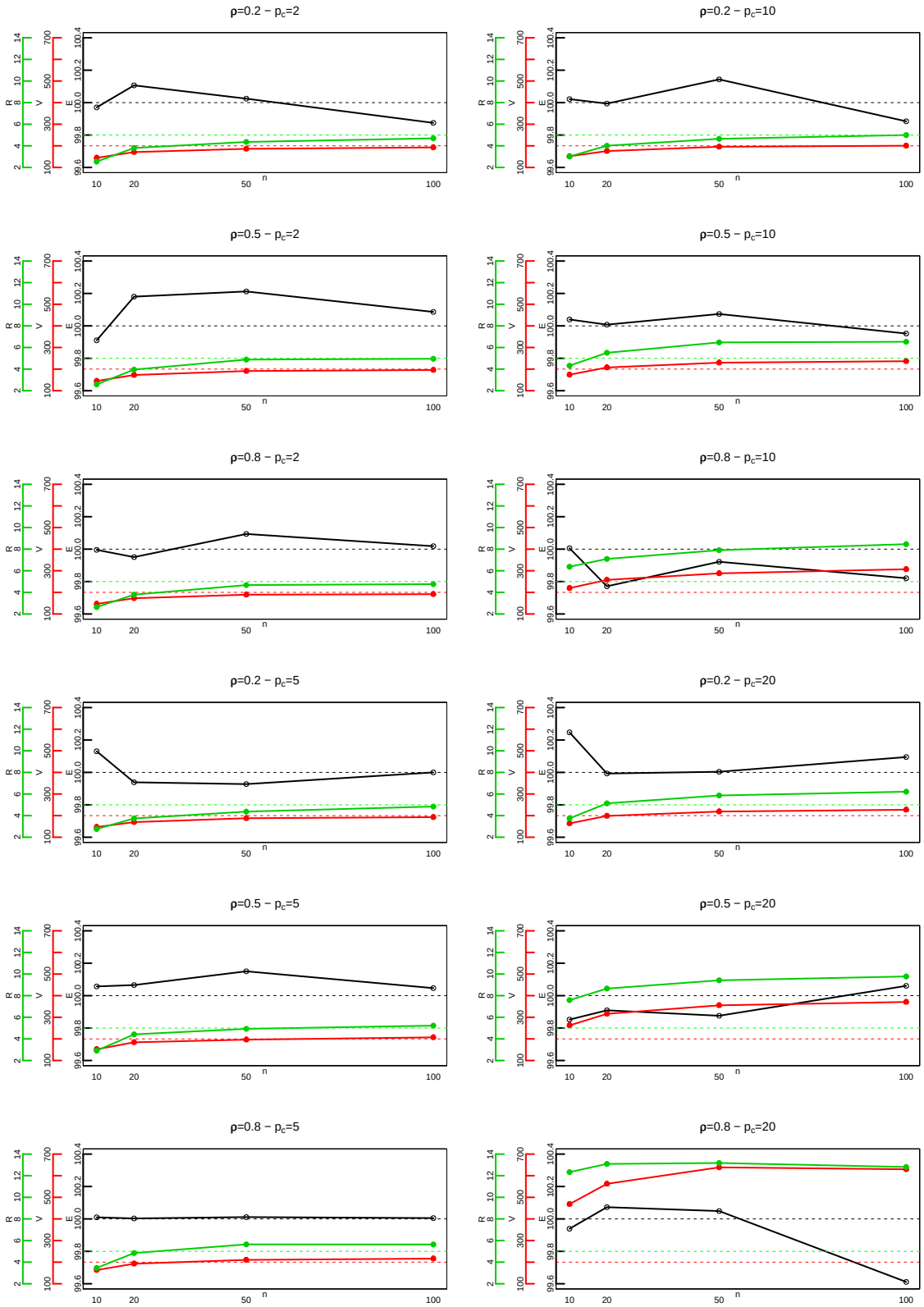


Figure A.3: Simulations under $\mathbf{H_0}$ with correlation - E is the empirical mean (in black), V is the unbiased empirical variance (in red) and R is the percentage of rejected hypotheses under $\mathbf{H_0}$ at the 5% level (in green) computed over 20000 simulations. Horizontal dashed lines represent the expected value for E (100), V (200) and R (5%) under $\mathbf{H_0}$ using the same color code. n corresponds to the number of observations, ρ is the correlation value, and p_c is the number of correlated variables. The number of variables has been set at 100.

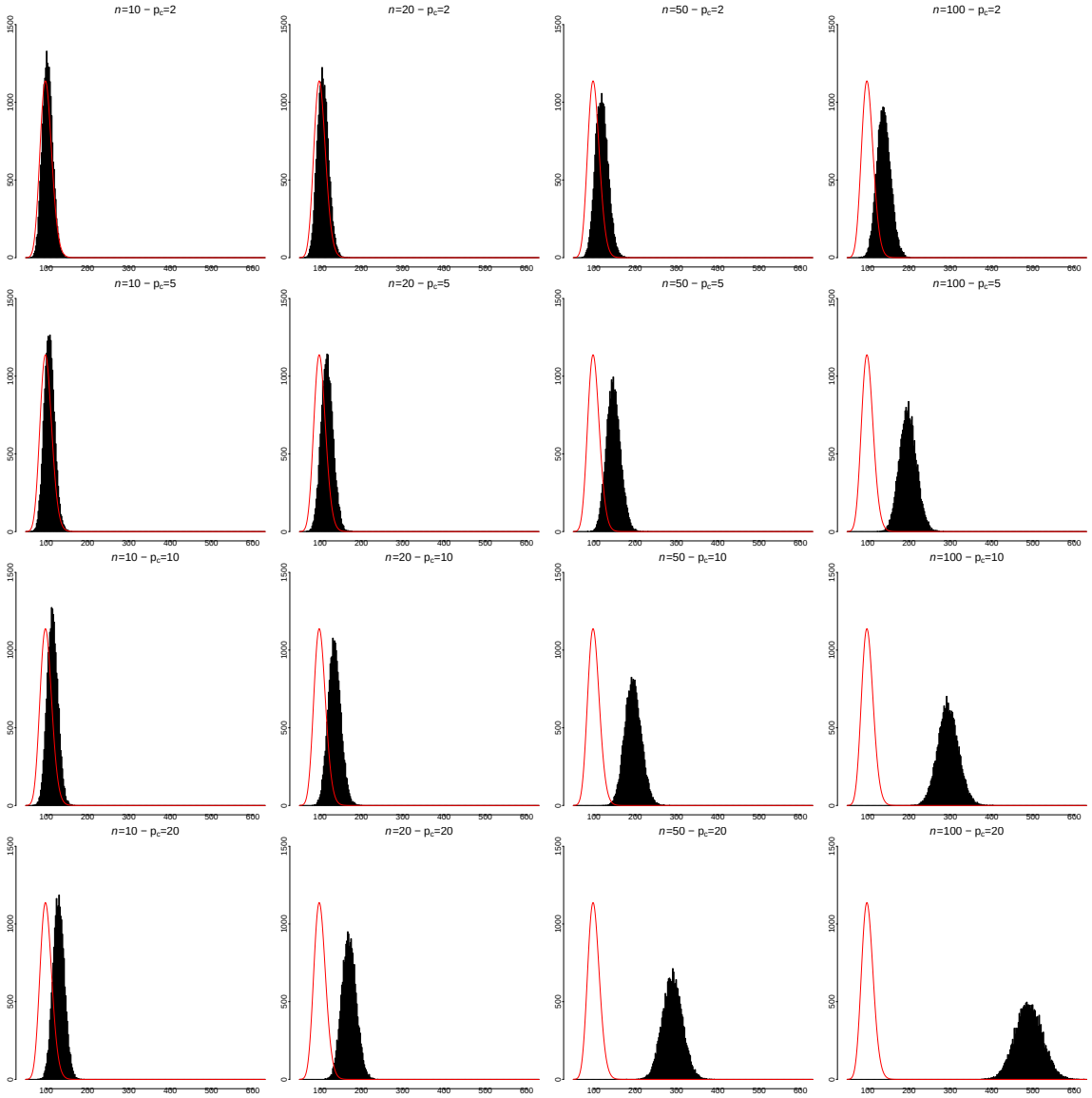


Figure A.4: Histograms for simulations under $\mathbf{H_1}$ without correlation - Histograms of the statistical criterion over 20000 simulations. The data have been simulated using the normal distribution. n corresponds to the number of observations and p_c is the number of variables with a class effect. The number of variables has been set at 100. The $\chi^2(100)$ probability density function is displayed as a red line.

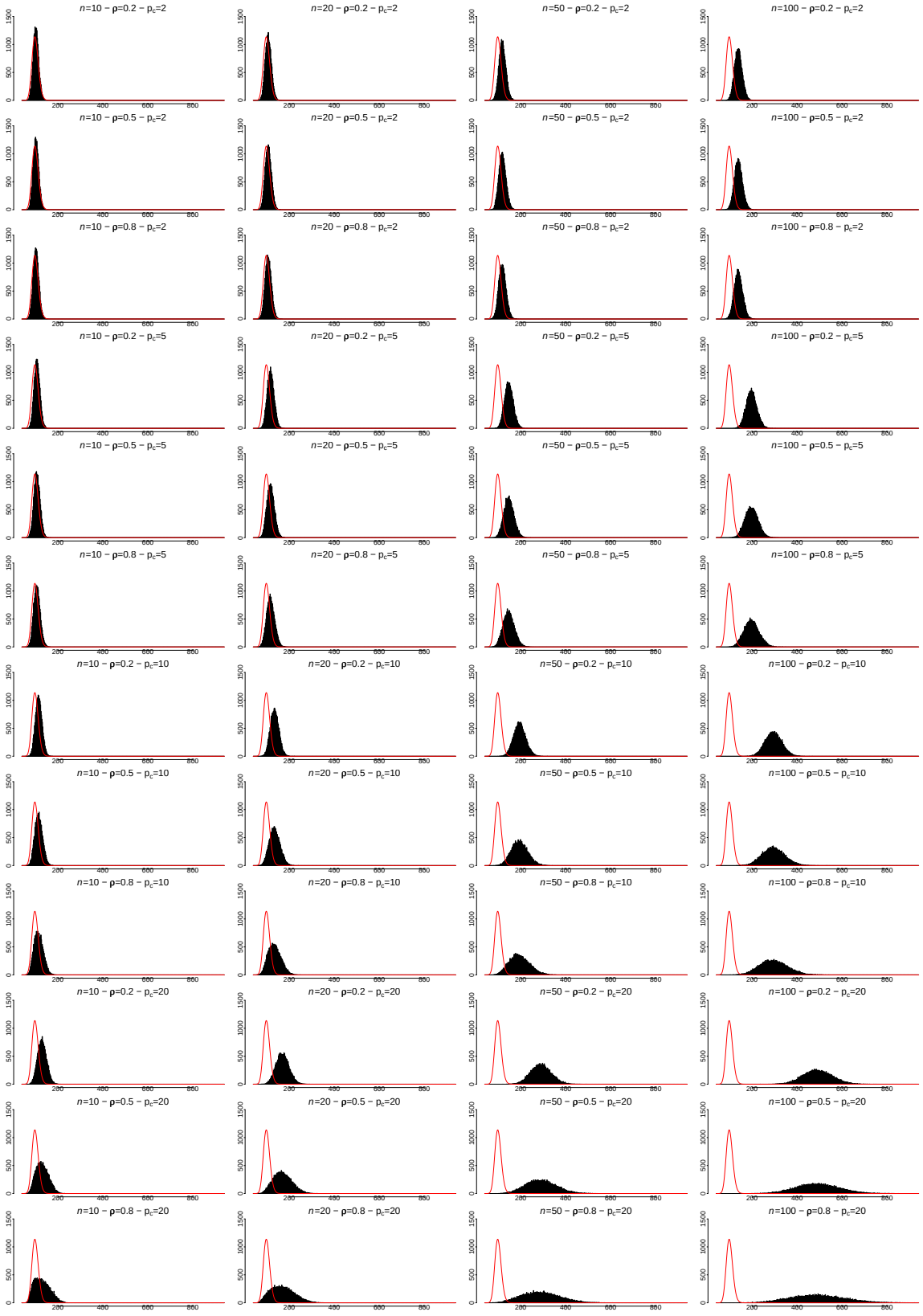


Figure A.5: Histograms for simulations under $\mathbf{H_1}$ with correlation - Histograms of the statistical criterion over 20000 simulations. The data have been simulated using the normal distribution. n corresponds to the number of observations, ρ is the correlation value, and p_c is the number of variables with a class effect and correlated. The number of variables has been set at 100. The $\chi^2(100)$ probability density function is displayed as a red line.

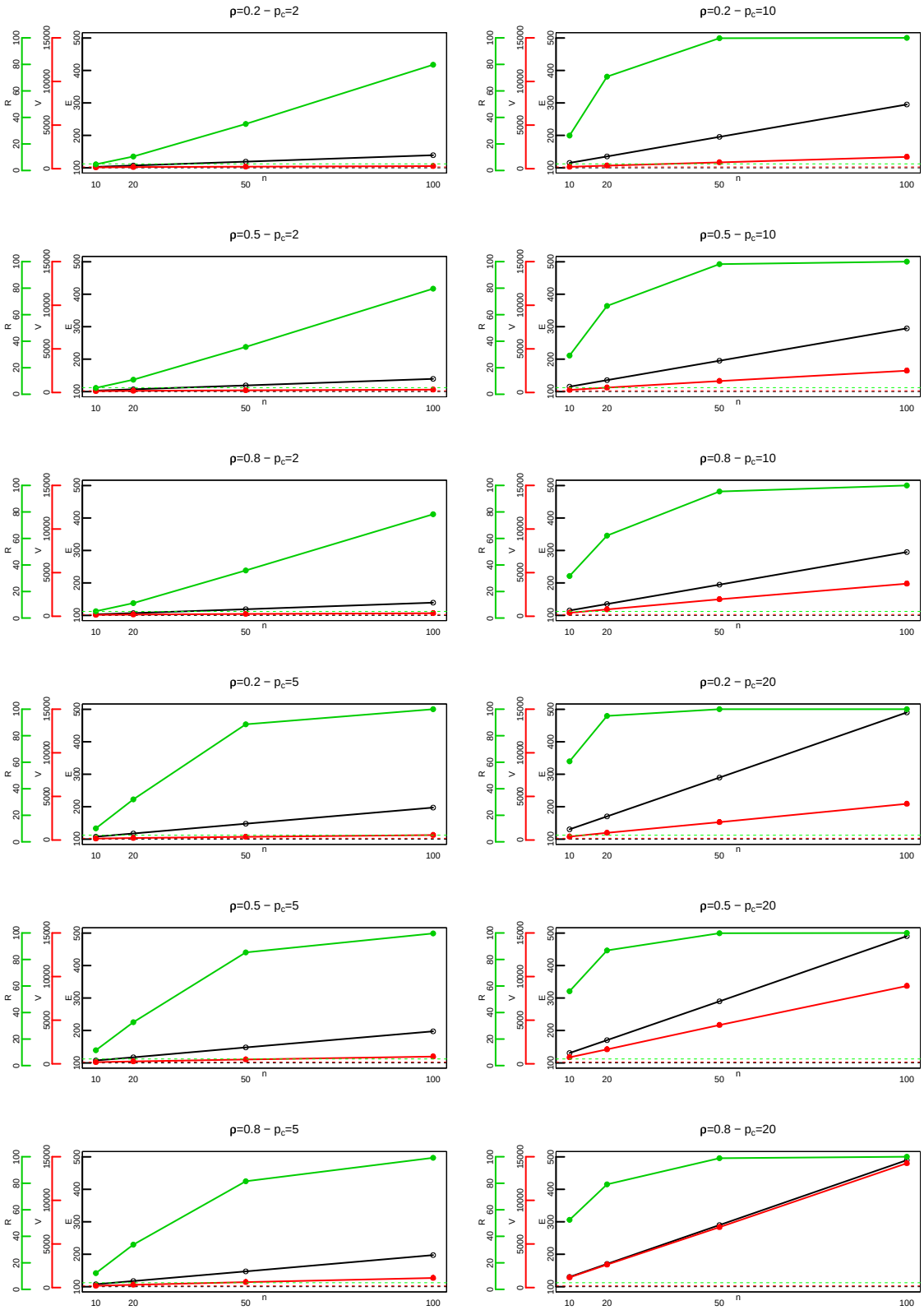


Figure A.6: Simulations under $\mathbf{H_1}$ with correlation - E is the empirical mean (in black), V is the unbiased empirical variance (in red) and R is the percentage of rejected hypotheses under $\mathbf{H_0}$ at the 5% level (in green) computed over 20000 simulations. Horizontal dashed lines represent the expected value for E (100), V (200) and R (5%) using the same colour code. n corresponds to the number of observations, ρ is the correlation value, and p_c is the number of variables with a group effect and correlated. The number of variables has been set at 100.

A.4 Supplementary tables of Chapter 3

P	M	m	n	p	1	2	3	4	5	6	7	8	9	10	11
g	C	LM	1	74.7											
g	C + A	LM	2	88.1	**										
a	C	LM	3	74.6	-	**									
a	C + A	LM	4	88	**	-	**								
g	C	P1	5	74.7	-	**	-	**							
g	C + A	P1	6	88.1	**	-	**	-	**						
a	C	P1	7	74.7	-	**	-	**	-	**					
a	C + A	P1	8	87.8	**	**	**	-	**	+	**				
g	C	P2	9	74.7	-	**	-	**	-	**	-	**			
g	C + A	P2	10	88.1	**	-	**	-	**	-	**	**	**		
a	C	P2	11	74.5	-	**	-	**	-	**	-	**	-	**	
a	C + A	P2	12	87.9	**	-	**	-	**	-	**	-	**	-	**

Table A.1: Comparison of performance in the additive case with the same number of observations in each subclass - The significance of the McNemar test is given: + indicates a p-value below 5% and ** a p-value below 1%. Either the two genes (g) or the two genes with the 100 random variables (a) have been used as continuous predictor variables (column P). The effects used in the MANOVA model are given in column M. Column m indicates what method has been used: LM, P1=PLS-LM (1) and P2=PLS-LM (2). Column n is the classifier number. Column p gives the prediction performance of each classifier. The log-linear model has not been used.

P	M	m	n	p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
g	C	LM	1	74.7																	
g	C + A	LM	2	89.7	**																
g	C×A	LM	3	91	**	**															
a	C	LM	4	74.6	-	**	**														
a	C + A	LM	5	89.5	**	**	**	**													
a	C×A	LM	6	91	**	**	-	**	**												
g	C	P1	7	74.8	-	**	**	-	**	**											
g	C + A	P1	8	88.9	**	**	**	**	**	**	**										
g	C×A	P1	9	88.9	**	**	**	**	**	**	**	-									
a	C	P1	10	74.6	-	**	**	-	**	**	-	**	**								
a	C + A	P1	11	88.5	**	**	**	**	**	**	**	**	**	**							
a	C×A	P1	12	88.5	**	**	**	**	**	**	**	**	**	**	-						
g	C	P2	13	74.7	-	**	**	-	**	**	-	**	**	-	**	**					
g	C + A	P2	14	89.7	**	-	**	**	**	**	**	**	**	**	**	**	**				
g	C×A	P2	15	91	**	**	-	**	**	-	**	**	**	**	**	**	**	**			
a	C	P2	16	74.5	-	**	**	-	**	**	-	**	**	-	**	**	-	**	**		
a	C + A	P2	17	89.2	**	**	**	**	**	**	**	-	-	**	**	**	**	**	**	**	
a	C×A	P2	18	90.4	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**

Table A.2: Comparison of performance in the interaction case with the same number of observations in each subclass - The significance of the McNemar test is given: + indicates a p-value below 5% and ** a p-value below 1%. Either the two genes (g) or the two genes with the 100 random variables (a) have been used as continuous predictor variables (column P). The effects used in the MANOVA model are given in column M. Column m indicates what method has been used: LM, P1=PLS-LM (1) and P2=PLS-LM (2). Column n is the classifier number. Column p gives the prediction performance of each classifier. The log-linear model has not been used.

P	M	L	n	p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
g	C		LM	1	84.8																		
g	C + A		LM	2	88.5	**																	
g	C + A	C×A	LM	3	89.6	**	**																
a	C		LM	4	84.7	-	**	**															
a	C + A		LM	5	88.3	**	+	**	**														
a	C + A	C×A	LM	6	89.5	**	**	+	**	**													
g	C		P1	7	84.8	-	**	**	-	**	**												
g	C + A		P1	8	88.5	**	-	**	**	**	**	**											
g	C + A	C×A	P1	9	89.6	**	**	-	**	**	+	**	**										
a	C		P1	10	84.8	-	**	**	-	**	**	-	**	**									
a	C + A		P1	11	88.1	**	**	**	**	-	**	**	**	**	**								
a	C + A	C×A	P1	12	89.4	**	**	+	**	**	-	**	**	+	**	**							
g	C		P2	13	84.8	-	**	**	-	**	**	-	**	**	-	**	**						
g	C + A		P2	14	88.5	**	-	**	**	+	**	**	-	**	**	**	**	**					
g	C + A	C×A	P2	15	89.6	**	**	-	**	**	+	**	**	-	**	**	+	**	**				
a	C		P2	16	84.8	-	**	**	-	**	**	-	**	**	-	**	**	-	**	**			
a	C + A		P2	17	88.2	**	+	**	**	-	**	**	+	**	**	-	**	**	+	**	**		
a	C + A	C×A	P2	18	89.3	**	**	**	**	**	**	**	**	**	**	**	-	**	**	**	**	**	
		C×A		19	70	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**

Table A.3: Comparison of performance in the additive case with a different number of observations in each subclass - The significance of the McNemar test is given: + indicates a p-value below 5% and ** a p-value below 1%. Either the two genes (g) or the two genes with the 100 random variables (a) have been used as continuous predictor variables (column P). The effects used in the MANOVA model are given in column M. The effects used in the log-linear model are given in column L. Column m indicates what method has been used: LM, P1=PLS-LM (1) and P2=PLS-LM (2). Column n is the classifier number. Column p gives the prediction performance of each classifier.

P	M	L	m	n	p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
g	C		LM	1	85.0																														
g	C + A		LM	2	89.9	**																													
g	C×A		LM	3	91.1	**	**																												
g	C + A	C×A	LM	4	91.0	**	**	-																											
g	C×A	C×A	LM	5	92.0	**	**	**	**																										
a	C		LM	6	85.0	-	**	**	**	**																									
a	C + A		LM	7	89.8	**	-	**	**	**	**																								
a	C×A		LM	8	91.0	**	**	+	-	**	**	**																							
a	C + A	C×A	LM	9	90.8	**	**	-	+	**	**	**	-																						
a	C×A	C×A	LM	10	91.8	**	**	**	**	+	**	**	**	**																					
g	C		P1	11	84.8	-	**	**	**	**	-	**	**	**	**																					
g	C + A		P1	12	88.2	**	**	**	**	**	**	**	**	**	**	**																			
g	C×A		P1	13	88.2	**	**	**	**	**	**	**	**	**	**	**	-																		
g	C + A	C×A	P1	14	89.4	**	**	**	**	**	**	-	**	**	**	**	**	**																	
g	C×A	C×A	P1	15	89.4	**	**	**	**	**	**	-	**	**	**	**	**	**	-																
a	C		P1	16	84.7	+	**	**	**	**	+	**	**	**	**	-	**	**	**	**															
a	C + A		P1	17	87.9	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**														
a	C×A		P1	18	87.8	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	-													
a	C + A	C×A	P1	19	89.1	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**												
a	C×A	C×A	P1	20	89.1	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	-											
g	C		P2	21	85.0	-	**	**	**	**	-	**	**	**	**	-	**	**	**	**	+	**	**	**	**	**									
g	C + A		P2	22	89.9	**	-	**	**	**	**	-	**	**	**	**	**	**	**	**	**	**	**	**	**	**									
g	C×A		P2	23	91.1	**	**	-	-	**	**	**	+	-	**	**	**	**	**	**	**	**	**	**	**	**	**								
g	C + A	C×A	P2	24	91.0	**	**	-	-	**	**	**	-	+	**	**	**	**	**	**	**	**	**	**	**	**	**	**	-						
g	C×A	C×A	P2	25	92.0	**	**	**	**	-	**	**	**	**	+	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**					
a	C		P2	26	85.0	-	**	**	**	**	-	**	**	**	**	-	**	**	**	**	-	**	**	**	**	-	**	**	**	**					
a	C + A		P2	27	89.6	**	**	**	**	**	**	**	**	**	**	**	**	**	-	-	**	**	**	+	+	**	**	**	**	**	**				
a	C×A		P2	28	90.6	**	**	**	+	**	**	**	**	-	**	**	**	**	**	**	**	**	**	**	**	**	**	**	+	**	**	**	**		
a	C + A	C×A	P2	29	90.7	**	**	+	**	**	**	**	-	-	**	**	**	**	**	**	**	**	**	**	**	**	**	+	**	**	**	**	-		
a	C×A	C×A	P2	30	91.5	**	**	+	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	+	**	**	**	**	**	**	
		C×A		31	70.0	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**

Table A.4: Comparison of performance in the additive case with a different number of observations in each subclass - The significance of the McNemar test is given: + indicates a p-value below 5% and ** a p-value below 1%. Either the two genes (g) or the two genes with the 100 random variables (a) have been used as continuous predictor variables (column P). The effects used in the MANOVA model are given in column M. The effects used in the log-linear model are given in column L. Column m indicates what method has been used: LM, P1=PLS-LM (1) and P2=PLS-LM (2). Column n is the classifier number. Column p gives the prediction performance of each classifier.