Acquisition et traitements statistiques des données de génomique

Carine Bernard, Bruno Meunier, Isabelle Cassar-Malek, Jean-François J.-F. Hocquette

Cah. Techn. I.N.R.A., 2004 52, 29-44

ACQUISITION ET TRAITEMENTS STATISTIQUES DES DONNEES DE GENOMIQUE

Carine Bernard [1,2], Bruno Meunier [2,3], Isabelle Cassar-Malek [2], Jean-François Hocquette [2]

ABREVIATIONS

ADNc : Acide désoxyribonucléique complémentaire
ARNm : Acide ribonucléique messager
Bdf : Bruit de fond
FDR : False Discovery Rate
PCR : Polymerisation Chain Reaction
SAM : Significance Analysis of Microarray
SCE : Somme des carrés des écarts à la moyenne
SNR : Signal to Noise Ratio

RESUME

Le développement des nouvelles technologies de génomique, en particulier celles permettant l'étude du transcriptome, soulèvent de nombreuses difficultés techniques pour l'acquisition et le traitement statistique des données. Ces difficultés sont notamment rencontrées par les biologistes moléculaires qui n'ont pas de formation spécifique en analyse d'images ou en analyse de données. Une première étape essentielle est en effet l'analyse des images résultant de l'hybridation moléculaire entre les gènes (représentés par des fragments d'ADNc préparés au laboratoire) et leurs produits d'expression dans les tissus étudiés. Après l'étape de localisation des spots, la segmentation des images (séparation des pixels en deux classes : bruit de fond non spécifique et signaux spécifiques) peut être effectuée selon différentes méthodes, puis les données caractéristiques à chaque spot (moyenne, médiane, écart-type) sont extraites. Plusieurs méthodes sont également disponibles pour le calcul de l'intensité de chaque spot (en retirant les signaux non spécifiques) mais aussi pour le filtrage des données de façon à ne conserver que les spots de qualité et d'intensité supérieure à celle du bruit de fond. Il existe également différentes méthodes de normalisation des données permettant d'éliminer toute variabilité technique et ainsi de comparer les données issues de différents échantillons, de différentes expériences d'hybridations ou de plusieurs réseaux. Enfin, les méthodes d'analyses statistiques classiquement utilisées en biologie doivent être adaptées à l'étude du transcriptome en raison du très grand nombre de gènes analysés de façon simultanée et compte tenu des caractéristiques des schémas expérimentaux et des facteurs de variabilité technique et biologique de ce type d'approche. Les choix méthodologiques de l'expérimentateur tout au long de cette chaîne de traitement dépendent des choix techniques préalables, interagissent entre eux et déterminent ensemble la fiabilité et la qualité des résultats obtenus.

MOTS CLES : transcriptome, analyse d'image, expression différentielle, statistique

[1] Carine Bernard a préparé un DESS de Bioinformatique à l'Université Blaise-Pascal de Clermont-Ferrand en 2003 et a effectué son stage de DESS à l'INRA.
[2] INRA, Unité de Recherches sur les Herbivores, Equipe Croissance et Métabolisme du Muscle, Centre de Clermont-Ferrand/Theix, 63122 Saint Genès-Champanelle
[3] Pour toute correspondance: bruno.meunier@clermont.inra.fr

1. INTRODUCTION

Les réseaux et les puces d'ADN constituent une nouvelle technologie à haut débit d'analyse d'un grand nombre de gènes de façon simultanée. Cette technologie permet d'étudier en une seule expérimentation le transcriptome, c'est-à-dire les niveaux des ARNm (ou transcrits) de l'ensemble des gènes étudiés dans les échantillons biologiques analysés. Elle donne donc accès aux profils d'expression des gènes dans une cellule, un tissu ou un organe d'intérêt, et de ce fait, à l'identification de gènes différentiellement exprimés selon différentes conditions physiologiques ou pathologiques. Il existe toutefois d'autres types de puces qui permettent le séquençage de l'ADN par l'hybridation, l'analyse de mutations et la recherche de polymorphismes. Toutes ces technologies concourent au même objectif : la recherche de gènes (mutés, polymorphes ou différentiellement exprimés) responsables des phénotypes en physiologie et en pathologie, et en particulier des gènes responsables des maladies humaines. Les applications qui en découlent sont nombreuses : le diagnostic de maladies et notamment de cancers, la pharmacogénomique où elles constituent un outil de choix dans la recherche et la caractérisation de nouvelles molécules à visée thérapeutique, la toxicogénomique, l'agroalimentaire, l'environnement…(Schena *et al.,* 1998). De nombreuses applications sont également envisagées chez les animaux de rente comme par exemple l'évaluation de la qualité de la viande bovine (Hocquette *et al.,* 2003).

Le principe des réseaux et des puces repose sur la technique d'hybridation entre des sondes fixées sur un support (membrane de nylon, lame de verre) et des cibles marquées préalablement, préparées à partir d'ARN extraits des tissus étudiés. Les sondes correspondent à des ADNc connus ou à des séquences oligonucléotidiques spécifiques de gènes d'intérêt. Elles sont déposées sur le support sous forme de spots à l'aide d'un robot. Les cibles représentent les ARNm des tissus rétro-transcrits en ADNc. Les signaux d'hybridation sont détectés selon le marquage des cibles, soit par autoradiographie dans le cas d'un marquage radioactif au ^{33}P, soit par fluorescence dans le cas de l'utilisation de fluorochromes tels que la cyanine 3 (Cy3 - Rouge) et la cyanine 5 (Cy5 - Vert). Les signaux obtenus (intensités des spots résultant de l'hybridation) sont ensuite quantifiés par un logiciel d'analyse d'images. Toutes ces étapes sont illustrées Figure 1. De nombreux auteurs ont montré la nécessité de déposer chaque sonde plusieurs fois sur le même support (réplicats intra-support) et d'hybrider plusieurs supports avec la même cible (réplicats inter-support) afin de valider statistiquement les données d'expression des gènes (Lee *et al.,* 2000).

Il existe différents types de réseaux et de puces qui diffèrent par la nature du support utilisé et/ou la densité de dépôt et les caractéristiques des sondes :
(i) les macro-réseaux (ou macro-arrays) : ce sont des membranes de nylon, ou des supports plastique, sur lesquels sont généralement déposés des ADNc issus de clones bactériens et préalablement amplifiés par PCR. Le marquage des cibles est radioactif, faute de pouvoir utiliser une détection fluorescente à cause de l'autofluorescence du nylon. Les macro-arrays 8 x 12 cm^2 permettent en moyenne d'effectuer jusqu'à 5000 dépôts.
(ii) les micro-réseaux (ou micro-arrays) : ils peuvent être réalisés soit sur nylon, soit sur lames de verre avec l'utilisation de marqueurs radioactifs ou de fluorochromes selon le type de support choisi. Les micro-arrays reposent sur le développement des membranes à haute densité, à dépôts d'ADN (oligomères ou ADNc) permettant d'étudier environ 10000 gènes.
(iii) les puces à oligonucléotides (oligo-chips) : elles font partie des puces les plus petites et les plus denses avec une synthèse des oligomères directement sur le support. Les capacités

actuelles de dépôt atteignent 250000 oligonucléotides sur une surface de 1,28 x 1,28 cm^2 (Le technoscope de BIOFUTUR 206, 2000). Les différentes méthodes d'analyse d'image, de quantification et filtrage des signaux, de normalisation des données et d'analyses statistiques qui vont être présentées doivent tenir compte des caractéristiques propres à chaque type de réseau.

Cet article a pour objectif de présenter les principales méthodes (i) d'analyse des images d'hybridation de réseaux et puces, (ii) de quantification et filtrage des signaux obtenus, (iii) de normalisation des données et (iii) d'analyses statistiques des résultats, méthodes que l'on peut trouver dans la littérature et qui sont adaptées à l'étude du transcriptome. Ces méthodes seront présentées de façon simple pour pouvoir répondre au mieux aux attentes des biologistes moléculaires qui ne sont pas familiarisés avec celles-ci.

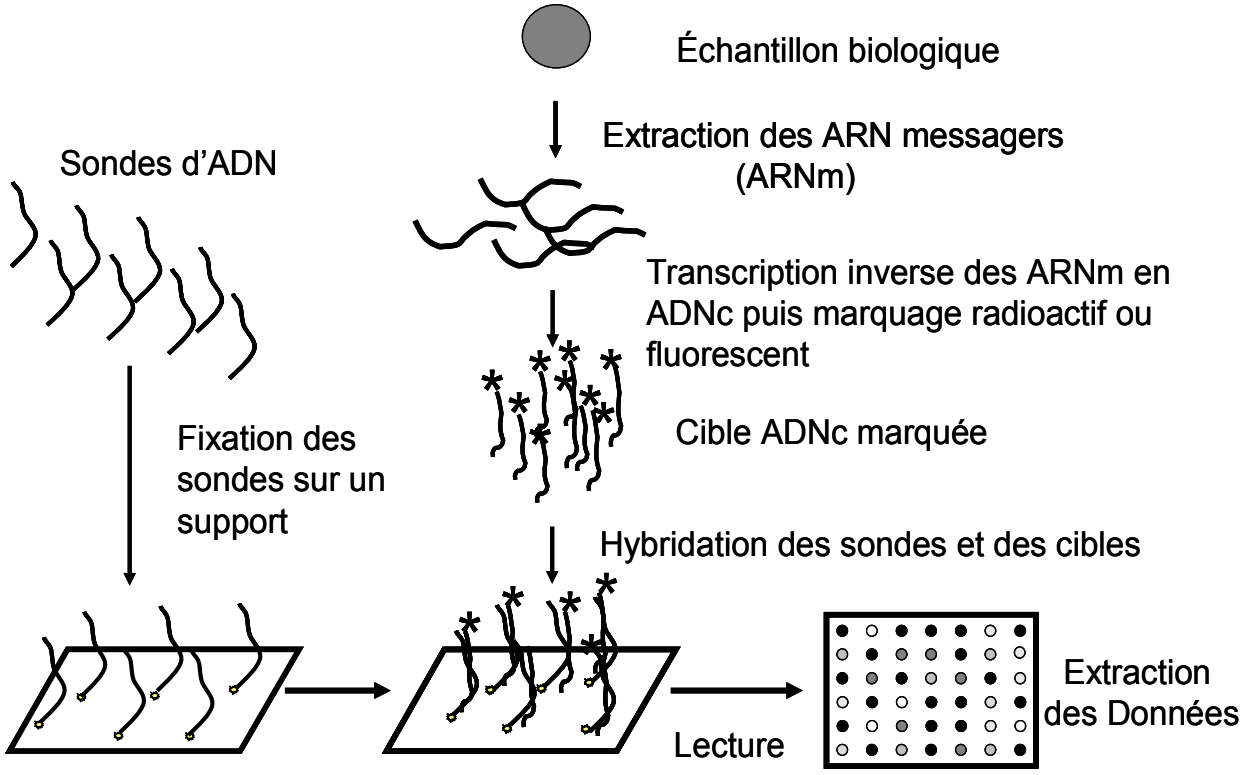

Figure 1 : *Les différentes étapes de l'étude du transcriptome qui précèdent l'analyse des données : réalisation du réseau d'ADN ou de la puce, hybridation avec l'échantillon biologique et quantification des signaux par analyse d'image*

2. ANALYSE DES IMAGES

Une étape essentielle pour l'étude du « transcriptome » est l'analyse des images qui résultent de l'hybridation des sondes et des cibles. Son but est de quantifier, de manière relative, le niveau d'expression des gènes. Pour cela, l'analyse se décompose généralement en trois étapes (Yang *et al.,* 2001) : la **localisation** des spots, la **segmentation** de l'image et l'**extraction des données**.

2.1. Localisation

Cette étape de localisation permet de définir une grille constituée d'un certain nombre de lignes et de colonnes. Ces paramètres sont fixés par l'utilisateur en fonction des conditions de dépôt des sondes sur le support (organisation éventuelle en bloc, diamètre des spots, espacement entre spots), afin de préciser les coordonnées de chaque spot.

2.2. Segmentation

La segmentation des spots consiste à partager une image en deux régions différentes, ayant chacune ses propriétés. Dans le cas des réseaux et puces à ADN, les pixels représentent soit un signal spécifique, soit un bruit de fond non spécifique. Ce dernier étant rarement uniforme sur l'ensemble du support, il est nécessaire de l'évaluer localement pour chaque spot. Les méthodes de segmentation peuvent être classées en quatre catégories, chacune dépendant de la géométrie des spots (taille, forme, irrégularités) :

✓ *Segmentation en cercle fixe* : cette méthode est basée sur la localisation et la taille des spots. Le contour du spot est défini par un diamètre constant pour l'ensemble des spots de l'image. Cette méthode est intéressante lorsque les spots sont diffus et que leur contour est difficilement détectable comme c'est parfois le cas en radioactivité (Figure 2b).

✓ *Segmentation en cercle adaptatif* : dans cette approche, chaque diamètre de spot est estimé automatiquement. Ce mode de segmentation combine analyse spatiale et intensité. La zone du spot est délimitée par un cercle, ainsi les pixels à l'intérieur du masque correspondent au signal, ceux à l'extérieur correspondent au bruit de fond. Une zone tampon (couronne d'exclusion) est définie autour du cercle afin de pallier les irrégularités du spot qui induisent des erreurs quant à la classification des pixels en signal ou bruit de fond (Figure 2)

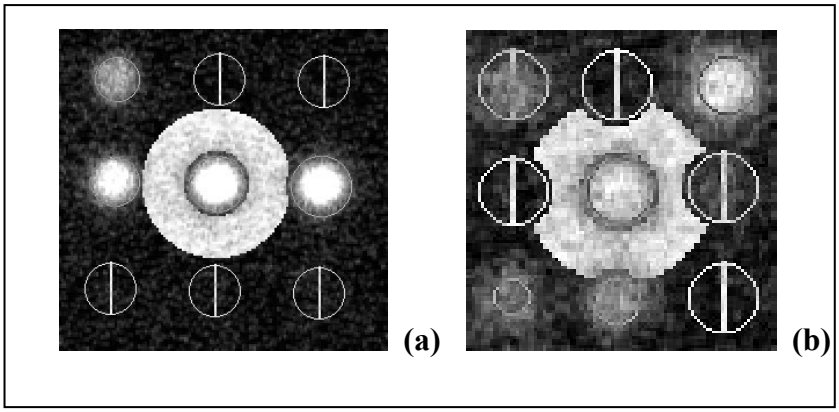

Figure 2 : ***Exemple de localisation et segmentation en cercle adaptatif.*** *Les spots d'hybridation représentent les petits cercles, noirs, gris ou blancs. Les cercles barrés indiquent que les spots ne sont pas différents du bruit du fond. Le bruit de fond local du spot central est calculé à partir des pixels contenus dans l'anneau lumineux entourant ce spot* ***(a)****. Une exclusion des spots adjacents peut être nécessaire en raison d'un espace entre spots insuffisant. Certains spots (en bas et en haut à gauche) sont mal segmentés s'ils sont trop diffus* ***(b)****.*

✓ *Segmentation par analyse de l'histogramme* : cette méthode utilise un masque choisi pour être plus grand que n'importe quel spot. Le signal et le bruit de fond sont estimés à partir de valeurs seuils de l'histogramme de distribution des intensités des pixels présents dans le masque.

✓ *Segmentation en contour adaptatif* : cette méthode est définie pour pallier aux problèmes d'irrégularité des spots. Deux algorithmes de segmentation prenant en considération la forme de chaque spot sont utilisés, chacun nécessitant la définition du point de départ du spot pour définir son contour et analyser de proche en proche les niveaux d'intensité. Cette segmentation est intéressante en fluorescence où le contour des spots est généralement bien dessiné.

2.3. Extraction des données

Après la segmentation, les valeurs des intensités des pixels dans les spots (F) et leur bruit de fond local (B) sont déterminées. Le nombre de pixels dépend de la résolution du scanner utilisé, qui lui-même dépend du type de marquage des cibles. Chaque spot est caractérisé entre autres par la moyenne (F_{moy}), la médiane (F_{med}) et le volume de l'intensité des pixels du spot en niveaux de gris ainsi que par la moyenne (B_{moy}) et la médiane (B_{med}) de l'intensité des pixels dans le bruit de fond local. Le volume calcule l'intensité d'hybridation en fonction de la taille du spot qui peut varier en cas d'irrégularités. L'écart-type de l'intensité des pixels est également calculé pour le spot ($F_{écart\text{-}type}$) et son bruit de fond ($B_{écart\text{-}type}$). Ces données brutes sont exportées vers un tableur, comme par exemple Excel.

3. QUANTIFICATION ET FILTRAGE DES SIGNAUX

3.1. Quantification

Il existe principalement quatre méthodes de quantification de l'intensité (I) de chaque spot utilisant généralement la médiane du bruit de fond (Bdf) local (B_{med}), celle-ci étant moins sensible aux variations artéfactuelles des niveaux de gris que la moyenne du Bdf (Yang *et al.,* 2002).

1) $I_1 = F_{moy} - B_{med}$ - La soustraction d'un bruit de fond local permet de mieux s'affranchir d'un niveau moyen d'hybridation parfois variable selon la localisation sur le support.

2) $I_2 = F_{med} - B_{med}$ - Cette méthode est souvent préconisée en fluorescence où la saturation des pixels est plus fréquente: ainsi, seules les valeurs médianes ($F_{med)}$ autorisent une quantification fiable de spots présentant un taux de pixels saturés inférieurs à 50%. De plus, la soustraction d'un bruit de fond local permet de s'affranchir de l'hétérogénéité d'hybridation conduisant par exemple à des dérives du bruit de fond de l'ordre de 20% entre le haut et le bas de lames (Meunier *et al.,* 2003).

3) $I_3 = F_{med} - B_{glob}$ où B_{glob} représente la moyenne globale des B_{med} sur l'ensemble de l'image ou sur l'ensemble des voisins du spot considéré. Cette méthode est recommandée notamment dans des conditions de densité trop importante des spots, conduisant à un espacement réduit entre ces derniers et donc une estimation biaisée du bruit de fond local de chaque spot.

4) I_4 = Volume – B_{med} - Cette méthode est rarement utilisée.

Dans le cas d'un marquage fluorescent, la quantification est réalisée sur les intensités obtenues pour chaque fluorophore (Smyth *et al.,* 2003) et elle est généralement notée R et G respectivement pour le rouge et le vert. Finalement, on choisira la méthode de quantification qui minimisera l'écart de quantification entre réplicats intra voire inter-support.

3.2. Filtrage

Le filtrage consiste à exclure des traitements ultérieurs, les intensités des spots considérés comme non fiables. La qualité des spots est déterminée en fonction de différents critères (Wang *et al.,* 2001) :

1) Rapport signal sur bruit (SNR) = (F_{moy} – B_{moy}) / $B_{écart\text{-}type}$. Ce critère permet d'évaluer si l'intensité d'un spot est significativement différente du bruit de fond. Ainsi, des rapports de 2 et 3 signifient respectivement que 50% et 80% des intensités des pixels du spot sont différentes du bruit de fond.

2) Niveau de saturation du spot exprimé en pourcentage du nombre de pixels saturés. Il est admis que, dans le cas de l'utilisation de la médiane de l'intensité du spot (F_{med}), un seuil maximal de 50% est toléré.

3) Homogénéité du spot (forme, dimension) et faible dispersion de l'intensité de ses pixels.

4) Homogénéité du bruit de fond (poussières, comètes).

Le calcul d'un score global pour chaque spot en pondérant ces critères a permis par exemple de ne déclarer fiables que 37,7% des spots des lames ou 38.8% des spots des membranes d'une expérimentation avec des échantillons musculaires visant à comparer les deux supports (Meunier *et al.,* 2003).

De plus, un gène est retenu si ses réplicats intra-support sont fiables et si leur intensité ne varie pas au delà d'un facteur 2 (Manduchi *et al.,* 2000). L'intensité des gènes est alors représentée par la moyenne des intensités de leurs réplicats.

3. NORMALISATION ET CORRECTION DES DONNEES

Les données issues d'expériences de réseaux et puces à ADN (niveaux des transcrits) sont sujettes à de multiples sources de variation d'origine biologique mais aussi technique (Novak *et al.,* 2002). Afin de pouvoir analyser ces données et mettre en évidence la variabilité biologique, il est nécessaire d'éliminer ou de contrôler la variabilité technique. La méthode de normalisation des données permettant de comparer plusieurs réseaux traités de façon identique s'avère donc importante, assurant ainsi la qualité et la fiabilité des résultats.

Il existe différentes techniques de normalisation basées soit sur l'ensemble des signaux du support, soit sur un nombre réduit de gènes (gènes domestiques exprimés de façon ubiquitaire, gènes de plante, ADN génomique) dont l'expression est stable (Yang *et al.,* 2002). Quelle que soit la méthode utilisée, les intensités normalisées des gènes sont ensuite corrigées par soustraction de l'intensité d'un témoin négatif (qui correspond par exemple à un spot sans ADNc mais avec seulement du tampon déposé sur le support).

Dans tous les cas, différentes représentations graphiques des données sont utilisées pour visualiser les gènes différentiellement exprimés en fonction du niveau d'expression. Ce sont les MA-plot et RI-plot (Figure 3)

$M = R = \log_2 \frac{R}{G}$ → représente le rapport d'expression pour le gène considéré, préalablement converti en logarithme base 2.

$A = \log_2 \sqrt{RG}$ $I = \log_{10}(RG)$ → représente le niveau moyen d'hybridation pour le gène considéré.

Ces représentations décrites ici pour un marquage fluorescent (R,G) s'appliquent également en marquage radioactif où R sera remplacé par I_1 (intensité dans la condition 1) et V par I_2 (intensité dans la condition 2).

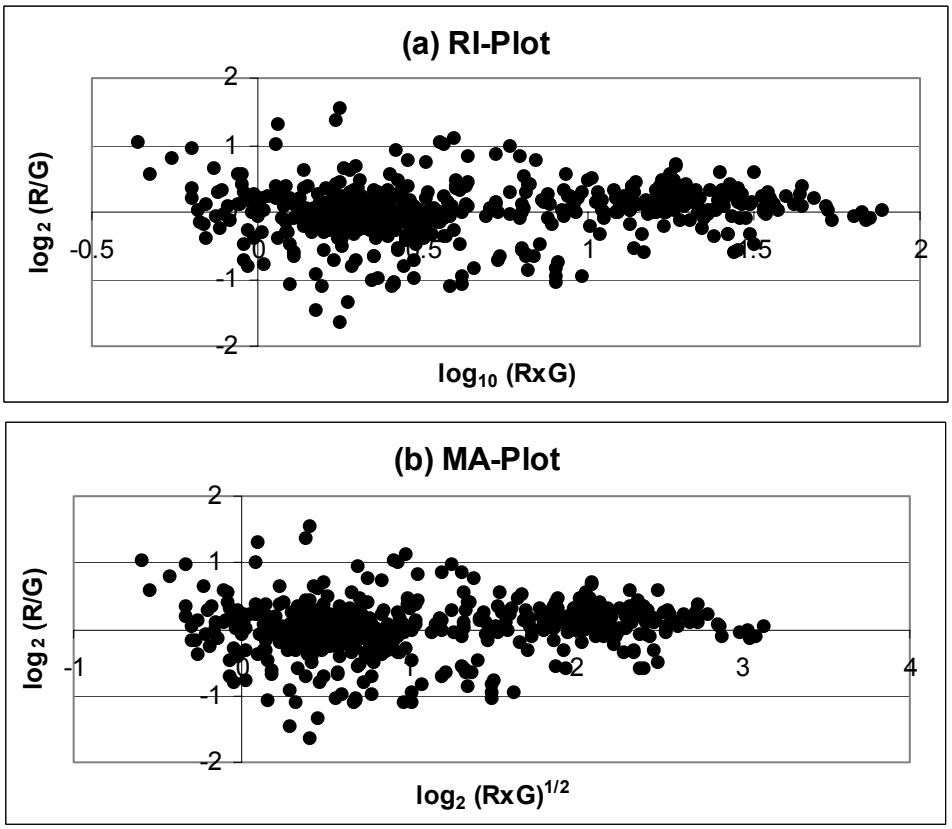

Figure 3 : ***Exemples de représentations graphiques des données transcriptomiques****. (a) RI Plot avec R=log₂ (R/G) et I=log₁₀RG. (b) MA Plot avec M=log₂ (R/G) et A=log₂(RG)^1/2 (dans les deux représentations graphiques, R et G représentent respectivement les intensités avec les fluorochromes rouge et vert).*

3.1. Normalisation globale par la moyenne ou la médiane.

Dans des conditions de marquage fluorescent, le facteur de normalisation k est calculé en faisant le rapport des intensités moyennes ou médianes des spots pour les deux fluorophores Cy3 (rouge) et Cy5 (vert) : $k = \frac{\overline{R}}{\overline{G}}$

où $\overline{R}$ et $\overline{G}$ représentent respectivement les intensités moyennes ou médianes des spots dans le rouge et dans le vert.

Ainsi, l'expression normalisée du rapport pour chaque spot testé est R/kG, ajustant chaque rapport de manière à ce que le rapport moyen soit égal à 1. Exprimé sous la forme logarithmique, cela revient à soustraire une constante c = $\log_2$ k

$$\log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{kG} = \log_2 \frac{R}{G} - c$$

Lors d'un marquage radioactif, cette méthode de normalisation consiste à diviser l'intensité *I* du spot considéré par l'intensité moyenne $\bar{I}$ sur l'ensemble des spots du support: $I \rightarrow \frac{I}{\bar{I}}$ et ce pour chaque condition (I_1 et I_2).

3.2. Normalisation globale pondérée par l'intensité :

Cette méthode de normalisation appelée « lowess normalisation » (pour « locally weighted regression scatterplot smoothing ») est généralement indispensable dans des conditions de marquage fluorescent. Il s'agit d'une normalisation permettant de supprimer les effets dépendant du niveau de l'intensité A, effets modélisés par la fonction c(A) issue de la régression (lowess) sur le M-A plot.

$$\log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - c(A) = \log_2 \frac{R}{k(A)G} \qquad \text{avec } A = \log_2 \sqrt{RG} \text{ et } c(A) = \log_2[k(A)].$$

où R et G représentent respectivement les intensités obtenues à partir des fluorochromes Cy5 et Cy3 pour le spot étudié, k correspond au rapport des intensités moyennes de l'ensemble des spots comme défini précédemment.

Dans le cas des marquages radioactifs, ce type de normalisation peut s'avérer également utile. Il conduit, d'après la formule précédente, à corriger les intensités d'une condition par le facteur de correction $2^{-c(A)}$. Toutefois, il est possible de corriger les intensités dans les deux conditions de façon symétrique comme suit :

$I_1 \rightarrow I_1 * \sqrt{2}^{-c(A)}$; $I_2 \rightarrow I_2 / \sqrt{2}^{-c(A)}$ avec $A=\log_2\sqrt{(I_1 I_2)}$

3.3. Normalisation globale de type print-tip :

Cette méthode de normalisation permet de supprimer les effets dépendant de l'aiguille i de spottage et du niveau de l'intensité (Figure 4).

$$\log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - c_i(A) = \log_2 \frac{R}{k_i(A)G} \qquad \text{avec } A = \log_2 \sqrt{RG}$$

où R et G représentent les intensités obtenues respectivement dans le rouge et le vert pour le spot étudié et k_i le rapport des intensités moyennes ($\bar{R}$ et $\bar{G}$) des spots réalisés avec l'aiguille i.

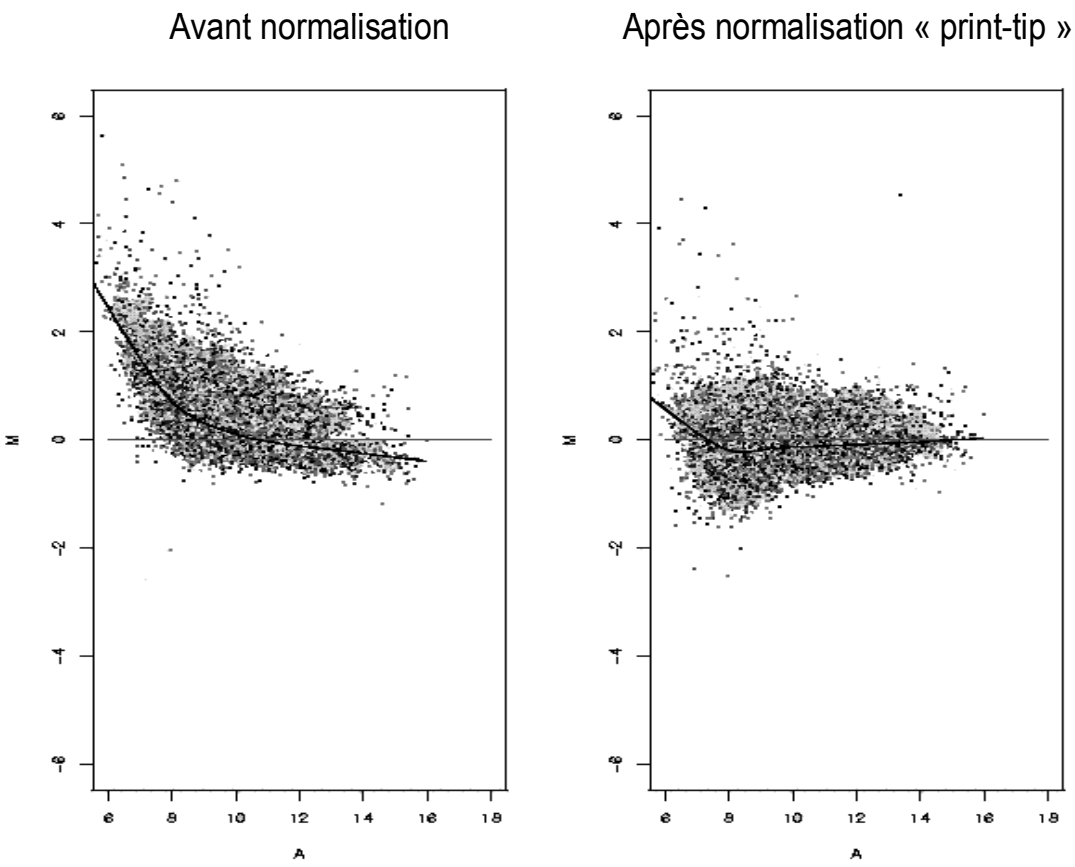

Figure 4 : ***Exemple de normalisation globale pondérée par l'intensité****. MA Plot de 4948 gènes déclarés fiables issus d'une puce avec 16254 oligonucléotides (données non publiées / C. Bernard, 2004). MA Plot avant et après normalisation réalisée avec l'outil informatique MADSCAN (http://cardioserve.nantes.inserm.fr/mad/)*

4. ANALYSES STATISTIQUES

Les données issues des expériences de micro-arrays sont nombreuses et nécessitent des méthodes statistiques appropriées, c'est-à-dire adaptées à leur grand nombre et capables de quantifier les différentes sources de variabilité technique et biologique.
Le premier objectif de l'analyse statistique est de déterminer quels sont les gènes significativement exprimés dans l'échantillon étudié. Pour cela, il faut d'abord définir le seuil au-delà duquel les signaux sont significativement différents du bruit de fond. Les signaux de faible intensité (en dessous du seuil) étant considérés comme des signaux non spécifiques, il est également nécessaire de vérifier que leur distribution suit bien une loi normale (Piétu *et al.,* 1996).

Le second objectif de l'analyse statistique est de déterminer quels sont les gènes différentiellement exprimés entre deux échantillons 1 et 2. Quelle que soit la méthode statistique utilisée, il existera une probabilité non nulle (risque de première espèce α) de détecter des **faux-positifs** (gènes déclarés différentiellement exprimés alors qu'ils ne le sont pas) et une autre probabilité non nulle de ne pas être capable de détecter des gènes réellement différentiellement exprimés (**faux-négatifs**). Il est bien entendu souhaitable de minimiser ces deux probabilités d'erreur sachant que la seconde augmente quand la première diminue et réciproquement. Le schéma expérimental et le modèle statistique approprié doivent donc être

réfléchis ensemble avant la réalisation des expérimentations de façon à avoir des dispositifs puissants (peu de faux-négatifs) et fiables (peu de faux-positifs). Cela nécessite notamment d'analyser un nombre d'échantillons suffisants qui se calcule en fonction de la variabilité des données et des différences que l'on souhaite mettre en évidence (Yang *et al.,* 2003).
Selon les méthodes statistiques utilisées, la transformation logarithmique des données peut s'avérer nécessaire notamment pour les rapports d'intensité entre deux échantillons étudiés car ce type de données n'est pas distribué de façon normale.

4.1. Tests statistiques pour comparer les résultats issus de deux échantillons

La méthode Fold Change
L'approche la plus simple, mais qui ne constitue pas en soit un test statistique, est la méthode « Fold Change ». A partir de cette méthode, un gène est déclaré comme étant exprimé différentiellement si son niveau moyen d'expression varie par plus d'un facteur constant entre les deux échantillons testés. Classiquement, la valeur seuil choisie est 2. Cependant cette méthode est sujette à des biais si les données n'ont pas été proprement normalisées. Par ailleurs, elle ne tient pas compte de la variabilité des résultats pour chaque échantillon. Ainsi, comme les gènes de faibles intensités ont un coefficient de variation plus important que ceux de fortes intensités, cette méthode peut générer un grand nombre de faux-positifs surtout pour les faibles intensités.

Test t
Une autre méthode permettant d'identifier des gènes exprimés différentiellement est l'utilisation du test de Student ou test *t,* sur les niveaux des intensités d'hybridation.

$$t = \frac{m_1 - m_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

m_1 et m_2 : moyennes des intensités des signaux d'un gène donné dans chaque condition 1 et 2.
S_1^2 et S_2^2 : variances des intensités des signaux d'un gène donné dans chaque condition 1 et 2.
n_1 et n_2 : nombre d'analyses pour les conditions 1 et 2 correspondant au nombre de réseaux (variabilité technique) ou au nombre d'individus (variabilité biologique), analysés par condition.

Ainsi quand t excède un certain seuil, dépendant du risque de première espèce α choisi (généralement 5%), les niveaux d'expression du gène étudié entre les deux populations testées sont considérés comme significativement différents.
Le problème fondamental avec ce test concerne le nombre de répétitions de l'expérience. Celui-ci étant très souvent petit, de l'ordre de 1 à 3, les estimations de variance sont sujettes à des fluctuations irrégulières. C'est pourquoi certains auteurs préfèrent utiliser des tests non paramétriques adaptés aux petits échantillons. D'autres auteurs éliminent d'emblée les signaux qui ont une variance trop élevée en considérant qu'il s'agit d'un critère de non fiabilité de la mesure. Ainsi, par exemple, certains auteurs préconisent de réaliser les hybridations sur quatre réseaux ou puces différents et de ne conserver que 3 ou 4 données à condition que celles-ci ne s'éloignent pas plus de 25% de la moyenne (Piétu *et al.,* 1996 ; Sudre *et al.,* 2003). Enfin, prendre un risque de 5% dans une expérimentation où 10000 gènes sont étudiés simultanément peut conduire à obtenir 500 (0.05*10000) faux-positifs, ce qui est parfaitement inacceptable. C'est pourquoi certains statisticiens ont proposé des modifications du test *t* adaptées à l'analyse du « transcriptome » comme décrit ci-après.

Modèle Bayesien

Afin d'obtenir des estimations de variances des niveaux d'expression plus fiables, il existe une approche probabiliste basée sur une régularisation du test *t* : le modèle probabiliste Bayesien. Cette méthode prend en compte les gènes ayant le même niveau d'expression que celui analysé car il a été observé que des gènes exprimés à des niveaux similaires ont une variance similaire.

Ainsi l'estimation de la variance est :

$$\sigma^2 = \frac{v_0\sigma_0^2 + (n-1)s^2}{v_0 + n - 2}$$

n : nombre de gènes de même niveau d'expression que le gène analysé.

s^2 : variance des intensités des signaux analysés.

v_0 : degré de confiance dans la variance du bruit de fond σ_0^2. Ce paramètre est réglable en respectant la condition v_0=K-n, avec K une constante supérieure à 2 pour estimer correctement l'écart-type.

σ_0^2 : variance du bruit de fond de tous les gènes ou des gènes d'expression voisine du gène considéré.

d'où $$t = \frac{m_1 - m_2}{\sqrt{\sigma^2}}$$

m_1 et m_2 : moyennes des intensités des signaux dans chaque condition 1 et 2.

Cette approche donne de meilleures estimations de la variance (σ^2) que le calcul classique de la variance pour le test *t*. Elle permet également de réduire les taux de faux-positifs dans le cas d'un petit nombre de réplications (de l'ordre de 2 ou 3 réplications). Toutefois lorsque le nombre de réplications est suffisamment élevé (de l'ordre de 5), les résultats obtenus à partir d'un simple test *t* ou de l'approche Bayesienne sont similaires (Baldi et Long, 2001).

Cette méthode est implémentée dans le logiciel Cyber T utilisé sous UNIX/LINUX (http://www.igb.uci.edu/servers/cybert/).

Méthode Significance Analysis of Microarrays (SAM) ou test S

La méthode SAM est basée sur un test *t* gène spécifique. Elle utilise une méthode de permutation consistant, pour un gène donné, à permuter entre les échantillons les mesures d'intensité des réplicats techniques. La statistique de *t* est alors calculée pour les données vraies (c'est-à-dire non permutées) et les données permutées. Les gènes déclarés différentiellement exprimés de façon significative avec les données permutées sont donc des faux-positifs. Cette méthode permet donc d'estimer le taux de fausse découverte (FDR pour "**False Discovery Rate**") (Tusher *et al*., 2001). Le principal avantage de cette approche est de calculer le risque d'erreur non plus par rapport à l'ensemble des gènes étudiés mais par rapport à ceux déclarés significativement différents. Ainsi on pourra parfaitement accepter de prendre un risque plus important. Par exemple, fixer le FDR à 10% n'induit qu'un seul faux-positif potentiel lorsque 10 gènes sont déclarés significativement différents, ce qui peut s'avérer tout à fait acceptable.

Dans le cas de données réparties selon deux conditions d'étude (exemple : traité, non traité), la différence relative d_i dans l'expression de chaque gène est calculée selon la formule suivante :

$$d_i = \frac{r_i}{s_i + s_0}$$

avec $r_i = m_{i1} - m_{i2}$

et $$s_i = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2}}$$

m_{i1} et m_{i2} : niveaux moyens d'expression du gène i dans les conditions 1 et 2.
SCE_1 et SCE_2 : somme des carrés des écarts à la moyenne dans les conditions 1 et 2.
s_i : écart-type des mesures d'expression répétées.
n_1 et n_2 : nombres de mesures dans chacune des conditions.
s_0 : constante positive calculée pour minimiser le coefficient de variation de d_i ; s_0 revient à considérer l'écart-type moyen sur tous les gènes.

La méthode SAM s'avère donc très efficace pour éliminer les faux-positifs, dont le taux est très faible par rapport à la méthode « Fold Change ». Cependant SAM nécessite un nombre important de réplications et induit alors un grand nombre de permutations, ce qui peut augmenter le temps de calcul de l'analyse.
Elle est implémentée dans le logiciel SAM (http://www-stat.stanford.edu/%7Etibs/SAM) et propose des représentations graphiques (SAM-plot) comme celle de la Figure 5.

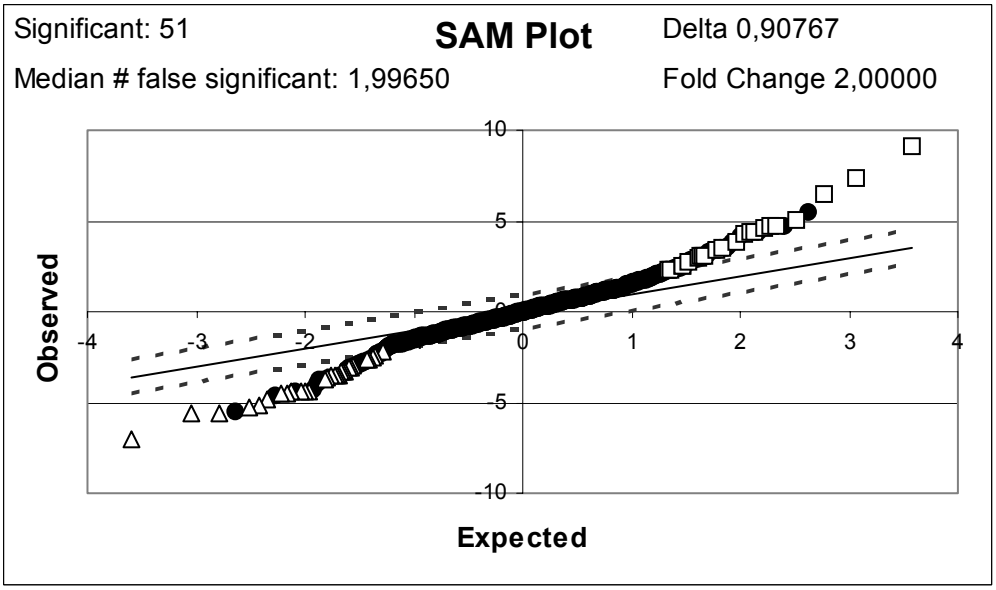

Figure 5 : ***Représentation graphique (SAM Plot) de l'ensemble des gènes analysés au cours d'une expérience représentative.*** *Les carrés représentent les gènes sur-exprimés, les triangles représentent les gènes sous-exprimés. Les cercles pleins noirs représentent les gènes non différentiellement exprimés. Le seuil delta choisi est matérialisé par les droites en pointillés.*

Une analyse comparative de ces méthodes (Hocquette *et al.,* 2003) a été conduite sur des membranes contenant 1060 spots d'ADNc. Parmi les 318 spots déclarés fiables, 9 ont été retenus avec la méthode du Fold Change (FC=1.5) comme différentiellement exprimés entre les deux échantillons étudiés, 7 ont été retenus avec la méthode du Test *t* et 8 avec la méthode SAM. Seule la méthode SAM a été capable de déclarer comme faux-positif un témoin négatif parmi ces spots retenus.

4.2. Tests statistiques pour comparer plus de 2 conditions.

Analyse de variance ANOVA ou test F

En raison de l'existence de trois sources importantes de variabilité des données issues de l'analyse du transcriptome (technique, physiologique et liée à l'échantillonnage ; Novak *et al.,* 2002), il est théoriquement indispensable de construire des schémas expérimentaux complexes faisant intervenir plusieurs individus par groupe étudié, de réaliser plusieurs prises d'échantillon par individu et d'hybrider chaque échantillon au moins sur trois réseaux sur lesquels les sondes sont déposées chacune plusieurs fois. Un modèle d'analyse de variance (ANOVA) est alors recommandé pour analyser ces différentes sources de variabilité (Kerr *et al.,* 2000 ; Cui et Churchill, 2003). Cela suppose de prévoir suffisamment de répétitions pour pouvoir étudier chacun de ces facteurs. Pour ce type d'analyse, une transformation préalable des données (par une fonction logarithmique par exemple) est nécessaire pour se placer dans la situation où tous les effets testés sont additifs.

Le modèle d'analyse de variance utilisé dépend, bien entendu, des différentes sources de variabilité appréhendées par le schéma expérimental.

Voici, comme exemple, le modèle suivant applicable lors d'une série de doubles marquages en fluorescence permettant d'hybrider deux échantillons sur le même réseau (Kerr et Churchill, 2001).

$$Y_{ijkgr} = \mu + A_i + D_j + (AD)_{ij} + G_g + (AG)_{ig} + (DG)_{jg} + (VG)_{kg} + S_{r(ij)} + \varepsilon_{ijkgr}$$

Y_{ijkgr} : logarithme de l'intensité lue sur le ijr^{ième} spot pour le gène g, sur le réseau i, pour le marquage j, le réplicat r et l'échantillon k.

μ : signal moyen parmi tous les facteurs de l'expérience.

A_i, D_j et $(AD)_{ij}$: effets rendant compte de la variabilité induite par le réseau (A pour array), du marquage (D pour Dye, c'est-à-dire rouge ou vert) et de la réaction de marquage proprement dite qui est propre à chaque réseau (il s'agit donc de l'interaction AD, correspondant au marquage rouge ou vert sur le réseau i).

G_g : effet du niveau d'expression du gène g moyenné sur les autres facteurs (réseau, marquage, etc).

$(AG)_{ig}$: effet "spot", c'est-à-dire du gène g sur le réseau, il s'agit donc de l'interaction AG

$(DG)_{jg}$: effet marquage gène spécifique permettant de contrôler un artefact technique courant qui est la réponse spécifique d'un fragment d'ADN aux conditions d'hybridation en fonction du type de marquage.

$(VG)_{kg}$: interaction gène x échantillon rendant compte de l'expression du gène g spécifiquement attribuée à l'échantillon k. Il s'agit de l'effet principal d'intérêt biologique. Cet effet peut lui-même être structuré de façon à apprécier les relations entre les échantillons (appartenance à un même groupe par exemple).

$S_{r(ij)}$: effet de la différence entre réplicats r sur le même réseau i et donc hybridés lors du même marquage j.

ε_{ijkgr} : erreur aléatoire résiduelle.

Le test F est utilisé pour détecter des expressions différentielles [effet $(VG)_{kg}$] de gènes G entre plusieurs échantillons V en tenant compte de la variation entre spots, marquages et réseaux. Le test F peut être considéré comme une généralisation du test *t* et présente également différentes versions adaptées à l'étude du transcriptome notamment le test F gène spécifique (F1), le test F de variance globale (F3) et le test F régularisé (F2). Ces 3 tests sont basés sur les sommes résiduelles des carrés (rss) et sur les degrés résiduels de liberté (df). Les tests d'hypothèses impliquent la comparaison de 2 modèles et les tests statistiques sont effectués gène par gène. Ainsi, l'hypothèse nulle constitue l'absence d'expression différentielle (la valeur VG est égale à zéro) et l'hypothèse alternative montre une différence d'expression (la valeur VG est différente de zéro).
Ainsi les différents tests statistiques sont :

$$F_1 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{rss_1 / df_1} \qquad F_2 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{\sigma^2_{pool}} \qquad F_3 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{(rss_1 / df_1 + \sigma^2_{pool})/2}$$

rss_0 et rss_1 : somme résiduelle des carrés du modèle (rss_0) et variance résiduelle (rss_1) pour un gène donné.
df_0 et df_1 : degrés résiduels de liberté du modèle df_0 et de la variance résiduelle df_1.
σ^2_{pool} : variance résiduelle commune sur l'ensemble des gènes.

Le test F_1 est le test statistique F courant qui est calculé si les données sont disponibles pour seulement un seul gène. Il ne nécessite pas la supposition d'une variance résiduelle commune, cependant il a une puissance faible en raison de la petite taille des échantillons et il peut être sensible aux variations dans les estimations de variance résiduelle rss_1.
Contrairement au test précédent, le test F_3 suppose une variance résiduelle commune sur l'ensemble des gènes (σ^2_{pool}). Ce test est le plus puissant et peut être appliqué à de petites expériences, toutefois il est sujet aux mêmes biais que le test « Fold Change ».
Le test F_2 est une combinaison des deux autres tests. Il est notamment performant dans le cas d'expériences répliquées indépendamment. De plus, il a un taux de fausse découverte plus faible que soit le test F_1, soit le test F_3 (Wu *et al.,* 2002). Il est au test F ce que la méthode SAM est au test *t*. En effet, comme la méthode SAM, la variance gène-spécifique est pondérée par un terme constant qui tient compte de la variance moyenne pour tous les gènes.
Ces différents tests F sont intégrés dans le modèle ANOVA à effets fixes qui suppose l'indépendance entre toutes les observations et seulement une source de variation aléatoire soit technique, soit biologique.
Les tests statistiques (F1, F2 et F3) sont implémentés dans le logiciel MAANOVA (http://www.jax.org/staff/churchill/labsite/software/anova/rmaanova).

Modèle mixed-ANOVA

Ce modèle présente la même structure que le précédent avec seulement une différence dans l'interprétation des termes qui sont traités comme des effets aléatoires, et non comme des effets fixes. Cette alternative de l'analyse de variance est aujourd'hui couramment employée en biologie (Littell *et al*., 1998). De façon générale, le terme AG est considéré comme un effet aléatoire et il est supposé avoir une distribution normale avec une moyenne de zéro. Des facteurs aléatoires supplémentaires pour les effets spot et marquage peuvent également être nécessaires dans le cas de réplication de spots sur un même réseau.
Dans des expériences multi-facteurs, les termes VG peuvent être décomposés à la fois en effets aléatoires et fixes. Les réplications biologiques sont traitées comme un effet aléatoire.

Ce modèle mixed-ANOVA fournit une approche générale et puissante pour permettre une complète utilisation des informations disponibles dans les expériences de micro-arrays avec de multiples facteurs de variabilité et un ensemble de sources de variation des résultats (Cui et Churchill, 2003).

CONCLUSION

Dans le domaine de la transcriptomique, en constante évolution, les biologistes ont souvent peu de recul et peu de compétences en matière d'analyse d'images et d'analyses statistiques des données issues des expériences de macro-arrays. Ceci est d'autant plus vrai que les méthodes d'analyse du transcriptome évoluent tous les jours en raison des difficultés techniques rencontrées régulièrement par les biologistes en fonction notamment de la diversité de leurs conditions expérimentales.
Il apparaît donc indispensable de trouver les méthodes d'analyse d'images, d'analyse de données et de traitements statistiques les plus appropriées aux conditions expérimentales propres à chaque laboratoire afin d'étudier les données issues d'expériences utilisant des réseaux ou des puces à ADN. Il est cependant tout aussi nécessaire d'homogénéiser ces méthodes d'analyse de façon à pouvoir comparer les données d'expression entre laboratoires.
Les exemples présentés dans cet article indiquent que la localisation des spots et la segmentation des images par la méthode en cercle adaptatif est probablement la plus adaptée. De même, la soustraction d'un bruit de fond local, le filtrage des données en fonction de différents critères de qualité des spots ainsi que leur normalisation sont également des étapes primordiales pour la qualité des résultats. Enfin, parmi les méthodes statistiques actuellement disponibles, la méthode SAM est probablement la plus utilisée et parmi les mieux adaptées pour comparer deux échantillons biologiques.

REFERENCES

1. Baldi, P., and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 17, 509-519.
2. Cui, X., and Churchill G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* 4, 210.
3. Hocquette, J.F., Cassar-Malek, I., Listrat, A., and Picard B (2003) Ce que la génomique fonctionnelle peut apporter à la filière viande bovine. *Renc. Rech. Ruminants* 10, 25-32.
4. Hocquette, J.F., Barnola, I., Bernard, C., Meunier, B., Sudre, K., Listrat, A., and Cassar-Malek, I. (2003) Importance de la méthode d'analyse statistique des données du transcriptome. *Renc. Rech. Ruminants* 10, 65.
5. Kerr, M.K., and Churchill, G.A. (2001). Statistical Design and the Analysis of Gene Expression Microarray Data. *Genetical Res.* 77, 123-128.
6. Kerr, M.K., Martin, M., and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. *J Comput Biol*. 7, 819-837.
7. Le technoscope. (2000) *Biofutur* n°206, cahier n°128
8. Lee, M.L., Kuo, F.C., Whitmore, G.A., and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A* 97, 9834-9839.

9. Littell, R.C., Henry, P.R., and Ammerman, C.B. (1998) Statistical analysis of repeated measures data using SAS procedures. *J Anim Sci*. 76, 1216-1231.
10. Manduchi, E., Grant, G.R., McKenzie, S.E., Overton, G.C., Surrey, S., and Stoeckert, C.J.Jr. (2000) Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics.* 16, 685-698.
11. Meunier, B., Sudre, K., Cassar-Malek, I., Listrat, A., and Hocquette, J.F. (2003) Analyse d'image de filtres à hautes densité : méthodes d'acquisition et de traitement. *Renc. Rech. Ruminants* 10, 65.
12. Novak, J.P., Sladek, R., and Hudson, T.J. (2002) Characterization of variability in large-scale gene expression data: implications for study design. *Genomics* 79, 104-113.
13. Pietu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage-Sampson, R., Houlgatte, R., Soularue, P., and Auffray, C. (1996) Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res.* 6, 492-503.
14. Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E., and Davis, R.W. (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* 16, 301-306.
15. Smyth, G.K., Yang, Y.H., and Speed, T. (2003) Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.* 224, 111-136.
16. Sudre, K., Leroux, C., Piétu, G., Cassar-Malek, I., Petit, E., Listrat, A., Auffray, C., Picard, B., Martin, P., and Hocquette, J.F. (2003) Transcriptome analysis of two bovine muscles during ontogenesis..*J Biochem*. 133, 745-756..
17. Tusher, V.G., Tibshirani, R., and Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 98, 5116-5121.
18. Wang, X., Ghosh, S., and Guo, S.W. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.* 29, E75-5.
19. Wu, H., Kerr, M.K, Cui, X., and Churchill, G.A. (2002) MAANOVA : A Software Package for the Analysis of Spotted cDNA Microarray Experiments. (http://www.jax.org/staff/churchill/labsite/pubs/Wu_maanova.pdf)
20. Yang, Y.H., Buckley, M.J., and Speed, T.P. (2001) Analysis of cDNA microarray images. *Briefings in bioinformatics.* 2, 341-349.
21. Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. 30, e15.
22. Yang, M.C.K., Yang, J.J., McIndoe, R.A., and She, J.X. (2003) Microarray experimental design: power and sample size consideration. *Physiol. Genomics* 16, 24-28.