Analyse génétique d'un caractère complexe à l'aide de données transcriptomiques

Yuna Blum

N° ordre : 2012-18
N° série : B-228

THESE / AGROCAMPUS OUEST

Sous le sceau de l'Université Européenne de Bretagne

pour obtenir le diplôme de :

DOCTEUR DE L'INSTITUT SUPERIEUR DES SCIENCES AGRONOMIQUES, AGRO-ALIMENTAIRES, HORTICOLES ET DU PAYSAGE

Spécialité : Mathématiques appliquées

Ecole Doctorale : Vie Agronomie Santé

présentée par

Yuna BLUM

Analyse génétique d'un caractère complexe à l'aide de données transcriptomiques : apport de la modélisation de réseaux de gènes

Soutenance prévue le **24 Août 2012** devant la commission d'Examen :

Christophe Ambroise	Genopole Evry (France)	Rapporteur
Alain Vignal	INRA Toulouse (France)	Rapporteur
Steve Horvath	UCLA (USA)	Examinateur/Rapporteur
Jake Lusis	UCLA (USA)	Examinateur/Rapporteur
Jean Mosser	CNRS Université de Rennes 1 (France)	Examinateur
Sandrine Lagarrigue	INRA/Agrocampus Ouest, Rennes (France)	Directrice de thèse
David Causeur	Agrocampus Ouest, Rennes (France)	Directeur de thèse

Résumé

Depuis une dizaine d'années, de nombreux projets de génomique fonctionnelle se sont développés, avec pour objectif de mieux comprendre des caractères complexes d'intérêt socio-économique en vue de mieux les maîtriser. Ces caractères sont dits complexes car contrôlés par de multiples facteurs : génétique, alimentation, état de santé... Une stratégie couramment utilisée pour l'étude de tels caractères consiste à localiser des QTL (*Quantitative Trait Loci*), c'est-à-dire des régions chromosomiques contrôlant leur variabilité. Parallèlement au développement de ces travaux, de nouvelles technologies ont émergé (puces à ADN) permettant de mesurer à haut débit l'expression de l'ensemble des gènes d'un organisme via la quantification des transcrits (données transcriptomiques). Des stratégies dites de "génétique génomique" combinant des approches de génomique fonctionnelle et de cartographie de QTL ont alors été développées avec comme objectif de faciliter l'identification des mutations causales sous-jacentes aux QTL détectés. Dans ce contexte nouveau, une originalité de la thèse est de prendre en compte l'hétérogénéité existante dans les données transcriptomiques et causée par des facteurs connus ou inconnus indépendamment au caractère d'intérêt. Au travers de plusieurs études, on montre que l'hétérogénéité du signal d'expression ou des profils d'expression masque bien souvent la détection des gènes et des régions du génome liés au caractère d'intérêt. Un deuxième volet de la thèse concerne l'inférence de réseaux de gènes à partir de données transcriptomiques. La modélisation de réseaux géniques semble être une solution prometteuse pour mieux comprendre les mécanismes de régulation des gènes impliqués dans la variabilité d'un caractère. Nous développons ici de nouvelles méthodes pour l'estimation de telles structures basées sur un modèle à facteurs. Ces méthodes sont appliquées dans le cadre de l'analyse génétique d'un caractère complexe, et permettent de caractériser les régulateurs clés et les processus biologiques sous-jacents, apportant de nouvelles informations fonctionnelles quant aux mutations causales recherchées.

Mots-clés : *données transcriptomiques, hétérogénéité d'expression, analyses en facteurs, réseau de gènes.*

Abstract

For the past ten years, many projects on functional genomics have been developed with the aim of better understanding complex traits of socio-economical interest in order to better control them. These traits are called complex traits because they are controlled by multiple factors: genetics, food, health status... One strategy commonly used to analyze such traits involves localizing QTL (*Quantitative Trait Loci*), i.e. chromosomic regions controlling their variability. In parallel to this work, new technologies (microarrays) have emerged, which allow the high throughput measurement of gene expression through the quantification of transcripts (transcriptomic data). Genetical genomic approaches combining functional genomic methods and QTL mapping have been developed with the aim of facilitating the identification of causal mutations underlying detected QTL. In this new context, an original aspect of my thesis is to take into account the heterogeneity existing in transcriptomic data and due to known or unknown factors independently of the trait of interest. Through several studies, we show that signal heterogeneity or expression profile heterogeneity most of the time hides the detection of genes or genome regions associated with the trait of interest. A second aspect of the thesis concerns gene network inference using transcriptomic data. Gene network modeling is a promising solution to better understand regulatory mechanisms of genes involved in the trait variability. We develop here new methods to estimate such structures based on a factor model. These methods are applied in the context of the genetic analysis of a complex trait. They allow characterizing key regulators and biological processes underlying the trait variability, giving new functional information about the sought causal mutations.

Key words: *transcriptomic data, expression heterogeneity, factor model, gene network.*

Thèse réalisée au :

Laboratoire de Génétique,
UMR INRA Agrocampus Ouest PEGASE, Rennes

et au :

Laboratoire de Mathématiques Appliquées,
Agrocampus Ouest, Rennes

Ecole Doctorale : Vie Agronomie Santé

Financement : bourse MENRT

A Omi

Remerciements

Je tiens tout d'abord à remercier chaleureusement mes deux directeurs de thèse, Sandrine Lagarrigue et David Causeur pour leur encadrement de très grande qualité et avec qui j'ai eu énormément de plaisir à travailler. Un immense MERCI pour tout ce que vous m'avez apporté durant ces années du point de vue professionnel et aussi humain, ainsi que pour la confiance que vous m'avez accordée.

Je remercie par ailleurs les rapporteurs de cette thèse, Alain Vignal, Christophe Ambroise pour l'intérêt qu'ils ont porté à mon travail. Un grand merci également à Steve Horvath, Jake Lusis et Jean Mosser pour avoir accepté de faire partie du jury.

Je souhaite remercier aussi Yann Audic et Mathieu Emily, membres de mon comité de thèse, pour leur suivi et leur conseils précieux durant ces trois années.

Je remercie ensuite l'ensemble des membres de mes deux laboratoires d'accueil :
Un grand merci à l'équipe de Génétique & Génomique de PEGASE à Agrocampus Ouest : Sandrine Lagarrigue, Pascale Le Roy, chef d'équipe, Christian Diot, Madeleine Douaire, Jean-Marc Fraslin, Olivier Demeure, Frédéric Lecerf, Sophie Allais, Colette Désert, Olivier Filangi, Gregory Guernec, Magalie Houée, Frédéric Hérault, Cécile Duby, Guillaume Le Mignon, Yvan Le Bras, Pierre Blavy, Xiao Wang, Marion Ouédraogo, Charles Bettembourg, Pierre-François Roux et Christine Gourbe ! Merci beaucoup à toi Guillaume pour ton aide précieuse au démarrage de ma thèse ! Merci aussi Yvan pour toutes nos discussions QTLiennes et pour avoir apporté de la bonne humeur (sonore) au bureau ! Pascale et Fifi (Philippé) merci pour vos lumières techniques sur QTLmap ! Merci beaucoup Colette pour tout ce que tu m'as appris à la paillasse. Ce fut un réel plaisir de travailler avec toi ! Olive, merci pour ton aide, tes conseils et le travail réalisé ensemble (semaine record Mcdo !), de bons souvenirs resteront (Gratouille a soif je crois). Un très grand merci aussi à Magalie, avec qui j'ai beaucoup aimé travailler ! Je suis très contente de continuer la collaboration ! Merci Pierre pour nos discussions intéressantes sur les réseaux de gènes et aussi pour ta relecture et tes conseils sur ma partie biblio ! Enfin mille mercis à mon équipe de choc «PICHAMA» (Pef, Charles et Marion) : merci pour tous ces moments inoubliables, au bureau (et en dehors aussi !), vous allez terriblement me manquer ! Un mot restera... «Knockback» !

Un grand merci également à l'équipe du laboratoire de mathématiques appliquées à Agrocampus Ouest. Merci à David Causeur naturellement, à Jérôme Pagès, directeur du laboratoire pour ses conseils et avec qui j'ai eu beaucoup de plaisir à discuter. Merci à François Husson, Julie Josse, Sebastien Lê pour leur accueil chaleureux et aussi à Karine Bagory et Elisabeth Lenault pour leur disponibilité ! Un merci particulier à Marine Cadoret et Malea Kloareg, de supers coachs pour les TD de stat' ! Merci énormément à Marie Verbanck et Thibaut Dutrion, mon autre équipe de choc «16h16» avec qui j'ai partagé de supers moments au bureau ! Enfin, je tiens à remercier tout particulièrement Chloé Friguet avec qui j'ai eu beaucoup de plaisir à travailler et qui m'a beaucoup apporté. J'espère que nous aurons l'occasion de continuer à travailler ensemble !

Un grand merci à Núria Mach, qui est venue travailler durant son post-doctorat quelques mois au labo de Génétique et avec qui j'ai collaboré. Je suis vraiment très heureuse de t'avoir rencontrée. J'ai beaucoup aimé travailler avec toi et j'espère que nous aurons d'autres opportunités de collaboration !

Je souhaite également remercier Hervé Guillou, CR à l'INRA de Toulouse qui m'a accueillie un mois dans son équipe afin d'acquérir les données lipidomiques. Merci Vévé pour tout ce bon temps et ce que tu m'as appris, vive les lipides ! Merci aussi à tout le reste de l'équipe en particulier à Simon, Arnaud, Alice, Afifa, Frédéric et Pascal pour m'avoir si bien accueilli.

J'aimerais remercier aussi deux équipes avec lesquelles j'ai travaillé durant ma mobilité de 4 mois à UCLA à Los Angeles. *First, I would like to thank Jake Lusis and his team, particularly, Rich', Mete, Calvin, Lisa, Atila, Elin, Christoph, Rosa, Melenie and Nam for their welcome and help during my stay. It was a real pleasure to work in such a high quality environment ! I'm also really grateful to Steve Horvath and his team, in particular Roberto Visintainer, for their help and interesting discussions. I hope we will continue to collaborate.*

J'ajoute un remerciement à Hervé Le Bris, Maria Manzanares et Françoise Prigent, de la Coordination de l'École Doctorale VAS à Agrocampus Ouest, pour la gestion logistique et administrative.

Ma thèse a été aussi l'occasion de rencontres avec d'autres doctorants et de retrouvailles, au travers d'associations comme DocAIR ou encore de congrès. Je pense en particulier à Cécile, Bertrand, Alice, Melen et Man ! Merci à vous pour les bons moments passés !

J'aimerais terminer ces remerciements par une note plus personnelle pour mes proches.
Merci à tous mes amis pour leur soutien et leur aide. En particulier, Elsa, je ne te remercierai jamais assez ! Merci à toi et à Clayton pour vos corrections en anglais. Un très grand merci aussi Gwenaëlle ! Ta relecture m'a été très utile ! Merci aussi à Julia, Ameline, Anna, Chiara, Silvia et Martina ! Je n'oublie pas mon groupe de musique, merci Nico, Paco et Fifi (encore !) pour ces moments de décompression !
Merci également à ma famille pour son soutien et son amour ! Je remercie tout particulièrement ma maman à qui je dois tant et aussi mon papa qui a toujours été là pour m'encourager et me conseiller. Vous êtes des parents extraordinaires, je vous admire et suis si fière de vous ! Merci aussi à mes frères et sœurs, ma mamma italiana et tout le reste de ma famille en Autriche, *danke meiner lieben Familie* ! Enfin, merci à Arthur l'enfant le plus sage au monde, si tranquille pendant ces longs weekends studieux !
Pour finir, merci mon choubitchou pour TOUT : ton soutien, ton aide, ta patience presque constante (...) et bien sûr, pour tout ton amour !

Table des matières

Table des figures

Introduction

Depuis une dizaine d'années, le nombre d'études utilisant les technologies du transcriptome est en constante augmentation, offrant un nouvel angle d'étude du fonctionnement du vivant. De nombreux projets de génomique fonctionnelle se sont développés avec pour objectif de mieux comprendre des caractères complexes d'intérêt socio-économique en vue de mieux les maîtriser. Ces caractères sont dits complexes car contrôlés par de multiples facteurs : génétique, alimentation, état de santé... L'équipe de Génétique Génomique de l'UMR PEGASE INRA/Agrocampus Ouest où j'ai effectué en partie mon doctorat, s'intéresse particulièrement à la variabilité d'engraissement chez le poulet de chair, ce caractère étant non valorisé et donc défavorable, tant pour les producteurs que pour les consommateurs (Alleman et al. (1999)). D'autres projets peuvent concerner des maladies humaines (maladies cardiovasculaires, diabète...) comme ceux développés au laboratoire de Jake Lusis à UCLA où j'ai effectué une mobilité de 4 mois durant mon doctorat.

Une stratégie couramment utilisée pour l'étude d'un caractère complexe consiste à localiser les régions chromosomiques contrôlant la variabilité d'un caractère par des approches de cartographie fondées sur l'analyse d'association génétique ou de liaison génétique. De telles régions observées par cartographie, sont appelées QTL pour *Quantitative Trait Loci*. Le principe général de la cartographie de QTL par analyse de liaison, approche mise en œuvre durant ma thèse, est d'observer dans une famille issue d'un parent hétérozygote M1/M2 pour un marqueur donné, une différence significative de performance moyenne entre les deux groupes de descendants ayant reçu M1 ou ayant reçu M2 de ce parent (la performance étant la mesure associée au caractère). Si une telle différence existe, elle s'explique par la co-ségrégation des allèles au marqueur M et des allèles Q/q d'un gène affectant le caractère et génétiquement lié à M : la région au voisinage de M est alors considérée comme un QTL (voir figure 1).

Bien que de nombreux QTL aient été détectés pour de multiples caractères et dans diverses espèces, l'identification des mutations causales à ces QTL est difficile (moins de 1% des QTL cartographiés à l'heure actuelle) (Lagarrigue and Tixier-Boichard (2011)). Plusieurs facteurs peuvent expliquer cette difficulté :

- la taille des régions QTL qui peuvent contenir plusieurs dizaines de gènes,
- une méconnaissance de la fonction de nombreux gènes de la région,
- la nature des mutations : le polymorphisme peut affecter les régions régulatrices ou codantes d'un gène. De plus, un changement dans l'expression d'un gène peut être causé par des processus

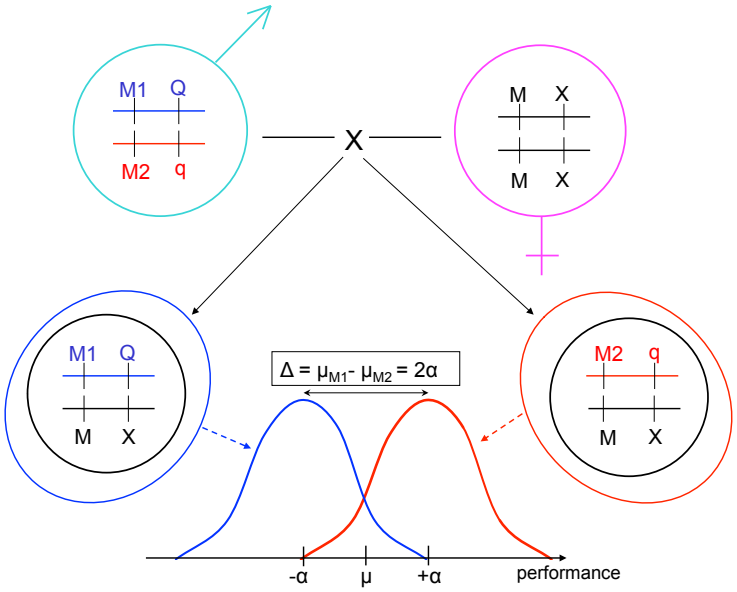


FIGURE 1 – Principe général de la détection d'un QTL

de méthylation de l'ADN sans modification de séquence.

- la nature du gène potentiellement impliqué : le gène peut être non codant à l'origine de transcrits non codants comme les miRNA qui ne sont pas traduits en protéines mais qui affectent l'expression de gènes,
- le type de causalité : une région QTL peut contenir non pas une seule mais plusieurs mutations en déséquilibre de liaison rendant plus complexe l'identification de la ou les mutations causales au caractère d'intérêt.

Parallèlement au développement des travaux basés sur la cartographie de QTL, de nouvelles technologies ont émergé dans les années 2000, comme les puces à ADN (ou puces à gènes), qui permettent de mesurer simultanément dans un tissu donné l'expression de l'ensemble des gènes d'un organisme via la quantification de leur ARNm (transcriptome). Les données issues de ces mesures sont appelées données transcriptomiques ou encore données d'expression. Bien que la mesure soit faite au niveau du transcrit, on parlera plutôt de "niveau d'expression du gène" que de "quantité de transcrits".

Dans ce contexte nouveau, l'objectif général de ma thèse est de proposer des stratégies utilisant des données transcriptomiques pour aider à l'analyse d'un caractère complexe et faciliter l'identification de la ou les mutations causales sous-jacentes à une région QTL. L'idée est d'identifier les gènes et les processus biologiques impactés par les régions QTL responsables du caractère d'intérêt et ainsi d'apporter de nouvelles informations fonctionnelles sur les mutations causales à ces régions. Les approches existantes dans la littérature sont détaillées dans une revue bibliographique à laquelle j'ai collaboré (Le Mignon et al. (2010)) qui figure à la fin de cette introduction générale. Ces approches sont décrites brièvement dans le paragraphe suivant.

Aussi appelée analyse différentielle, une stratégie classique consiste à identifier des ARNm dont les quantités diffèrent significativement entre des lignées divergentes ou entre des individus extrêmes au sein de familles ou populations utilisées pour la détection de QTL. Cette approche renseigne de manière globale sur les gènes impliqués dans la variabilité du caractère mais ne permet pas d'affecter spécifiquement ces gènes à différentes régions QTL.

Dans les années 2000 ont été développées de nouvelles stratégies dites de "génétique-génomique" qui combinent des approches de génomique fonctionnelle et de cartographie de QTL (Brem et al. (2002), Schadt et al. (2003)). Ces stratégies sont basées sur l'identification de gènes dont l'expression est régulée par une région chromosomique. Une telle région est alors appelée "région eQTL" pour région contrôlant un caractère quantitatif de type expressionnel. Une région eQTL est détectée par analyse génétique comme un QTL, mais au lieu d'utiliser des données de performances de l'individu, l'analyse repose sur les données d'expression de gènes dans un tissu donné. Cette démarche ne suppose aucun a priori sur les gènes, car l'ensemble des gènes de l'individu sont étudiés. Par contre, elle nécessite un a priori sur la fenêtre d'observation relative à un tissu : le choix du tissu repose sur des hypothèses biologiques concernant la localisation tissulaire des processus biologiques majeurs impactés par le caractère d'intérêt. Pour rechercher les gènes influencés par une région QTL, l'idée est alors de rechercher les gènes dont l'expression est contrôlée par une région eQTL co-localisant avec la région QTL d'intérêt (région appelée eQTL/QTL). Si les gènes régulés sont en lien avec le caractère, ils ont une forte chance d'être la cause des variations du caractère au niveau de cette région. Ils apportent alors une nouvelle information sur la fonction du gène causal recherché dans la région eQTL/QTL.

La majorité des études se focalisent sur les gènes ayant un eQTL de type *cis* co-localisant avec la région QTL, c'est-à-dire les gènes localisés dans la région eQTL qui contrôle leur propre expression (Schadt (2005)). En effet, l'analyse fonctionnelle de ces cis-eQTL permet de trouver rapidement des candidats à la mutation causale qui sont à la fois des candidats fonctionnels, de part leur fonction en lien avec le caractère, mais aussi positionnels, de part leur localisation dans la région QTL d'intérêt. De plus, cette démarche permet d'émettre des hypothèses sur la nature de la mutation recherchée : si elle se trouve dans un gène cis-eQTL, elle sera alors recherchée dans les régions régulatrices et non la région codante du gène. Une parfaite illustration peut être donnée avec l'identification de la mutation causale dans le gène BCMO1 responsable de la couleur de la viande. Une analyse des gènes localisés dans une région QTL contrôlant la couleur de la viande chez le poulet de chair (Nadaf et al. (2009)) a permis de mettre en évidence le gène BCMO1, ayant à la fois un cis-eQTL co-localisant avec cette région et une fonction en lien avec le caractère. Par la suite, deux mutations dans le promoteur du gène BCMO1 ont été identifiées et des expérimentations de biologie moléculaire ont montré que ces mutations avaient bien un effet sur l'expression du gène (Le Bihan-Duval et al. (2011)). Ainsi, l'existence dans une région QTL d'intérêt d'un gène cis-eQTL ayant une fonction en relation avec le caractère d'intérêt peut considérablement motiver l'adoption de cette première approche. Dans cette optique, j'ai participé au développement d'un outil nommé AnnotQTL (Lecerf et al. (2011)), permettant de récupérer rapidement

pour une région chromosomique donnée, la liste des gènes présents et l'annotation fonctionnelle disponible suivant plusieurs bases de données avec gestion de la redondance. De plus, une option permet de mettre en surbrillance les gènes ayant une annotation fonctionnelle supposée en lien avec le caractère d'intérêt, ce qui facilite l'identification des meilleurs candidats fonctionnels parmi des gènes cis-eQTL. L'article associé est en Annexe du manuscrit (Travaux complémentaires).
Il est important de remarquer que la recherche de cis-eQTL repose sur une hypothèse forte selon laquelle le gène causal contient la mutation affectant à la fois sa régulation transcriptionnelle ou la stabilité de ses ARN et la variabilité du caractère. Finalement, peu d'études ont révélé des cis-eQTL causaux, incitant les biologistes à élargir les analyses aux autres eQTL de type *trans* (régulés par une région eQTL mais non localisés dans la région). Sous l'hypothèse que la mutation dans la région eQTL est la même que celle contrôlant le caractère d'intérêt (dans la région QTL), ces gènes peuvent apporter de nouvelles informations fonctionnelles concernant la ou les mutations responsables du caractère. Néanmoins, comme indiqué dans la figure 2, l'identification de gènes intermédiaires entre la mutation recherchée et le caractère d'intérêt n'est pas si simple. Une région eQTL/QTL peut recouvrir diverses relations entre gènes régulés et caractère d'intérêt : gènes causaux, réactifs, indépendants ou encore contrôlés par une autre mutation proche en déséquilibre de liaison avec la mutation recherchée (Schadt et al. (2005)).

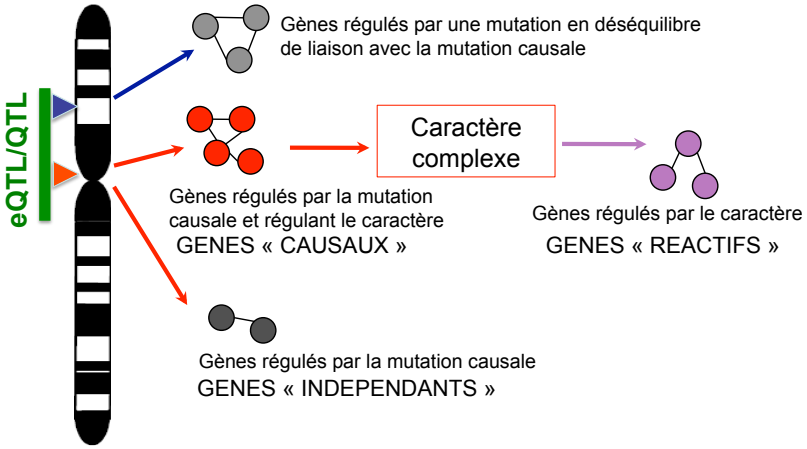


FIGURE 2 – Relations possibles entre "région eQTL/QTL", "gènes régulés" et "caractère". La mutation causale est représentée par un triangle rouge, celle en déséquilibre de liaison avec la mutation causale, par un triangle bleu.

Pour une meilleure appréhension de la complexité des mécanismes biologiques impliqués, la modélisation de réseaux de gènes paraît être une solution prometteuse. L'analyse de tels réseaux, en particulier la recherche de modules fonctionnels de gènes fortement dépendants, offre de nouvelles perspectives pour mettre en évidence les différentes relations entre "mutations", "gènes régulés" et "caractère" (voir figure 2) et caractériser les processus biologiques sous-jacents à une région QTL. De plus, la détection de *hub* gènes (gènes interagissant avec beaucoup d'autres gènes) permettrait d'identifier les régulateurs clés de la variabilité du caractère d'intérêt. De nombreuses approches

ont été développées pour la reconstruction de réseaux géniques à partir de données transcriptomiques (ces approches sont détaillées dans l'introduction du deuxième chapitre). Cependant, en raison de la grande dimension des données (nombre d'échantillons faible par rapport au nombre d'expressions géniques mesurées) et de l'hypothèse de parcimonie des réseaux biologiques (dans un ensemble de gènes, très peu d'entre eux sont réellement en interaction directe), une estimation efficace de telles structures est toujours un enjeu actuel qui suscite un intérêt croissant. Récemment, une étude comparative des méthodes d'inférence de réseaux géniques couramment utilisées dans la littérature (en particulier, les méthodes WGCNA (Langfelder and Horvath (2008)), GeneNet (Schäfer et al. (2006)) et SPACE (Peng et al. (2009)) qui sont détaillées dans le deuxième chapitre) a montré que chacune d'elles étaient performantes sur des critères spécifiques : certaines sont spécialisées dans la détection de modules (WGCNA) alors que d'autres sont plus performantes pour la détection de *hub* gènes (SPACE) (Allen et al. (2012)).

Dans ce contexte, un premier objectif de la thèse est d'améliorer les méthodes existantes utilisant des données transcriptomiques pour l'analyse génétique de caractères complexes. L'originalité de la thèse est de prendre en compte l'hétérogénéité d'expression dans les données transcriptomiques pour les analyses différentielles et les analyses eQTL. Comme évoqué par Leek and Storey (2007) ou encore par Friguet et al. (2009), des facteurs connus ou inconnus peuvent créer une dépendance génique et ainsi bruiter les niveaux d'expression, autrement dit, affecter l'intensité du signal. Il peut s'agir de facteurs génétiques, environnementaux, expérimentaux... On se propose ici d'utiliser une méthode récemment introduite par Friguet et al. (2009) appelée FAMT (Factor Analysis For Multiple Testing) qui permet de capturer cette hétérogénéité d'expression indépendante du caractère d'intérêt par un modèle d'analyse en facteurs. Cette méthode initialement conçue pour améliorer les procédures de tests multiples dans les analyses différentielles est étendue dans mon travail aux approches eQTL.
Un autre type d'hétérogénéité peut être pris en compte dans les données transcriptomiques afin de préciser un caractère et ainsi d'augmenter la puissance et la précision de détection des QTL. En effet, des animaux ayant de mêmes tendances phénotypiques peuvent avoir des profils transcriptomiques hétérogènes ce qui peut s'expliquer par une hétérogénéité du déterminisme génétique : des mutations différentes peuvent mener à un même phénotype et avoir un impact différent sur le transcriptome. Il a été montré que prendre en compte ces sous-types d'animaux ayant des profils transcriptomiques homogènes dans les analyses QTL peut permettre de préciser la localisation de certains QTL et même d'en révéler de nouveaux (Schadt et al. (2003), Le Mignon et al. (2009)). Cette approche peut être vue comme une décomposition du caractère en sous-types phénotypiques. Chacun de ces sous-types correspond à des animaux qui sont plus homogènes en termes de profils transcriptomiques et aussi probablement au niveau de leur génétique. L'ensemble de ces travaux fait l'objet de trois articles publiés et d'une étude réalisée lors de ma mobilité à l'étranger qui sont présentés dans le chapitre 1.

Un deuxième objectif est de proposer de nouvelles méthodes pour la construction de réseaux de

gènes. D'une façon générale, la modélisation de réseaux géniques est utile pour mieux comprendre au niveau génétique le fonctionnement global d'un système. Dans cette thèse, nous l'appliquons dans le cadre de l'analyse génétique d'un caractère complexe, afin d'identifier des modules de gènes impactés par les mutations recherchées. De tels modules permettent de mieux caractériser les régions QTL et ainsi d'apporter de nouvelles pistes pour la recherche des mutations causales au caractère d'intérêt.

On se focalise sur deux grandes approches associées aux modèles graphiques qui ont donné des résultats prometteurs : les *relevance networks* qui utilisent comme mesure de dépendance génique la corrélation de Pearson et les modèles graphiques Gaussiens qui sont basés sur les corrélations partielles. L'originalité de mes travaux est d'une part, d'utiliser un modèle à facteurs pour l'estimation des corrélations brutes et des corrélations partielles et d'autre part, d'introduire de nouvelles procédures permettant de tenir compte d'une structure parcimonieuse du réseau génique.

Les travaux sur le sujet font l'objet de deux articles en préparation et sont présentés dans le chapitre 2.

La figure 3 résume l'ensemble de la démarche suivie et les apports de la thèse.

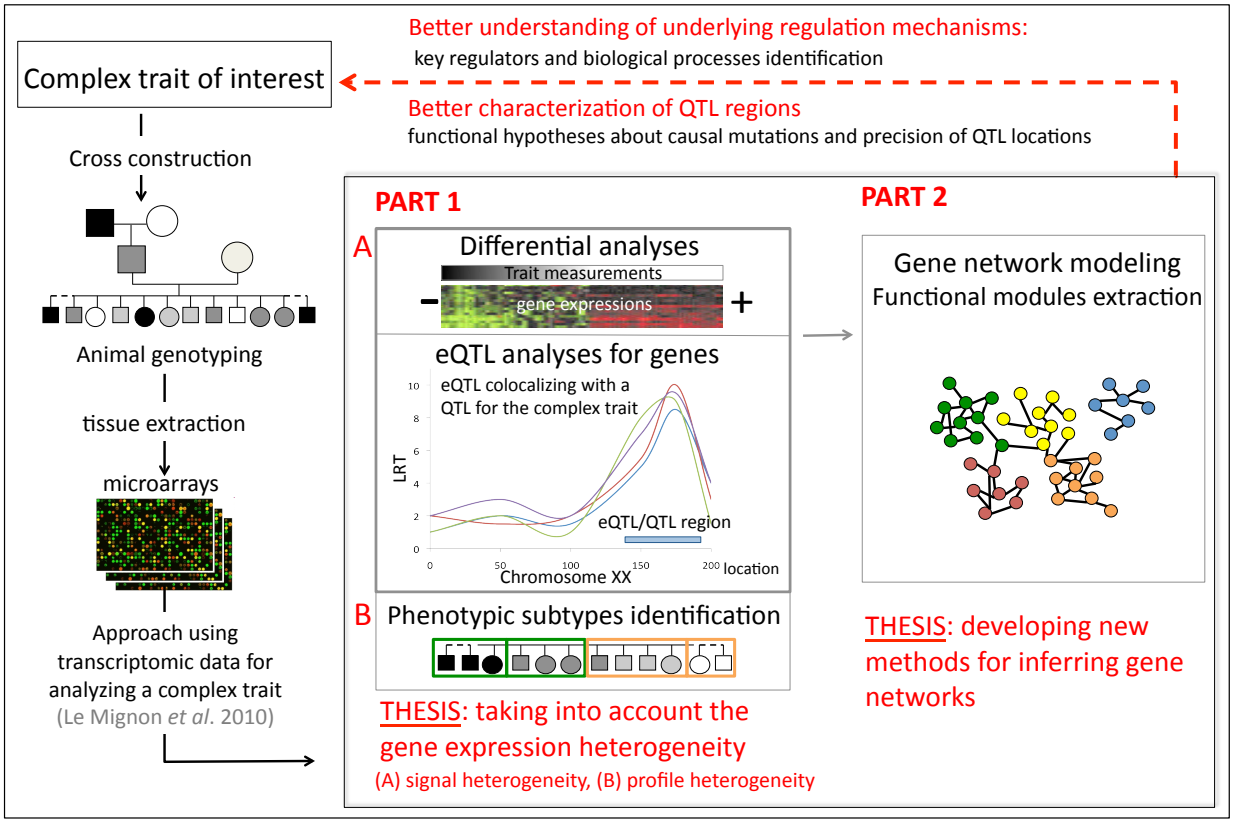


FIGURE 3 – Démarche globale et apports de la thèse

*

* *

Article 1 : Le Mignon *et al.*, INRA Prod Anim, 2010

Dans le prolongement de l'introduction générale, l'article suivant propose un état de l'art sur l'apport de la génomique fonctionnelle pour la cartographie de QTL :

Le Mignon G, **Blum Y**, Demeure O, Diot C, Le Bihan-Duval E, Le Roy P, Lagarrigue S. Contribution of functional genomics to the fine mapping of QTL. Inra Prod.Anim. 2010, 23 (4), 343-358

Apport de l'article

Cette revue bibliographique est centrée sur les apports de la "génétique génomique" dans le contexte de la détection de QTL. Après avoir défini la notion de QTL d'expression (eQTL), cet article propose dans un premier temps un bilan des différents programmes de cartographie de QTL d'expression décrits avant 2010 dans la littérature. Sont ensuite détaillées les différentes approches introduites plus haut qui utilisent des données d'expression pour préciser la localisation de régions QTL responsables de la variation d'un caractère d'intérêt ou encore pour caractériser fonctionnellement ces régions :

- analyse différentielle,
- co-localisation de régions eQTL et QTL et recherche de cis-eQTL ayant une fonction en lien avec le caractère d'intérêt,
- décomposition du caractère complexe en sous-types phénotypiques,
- recherche de modules géniques fonctionnels.

INRA Prod. Anim.,
2010, 23 (4), 343-358

Apports de la génomique fonctionnelle à la cartographie fine de QTL

G. LE MIGNON[1,2,3], Y. BLUM[1,2,4], O. DEMEURE[1,2], C. DIOT[1,2], E. LE BIHAN-DUVAL[5], P. LE ROY[1,2], S. LAGARRIGUE[1,2]

[1] INRA, UMR598 Génétique Animale, F-35000 Rennes, France

[2] Agrocampus Ouest, UMR598 Génétique Animale, F-35000 Rennes, France

[3] ITAVI, 28 rue du Rocher, F-75008 Paris, France

[4] Agrocampus Ouest, Laboratoire de Mathématiques Appliquées, F-35000 Rennes, France

[5] INRA, UR83 Recherches Avicoles, F-37380 Nouzilly, France

Courriel : Sandrine.Lagarrigue@agrocampus-ouest.fr

De nombreux programmes de recherche en génétique animale ont permis de localiser des régions QTL alors que les mutations causales sous-jacentes sont encore rarement identifiées. Après avoir introduit le concept de QTL d'expression, cet article présente les principales stratégies utilisant des données d'expression génique pour mieux caractériser ces régions QTL et en faciliter ainsi l'identification des mutations causales.

De nombreux programmes de recherche en génétique animale ont pour objectif de localiser des QTL (*Quantitative Trait Locus),* régions du génome responsables de la variabilité de caractères complexes d'intérêt économique et d'en identifier le(s) polymorphisme(s) causal(aux) sous-jacent(s). Depuis les années 2000, des technologies de génomique fonctionnelle se sont développées permettant de mesurer simultanément l'expression de l'ensemble des gènes d'un génome. Ces phénotypes plus élémentaires peuvent être des quantités d'ARN messager (ARNm), de protéines ou par extension de métabolites. Nous proposons dans cet article de présenter les principales stratégies utilisant des données d'expression dans le cadre de la cartographie de QTL. L'une de ces stratégies que nous appellerons «décomposition du caractère» peut se révéler efficace pour affiner voire même permettre la détection de nouveaux QTL. Une seconde stratégie, plus courante et appelée «eQTL» (pour QTL d'expression), peut apporter de nouvelles informations fonctionnelles sur la région QTL. D'autres approches variantes de la précédente peuvent contribuer à une localisation plus fine de la région QTL.

Dans un premier temps, nous introduirons le concept de QTL d'expression (eQTL) et les principaux résultats de cartographie d'eQTL que l'on peut extraire de la bibliographie indépendamment du contexte de recherche de QTL (parties 1 et 2). Nous exposerons ensuite les principales stratégies utilisant des données transcriptomiques dans le cadre de la détection de QTL (partie 3).

1 / QTL d'expression

1.1 / Principes généraux

Un gène est une séquence du génome qui est exprimée et va, dans un bon nombre de cas, conduire à la synthèse d'une ou de protéines qui lui sont spécifiques, les protéines représentant des entités fonctionnelles de la cellule. On estime aujourd'hui que les génomes des animaux d'élevage sont composés d'environ 40 000 gènes, eux-mêmes à l'origine de plusieurs millions de protéines (http://www.ensembl.org/species/Info/StatsTable?db=core). En raison de leurs fonctions très variées les protéines jouent un rôle majeur dans l'établissement des caractères d'intérêt agronomiques visibles à l'échelle de l'animal. Certaines sont des protéines de structure, d'autres des enzymes catalysant des réactions biochimiques, d'autres encore sont des régulateurs des différents mécanismes cellulaires permettant le décodage de l'information génétique depuis les gènes situés dans le noyau de nos cellules jusqu'aux protéines qu'ils codent. En effet, le passage du gène à la protéine est un processus complexe comprenant deux étapes majeures (encadré 1). La première étape, la transcription, consiste à transcrire le gène en un certain nombre de copies, dites transcrits ou ARN messager. Selon le nombre de copies générées, un gène sera dit plus ou moins exprimé dans un tissu considéré. La seconde étape, la traduction, est le décodage de chaque copie d'ARNm en protéine. A chaque étape, transcrits et protéines peuvent être dégradés.

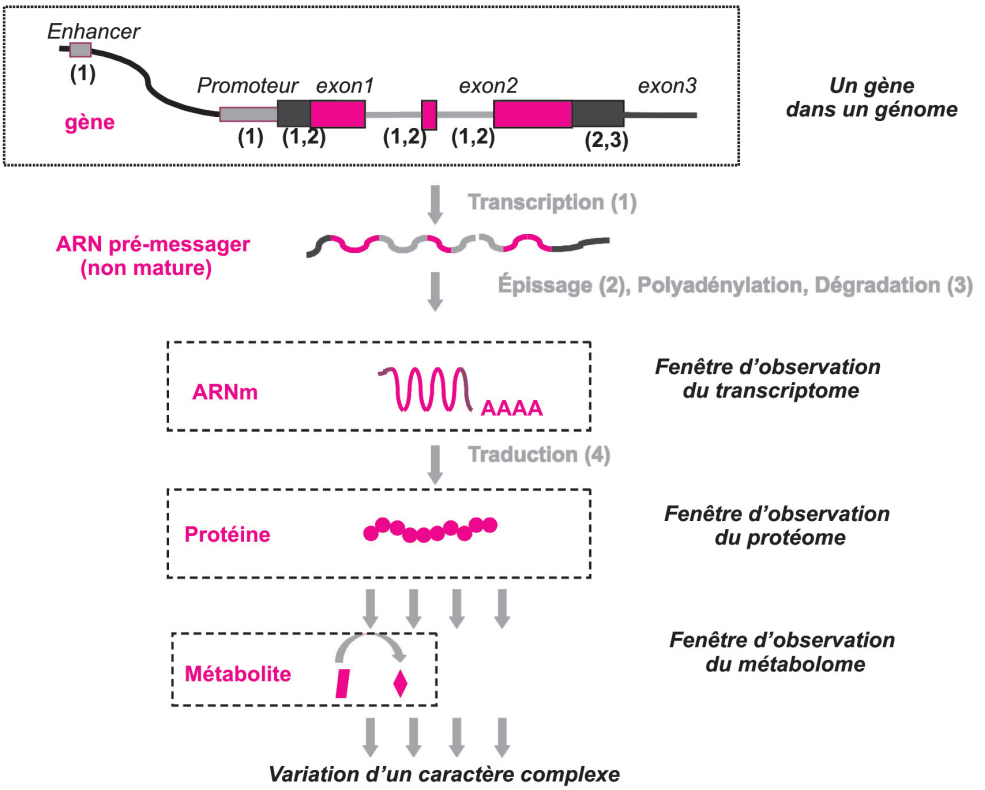
Ainsi, la mesure de la quantité d'un transcrit (ou ARNm) ou d'une protéine résulte des mécanismes de synthèse mais aussi de dégradation de ces molécules. A cause de leur complexité, les voies de contrôle par la cellule de ces mécanismes sont encore loin d'être élucidées. Aussi, la quantité d'ARNm d'un gène donné dans un tissu (ou des protéines associées) est probablement régulée par de nombreuses protéines et donc de gènes les codant.

Etant donné sa structure, un gène peut se décomposer en différentes régions. Les régions codantes (boîtes rouges sur l'encadré 1) portent l'information de la future protéine. Les autres régions dites régulatrices (boîtes noires ou grises ou traits gris sur l'encadré 1) participent au contrôle de la transcription du gène en ARNm (*enhancer*, promoteur, introns…) ou à la stabilité/dégradation de ces ARNm (régions transcrites mais non codantes en fin de gène).

Encadré 1. *Rappel des différentes étapes de transmission de l'information du gène à la protéine correspondante.*

Exons (régions transcrites dans l'ARNm) :
Régions codantes (traduites en protéine)
Régions non codantes (non traduites en protéines)
Introns (régions inter-exoniques, excisées de l'ARNm pré-messager)

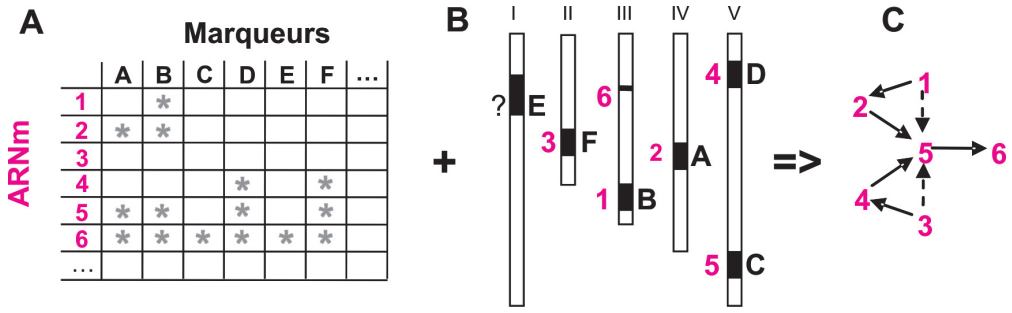
(1), (2) et (3) : Régions du gène intervenant dans la régulation de sa transcription (1), dans l'épissage de ses ARNm non matures (2) ou encore dans la stabilité de ses ARNm (3)



Les gènes sont dans un premier temps transcrits en ARN pré-messagers. La maturation des pré-messagers implique principalement l'excision des séquences introniques (on parle alors d'épissage) et l'addition d'une succession de plusieurs ribonucléotides de types Adénosine (queue poly A) en extrémité 3' de la molécule (ou polyadénylation). L'ARN acquiert alors une meilleure stabilité, nécessaire pour limiter les phénomènes de dégradation dus en majeure partie aux enzymes de types RNAse. C'est à ce stade, en aval des nombreux mécanismes ci-dessus mentionnés, que l'on peut mesurer le niveau d'expression d'un gène. Cet ARNm mature est ensuite traduit en protéine lors de la traduction. Certaines protéines, les enzymes, peuvent conduire à la transformation de métabolites. Un ensemble de protéines contribue par la suite aux variations d'un caractère complexe visible à l'échelle de l'animal.

Aujourd'hui, grâce aux technologies de la génomique fonctionnelle, il devient possible de mesurer simultanément dans un tissu donné l'expression de plusieurs gènes au niveau de leur ARNm (transcriptome) ou de leurs protéines (protéome). Il est également possible de mesurer différentes métabolites d'un tissu (métabolome).

Figure 1. *Interprétation conjointe des positions chromosomiques des régions eQTL et des gènes qu'elles contrôlent en vue de reconstituer une voie métabolique (d'après Jansen et Nap 2001).*



L'identification de régions eQTL (au niveau des marqueurs A, B, C,…) régulant des gènes (gène 1, 2, 3…) (indiqué en A) couplée à la localisation de ces gènes sur le génome, en particulier dans les régions eQTL(B) permet de reconstituer en théorie une voie métabolique ou voie de régulation (C). Par exemple, l'expression du gène 6 est contrôlée par les 6 régions indiquées, il est donc probable que ce gène code pour une protéine en fin de voie métabolique ou de régulation. Le gène 5 partageant 4 de ces 6 régions et se trouvant être localisé dans la région C contrôlant le gène 6, serait donc régulateur du gène 6. Par un raisonnement similaire généralisé à l'ensemble des gènes et régions indiquées à gauche de la figure, une voie métabolique ou de régulation est proposée à droite de la figure.

Une mutation dans les régions régulatrices d'un gène peut donc conduire à une variation de la quantité de ses transcrits, une mutation dans la région codante peut affecter la fonctionnalité de la protéine codée, et éventuellement affecter les quantités d'ARNm d'autres gènes. Dans ce cas, une mutation dans un gène donné peut avoir des impacts sur la quantité d'ARNm d'autres gènes sans que la quantité de ses propres ARNm ne soit affectée.

Grâce à l'essor des technologies de génomique fonctionnelle dans les années 2000, il est devenu possible de quantifier dans un tissu donné les niveaux d'ARNm de l'ensemble des gènes contenus dans un génome. Ces niveaux de transcrits constituent le «transcriptome» du tissu considéré. De même, il est possible de quantifier quelques centaines de protéines en une seule expérimentation (appelé alors protéome) ou quelques dizaines à centaines de métabolites (appelé alors métabolome) (encadré 1). Concernant les métabolites, une partie d'entre eux sont le produit de réactions biochimiques effectuées par des proteines (et donc plus en amont par des gènes codant ces protéines). Aussi, la fenêtre d'observation des métabolites est intéressante à la fois pour préciser un phénotype d'intérêt agronomique mais aussi pour apprécier l'activité fonctionnelle/expressionnelle d'un génome. L'accès à ces quantités d'ARNm, de protéines ou de métabolites permet donc d'observer à grande échelle les phénotypes intermédiaires entre les polymorphismes du génome et les caractères d'intérêt agronomique. Il est intéressant de noter que les technologies du transcriptome sont les seules à permettre l'analyse de l'expression de l'ensemble des gènes d'un génome. Aussi sont-elles plus utilisées que les autres technologies du protéome ou du métabolome. Les références bibliographiques et les illustrations mentionnées dans cet article seront donc centrées sur les seules données du transcriptome.

Selon les mêmes principes que ceux concernant la cartographie de QTL classique pour des caractères visibles au niveau de l'animal, il devient maintenant possible de détecter des régions du génome contrôlant le niveau d'ARNm d'un ou de plusieurs gènes. Ces régions sont appelées eQTL pour QTL d'expression. Le niveau d'ARNm d'un gène est alors considéré comme un caractère complexe à part entière. En 2001, Jansen et Nap proposent de nommer ce nouveau concept «*genetical genomics*» (pour génétique de la génomique ou encore «génétique génomique») pour lequel une analyse de la liaison génétique entre marqueurs et expression d'un gène permet de mettre en évidence des locus responsables d'une part de la variation de son expression. Comme nous le verrons plus loin, il est ainsi possible d'identifier des gènes dont les expressions sont affectées par plusieurs locus ou encore un ensemble de gènes partageant les mêmes locus de contrôle. Comme indiqué dans la figure 1, l'analyse de ces différentes liaisons génétiques entre locus et gènes régulés combinées à la localisation des gènes dans le génome devrait en théorie permettre selon Jansen et Nap de reconstituer des voies de régulation ou voies métaboliques en mettant en évidence le gène le plus aval de la voie, celui dont l'expression est gouvernée par le plus grand nombre de locus.

La «génétique génomique» constitue une nouvelle manière d'observer les événements de régulation génique à une échelle encore jamais explorée.

Un an plus tard, le concept de la «génétique génomique» est validé par des premières expériences chez la levure où une cartographie de régions contrôlant la variabilité de l'expression de gènes est réalisée à l'échelle du génome (Brem *et al* 2002). Depuis, de nombreuses études ont été menées visant à cartographier les eQTL de gènes chez la levure (Yvert *et al* 2003), chez des espèces modèles comme la souris (Schadt *et al* 2003), la drosophile (Wayne et McIntyre 2002), le rat (Hubner *et al* 2005) ou encore chez l'Homme (Monks *et al* 2004, Morley *et al* 2004). On note également des études similaires chez les végétaux comme le maïs (Schadt *et al* 2003), l'*Eucalyptus* (Kirst *et al* 2004), *Arabidopsis thaliana* (DeCook *et al* 2006) et l'orge (Potokina *et al* 2008).

Le concept de la cartographie de régions eQTL peut également être utilisé pour des caractères de type «quantité d'une protéine (pQTL)» (Zivy et de Vienne 2000) ou encore «quantité d'un métabolite (mQTL)» (Ferrara *et al* 2008). Les premiers travaux visant à cartographier des régions responsables de la variabilité d'un taux protéique sont d'ailleurs plus anciens que la cartographie de QTL d'expression au niveau des ARNm (De Vienne *et al* 1999).

Comparé à un phénotype d'intérêt agronomique, les phénotypes élémentaires que sont les transcrits ou les protéines d'un gène diffèrent par le nombre plus faible de mécanismes impliqués dans leur variation. Malgré tout, ces mécanismes sont divers comme indiqué plus haut (mécanismes de contrôle de l'activité transcriptionnelle du gène ou de la dégradation de ces ARNm ou protéines). Ainsi, la variation des quantités d'ARNm d'un gène peut être la résultante d'un polymorphisme présent

dans le gène lui-même ou présent dans un autre gène qui serait alors impliqué dans l'un des mécanismes de contrôle précédemment cité. Le vocabulaire qualifiant les régions eQTL s'est donc enrichi par rapport à celui qualifiant les QTL.

1.2 / Les *cis* et *trans* eQTL

Une région eQTL sera qualifiée de «*cis* eQTL» si sa localisation est proche de la position du gène dont elle gouverne l'expression. Au contraire, si ce gène est positionné ailleurs dans le génome par rapport à la position de la région eQTL, celle-ci sera qualifiée de «*trans* eQTL». Ce vocabulaire de *cis* et *trans* eQTL a été emprunté au domaine de la biologie moléculaire. En biologie moléculaire, le terme *trans* est souvent associé aux facteurs de transcription régulant un gène et se fixant en *trans* sur des séquences en général promotrices du gène régulé appelées, elles, séquences-*cis*. Ces termes *trans* et *cis* ont donc une connotation mécanistique. Ils ne sont donc pas toujours appropriés aux eQTL, les études de cartographie ne renseignant pas les mécanismes moléculaires mis en jeu dans les régulations. Certains auteurs estiment préférable d'adopter les termes d'eQTL «locaux» et «distants» (figure 2) faisant référence à la position des marqueurs génétiques par rapport à la position des gène régulés (Rockman et Kruglyak 2006) : le terme d'eQTL local est communément employé lorsque la position du marqueur est comprise dans une fenêtre de taille arbitraire (le plus souvent 5 à 10 Mb) autour de la position du gène régulé, dans le cas contraire, on parle d'eQTL «distant».

D'après les connaissances que nous avons de la régulation des niveaux d'ARNm d'un gène, les *cis* eQTL peuvent être causés par des polymorphismes dans les régions régulatrices des gènes eux mêmes : dans leurs séquences promotrices, introniques, exoniques non codantes, ou *enhancer* (cf. encadré 1). Cependant, la démonstration de l'effet d'un polymorphisme dans des régions régulatrices sur l'expression d'un gène est difficile du fait des localisations imprécises de ces régions eQTL ; elle nécessite *in fine* des expérimentations de biologie moléculaire lourdes à mettre en place.

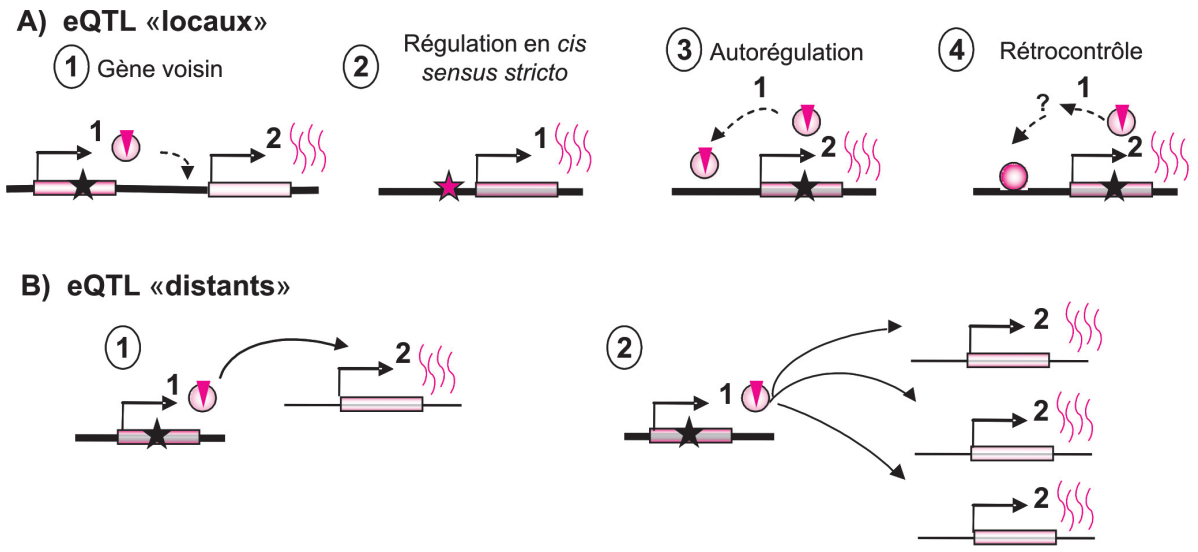
Les régions *trans* eQTL seraient quant à elles causées par des mutations affectant l'activité de gènes de la région qui réguleraient alors le niveau transcriptionnel d'autres gènes localisés ailleurs dans le génome (Farrall 2004). On peut donc imaginer des facteurs de transcription de toutes sortes. Néanmoins, différents auteurs ayant détecté plusieurs régions *trans* eQTL n'observent pas systématiquement la présence de gènes codant des facteurs de transcription dans ces régions (Yvert *et al* 2003, Bing et Hoeschele 2005). Il semblerait donc que l'intervention de gènes participant autrement à la machinerie transcriptionnelle (protéine d'épissage, protéine chaperonne, import/export nucléaire…) ou intervenant dans la dégradation des ARNm soit beaucoup plus fréquente que prévue.

1.3 / Les master eQTL

Les régions eQTL contrôlant l'expression d'un grand nombre de gènes sont communément appelées régions master eQTL (Morley *et al* 2004). Certains auteurs utilisent ce terme lorsque le nombre de gènes régulés par une telle région excède 25 (Gibson et Weir 2005).

A l'instar des régions *trans* eQTL, bien qu'on puisse imaginer que les master eQTL soient le reflet de l'action de gènes codant des facteurs de transcription, des analyses chez la levure avec une densité de marqueurs raisonnable permettent de conclure que la plupart de ces eQTL ne possèdent pas de gènes codant des facteurs de transcription, suggérant donc d'autres types de régulateurs (Yvert *et al* 2003).

Figure 2. *Régions eQTL contenant une mutation agissant localement (A) ou à distance (B).*



La région eQTL est indiquée par un trait gras (—). Le gène porteur de la mutation est indiqué par une boite rouge (▭). La mutation est indiquée par une étoile rouge (★) quand elle est située dans les régions régulatrices du gène (région promotrice ici) provoquant alors une variation de quantité d'ARNm ( ) et par une étoile noire (★) quand elle est située dans la région codante provoquant une modification de l'activité de la protéine associée ( ).

2 / Cartographie de QTL d'expression : principaux résultats extraits de la littérature

Les études de la régulation génétique à l'échelle d'un ou quelques gènes existent depuis longtemps (Jacob et Monod 1961). En revanche, les nouvelles méthodologies, plus exhaustives et dites à haut débit, reposant en partie sur l'utilisation de puces à ADN n'ont réellement débuté qu'au début des années 2000. La technologie de puce à ADN s'est depuis démocratisée avec des coûts d'expérimentation (hors analyse) raisonnable de l'ordre de 150 euros par échantillon permettant d'analyser le profil transcriptomique de milliers de gènes d'un tissu pour des dizaines voire quelques centaines d'animaux. Par ailleurs, le développement de cartes génétiques relativement denses en marqueurs, disponibles pour la majorité des espèces modèles ou d'intérêt économique, ainsi que le savoir-faire des équipes qui ont dans un premier temps travaillé sur la cartographie de QTL, ont conduit à un essor des travaux de cartographie de QTL d'expression dont une revue est présentée ci-après.

2.1 / Proportion de gènes régulés par un QTL d'expression

a) Proportion de gènes ayant au moins un eQTL

Parmi les études de «génétique génomique» conduites essentiellement chez la levure, la souris et l'Homme, la proportion de gènes dont l'expression est détectée comme régulée par au moins un eQTL peut largement varier. Les différences sont dues à différents facteurs que sont la puissance de l'analyse reposant essentiellement sur l'effectif du dispositif d'animaux et le nombre de marqueurs utilisés ou encore le choix des seuils de signification d'un eQTL. Par exemple, seulement 0,6% des 374 gènes analysés dans l'étude de Stranger *et al* (2005) présentent au moins un eQTL. Ce pourcentage, un des plus bas observé, est en partie dû au manque de puissance de l'analyse d'association qui a été effectuée sur seulement 60 hommes non apparentés ; à noter également que le seuil de signification a été corrigé pour les tests multiples par la méthode de Bonferroni, méthode la plus drastique. A l'opposé, 59% des 5700 gènes analysés dans l'étude de Brem et Kruglyak (2005) présentent au moins un eQTL, cette étude a été réalisée sur un dispositif plus puissant, à savoir par analyse de liaison effectuée sur une structure en ségrégation de 112 levures. Ce pourcentage est de 40% dans l'étude de Yvert *et al* (2003) fondée sur 86 levures et diminue à 9% dans l'étude de Brem *et al* (2002) n'utilisant plus que 40 levures, ces études utilisant des seuils de signification similaires pour détecter un eQTL. Comme l'illustrent ces exemples, la puissance des dispositifs est un élément majeur dans les différences de résultats observées. Dans la majorité des études qu'ont recencées Williams *et al* (2007), les proportions de gènes ayant au moins un eQTL sont de l'ordre de 10 à 30%.

b) Proportion de cis et trans eQTL

D'après le tableau 1, le pourcentage de gènes pour lesquels une région eQTL est considérée comme agissant en *cis* est très variable d'une étude à l'autre (0,8 à 98%). Ce pourcentage peut varier en fonction de la taille de la fenêtre séparant les positions du gène régulé et du marqueur (= fenêtre *cis*) ainsi que du niveau de signification fixé pour définir une liaison génétique et/ou une association entre marqueur et expression d'un gène. Selon les études, la définition de la taille des fenêtres *cis* varie de 10kb (Brem *et al* 2002) à 20Mb (Bystrykh *et al* 2005). Certains auteurs font même référence aux *cis* eQTL lorsque l'eQTL régulant le niveau d'expression du gène correspond au marqueur le plus proche de l'oligonucléotide (Chesler *et al* 2005, Emilsson *et al* 2008). Des éléments *cis* régulateurs sont connus pour agir à plus de 1Mb du gène qu'il régule (Pfeifer *et al* 1999) ; aussi des tailles des fenêtres qui peuvent parfois paraître démesurées ne sont pas invraisemblables. De façon logique, le pourcentage de gènes régulés en *cis* augmente lorsque la fenêtre *cis* est élargie. Par ailleurs, les régions *cis* eQTL ont un effet plus prononcé sur les variations des niveaux d'ARNm des gènes comparé aux régions *trans*, du fait probablement d'un nombre d'événements biologiques moins élevé entre le polymorphisme et son effet. Ces régions sont en conséquence identifiées avec des statistiques de test plus élevées. Pour illustrer ce point, les études de Schadt *et al* (2003) menées chez la souris identifient comme *cis* eQTL 34% des 3701 régions eQTL localisées avec un LOD score > 4,3 et 71% des 784 régions eQTL identifiées avec un LOD score > 7. La proportion de régions eQTL agissant en *cis* tend donc à augmenter lorsque le seuil de signification des eQTL augmente (Schadt *et al* 2003, Doss *et al* 2005).

c) Proportion de master-trans eQTL

Il existe dans la littérature différentes façons de définir les gènes dont le niveau d'expression est régulé par une même région eQTL. Plusieurs études comptent le nombre de transcrits qui sont cartographiés pour un même marqueur (Bystrykh *et al* 2005, Chesler *et al* 2005, Hubner *et al* 2005, Cotsapas *et al* 2006). D'autres études comptent le nombre de gènes ayant un eQTL cartographié dans une fenêtre de taille prédéfinie (Brem *et al* 2002, Schadt *et al* 2003, Yvert *et al* 2003, Morley *et al* 2004).

Pour certains dispositifs eQTL actuellement décrits dans les deux espèces Homme et souris (tableau 1), le nombre de régions *master* eQTL varie de 1 (Cotsapas *et al* 2006) à 17 (Bystrykh *et al* 2005). En revanche, Monks *et al* (2004) et Emilsson *et al* (2008) n'en détectent aucune et suggèrent alors que ces régions ne seraient pas universelles dans le règne animal.

2.2 / Proportion de gènes régulés par plusieurs QTL d'expression

La plupart des analyses eQTL effectuées aujourd'hui utilise des méthodes «simple *locus*» où chaque locus est analysé indépendamment des autres locus pour détecter des liaisons ou associations avec les données transcriptomiques. Des analyses de liaison multi-QTL ou des tests d'associations multiples sont encore peu utilisés du fait de la complexité des tests statistiques, de l'importance des ressources informatiques nécessaires pour les effectuer et également et surtout de la taille requise des dispositifs. Néanmoins, des transcrits régulés par plusieurs endroits du génome peuvent être identifiés (Brem *et al* 2002, Schadt *et al* 2003, Morley *et al* 2004, Monks *et al* 2004, Brem et Kruglyak 2005, Cheung *et al* 2005, Hubner *et al* 2005, Stranger *et al* 2005, Cotsapas *et al* 2006). Ces études montrent que seulement 3% des phénotypes d'expression seraient contrôlés par un seul locus alors que plus de 50% seraient sous l'influence d'au moins 5 régions eQTL. Il n'y aurait que 23% des transcrits qui seraient régulés par un eQTL expliquant plus de 50% de la variance génétique. Ces observations sont cohérentes avec les mécanismes multiples de régulation de la quantité des ARNm d'un gène (cf. § 1.1). Tout comme les caractères complexes visibles à l'échelle de l'animal, les quantités d'ARNm sont à juste raison considérées comme des caractères quantitatifs dont les variations reposent sur un modèle polygénique additif avec quelques QTL à effets moyens à forts.

2.3 / Les interactions épistasiques

L'épistasie entre locus peut se définir comme étant l'interaction entre deux locus (ou plus) avec pour conséquence

Tableau 1. *Pourcentage de cis eQTL observés dans différentes études.*

Etude	Espèce (effectif de la population analysée)	Tissu	Nombre de gènes analysés	% de *cis eQTL*	Taille fenêtre *cis*	Seuil de signification	Nombre de *master eQTL*
Brem *et al* (2002)	Levure (40)	-	6215	36	10kb	P < 5 ×10 - 5	8
Yvert *et al* (2003)	Levure (86)	-	6215	25	10kb	P < 3,4 ×10 - 5	13
Schadt *et al* (2003)	Souris (111)	Foie	7861	34	2cM	LOD > 4,3	7
Monk *et al* (2004)	Homme (167)	Lignées cellulaires lymphoblastoïdes	2430	39	5Mb	P < 5 × 10 - 5	0
Morley *et al* (2004)	Homme (195)	Lignées cellulaires lymphoblastoïdes	3554	22	5Mb	P < 4, 3×10 - 7	2
Hubner *et al* (2005)	Rat (22)	Rein	15923	32	10Mb	P < 0,05	2
Chesler *et al* (2005)	Souris (32)	Prosencéphale	608	94	eQTL= marqueur le plus proche de l'oligo cible	FDR = 0,05	7
Bystrykh *et al* (2005)	Souris (22)	Cellules souches hématopoiétiques	12422	13	20Mb	P ≤ 0,005	17
Lan *et al* (2006)	Souris (60)	Foie	45000	12	10cM	LOD > 3,4	15
Wang *et al* (2005)	Souris (312)	Foie	23574	31	20Mb	P < 5 × 10 - 5	7
Cotsapas *et al* (2007)	Souris (31)	Cellules adipeuses	-	29	5Mb	Bonferroni alpha = 0,05	1
		Cerveau	17706	0,8			2
		Rein	9237	5,5			2
		Foie	10728	3,4			3
Bhasin *et al* (2008)	Souris (203)	Macrophage	17632	20	20Mb	LOD > 3	11
Schadt *et al* (2008)	Homme (427)	foie	39280	87	1Mb	Bonferroni alpha = 0,05	-
Emilsson *et al* (2008)	Homme (1002)	Sang	23720	98	eQTL= marqueur le plus proche de l'oligo cible	FDR = 0,05	0
Ghazalpour *et al* (2008)	Souris (110)	Foie	24048	40	10Mb	FDR = 0,1	4
Ponsuksili *et al* (2008)	Porc (74)	*Longissimus dorsi*	23256	7	Colocalisation QTL	P ≤ 0,05	-

un effet sur un caractère. En terme qualitatif (aussi appelé Mendélien), les interactions entre deux locus vont aboutir à l'atténuation ou disparition des effets de certains allèles d'un des deux locus selon la présence d'allèles à l'autre locus. En terme quantitatif, l'épistasie se réfère à la part de la variance génétique qui ne peut être expliquée ni par les effets additifs des allèles en présence, ni par les effets de dominance. L'effet de la coségrégation de plusieurs marqueurs sur la variabilité d'expression d'un gène a été testé chez la levure (Storey *et al* 2005). Ces travaux montrent que les niveaux de 37% des transcrits seraient régulés par au moins deux locus dont 14% seraient sous le contrôle de régulations épistasiques. Chez la drosophile, Anholt *et al* (2003) démontrent également l'importance des interactions épistasiques entre eQTL. Ces interactions entre eQTL régulant un même transcrit ne sont pas surprenantes au regard des connaissances que l'on a sur la régulation génique. En effet, la régulation des transcrits est généralement due à l'action de protéines agissant en combinaison. Ces premières études restent à être enrichies par de futurs travaux réalisés sur des dispositifs avec des effectifs plus élevés permettant de gagner en puissance. Par ailleurs, les algorithmes permettant de tester l'épistasie entre eQTL font l'objet de recherches importantes en particulier pour les adapter au nombre très élevé de variables à étudier (dizaines de milliers de gènes) (Carlborg *et al* 2005).

En conclusion de cette partie 2, les études évoquées conduisent à des résultats parfois disparates concernant les nombres d'eQTL de type *cis*, d'eQTL recencés par gène, d'eQTL en interaction. Ces disparités résultent probablement de la disparité des dispositifs expérimentaux analysés et des méthodes d'analyse utilisées. On peut ainsi recencer entre les études des différences dans le nombre d'individus analysés, la complexité génétique de la population en ségrégation étudiée (population F2 issus du croisement entre lignées consanguines ou non), les différentes sources biologiques d'où sont extraits les ARNm (tissus, cellules), les méthodes de quantification des ARNm à base de puces à ADN plus ou moins exhaustives en terme de gènes déposés ou de qualité de fabrication, le type et le nombre de marqueurs génétiques (microsatellites *vs* SNP, 100 *vs* 100 000 marqueurs analysés), les «fenêtres» sur le chromosome prises en compte pour la définition des *cis/trans*-eQTL, les

Tableau 2. *Principales études ayant, par une approche eQTL, identifié un gène candidat sous-jacent à un QTL responsable d'un caractère d'intérêt.*

Etude	Echantillon	Source des marqueurs génétiques	Technologie d'hybridation utilisée	Gène candidat identifié	Caractère ou maladie	Méthode d'identification	Mutation causale
Schadt *et al* (2003) (papier princeps)	111 souris F2 (B×D). Foie	Microsatellite. 1 marqueur tous les 13 cM	Puce Agilent bicouleur. 23574 oligos.	dolichyl-diospho-oligosaccharide-protein glycosyltransferase	Obésité	cis-eQTL	?
				cation-transporting ATPase	Obésité	cis-eQTL	?
Hubner *et al* (2005)	22 rats RI B×H/H×B. Rein et cellules adipeuses	Puce Affymetrix rat. 15923 oligos	1011 marqueurs microsatellites autosomaux provenant de WebQTL	73 gènes à tester dans populations humaines	Hypertension	cis-eQTL + co-localisation QTL	?
Schadt *et al.* (2005)	111 souris F2 (B×D). Foie	Microsatellite. 1 marqueur tous les 13 cM	Puce Agilent bicouleur. 23574 oligos.	Hsd11b1	Obésité	LCMS+Knockout	?
Mehrabian *et al* (2005)	111 souris F2 (B×D). Foie	Microsatellites. 1 marqueur tous les 13 cM	Puce Agilent bicouleur. 23574 oligos.	Alox5	Obésité	cis-eQTL + co-localisation QTL+ Knockout	?
Yaguchi *et al* (2005)	113 souris F2 (B×D). Foie et tissus adipeux	Microsatellite. 227 couvrant le génome	RT-PCRq pour 76 parmi 106 gènes d'une région QTL	17 gènes candidats	Diabète	cis/trans-eQTL + co-localisation QTL	?
Mootha *et al* (2006)	54 hommes d'âge similaires avec des degrés de tolérance au glucose différents	-	Puce Affymetrix HG-U133A. 39000 transcripts.	PGC-1α	Diabète type 2	GSEA	? (Mutations non sens reportées dans 2 autres études)
Lum *et al* (2006)	300 souris F2 (B×D). Brain	SNP. 1200 couvrant le génome	Puce Agilent bicouleur. 23574 oligos.	Pttg1	Obésité/Diabète	cis-eQTL	?
Meng *et al* (2007)	111 souris F2 (B×D) 334 souris F2 (B×H)		RT-PCRq	Abbc6	calcification cardiaque dystrophique	cis-eQTL+Knockout	Délétion de 10 pb en 3'UTR du gène
Bao *et al* (2007)	42 souris RIL (B×D). Brain	SNP. 3795 couvrant le génome	Puce Affymetrix M430. 39000 transcripts	Adcy2	Résistance douleur induite par la chaleur (nociception)	trans-eQTL + co-localisation QTL	? (2 mutations non sens dans la séquence codante (effets non validés))
				Myo7a	Trouble neurologiques et comportementaux	cis-eQTL + co-localisation QTL	?
				Ttc8	Syndrome Bardet-Biedl	cis-eQTL + co-localisation QTL	?
				Ank2	Troubles activités locomotrice	cis-eQTL + co-localisation QTL+ Knockout (dans autre design)	?
Schadt *et al* (2008)	427 hommes caucasiens. Foie	SNP. 782476 couvrant le génome	Puce Agilent bicouleur. 39280 oligos.	Rps26	Diabète type I	eQTL + GWAS	?
				Sort1	Maladie artères coronariennes	eQTL + GWAS	?
				Celsr2	Concentration LDL cholestérol dans le sang	eQTL + GWAS	?
Ponsuksili *et al* (2008)	74 porc F2. Muscle	Microsattelite. 116 couvrant le génome	Puce Affymetrix porcine. 23937 oligos.	Ahnak	Capacité de rétention en eau du muscle	cis-eQTL	?

méthodes d'estimation de liaison génétique ou d'association, le choix des seuils de signification, la prise en compte des tests multiples. Malgré ces différences, les approches combinant des données génétiques et génomiques permettent un nouvel angle d'étude de la variabilité d'expression des gènes et de leurs régulations que l'on sait complexes et aujourd'hui très partiellement connues. Notons que les méthodes permettant d'observer des modules géniques (présentées dans le § 3.3) peuvent également contribuer à cette connaissance. Enfin la prise en compte des effets épistasiques entre eQTL est également un élément important dans le décryptage de ces régulations.

3 / Apport des données transcriptomiques à la cartographie de QTL – Etat des lieux

Dans le cadre de l'identification de QTL responsables de la variation de caractères complexes, ces approches de génétique génomique offrent également une opportunité nouvelle pour caractériser ces régions QTL et faciliter l'identification des gènes causaux.

Les études utilisant les données transcriptomiques pour mieux caractériser des QTL contrôlant un caractère complexe sont en augmentation constante. La majorité de ces études, recencées dans le tableau 2, consiste à identifier des eQTL co-localisant avec la région QTL d'intérêt, donnant ainsi des informations fonctionnelles sur la région. Celles-ci peuvent dans certains cas permettre d'identifier le gène candidat positionnel et fonctionnel recherché. Par ailleurs, quelques auteurs ont utilisé les données transcriptomiques pour décomposer le caractère d'intérêt grâce aux nombreux phénotypes élémentaires que sont les niveaux de transcrits. Nous commencerons donc par exposer cette approche pour ensuite aborder l'approche eQTL plus couramment utilisée. Nous aborderons enfin différentes variantes visant à cartographier plus finement une région QTL.

3.1 / Décomposition d'un caractère complexe en phénotypes plus élémentaires

Cette stratégie consiste à identifier dans la population où des QTL ont été détectés, des sous-groupes de descendants partageant des profils transcriptomiques similaires. L'hypothèse alors faite est que ces sous-groupes d'individus homogènes transcriptionnellement le sont également génétiquement, conséquence possible de différentes mutations dont celles influençant le caractère d'intérêt. C'est d'autant plus vrai lorsque ces sous-groupes correspondent à des sous-groupes du caractère d'intérêt révélant ainsi la multiplicité des processus biologiques et donc des déterminants génétiques pouvant conduire à un même phénotype : par exemple des individus «gras» pourraient se diviser en deux sous-groupes, l'un caractérisé par la présence d'allèles codant des enzymes peu efficaces pour brûler les graisses, l'autre caractérisé par la présence d'allèles codant des enzymes très efficaces pour synthétiser des graisses à partir des sucres alimentaires. Une analyse de liaison sur le caractère d'intérêt en utilisant une partie de ces sous-groupes, peut alors révéler de nouveaux QTL, non observés de façon significative sur l'ensemble du dispositif, du fait de l'hétérogénéité du déterminisme génétique dans l'échantillon d'origine. Cette hétérogénéité peut être due aux influences polygéniques, aux interactions entre des QTL, ou peut-être encore aux effets environnementaux sur certains gènes.

Schadt *et al* (2003), ont été les premiers à développer ce concept. Ils l'ont appliqué à un dispositif de souris F2 issues d'un croisement de deux lignées consanguines en vue de mieux caractériser les QTL responsables du poids de tissu adipeux. Les transcriptomes hépatiques correspondant à 23 574 gènes ont été analysés. Après avoir identifié 208 transcrits corrélés au caractère et conservé une quarantaine de souris ayant les valeurs de caractère les plus extrêmes, les auteurs ont alors effectué une classification ascendante hiérarchique permettant de classer les individus selon leur profil transcriptome. Cette classification a permis de discriminer les animaux gras des maigres mais aussi de distinguer au sein de ces deux groupes d'extrêmes deux sous-groupes d'individus gras et maigres. Une nouvelle analyse QTL réalisée sur ces sous-groupes pris séparément a permis à la fois de confirmer de façon plus significative un QTL mais également de détecter un autre QTL non observé sur la population entière d'animaux (Schadt *et al* 2003).

Ce concept a été repris et appliqué à une famille de 50 poulets de chair, descendants d'un père connu pour être hétérozygote à un QTL pour le poids de tissu adipeux abdominal, localisé sur le chromosome 5 à environ 170 cM. Une analyse des corrélations prenant en compte la dépendance entre gènes, a permis d'identifier 688 transcrits corrélés au caractère «poids de gras» (Blum *et al* 2010). Une double classification sur ces gènes et individus (figure 3) met en évidence cinq groupes d'animaux présentant des profils transcriptomiques différents, en particulier deux sous-goupes pour chacun des groupes extrêmes gras et maigres. Une nouvelle analyse QTL après exclusion des huit animaux du sous-groupe n°5, un sous-groupe d'individus maigres, permet de détecter un autre QTL significatif à environ 100 cM, non observé sur la population entière d'animaux (figure 3). Une analyse fine des haplotypes des huit individus de ce sous-groupe montre qu'ils possèdent tous l'haplotype «q» au QTL à 170cM, suggérant une interaction entre les deux QTL, interaction qui a depuis été démontrée.

Ces études menées chez la poule et chez la souris, démontrent l'intérêt de cette approche pour mieux caractériser les QTL responsables d'un caractère complexe, et ce dans deux dispositifs d'animaux d'effectif réduit (environ 50 animaux) et de structure génétique assez différente.

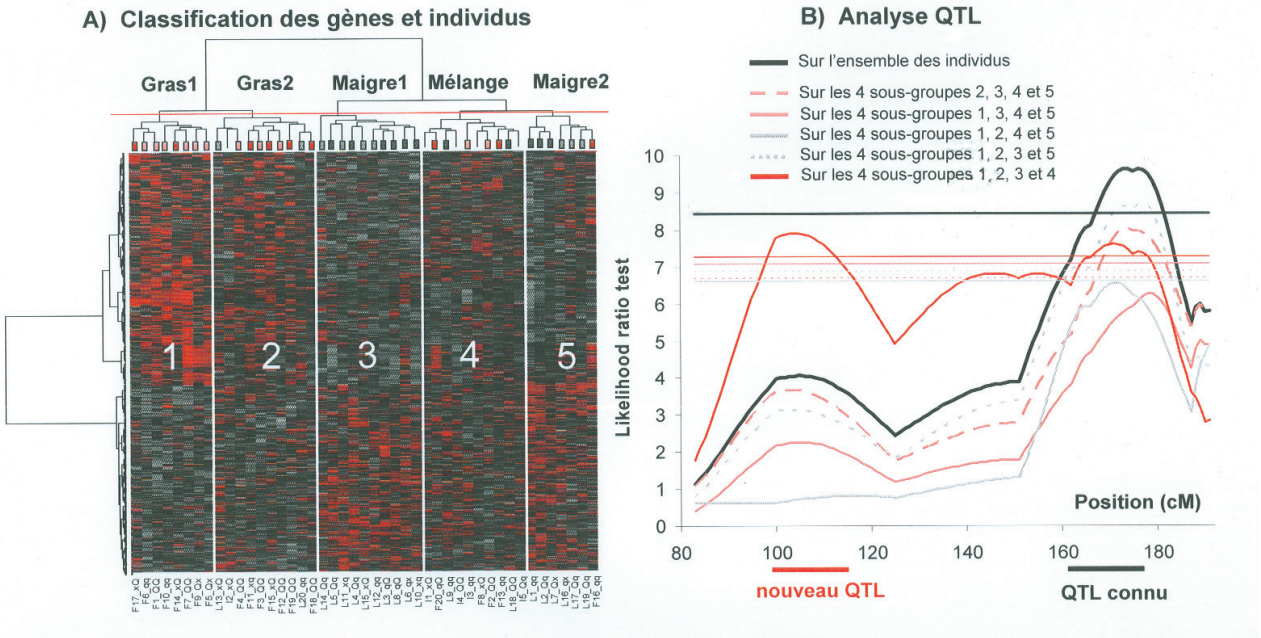
3.2 / Identification de régions eQTL co-localisant avec les régions QTL d'intérêt

L'objectif final de la cartographie de régions QTL est d'identifier le ou les gènes responsables de la variabilité d'un caractère quantitatif. Néanmoins, comme déjà rappelé dans l'introduction, les régions QTL comprennent pour la plupart des centaines de gènes candidats positionnels. La cartographie de régions eQTL a donc pour objectif de faciliter l'identification des meilleurs gènes candidats positionnels en apportant une information fonctionnelle les concernant. Cette approche peut également aider à affiner la localisation de régions QTL préalablement détectées. Cependant, on attend beaucoup plus des méthodologies de génotypage haut débit et de reséquençage concernant ce dernier point.

a) Principes et limites

Le principe est d'identifier les régions eQTL qui colocalisent avec les régions QTL d'intérêt (que nous appelerons régions eQTL/QTL), les gènes régulés pouvant eux se trouver n'importe où dans le génome. L'idée est alors de considérer que parmi les gènes régulés par la région, il y a ceux qui le sont par la mutation causale recherchée, apportant ainsi des informations fonctionnelles sur la mutation, de par les gènes qu'elle régule. Néanmoins l'identification de tels gènes, intermédiaires entre la mutation recherchée et le caractère d'intérêt, n'est pas si simple car, comme indiqué dans la figure 4, une région eQTL/QTL peut recouvrir bien d'autres

Figure 3. *Approche «décomposition d'un phénotype complexe en phénotypes plus élémentaires».*



(A) Classification des gènes et individus : Les individus en colonnes ont été classés selon leur profil transcriptomique sur la base des 688 gènes (situés en lignes) ayant une expression corrélée au caractère «poids de gras». Les individus en rouge et noir (plus ou moins intense selon le poids de gras) correspondent aux 20 individus extrêmes gras et maigres respectivement (F11 à F20 et L11 à L20).

(B) analyse QTL du poids de gras sur le chromosome 5, en utilisant l'ensemble des individus (courbe noire) ou en enlevant un des 5 sous-groupes observés après classification (autres couleurs). Dans le premier cas un QTL est observé à droite du chromosome 5 ; l'élimination des individus du sous-groupe 5 fait apparaître un QTL à gauche du chromosome 5 (courbe rouge). Les distances génétiques (cM) et la statistique de test de présence d'un QTL à une position donnée (LRT) sont indiquées sur les axes X et Y respectivement.

relations entre gènes régulés et caractère d'intérêt. On peut distinguer 4 types de relations :

– l'expression d'un ou plusieurs gènes contrôlée par la région eQTL/QTL peut être régulée par une mutation proche de la mutation contrôlant le caractère d'intérêt (cas 1 de la figure 4). En corollaire, plus ces deux mutations sont proches, plus elles sont en déséquilibre de liaison et plus les deux caractères «gène régulé» et «caractère d'intérêt» sont corrélés (Georges 2007). En conséquence l'identification de tels gènes apporte, par leurs corrélations avec le caractère, une information nouvelle sur la position la plus vraisemblable du QTL et devrait ainsi permettre par des méthodes multivariées de cartographie (Gilbert et Le Roy 2003) de réduire son intervalle de localisation. Cependant, de tels gènes n'apportent aucune information fonctionnelle sur la mutation causale recherchée ;

– l'alternative à la situation précédente est que l'expression d'un ou plusieurs des gènes contrôlés par la région eQTL/QTL soit régulée par la même mutation que celle contrôlant le caractère d'intérêt. Comme dans le cas précédent, de tels gènes nous apportent une information de position sur la mutation causale du QTL mais apportent aussi une information fonctionnelle sur les perturbations biologiques provoquées par cette mutation. Une telle information peut ainsi conduire à la mise en évidence d'un gène candidat positionnel et fonctionnel comme responsable du caractère d'intérêt. Ce cas idéal dans notre contexte correspond au cas 2 de la figure 4 mais ce n'est malheureusement pas la seule situation à envisager dans le cadre de gènes et caractère régulés par une même mutation ;

– la mutation causale au QTL peut contrôler les niveaux d'ARNm de gènes impliqués dans d'autres processus biologiques que ceux responsables de l'établissement du caractère complexe (cas 3 de la figure 4) ;

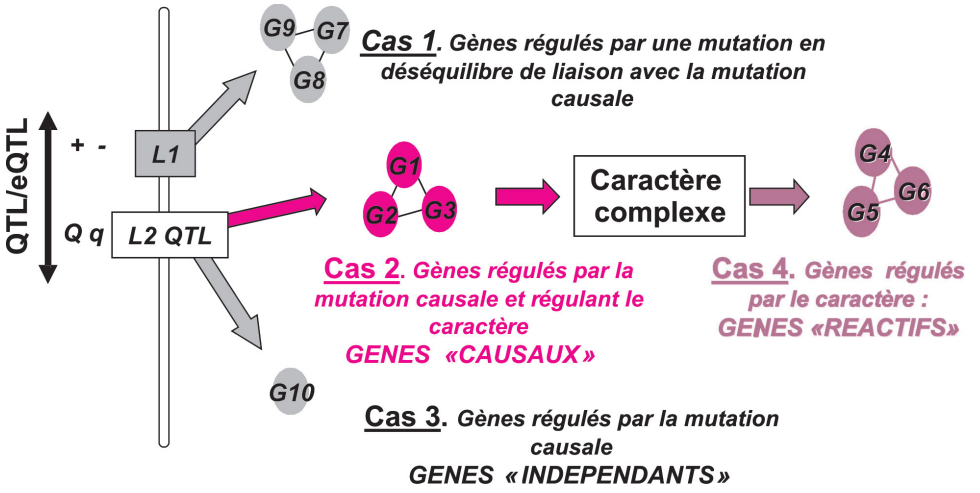
– enfin le caractère peut également réguler *a posteriori* le niveau d'autres ARNm (cas 4 de la figure 4).

Ces trois dernières situations faisant intervenir différentes relations entre caractère d'intérêt et gènes régulés par une région eQTL/QTL ont conduit certains auteurs à adopter une nomenclature pour les qualifier. Schadt *et al* (2005) ont ainsi introduit les termes de gènes «causaux», «indépendants» et «réactifs» par rapport au caractère, termes repris dans la figure 4**.**

Une spécificité de ces gènes «causaux», «réactifs» et «indépendants» est qu'ils sont tous régulés par la même région eQTL/QTL d'intérêt compliquant l'identification des gènes «causaux» qui sont les seuls donnant des informations fonctionnelles directes sur le gène et la mutation causale sous-jacente au QTL d'intérêt. Généralement, les chercheurs procèdent à une analyse plus fine des fonctions des gènes régulés par la région eQTL/QTL en espérant identifier des gènes dont la fonction biologique ait un lien évident avec le caractère d'intérêt. Ces gènes sont alors considérés comme des gènes «causaux» et peuvent ouvrir des pistes quant au meilleur gène candidat positionnel et fonctionnel de la région QTL qui à la fois contrôle leur expression et contrôle le caractère d'intérêt.

Concernant les espèces d'élevage, quelques études ont mis en œuvre cette approche et ont ainsi identifié un ou des gènes en lien avec le caractère d'intérêt et régulé(s) par la région QTL, en faisant donc une bonne signature fonctionnelle du gène candidat positionnel et fonctionnel recherché. On peut ainsi citer une étude conduite par Ponsuksili *et al* (2008) sur le pouvoir de rétention d'eau de la viande chez le porc, une autre sur le poids de gras chez le poulet de chair (Le Mignon *et al* 2009, Blum *et al* 2010) ou encore sur la couleur de la

Figure 4. *Différentes relations pouvant exister entre le caractère d'intérêt et les gènes «contrôlés» par une région QTL/eQTL.*



Quatre types de relations entre «région eQTL/QTL», «gènes» et «caractère» peuvent être distingués :

Cas 1 : l'expression d'un ou plusieurs gènes est contrôlée par la région eQTL/QTL par une mutation proche de la mutation causale recherchée contrôlant le caractère d'intérêt.

Cas 2, 3 et 4 : l'expression d'un ou plusieurs gènes est contrôlée par la région eQTL/QTL par la même mutation que celle contrôlant le caractère. Cependant dans le **cas 2**, les gènes sont responsables d'une part de la variation du caractère. Dans le **cas 3**, ils sont régulés par les variations du caractère (reflet bien souvent de boucles de rétrocontrôle). Enfin, dans le **cas 4**, les gènes sont contrôlés par la mutation causale indépendamment du caractère d'intérêt.

viande chez le poulet de chair (LeBihan-Duval *et al* communication personnelle). La dernière étude est la seule à notre connaissance qui ait abouti à l'identification des mutations causales au QTL d'intérêt ; cette étude correspond à la situation favorable des *cis* eQTL (voir ci-après).

b) Cas particulier des gènes régulés par un cis eQTL

Les gènes dont le niveau d'ARNm est régulé par une région eQTL de type *cis* co-localisant avec une région QTL sont des gènes particulièrement intéressants.

Tout d'abord, comme déjà indiqué précédemment, les régions *cis* eQTL sont plus facilement détectables que les régions *trans* eQTL (LOD score élevé) probablement dû à un nombre réduit d'événements biologiques séparant la mutation dans un gène et sa propre régulation transcriptionnelle. Afin de mettre en évidence expérimentalement cette caractéristique donnée aux régions *cis* eQTL, Schadt *et al* (2003) ont mené un programme de cartographie d'eQTL sur un croisement de lignées de souris DBA et B6 connues pour différer par une délétion de 2 paires de bases dans le gène C5, cette délétion affectant la dégradation de ses transcrits. Comme attendu, ils ont identifié au niveau du gène C5 un *cis*-eQTL avec un LOD score hautement significatif, et même un des plus élevés de l'étude (LOD score > 27,4).

Par ailleurs, l'analyse de la fonction de ces *cis* eQTL permet, si cette fonction est en relation avec le caractère d'intérêt, d'émettre une hypothèse simple sur le meilleur gène candidat positionnel et fonctionnel de la région QTL d'intérêt ainsi que sur la position de la mutation causale dans le gène candidat : ainsi le gène recherché est le gène *cis* eQTL qui est alors responsable du caractère et régulé par la mutation eQTL/QTL située dans ses régions régulatrices.

Prenons l'exemple des travaux de Le Bihan-Duval *et al* (communication personnelle) dans lesquels moins de deux années se sont écoulées entre la primo-localisation d'une région QTL contrôlant la couleur de la viande (Nadaf *et al* 2007) et la détection de la mutation causale associée, montrant ainsi l'efficacité de l'approche. Une analyse des gènes présents dans la région QTL a révélé l'existence d'un gène, *BCMO1*, qui code une enzyme clef de la dégradation du β-carotène, pigment dont l'effet sur la couleur des tissus est bien connu. Ce gène représente donc un bon candidat à la fois positionnel et fonctionnel pour le QTL d'intérêt. Par ailleurs, l'expression de ce gène est contrôlée par une région eQTL qui co-localise avec la région QTL affectant la couleur de la viande. Ces résultats suggéraient donc que *BCMO1* était le gène causal et que la mutation causale recherchée se trouvait dans les parties régulatrices du gène *BCMO1* ; cette mutation contrôlait ainsi la variation de l'expression de *BCMO1*, cette variation impactant alors la couleur de la viande. Le promoteur de ce gène a donc été séquencé sur des animaux de génotypes variés au QTL. Deux mutations dans le promoteur du gène *BCMO1* ont été identifiées et des expérimentations de biologie moléculaire ont montré que ces mutations avaient bien un effet sur l'expression du gène. Notons que la différence d'expression musculaire du gène *BCMO1* entre les deux génotypes au QTL est de deux écart-types contre seulement un écart-type pour celle de la couleur de la viande, illustrant ainsi que la différence entre génotypes d'un phénotype élémentaire (ici l'expression d'un gène) est supérieure à celle d'un phénotype plus complexe, ce qui rend les analyses de cartographie de eQTL/QTL plus puissante et plus précise.

Ainsi, après l'identification de gènes régulés par des régions eQTL co-localisant avec un QTL responsable d'un caractère complexe, la majorité des études se focalise sur les gènes ayant un eQTL de type *cis* (Binget Hoeschele 2005, Doss *et al* 2005, Yamashita *et al* 2005, GuhaThakurta *et al* 2006, Lum *et al* 2006, Ponsuksili *et al* 2008). Cependant, contrairement à l'exemple évoqué plus haut, peu d'études ont abouti à la découverte de la mutation causale, pour différentes raisons : *i)* des gènes présents dans la région eQTL/QTL (dont le gène causal régulé en *cis* recherché) peuvent avoir une

fonction partiellement connue voire même inconnue, ne permettant donc pas de faire le lien avec le caractère d'intérêt ; un grand nombre de gènes sont encore dans ce cas. De plus, tous les gènes de l'intervalle peuvent ne pas avoir été déposés sur la puce et ne sont par conséquent pas analysables. *ii)* la démonstration que le gène régulé en *cis* est le gène réellement responsable de la variabilité du caractère complexe n'est pas des plus simples. Une telle démonstration nécessite des expérimentations supplémentaires de biologie moléculaire qui sont assez longues à mettre en œuvre ; à défaut il est nécessaire que la fonction du gène soit clairement décrite comme étant en lien étroit avec le caractère d'intérêt (cf. point précédent). *iii)* Le nombre de gènes candidats régulés en *cis* présents dans l'intervalle de localisation du QTL peut parfois être élevé. *iv)* des régions *cis* eQTL peuvent être des faux positifs. En effet, Alberts *et al* (2005) ont observé que des polymorphismes de séquences dans la région codante de certains gènes peuvent plus ou moins influencer la qualité de fixation de l'ARNm cible à la sonde déposée sur la puce. Cette différence de fixation de l'ARNm selon les polymorphismes de séquence est ensuite interprétée à tort par l'expérimentateur comme une variation de la quantité d'ARNm. *v)* On peut aussi considérer la situation où des gènes régulés en *cis* dans la région QTL seraient également régulés par un autre endroit du génome (en *trans*) avec des phénomènes d'interaction, perturbant ainsi la détection du *cis* eQTL par des analyses classiques. Yaguchi *et al* (2005) proposent la construction de lignées congéniques qui seraient hétérozygotes uniquement pour la région eQTL/QTL d'intérêt privilégiant ainsi les interactions en *cis*. Cependant, la création d'une lignée congénique est souvent coûteuse et ne garantit pas de reproduire le phénotype assigné par le QTL.

Notons que la co-localisation eQTL/QTL peut également permettre d'exclure des gènes candidats fonctionnels. En effet, Lan *et al* (2004) ont identifié un *cis* eQTL très significatif (LOD 30) pour un gène candidat fonctionnel au diabète (*Pdi*) rapidement exclu puisque ce dernier ne co-localisait pas avec les régions QTL identifiées en parallèle dans le même dispositif.

Quelles que soient les études, le nombre de gènes régulés par une région eQTL/QTL est élevé. Différents facteurs concourent à ces observations : d'une part, les régions QTL ne sont pas localisées avec précision ; d'autre part, les études transcriptomiques permettent l'analyse de dizaines de milliers de gènes dont les processus de régulations sont complexes ; enfin le polymorphisme dans les dispositifs analysés (en général une population issue du croisement entre lignées divergentes) peut être elevé, suggérant de nombreuses mutations en déséquilibre de liaison dans une région eQTL/QTL. Afin de sélectionner les gènes les plus en lien fonctionnellement avec le caractère parmi ceux, parfois nombreux, présents dans une région eQTL/QTL, une analyse bibliographique peut etre menée. Néanmoins, celle-ci se révèle la plupart du temps très chronophage et parfois non informative, les connaissances sur la fonction des gènes étant partielles.

Aussi, des méthodes se sont développées pour diriger plus rapidement l'expérimentateur vers la mutation causale. Certaines n'utilisent d'ailleurs aucune information extérieure au jeu de données (génotype et transcriptome) et vise à distinguer les modules géniques de type «causaux» de ceux considérés comme «réactifs» ou «indépendants», les premiers étant particulièrement utiles dans notre contexte d'étude. D'autres méthodes au contraire utilisent des bases de données associant des informations fonctionnelles aux différents gènes d'une espèce et ont pour objectif d'identifier des modules géniques partageant une même fonction. De telles méthodes permettent dans notre contexte de sélectionner le module dont la fonction serait le plus en relation avec le caractère d'intérêt.

3.3 / Recherche de modules géniques

a) Sélection des modules géniques «causaux», «réactifs» et «indépendants»

Schadt *et al* proposent une méthode appelée LCMS pour *Likelihood-based Causality Model Selection* (Schadt *et al* 2005), permettant d'identifier la relation la plus probable qui existe entre un caractère complexe et l'expression d'un gène, tous deux contrôlés par une même région eQTL/QTL. La méthode considère les trois modèles présentés dans la figure 4 : le modèle «causal», le modèle «réactif» et le modèle «indépendant».

Pour chaque modèle, les paramètres sont estimés par maximisation du critère de vraisemblance et le modèle retenu correspond à celui minimisant le critère d'Akaike (critère couramment utilisé pour identifier le modèle le plus vraisemblable).

Ces auteurs ont appliqué leur méthode dans le cadre de l'identification des gènes causaux au gras viscéral dans un croisement de souris F2 (Schadt *et al* 2005). Après avoir identifié environ 100 gènes dont les niveaux d'ARNm sont à la fois corrélés au caractère et régulés par au moins deux régions eQTL co-localisant avec deux des 4 régions QTL contrôlant le caractère, les auteurs se sont concentrés sur les dix gènes ayant la plus forte probabilité d'être gènes «causaux». Ils ont ainsi pu identifier le gène *Hsd11b1* déjà connu pour son lien avec l'obésité (Masuzaki et Flier 2003). Drake *et al* (2006) ont poursuivi ces travaux en surexprimant ou réprimant chacun des 9 autres gènes chez des souris et ont étudié leur phénotype. Ils ont ainsi pu valider que 8 des 9 gènes ont un effet sur le caractère, deux d'entre eux ayant un effet opposé selon le sexe de l'animal. Ces gènes peuvent donc être considérés comme une signature fonctionnelle de la(des) mutations(s) des régions eQTL/QTL considérées.

Comme montré dans l'exemple précédent, les auteurs ont du procéder à des sélections de gènes pour obtenir les gènes «causaux» les plus vraisemblables. Sans sélection préalable, la méthode LCMS peut prédire des milliers de gènes «causaux» dont beaucoup sont des faux positifs. En effet, des facteurs de variabilité inconnus ou non mesurés peuvent influer sur l'expression des gènes ainsi que sur la variabilité du caractère d'intérêt de sorte à faire apparaître un faux lien de causalité entre les deux (Kruglyak et Storey 2009). De plus, Schadt *et al* supposent qu'un gène corrélé au caractère peut être soit «causal», soit «réactif» et non les deux à la fois. Or, il se peut qu'il y ait des mécanismes complexes de rétrocontrôle et que les hypothèses posées soient donc trop restrictives. Néanmoins, l'exemple cité plus haut démontre l'intérêt de cette méthode.

b) Sélection de modules géniques partageant une même fonction

Ce paragraphe expose des méthodes recherchant à tirer profit de l'information des bases de données associant une fonction à un gène.

Concernant les bases de données d'information biologique, il en existe principalement deux : la base de données des termes Gene Ontology (GO) (Ashburner *et al* 2000) et la base de données des termes KEGG (Kanehisa *et al* 2006). Ces bases de données proposent d'associer à chaque gène les termes le caractérisant le mieux selon les connaissances du moment. A noter que les associations entre termes fonctionnels et gènes peuvent différer entre ces deux bases. La base de données des termes KEGG concerne en majorité des réactions biochimiques et donc des enzymes et les associations «termes

KEGG-gène» sont inférées manuellement. La base de données des termes GO, elle, est scindée en trois classes d'ontologie : les processus biologiques, les fonctions moléculaires et les composants cellulaires. Les associations «termes GO-gènes» sont pour la majorité inférées automatiquement par bioinformatique par exemple en prédisant la(les) fonction(s) d'un gène à partir de sa séquence par recherche de motifs fonctionnels. Aussi, les associations trouvées entre gènes et termes fonctionnels ne sont pas parfaites puisque parfois erronées pour certaines ou encore partielles par rapport à ce qui est connu dans la littérature. Malgré tout, ces informations permettent l'analyse fonctionnelle de nombreux gènes à la fois. Ces deux bases de données sont très largement utilisées et sont considérées comme complémentaires. La base KEGG est plus fiable mais moins complète que la base GO. Par ailleurs, il faut garder à l'esprit que la proportion de gènes ayant des fonctions inconnues ou partielles est importante ; en conséquence ces bases de données reflètent seulement la connaissance partielle que l'on a aujourd'hui de la fonction des gènes.

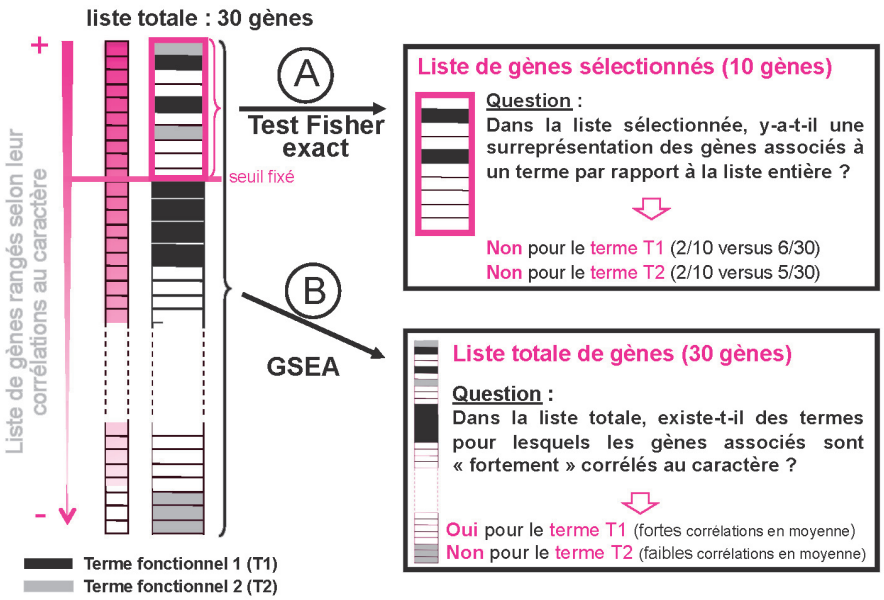
Concernant les méthodes permettant d'identifier des termes fonctionnels que ce soit des termes GO ou KEGG (et les gènes associés) en lien avec le caractère d'intérêt, il en existe deux couramment utilisées : il s'agit du test de Fisher exact et de la méthode GSEA pour *Gene Set Enrichment Analysis*, présentées ci-après et dans la figure 5.

Le test de Fisher exact consiste à mesurer, au sein d'une sous-liste de gènes d'intérêt, l'enrichissement de gènes associés à un terme fonctionnel particulier par rapport à la liste entière des gènes sur la puce. Autrement dit, il s'agit de tester la surreprésentation d'un terme fonctionnel dans une sous-liste de gènes. Ce test est effectué pour chaque terme fonctionnel permettant de recenser ceux qui caractérisent la sous-liste de gènes et qui sont donc impliqués dans la variabilité du caractère. En effet, la sous-liste de gènes d'intérêt est prédéfinie et correspond classiquement aux gènes corrélés au caractère ou encore aux gènes différentiellement exprimés entre individus extrêmes pour le caractère. Une des limites de cette méthode est de se focaliser seulement sur cette sous-liste, ce qui peut se révéler trop restrictif. Une fois la sous-liste d'intérêt sélectionnée, la méthode ne tient plus compte des valeurs de corrélation : tous les gènes sont considérés au même niveau alors que certains sont plus corrélés au caractère que d'autres. Par ailleurs, parmi les gènes non retenus, certains ont des corrélations en limite de signification (proche du seuil fixé) (figure 5).

Pour pallier ces problèmes, la méthode GSEA proposée par Mootha *et al* (2003), permet de trouver les termes fonctionnels en lien avec le caractère en considérant cette fois-ci la liste entière de gènes et leur corrélation avec le caractère (figure 5). D'un certain point de vue, il s'agit de réaliser une analyse différentielle non plus à l'échelle du gène mais à l'échelle du terme fonctionnel. Le principe de la méthode est de regarder au sein d'une liste de gène si un terme fonctionnel particulier est associé préférentiellement à des gènes fortement corrélés au caractère d'intérêt. Si c'est le cas, on pourra dire que ce terme est lié à la variabilité du caractère. Comme indiqué en figure 5, la méthode GSEA peut révéler des termes fonctionnels en lien avec le caractère d'intérêt non-identifiés par la méthode du Fisher exact lorsqu'ils sont associés à des gènes ayant une corrélation avec le caractère en limite de signification.

Différents logiciels permettent une mise en œuvre rapide de l'une ou l'autre de ces deux méthodes même si la date de mise à jour de la base de données d'annotations utilisée n'est pas toujours accessible. Les logiciels fondés sur le test du Fisher exact utilisent le plus souvent les termes GO : AmiGO (Carbon *et al* 2009), EasyGO (Zhou et Su 2007), GOtm (Zhang *et al* 2004) dont une version prend en compte les corrélations des gènes avec le caractère d'intérêt (GOrilla, Eden *et al* 2009), GOstats sous l'environnement R utilisant les termes GO et les termes KEGG (Falcon et Gentleman 2007). Quant à la méthode GSEA, elle est disponible sous l'environnement R avec le logiciel GSEA-P-R (Subramanian *et al* 2005), et plusieurs autres versions comme GSA dont certaines proposent leurs propres associations gènes-termes fonctionnels, les termes pouvant être la localisation chromosomique, le partage de motifs nucléotidiques (Kim et Volsky 2005, Efron et Tibshirani 2007, Jiang et

Figure 5. *Méthodes permettant de tester si un terme fonctionnel est associé significativement à un caractère complexe : en A, test Fisher exact, en B, méthode GSEA.*



Gentleman 2007, Subramanian *et al* 2007).

Les deux méthodes reposent donc sur la constitution de groupes de gènes selon leur appartenance à un terme fonctionnel, la méthode GSEA étant généralement plus puissante. Par ailleurs, selon la source d'information biologique utilisée, généralement la base de données des termes KEGG ou encore des termes GO, les associations entre termes fonctionnels et gènes peuvent parfois différer.

Dans le présent contexte de caractérisation de QTL, Ghazalpour *et al* (2005) ont utilisé ces deux méthodes afin de trouver les métabolismes liés au caractère gras subcutané dans un croisement de souris F2. Après avoir sélectionné les gènes différentiellement exprimés selon le poids en gras subcutané des individus, soit environ 5000 gènes, ils ont recherché les principaux métabolismes dans lesquels ces gènes interviennent. Pour cela, les auteurs ont appliqué à la fois la méthode GSEA et la méthode de Fisher exact en considérant comme liste entière de gènes la liste des 5000 gènes différentiellement exprimés et comme sous-liste de gènes les 20% les plus corrélés. Ces auteurs ont utilisé les termes KEGG pour constituer les groupes de gènes. Les deux méthodes se révèlent comparables, avec cependant une sensibilité légèrement supérieure pour la méthode GSEA : 13 métabolismes sont trouvés avec GSEA et 10 par le test de Fisher exact (tous inclus dans les 13). Les métabolismes trouvés correspondent à des métabolismes liés à l'énergie et aux lipides et concernent 150 gènes dont 68 sont corrélés au gras subcutané. Ce résultat rappelle le concept selon lequel la régulation d'un métabolisme n'implique le contrôle que d'une partie seulement des gènes qui y sont associés. Les auteurs ont alors voulu tester si ces 150 gènes étaient régulés par des régions communes du génome. Ils ont montré qu'il y avait une surreprésentation d'eQTL contrôlant ces gènes dans des régions QTL gras déjà identifiées (Drake *et al* 2001, Schadt *et al* 2003) apportant ainsi une information fonctionnelle sur ces dernières.

c) Cartographie fine d'un QTL par prédiction, sur la base d'une signature fonctionnelle, de l'allèle causal chez des animaux recombinants

Dans la continuité des paragraphes précédents, les niveaux d'expression des gènes identifiés comme signature fonctionnelle de la mutation causale d'intérêt (que nous appelerons gènes «signatures»), peuvent donc en théorie être utilisés comme prédicteurs de l'allèle à la mutation causale. Appliquée à des animaux recombinants, tous descendants d'un père par exemple, la prédiction de l'allèle paternel Q ou q reçu à la mutation causale, devrait permettre de réduire la région QTL d'intérêt. L'approche que nous avons développée en ce sens à l'UMR de génétique animale INRA Agrocampus Ouest, dans une famille d'une cinquantaine de descendants issus d'un père hétérozygote pour un QTL, peut être décomposée en 4 étapes :

– la première étape consiste à identifier, uniquement sur la base des marqueurs, les descendants ayant reçu la totalité de l'haplotype paternel Q (porteur de l'allèle causal Q) ou q (porteur de l'allèle causal q) de la région QTL (individus non recombinants) ;

– la seconde étape consiste alors à établir une fonction à partir du niveau d'expression des gènes «signatures» qui permet de discriminer au mieux l'haplotype Q de l'haplotype q ;

– cette fonction est ensuite utilisée pour prédire le statut de l'allèle causal (Q ou q) chez les animaux recombinants, sur la base du niveau d'expression des différents gènes «signatures» ;

– enfin en confrontant les fragments haplotypiques Q ou q transmis par le père avec l'allèle Q ou q prédit à la mutation causale des recombinants, l'intervalle de localisation du QTL peut être réduit. Le gain de réduction est d'autant plus important que le nombre de marqueurs dans la région est grand, permettant de déterminer avec précision le point de recombinaison chez chaque individu.

Nous avons ainsi réduit une région QTL responsable du poids de gras abdominal chez le poulet de chair de 31 cM à 7 cM (Le Mignon *et al* 2009). Cette approche peut donc se révéler très efficace sous réserve d'une signature fonctionnelle fiable de la mutation causale d'intérêt et de disposer d'animaux recombinants intéressants. A noter que pour les programmes ayant créé des recombinants maîtrisés, cette approche peut éviter de générer de nouveaux descendants pour identifier leur génotype au QTL, ce qui est particulièrement long et coûteux.

Conclusion

La génomique fonctionnelle vise à améliorer notre compréhension des fonctions et de la régulation de l'expression des gènes, de leurs transcrits et des protéines associées, à l'échelle globale du génome (pour revue, Hocquette *et al* 2009). Elle permet d'établir le pont entre le séquençage du génome et les phénotypes observés. Le développement de larges collections d'EST et surtout le séquençage de génomes complets couplé à la prédiction de séquences géniques a permis le développement de puces oligonucléotides composées de tout ou partie des gènes prédits ou connus d'un génome.

Toutes ces nouvelles ressources ouvrent des perspectives dans la compréhension des mécanismes de régulation transcriptionnelle, des réseaux de gènes ou bien des chemins métaboliques déterminant l'établissement d'un phénotype. La «génétique génomique», qui combine les données d'expression avec les données de polymorphisme génétique et qui recouvre d'autres approches que l'approche eQTL proprement dite, ouvre des perspectives intéressantes dans le cadre de l'identification des gènes responsables de la variabilité d'un caractère complexe. Il ne faut cependant pas perdre de vue ses limites, d'ordres économiques, techniques ou méthodologiques : *i)* La technologie reste onéreuse dès lors que des centaines d'animaux sont à analyser. *ii)* L'expression des gènes est mesurée dans un tissu précis à un temps précis, les résultats et les conclusions sont donc difficilement transposables et doivent ainsi être interprétés en conséquence. *iii)* La plupart des méthodes actuelles ne peuvent pas traiter les modèles non additifs, épistasiques ou résultant d'autres effets complexes sur la variabilité d'expression des gènes, de sorte que le nombre de gènes présentant des régions eQTL est probablement sous-estimé. *iv)* La plupart des analyses souffrent de puissance statistique limitée. Sans puissance statistique suffisante, les études se limitent aux gènes présentant de fortes variabilités d'expression. *v)* A l'heure actuelle, la proportion de gènes ayant des fonctions totalement inconnues ou partiellement connues est encore importante.

En revanche, la réduction des coûts des puces à ADN (puces qui seront remplacées dans un avenir plus ou moins proche par du séquençage haut débit), couplée au développement de puces à SNP à haute densité permettra d'améliorer dans un futur proche la puissance et la précision des analyses eQTL. Par ailleurs, les puces à SNP à très haut débit couplées à la possibilité maintenant de re-séquencer plusieurs individus devraient permettre de localiser très finement les régions QTL et d'en identifier tous les polymorphismes que l'on sait nombreux. Aussi, apporter de l'information fonctionnelle sur l'impact de la mutation causale dans une région QTL d'intérêt grâce aux approches de génétique génomique peut être un élément déterminant pour identifier cette mutation parmi les différents polymorphismes observés dans une région.

Références

Alberts R., Terpstra P., Bystrykh L.V., de Haan G., Jansen R.C., 2005. A statistical multiprobe model for analyzing *cis* and *trans* genes in genetical genomics experiments with short-oligonucleotide arrays. Genetics, 171, 1437-1439.

Anholt R.R., Dilda C.L., Chang S., Fanara J.J., Kulkarni N.H., Ganguly I., Rollmann S.M., Kamdar K.P., Mackay T.F., 2003. The genetic architecture of odor-guided behavior in Drosophila: epistasis and the transcriptome. Nat. Genet., 35, 180-184.

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25, 25-29.

Bing N., Hoeschele I., 2005. Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics, 170, 533-542.

Blum Y., Le Mignon G., Lagarrigue S., Causeur D., 2010. A factor model to analyze heterogeneity in gene expression. BMC Bioinformat., 11, 368.

Brem R.B., Kruglyak L., 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci., 102, 1572-1577.

Brem R.B., Yvert G., Clinton R., Kruglyak L., 2002. Genetic dissection of transcriptional regulation in budding yeast. Science, 296, 752-755.

Bystrykh L., Weersing E., Dontje B., Sutton S., Pletcher M.T., Wiltshire T., Su A.I., Vellenga E., Wang J., Manly K.F., 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using «genetical genomics». Nat. Genet., 37, 225-232.

Carbon S., Ireland A., Mungall C.J., Shu S., Marshall B., Lewis S., 2009. AmiGO: online access to ontology and annotation data. Bioinformatics, 25, 288-289.

Carlborg O., De Koning D.J., Manly K.F., Chesler E., Williams R.W., Haley C.S., 2005. Methodological aspects of the genetic dissection of gene expression. Bioinformatics, 21, 2383-2393.

Chesler E.J., Lu L., Shou S., Qu Y., Gu J., Wang J., Hsu H.C., Mountz J.D., Baldwin N.E., Langston M.A., 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat. Genet., 37, 233-242.

Cheung V.G., Spielman R.S., Ewens K.G., Weber T.M., Morley M., Burdick J.T., 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature, 437, 1365-1369.

Cotsapas C.J., Williams R.B., Pulvers J.N., Nott D.J., Chan E.K., Cowley M.J., Little P.F., 2006. Genetic dissection of gene regulation in multiple mouse tissues. Mamm. Genome, 17, 490-495.

De Vienne D., Leonardi A., Damerval C., Zivy M., 1999. Genetics of proteome variation for QTL characterization: application to drought-stress responses in maize. J. Exp. Bot., 50, 303-309.

DeCook R., Lall S., Nettleton D., Howell S.H., 2006. Genetic regulation of gene expression during shoot development in *Arabidopsis*. Genetics, 172, 1155-1164.

Doss S., Schadt E.E., Drake T.A., Lusis A.J., 2005. *Cis*-acting expression Quantitative Trait Loci in mice. Genome Res., 15, 681-691.

Drake T.A., Schadt E., Hannani K., Kabo J.M., Krass K., Colinayo V., Greaser L.E., Goldin J., Lusis A.J., 2001. Genetic loci determining bone density in mice with diet-induced atherosclerosis. Physiol. Genomics, 5, 205-215.

Drake T.A., Schadt E.E., Lusis A.J., 2006. Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. Mamm. Genome, 17, 466-479.

Eden E., Navon R., Steinfeld I., Lipson D., Yakhini Z., 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10, 48.

Efron B., Tibshirani R., 2007. On testing the significance of sets of genes. Ann. Appl. Statist., 1, 107-129.

Emilsson V., Thorleifsson G., Zhang B., Leonardson A.S., Zink F., Zhu J., Carlson S., Helgason A., Walters G.B., Gunnarsdottir S., 2008. Genetics of gene expression and its effect on disease. Nature, 452, 423-428.

Falcon S., Gentleman R., 2007. Using GOstats to test gene lists for GO term association. Bioinformatics, 23, 257-258.

Farrall M., 2004. Quantitative genetic variation: a post-modern view. Hum. Mol. Genet., 13, Spécial Issue, 1, R1-R7.

Ferrara C.T., Wang P., Neto E.C., Stevens R.D., Bain J.R., Wenner B.R., Ilkayeva O.R., Keller M.P., Blasiole D.A., Kendziorski C., 2008. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. PLoS Genet., 4, e1000034.

Georges M., 2007. Mapping, fine mapping, and molecular dissection of Quantitative Trait Loci in domestic animals. Ann. Rev. Genomics Hum. Genet., 8, 131-162.

Ghazalpour A., Doss S., Sheth S.S., Ingram-Drake L.A., Schadt E.E., Lusis A.J., Drake T.A., 2005. Genomic analysis of metabolic pathway gene expression in mice. Genome Biol., 6, R59.

Gibson G., Weir B., 2005. The quantitative genetics of transcription. Trends Genet., 21, 616-623.

Gilbert H., Le Roy P., 2003. Comparison of three multitrait methods for QTL detection. Genet. Sel. Evol., 35, 281-304.

GuhaThakurta D., Xie T., Anand M., Edwards S.W., Li G., Wang S.S., Schadt E.E., 2006. *Cis*-regulatory variations: a study of SNPs around genes showing cis-linkage in segregating mouse populations. BMC Genomics, 7, 235.

Hocquette J.F., Cassar-Malek I., Scalbert A., Guillou F., 2009. Contribution of genomics to the understanding of physiological functions. J. Physiol. Pharmacol., 60, numéro spécial, Suppl., 3, 5-16.

Hubner N., Wallace C.A., Zimdahl H., Petretto E., Schulz H., Maciver F., Mueller M., Hummel O., Monti J., Zidek V., 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet., 37, 243-253.

Jacob F., Monod J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol., 3, 318-356.

Jansen R.C., Nap J.P., 2001. Genetical genomics: the added value from segregation. Trends Genet., 17, 388-391.

Jiang Z., Gentleman R., 2007. Extensions to gene set enrichment. Bioinformatics, 23, 306-313.

Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., Hirakawa M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res., 34, D354-D357.

Keller A., Backes C., Lenhof H.P., 2007. Computation of significance scores of unweighted gene set enrichment analyses. BMC Bioinformatics, 8, 290.

Kim S.Y., Volsky D.J., 2005. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics, 6, 144.

Kirst, M., Myburg A.A., De Leon J.P., Kirst M.E., Scott J., Sederoff R., 2004. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiol., 135, 2368-2378.

Kruglyak L., Storey J.D., 2009. Cause and express. Nat. Biotechnol., 27, 544 545.

Lan H., Rabaglia M.E., Schueler K.L., Mata C., Yandell B.S., Attie A.D., 2004. Distinguishing covariation from causation in diabetes: a lesson from the protein disulfide isomerase mRNA abundance trait. Diabetes, 53, 240-244.

Le Mignon G., Desert C., Pitel F., Leroux S., Demeure O., Guernec G., Abasht B., Douaire M., Le Roy P., Lagarrigue S., 2009. Using transcriptome profiling to characterize QTL regions on chicken chromosome 5. BMC Genomics, 10, 575.

Lum P.Y., Chen Y., Zhu J., Lamb J., Melmed S., Wang S., Drake T.A., Lusis A.J., Schadt E.E., 2006. Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. J. Neurochem., 97 Suppl 1, 50-62.

Masuzaki H., Flier J.S., 2003. Tissue-specific glucocorticoid reactivating enzyme, 11 beta-hydroxysteroid dehydrogenase type 1 (11 beta-HSD1)-a promising drug target for the treatment of metabolic syndrome. Curr. Drug Targets Immune Endocr. Metabol. Disord., 3, 255-262.

Monks S.A., Leonardson A., Zhu H., Cundiff P., Pietrusiak P., Edwards S., Phillips J.W., Sachs A., Schadt E.E., 2004. Genetic inheritance of gene expression in human cell lines. Am. J. Hum. Genet,75, 1094-1105.

Mootha V.K., Lindgren C.M., Eriksson K.F., Subramanian A., Sihag S., Lehar J., Puigserver P., Carlsson E., Ridderstrale M., Laurila E., 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately

downregulated in human diabetes. Nat. Genet., 34, 267-273.

Morley M., Molony C.M., Weber T.M., Devlin J.L., Ewens K.G., Spielman R.S., Cheung V.G., 2004. Genetic analysis of genome-wide variation in human gene expression. Nature, 430, 743-747.

Nadaf J., Pitel F., Gilbert H., Duclos M.J., Vignoles F., Beaumont C., Vignal A., Porter T.E., Cogburn L.A., Aggrey S.E., Simon J., Le Bihan-Duval E., 2009. QTL for several metabolic traits map to loci controlling growth and body composition in an F2 intercross between high- and low-growth chicken lines. Physiol. Genomics, 38, 241-249.

Pfeifer D., Kist R., Dewar K., Devon K., Lander E.S., Birren B., Korniszewski L., Back E., Scherer G., 1999. Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to SOX9: evidence for an extended control region. Am. J. Hum. Genet., 65, 111-124.

Ponsuksili S., Jonas E., Murani E., Phatsara C., Srikanchai T., Walz C., Schwerin M., Schellander K., Wimmers K., 2008. Trait correlated expression combined with expression QTL analysis reveals biological pathways and candidate genes affecting water holding capacity of muscle. BMC Genomics, 9, 367.

Potokina E., Druka A., Luo Z., Wise R., Waugh R., Kearsey M., 2008. Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J., 53, 90-101.

Rockman M.V., Kruglyak L., 2006. Genetics of global gene expression. Nat. Rev. Genet., 7, 862-872.

Schadt E.E., Monks S. A., Drake T.A., Lusis A.J., Che N., Colinayo V., Ruff T.G., Milligan S.B., Lamb J.R., Cavet G., 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature, 422, 297-302.

Schadt E.E., Lamb J., Yang X., Zhu J., Edwards S., Guhathakurta D., Sieberts S.K., Monks S., Reitman M., Zhang C., 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet., 37, 710-717.

Storey J.D., Akey J.M., Kruglyak L., 2005. Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol., 3, e267.

Stranger B.E., Forrest M.S., Clark A.G., Minichiello M.J., Deutsch S., Lyle R., Hunt S., Kahl B., Antonarakis S.E., Tavare S., 2005. Genome-wide associations of gene expression variation in humans. PLoS Genet., 1, e78.

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., Mesirov J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA, 102, 15545-15550.

Subramanian A., Kuehn H., Gould J., Tamayo P., Mesirov J.P., 2007. GSEA-P: a desktop application for gene set enrichment analysis. Bioinformatics, 23, 3251-3253.

Wayne M.L., McIntyre L.M., 2002. Combining mapping and arraying: An approach to candidate gene identification. Proc. Natl. Acad. Sci. USA, 99, 14903-14906.

Williams R.B., Chan E.K., Cowley M.J., Little P.F., 2007. The influence of genetic variation on gene expression. Genome Res., 17, 1707-1716.

Yaguchi H., Togawa K., Moritani M., Itakura M., 2005. Identification of candidate genes in the type 2 diabetes modifier locus using expression QTL. Genomics, 85, 591-599.

Yamashita S., Wakazono K., Nomoto T., Tsujino Y., Kuramoto T., Ushijima T., 2005. Expression Quantitative Trait Loci analysis of 13 genes in the rat prostate. Genetics, 171, 1231-1238.

Yvert G., Brem R.B., Whittle J., Akey J.M., Foss E., Smith E.N., Mackelprang R., Kruglyak L., 2003. Transacting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat. Genet., 35, 57-64.

Zhang B., Schmoyer D., Kirov S., Snoddy J., 2004. GOTree Machine GOTM: a web-based platform for interpreting sets of inter-res ting genes using Gene Ontology hierarchies. BMC Bioinformatics, 5, 16.

Zhou X., Su Z., 2007. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics, 8, 246.

Zivy M., de Vienne D., 2000. Proteomics: a link between genomics, genetics and physiology. Plant Mol. Biol., 44, 575-580.

Résumé

De nombreux progrès ont été réalisés ces dernières années en génomique. Le développement de technologies à base de supports miniaturisés, permet aujourd'hui d'explorer les génomes tant au niveau de leur structure que de leur expression. Les puces à ADN permettent ainsi de génotyper plusieurs milliers de marqueurs SNP d'un génome ou encore de mesurer le niveau d'expression de plusieurs milliers de gènes d'un tissu. Combiner l'information génotypique avec des mesures phénotypiques élémentaires (ARNm, protéines ou encore métabolites) ouvre de nouvelles perspectives dans l'étude du fonctionnement du vivant et a donné naissance à un nouveau concept, la «génétique génomique». Cet article est centré sur les apports de la «génétique génomique» dans le contexte de la détection de QTL (*Quantitative Trait Locus*). Après avoir défini la notion de QTL d'expression (eQTL), cet article propose dans un premier temps un bilan des différents programmes de cartographie de QTL d'expression décrits dans la littérature. Sont ensuite présentés les différentes méthodes utilisant des données d'expression pour préciser ou caractériser fonctionnellement des régions QTL responsables de la variation de caractères d'intérêt avec des exemples concernant les animaux d'élevage.

Abstract

Contribution of Functional Genomics to the Fine Mapping of QTL

Much progress has been made in recent years in the genomics field. The development of technologies -based on miniaturized arrays makes it possible to explore genomes at both structural and functional levels. DNA microarrays allow to genotype several thousands of SNP in a genome, or measure the expression level of several thousands of genes in a tissue. Strategies combining genotypic information with elementary phenotypes (mRNA, proteins or metabolites) open new perspectives in biology research and are grouped under the new concept of «genetical genomics». We will focus here on the contributions of the «genetical genomics» in the context of QTL detection. Firstly, after defining the concept of expression QTL (eQTL), this article reports the main results on expression QTL mapping reported in the literature. Then, the different methods using expression data to refine or functionally characterize a QTL region are presented and illustrated through some examples on model and livestock species.

LE MIGNON G., BLUM Y., DEMEURE O., DIOT C., LE BIHAN-DUVAL E., LE ROY P., LAGARRIGUE S., 2010. Apports de la génomique fonctionnelle à la cartographie fine de QTL. Inra Prod. Anim., 23, 343-358.

Chapitre 1

Prise en compte de l'hétérogénéité d'expression dans les données transcriptomiques pour l'analyse génétique d'un caractère complexe

1.1 Introduction

Dans ce chapitre, nous proposons différentes approches pour prendre en compte l'hétérogénéité d'expression dans les données transcriptomiques de sorte à faciliter l'identification de gènes ou régions du génome impliqués dans la variabilité d'un caractère d'intérêt.
Dans une première partie, on s'intéresse à l'hétérogénéité d'expression due à des facteurs connus ou inconnus de type expérimentaux ou environnementaux, qui introduisent une variabilité supplémentaire dans les données indépendamment du caractère. Comme montré par Leek and Storey (2007) et Friguet et al. (2009), cette hétérogénéité peut nuire aux analyses différentielles. De même, Listgarten et al. (2010) suggèrent sa prise en compte pour les analyses eQTL. On propose d'utiliser la méthode FAMT récemment introduite par Friguet et al. (2009) afin d'extraire la variabilité indépendante du caractère d'intérêt à la fois pour les analyses différentielles et pour les analyses eQTL.
Dans une deuxième partie, on prend en compte un autre type d'hétérogénéité dans les données transcriptomiques. Cette fois-ci, grâce à la correction des données transcriptomiques pour l'hétérogénéité du signal, on va pouvoir s'intéresser à l'hétérogénéité des profils d'expression due à la complexité du caractère. En effet, il existe différentes voies biologiques menant à un même phénotype mesuré à l'échelle de l'animal (par exemple, différents sous-types d'animaux gras). Autrement dit, des mutations différentes peuvent conduire à un même phénotype et avoir un impact différent sur le transcriptome, créant ainsi des profils d'expression hétérogènes pour des animaux ayant un même phénotype. Prendre en compte cette hétérogénéité, c'est-à-dire ces sous-types phénotypiques pour le caractère, peut permettre de distinguer les différentes mutations et donc de faire

apparaître de nouveaux QTL ou encore de préciser la localisation de QTL déjà détectés (Schadt et al. (2003)).

1.2 Prise en compte de l'hétérogénéité du signal dans les données transcriptomiques à l'aide d'un modèle à facteurs

Cette partie correspond aux 2 articles :

- **Blum Y**, Le Mignon G, Lagarrigue S, and Causeur D. A factor model to analyze heterogeneity in gene expression. BMC Bioinformatics, 2010, 11:368.

- Mach N, **Blum Y**, Bannink A, Causeur D, Houée M, Lagarrigue S, Smith M.A. Pleiotropic effects of polymorphism of the gene diacylglycerol-o-transferase 1 (DGAT1) in the mammary gland tissue of dairy cows. Journal of Dairy Science, 2012, 95:1-12

1.2.1 Article 2 : Blum *et al.*, BMC Bioinformatics, 2010

Apport de l'article

Les méthodes classiques dédiées à l'analyse différentielle ne prennent pas en compte l'hétérogénéité du signal dans les données transcriptomiques qui peut être causée par des facteurs connus ou inconnus (Leek and Storey (2007)). Par exemple, des gènes peuvent faire partie d'un même processus biologique indépendamment du caractère d'intérêt et ainsi introduire une variabilité supplémentaire dans les données. De même, un facteur expérimental indépendant du caractère d'intérêt peut avoir un effet sur l'ensemble ou sur une partie des expressions géniques et empêcher la détection de certaines corrélations avec le caractère ou bien plus rarement, créer une corrélation artificielle avec le caractère. On se propose d'utiliser dans cette étude une méthode récemment introduite par Friguet et al. (2009) qui permet de capturer cette variabilité d'expression indépendante du caractère d'intérêt par un modèle d'analyse en facteurs. Dans ce modèle les facteurs capturent les composantes d'hétérogénéité. Friguet et al. (2009) montrent que leur méthode appelée FAMT pour Factor Analysis for Multiple Testing, permet d'augmenter la puissance des tests statistiques et de diminuer la variabilité des taux d'erreurs dans les analyses différentielles. On s'appuie sur cette méthode pour "nettoyer" le jeu de données de la variabilité indépendante du caractère d'intérêt dans le cadre de l'analyse différentielle et des analyses eQTL. La stratégie proposée est appliquée sur un jeu de données d'expression hépatique concernant une famille de 45 poulets provenant d'un croisement de deux lignées divergentes pour le gras abdominal (Le Mignon et al. (2009)). On montre alors la pertinence biologique de travailler sur des données ajustées par les facteurs d'hétérogénéité. En effet, de nouveaux gènes liés au caractère d'intérêt ou associés à une région eQTL/QTL sont détectés à l'aide des données ajustées, alors qu'ils étaient auparavant masqués par l'hétérogénéité du signal. De plus on interprète les facteurs d'hétérogénéité à l'aide de l'information extérieure disponible concernant le dispositif expérimental.

Blum *et al. BMC Bioinformatics* 2010, **11**:368
http://www.biomedcentral.com/1471-2105/11/368

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A factor model to analyze heterogeneity in gene expression

Yuna Blum*[1,2,3], Guillaume Le Mignon[1,2], Sandrine Lagarrigue[1,2] and David Causeur[3]

Abstract

Background: Microarray technology allows the simultaneous analysis of thousands of genes within a single experiment. Significance analyses of transcriptomic data ignore the gene dependence structure. This leads to correlation among test statistics which affects a strong control of the false discovery proportion. A recent method called FAMT allows capturing the gene dependence into factors in order to improve high-dimensional multiple testing procedures. In the subsequent analyses aiming at a functional characterization of the differentially expressed genes, our study shows how these factors can be used both to identify the components of expression heterogeneity and to give more insight into the underlying biological processes.

Results: The use of factors to characterize simple patterns of heterogeneity is first demonstrated on illustrative gene expression data sets. An expression data set primarily generated to map QTL for fatness in chickens is then analyzed. Contrarily to the analysis based on the raw data, a relevant functional information about a QTL region is revealed by factor-adjustment of the gene expressions. Additionally, the interpretation of the independent factors regarding known information about both experimental design and genes shows that some factors may have different and complex origins.

Conclusions: As biological information and technological biases are identified in what was before simply considered as statistical noise, analyzing heterogeneity in gene expression yields a new point of view on transcriptomic data.

Background

Microarray technology allows the analysis of expression levels for thousands of genes simultaneously and is a powerful tool to characterize mRNA level variation due to measured variables of interest (various phenotypes, treatments...). Typical approaches to find significant relationships between gene expressions and experimental conditions ignore the correlations among expression profiles and functional categories [1]. This dependence structure leads to correlation among test statistics which affects a strong control of the actual proportion of false discoveries [2]. Indeed, a number of unmeasured or unmodeled factors independent of the variables of interest may influence the expression of any particular gene [3,4]. These factors may induce extra variability in the expression levels and decrease the power to detect links with the variables of interest.

Recently, several works have introduced models for the common information shared by all the genes. Especially Friguet *et al* [4] propose to model this sharing of information by a factor analysis structure in a method called Factor Analysis for Multiple Testing (FAMT). The estimated factors in the model capture components of the expression heterogeneity. As well, Storey *et al* [3] introduce Surrogate Variable Analysis (SVA) to identify and estimate these extra sources of variation. The factors in FAMT and the surrogate variables in SVA are similarly designed to model dependence among tests by a linear kernel but they are estimated differently. Contrarily to the SVA model, independence between the factors and the experimental conditions of interest is explicitly assumed in FAMT in order to separate clearly the effects of the experimental conditions on the gene expressions and the nuisance variability due to unmodeled technological effects and other known or unknown effects that could be uncontrolled in the experimental design.

The major sources of expression variation are then assumed to be the experimental conditions of interest,

* Correspondence: Yuna.blum@rennes.inra.fr
[1] Agrocampus Ouest, UMR598, Animal Genetics, 35000 Rennes, France
Full list of author information is available at the end of the article

but also gene dependence and uncontrolled factors in the experimental design. Indeed, even after normalization, variation due to the experimental design still exists in expression data. The factors extracted in the residual part of the regression models explaining the gene expressions by the experimental conditions of interest are therefore analyzed to give more insight both on expression heterogeneity among sampling units and the contribution of some biological processes to gene dependence. First, factors are extracted from illustrative expression data sets with simple patterns of expression heterogeneity in order to show how they can straightforward be related to sources of heterogeneity. Henceforth, the same factor model approach is used to analyze an expression data set initially generated to map quantitative trait loci (QTL) for abdominal fatness (AF) in chickens, especially on chromosome 5 (GGA5) [5]. This data set concerns hepatic transcriptome profiles for 11213 genes of 45 half sib male chickens generated from a same sire. This sire was generated by successive inter-crossing of two experimental chicken lines divergently selected on AF and was known to be heterozygous for an AF QTL on the GGA5 chromosome around 175 cM (For more details, see [5]). The 45 half sib chickens show therefore variation on AF. According to the polygenic effect model of quantitative traits, this variation is probably due to multiple mutations and biological processes.

Two lists of genes significantly correlated to the AF trait are first generated using the raw and the factor-adjusted expression dataset. Then, the relevance of the two gene lists to characterize functionally fatness variation in the family are compared, regarding the frequencies of biological processes related to the AF trait in their functional annotations. Factor-adjusted expression data is finally used to identify a gene whose expression is controlled by the AF QTL region.

Furthermore, the extracted factors are interpreted using external information on the experimental design such as the hatch, dam and body weight and also gene information such as functional categories, oligonucleotide size and location on the microarray. It is deduced that some factors may have different and complex origins, which confirms the importance of taking into account these extra sources of variability to be more relevant in the transcriptomic analyses.

Results

Illustrative Examples

Similarly to Storey *et al* (2007) [3], three simple situations of heterogeneity are considered. For each one, independent expressions for 1000 genes on 20 arrays are simulated according to a standard normal distribution. The sample is split into two equal groups and a constant is added on the first 100 gene expressions to mimic a differential expression between these two groups.

Case 1: One independent variable affecting all genes

All genes are affected by an independent grouping variable marked by colors red and green on Figure 1A. A single factor is extracted by FAMT. Figure 2A helps interpreting this factor and shows that it clearly discriminates the two colored groups of individuals (P-value $\leq 2.2 \times 10^{-16}$). This shows a high association between the factor and the independent grouping variable. The genes representation does not show any particular structure. In this simple case the factor estimated by FAMT can therefore be easily interpreted through the individuals representation.

Case 2: One independent variable affecting a set of genes

Only genes 70-170 are affected by an independent grouping variable marked by colors red and green on Figure 1B. A single factor is also found using FAMT. As shown on Figure 2B, the factor discriminates the two groups of individuals (P-value $\leq 2.2 \times 10^{-16}$) and the two groups of genes (P-value $\leq 2.2 \times 10^{-16}$). In this case, the estimated factor can be interpreted through the individuals and genes representations.

Case 3: Two independent variables affecting two different sets of genes

Gene sets 70-170 and 171-271 are each affected by an independent grouping variable marked respectively by colors red and green and by colors orange and blue as illustrated by Figure 1C. Two factors are identified by FAMT which are now interpreted regarding the two external sources of heterogeneity (Figure 2C). The red-green variable seems to be highly associated with the first axis (P-value $\leq 2.2 \times 10^{-16}$ in both representation). On the contrary, the orange-blue variable is not associated with this axis considering a significance level of 0.05 (P-value = 0.7933 for the individuals representation, p-value = 0.1109 for the genes representation). The same strategy is implemented for the second factor. The red-green variable appears to be not associated with this factor (P-value = 0.7949 for the individuals representation, p-value = 0.1926 for the genes representation) whereas the orange-blue variable is highly associated (P-value $\leq 2.2 \times 10^{-16}$ in both representations). In this case, each of the two estimated factors can be explained by one of the two independent grouping variables.

Analysis of the AF expression data set

Classical approach

Examination of the Pearson coefficient correlation between hepatic transcript levels and AF trait shows that 287 genes are significantly correlated considering a significance threshold of 0.05 without any correction for multiple tests. This low amount of differentially expressed genes might be explained by a poor genetic variability

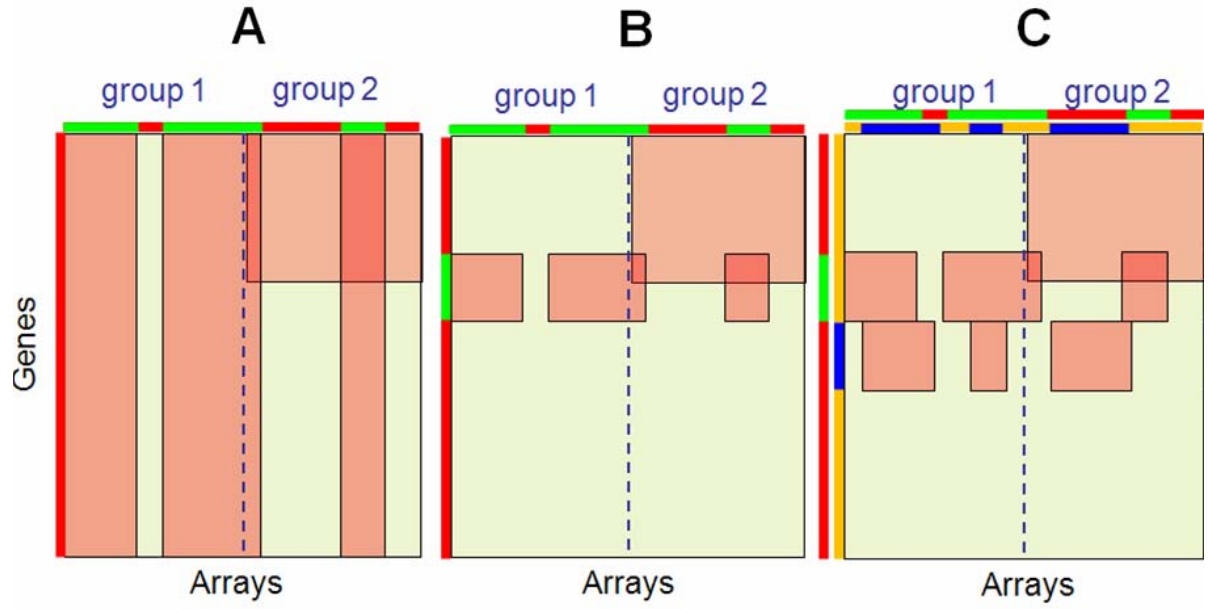


Figure 1 Structure of the illustrative data sets. Representation of three illustrative studies consisting of 1000 genes on 20 arrays divided between two groups. (A) Case 1: one independent grouping variable with red and green levels affecting all genes. (B) Case 2: one independent grouping variable with red and green levels affecting a gene set. (C) Case 3: two independent grouping variables with red and green levels and with blue and orange levels affecting each a different gene set.

between individuals which are half sib offsprings and could also be due to dependence between genes that can lead to under representation of the smallest p-values [6].

Heterogeneity analysis

Minimizing the variance inflation criterion proposed by Friguet *et al.* [4], six factors containing a common information shared by all genes and independent from the AF trait are extracted. Subtracting the linear dependence kernel defined by these factors from the raw expression data yields the factor-adjusted expression data. The significance analysis based on these expressions results in a list of 688 gene expressions significantly correlated to the AF trait. 93% of the 287 genes found with the classical approach are included in this list. This larger number of differentially expressed genes suggests that correlation between many gene expressions and the variable of interest is under estimated due to gene dependence. Considering the Gene Ontology (GO) terms and KEGG pathways, one enriched term related to the lipid metabolism is found in the gene list resulting from factor-adjustment (688 genes) whereas none is observed in the gene list obtained using the raw expressions (287 genes). This term concerns "Steroid biosynthesis process" with 3 genes associated (Table 1). More precisely, these genes are involved in the cholesterol metabolism or in conversion of cholesterol in steroids. Several works show relationships between cholesterol metabolism and obesity [7-9]. This result shows that the genes found after factor-adjustment are more related to the fatness trait. Furthermore, the impact of factor-adjustment is shown in Figure 3, where a principal component analysis (PCA) generated with the 688 factor-adjusted transcript levels of correlated genes (Figure 3B) separates much more fat and lean chickens than the same PCA generated with the raw expressions of the same 688 genes (Figure 3A). This observation displays that factor-adjustment has cleaned up the data from dependence, which highlights masked relationships with the AF trait.

We focus on one of the 3 genes involved in the "Steroid biosynthesis process", DHCR7, which is only observed in the list of 688 genes and known for encoding the last enzyme involved in the cholesterol synthesis. As shown in Figure 4, the analysis of the factor-adjusted expressions for this gene highlights an eQTL (P-value < 0.05) colocalizing with the AF QTL previously observed [5]. The same LRT curve based on the raw expressions does not point out any eQTL. This result shows that the expression of this gene is controlled by a mutation in the same GGA5 AF QTL region. Because of the function of this gene related to lipid metabolism, this result suggests that this mutation could be the same as the QTL mutation for fatness phenotype. Further investigations are necessary to refine these QTL and eQTL locations.

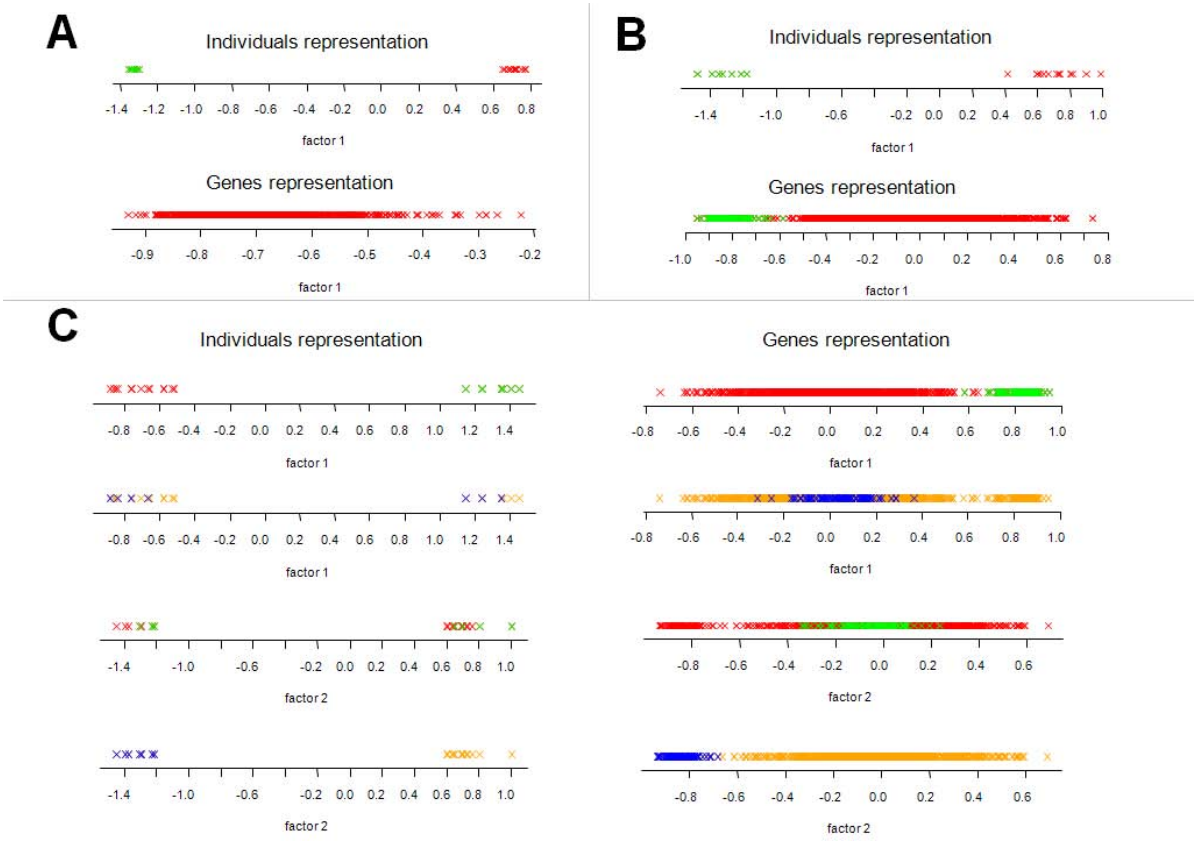


Figure 2 Illustrative data sets: Individuals and genes representations. Individuals and genes representation using respectively the Z matrix corresponding to the factors and the B matrix of the loadings found by FAMT. Individuals and genes are colored according to the independent variable they are affected by. (A) representations corresponding to case 1, (B) case 2 and (C) case 3.

Factor interpretation

In the present study, some external information about the experimental design and the genes is available. As we did in the simulated examples, we interpret the factors extracted from the AF expression dataset using this known information.

Using information on experimental design

The hatch, the dam and the body weight were previously measured for each bird and should be independent of the AF variation. For the body weight, the founder chicken lines were selected on AF criteria maintaining a constant body weight. The variables "hatch" and "dam" are both categorical with respectively, four and eight levels and the body weight is a continuous variable. We first focus on the "hatch" and for each factor we represent the individuals colored according to their hatch (Figure 5). Factor 1 seems to discriminate hatches 1 and 4, factor 3 hatch 2 from the others and factor 4 hatches 2 and 3 from hatch 1. The effect of the hatch on each factor is tested and the results given in Table 2 confirm our previous observations: factor 1, 3 and 4 can be partly explained by a hatch effect (the significant test for each hatch level is given in Additional file 1). We then calculate the association for the "dam" and "body weight" with each of the six factors. Table 2 shows no effect of the dam and a high correlation between the weight and factor 2. Contrarily to the illustrative cases where each factor could be interpreted by a unique variable, the factors found here seem to have more complex origins. Indeed, three of the six factors can be interpreted by an hatch effect and another one by a body weight effect. The same analysis is now performed after adjustment of the raw expression data for hatch and body weight. Interestingly, only five factors independent of the AF trait are extracted and still a hatch effect exists but only on the first factor and a weight effect on the second factor (Table 2). This persistence of both effects suggests that there exists an interaction involving hatch and body weight with other unmeasured and/or unknown variables. Therefore, taking into account the hatch and body weight in the statistical model seems to be not sufficient to remove a consequent part of the heterogeneity in gene expression.

Table 1: Enrichment tests for the list of 287 genes and 688 genes

LIST OF 287 GENES					
GOID	**GO Term**	**Size**	**Count**	**Pvalue**	**HGNC ID**
GO.0006470	*protein amino acid dephosphorylation*	56	5	0.015	ACP1, PTPN14, PTPRE, PTP4A3, PTPN6
GO.0006725	*cellular aromatic compound metabolic process*	38	4	0.017	PPME1, GART, MOCS1, ALDH6A1
GO.0007259	*JAK STAT cascade*	9	2	0.022	SOCS1, STAMBP
GO.0043543	protein amino acid acylation	9	2	0.022	NULL, ZDHHC17
GO.0044259	multicellular macromolecule metabolic process	10	2	0.027	ACE2, SERPINH1
GO.0008033	tRNA processing	26	3	0.0296	TSEN15, FARS2, NSUN2
GO.0033002	muscle cell proliferation	11	2	0.032	NOX1, BMP10
GO.0050730	regulation of peptidyl tyrosine phosphorylation	12	2	0.038	SOCS1, EGFR
Kegg ID	Kegg pathway	Size	Count	Pvalue	HGNC ID
map04320	*Dorso ventral axis formation*	9	3	2.38E-03	EGFR, SPIRE1, ETS1
LIST OF 688 GENES					
GOID	GO Term	Size	Count	Pvalue	HGNC ID
GO.0006470	*protein amino acid dephosphorylation*	56	10	1.80E-03	ACP1, PPM1E, PTPN14, PTPRE, PTP4A3, PPM1G, PTPRU, PPP3CB, PPM1L, PTPRF
GO.0046483	heterocycle metabolic process	33	7	3.21E-03	AMBP, GART, P4HA2, HMOX2, AFMID, MTHFS, ALDH6A1
GO.0051186	cofactor metabolic process	64	10	4.97E-03	AMBP, TXNRD3, NOX1, HMOX2, AFMID, GGT7, MTHFS, MOCS1, HMGCS1, ACO2
GO.0016202	regulation of striated muscle development	15	4	0.011	MBNL3, LEF1, NRG1, BMP4
GO.0007259	*JAK STAT cascade*	9	3	0.014	SOCS1, HCLS1, STAMBP
GO.0040011	locomotion	111	13	0.017	PRKG1, EDNRB, ACE2, NOX1, EGFR, NRG1, BMP10, ARAP3, JPH3, VHL, VAX1, DAB1, LAMA2
GO.0001932	regulation of protein amino acid phosphorylation	26	5	0.019	PDGFA, SOCS1, HCLS1, EGFR, BMP4
GO.0048585	negative regulation of response to stimulus	10	3	0.020	AMBP, PPP3CB, FABP7
GO.0006534	cysteine metabolic process	4	2	0.021	CBS, CDO1
GO.0002274	myeloid leukocyte activation	11	3	0.026	IRF4, LCP2, NDRG1
GO.0006725	*cellular aromatic compound metabolic process*	38	6	0.026	PPME1, GART, AFMID, MTHFS, MOCS1, ALDH6A1
GO.0007185	transmembrane receptor tyrosine phosphatase signaling	5	2	0.033	PTPRE, PTPRF
GO.0007271	synaptic transmission cholinergic	5	2	0.033	CHRNA4, LAMA2
GO.0000097	sulfur amino acid biosynthetic process	5	2	0.033	CBS, CDO1
GO.0006700	**C21 steroid hormone biosynthetic process**	5	2	0.033	STAR, CYP17A1
GO.0006787	porphyrin catabolic process	5	2	0.033	AMBP, HMOX2
GO.0001764	neuron migration	12	3	0.033	PRKG1, VAX1, DAB1
GO.0030509	BMP signaling pathway	21	4	0.036	SOSTDC1, BMP10, MSX2, BMP4
GO.0045321	leukocyte activation	64	8	0.040	SWAP70, CHRNA4, FKBP1B, IRF4, LCP2, PPP3CB, NDRG1, SFRS17A

Table 1: Enrichment tests for the list of 287 genes and 688 genes (Continued)

GO.0006790	sulfur metabolic process	32	5	0.043	CBS, CDO1, TXNRD3, GGT7, CHST1
GO.0018193	peptidyl amino acid modification	43	6	0.045	PDGFA, SOCS1, P4HA2, HCLS1, EGFR, MAP2
GO.0008211	glucocorticoid metabolic process	6	2	0.048	STAR, CYP17A1
GO.0006769	nicotinamide metabolic process	6	2	0.048	NOX1, AFMID
GO.0030111	regulation of Wnt receptor signaling pathway	14	3	0.050	SENP2, LEF1, SENP2
Kegg ID	Kegg pathway	Size	Count	Pvalue	HGNC ID
map00630	Glyoxylate and dicarboxylate metabolism	9	4	1.87E-03	GLYCTK, HYI, AFMID, ACO2
map00140	**C21 Steroid hormone metabolism**	6	3	5.11E-03	**DHCR7**, HSD11B1, CYP17A1
map04320	*Dorso ventral axis formation*	9	3	0.018	EGFR, SPIRE1, ETS1
map04012f	ErbB signaling pathway	35	6	0.026	PIK3R5, PLCG1, PAK3, EGFR, NRG1, PTK2

Enrichment tests were performed using an R program (see Methods section) with GO BP terms and Kegg pathways. The tests were done on the list of 287 genes found using the classical approach and the list of 688 genes found by FAMT. For each enriched term, the identifier (ID), the biological term, the size of the whole list of genes related to the term (size), the number of genes in the sub-list related to the term (count), the pvalue of the test and the HGNC Hugo abbreviations of the related genes are given. Italic terms are those which are present in both lists.

Using gene information

To interpret the estimated factors in terms of gene expressions, we use known information about genes as oligonucleotide size and location on the chip: block, row and column (genes representation on each factor is given in Additional file 2). For these variables, we test their association with each factors extracted from the raw data and the hatch and body weight adjusted data. As shown in Table 2, there is a strong oligonucleotide size effect and block effect captured by almost all the factors. We exhibit also a row and column effect associated with some factors. Moreover, the genes that most contribute to the first two factors are identified (score larger than 0.8). We obtain a set of 313 genes for factor 1 and a set of 175 genes for factor 2 and which were as expected not included for 95% of them in the list of 688 genes. We perform a term enrichment test for this two sets (Table 3). As we expect, there are essentially biological terms independent of the lipid metabolism. Factor 1 is mainly characterized by genes involved in cell division metabolism and interestingly also to pigmentation. Factor 2 is more characterized by genes involved in the nucleotide metabolism. The enriched terms found are thus not implicated in the metabolim changes induced by the AF variability. We previously highlighted a hatch and body weight effects on the factors. As in PCA, individuals and genes representa-

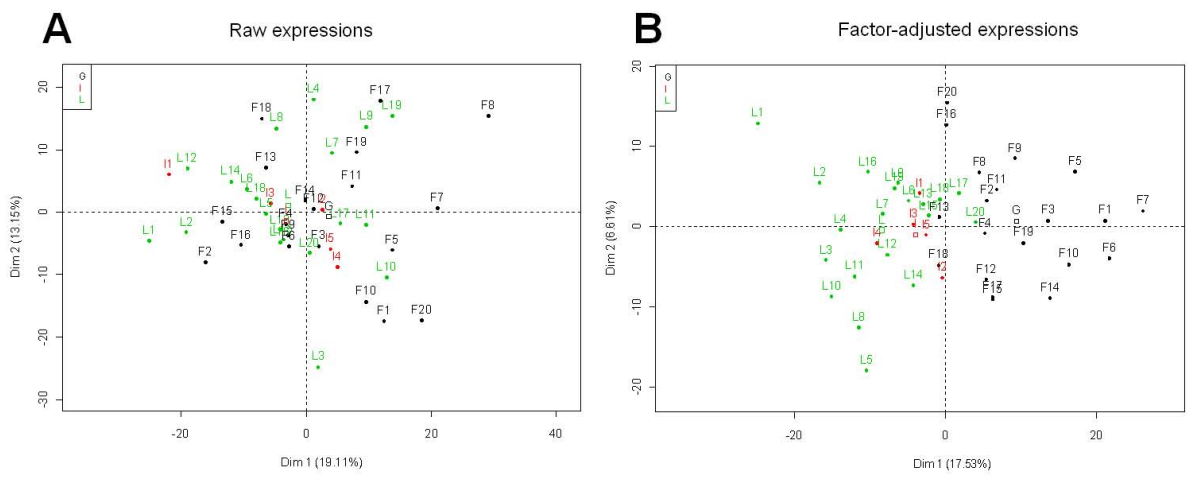


Figure 3 Principal component analysis: individuals representation. Birds are colored according to their weight split into 3 classes: lean (green), intermediate (red) and fat (black). (A) PCA generated with the raw expression of the 688 differentially expressed genes. (B) PCA generated with the factor-adjusted expression of the 688 differentially expressed genes.

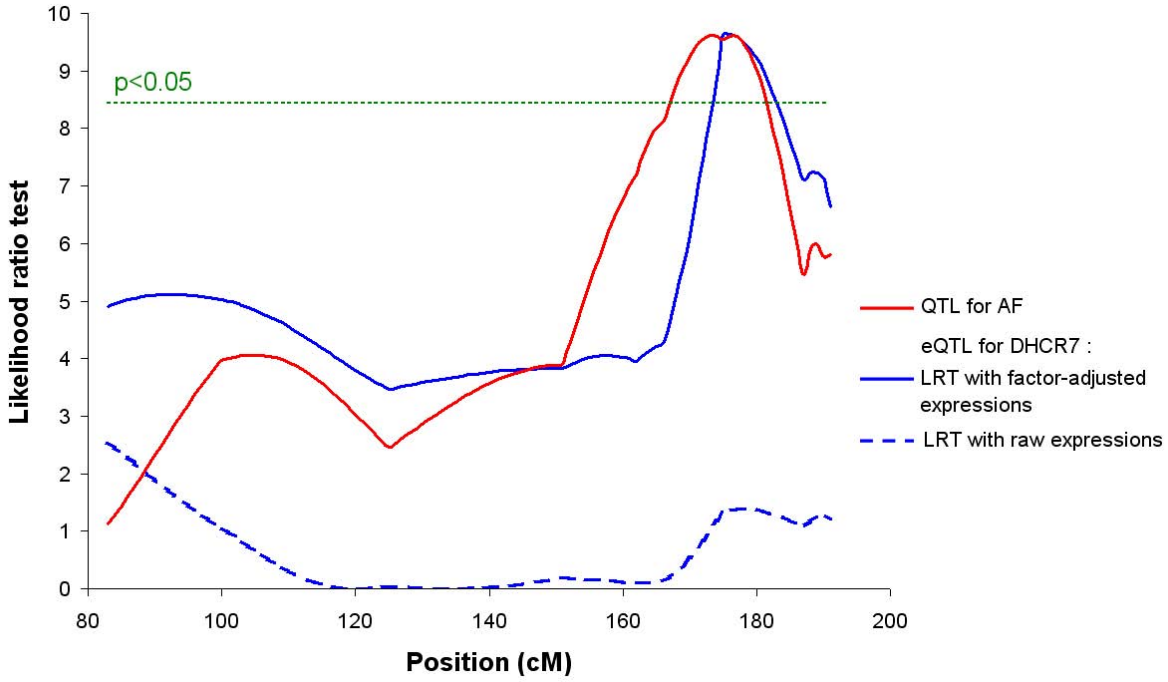


Figure 4 eQTL mapping for DHCR7 on chromosome 5. The LRT curves for the gene DHCR7 are represented in blue (plain line for the factor-adjusted analysis and dotted line for the raw analysis). The LRT curve for the AF trait is represented in red. The plain curves reveal the existence of a QTL/eQTL for the 2 traits in the same region, around 175 cM on the GGA5 chromosome. This colocalization is not revealed by the raw analysis. The Significance level of 5% is represented by the horizontal green line. The genetic distances (cM) and likelihood ratio (LR) are shown on the X-axis and Y-axis, respectively.

tions can be interpreted commonly. Hatch effect could therefore be related to the particular metabolisms characterizing factor 1.

Discussion and Conclusion

The model used in the present study assumes that the gene expressions are uncorrelated given a set of hidden variables called *factors*. In comparison to classical methods which do not take into account the dependence between genes, this approach provides a list of genes more correlated to the variable of interest. Moreover, factor-adjustment of the expression dataset turns out to give more insight to subsequent analyses such as QTL characterization. As a result, a gene is identified as correlated to the AF trait and related to the cholesterol metabolism having a trans-eQTL colocalizing with GGA5 AF QTL. Because several works show a link between cholesterol and obesity, this gene could be considered as a signature of the mutation underlying this AF QTL rather than a mutation close to it. This result provides functional hypothesis about genes whose expression could be impacted by the QTL of interest.

Factor analysis was introduced in the psychometric field in 1904 by Spearman [10] in order to extract the common factors in intelligence and personality. In this particular domain, the individuals are explained by their responses to different subsets of tests. The method usually furnished at least five factors which were interpreted as follows: neuroticism, extraversion, conscientiousness, agreeableness and openness to ideas. In our study, the factors were found using an EM algorithm presented by [4]. Our purpose was first to interpret the estimated factors and consequently to investigate which kind of information present in this factor structure could generate heterogeneity of the gene expressions. External information concerning the experimental design and functional annotations of the genes were used to analyse the factors. It is deduced that some factors seem to have a complex explanation with at least 2 variables associated to them. For factor 1, the individuals variability independent of the trait of interest is for instance shown to be related to the hatch. Enrichment tests also give a characterization of this factor by specific metabolisms.

To remove expression heterogeneity from the data for the subsequent statistical analyses, the basic idea consists in adjusting the raw expression data from the common factor structure. As we extract uncontrolled effects and technological biases from what was before simply consid-

Table 2: Description of the factors extracted from the raw data and from the hatch and weight adjusted data

	Factors extracted from the raw AF expression dataset						
	Individual information			Gene information			
	hatch	**dam**	**weight**	**oligo size**	**chip block**	**chip row**	**chip column**
Factor 1	**8.92E-05**	0.139	0.129	**2.20E-16**	**2.20E-16**	0.074	0.179
Factor 2	0.074	0.913	**4.70E-03**	**2.20E-16**	**2.20E-16**	0.041	0.857
Factor 3	**1.90E-02**	0.848	0.489	**2.55E-14**	**2.20E-16**	0.716	0.376
Factor 4	**6.00E-03**	0.127	0.959	**1.41E-07**	**2.20E-16**	0.707	0.167
Factor 5	0.435	0.217	0.884	0.529	**2.20E-16**	**4.97E-03**	**9.99E-05**
Factor 6	0.946	0.412	0.615	**1.79E-07**	**2.20E-16**	0.876	**5.11E-07**
	Factors extracted from the AF expression dataset adjusted for the hactch and body weight effects						
	Individual information			Gene information			
	hatch	**dam**	**Weight**	**oligo size**	**chip block**	**chip row**	**chip column**
Factor 1	**1.13E-04**	0.219	0.156	**2.20E-16**	**2.20E-16**	0.078	0.209
Factor 2	0.052	0.841	**3.40E-03**	**2.20E-16**	**2.20E-16**	0.036	0.814
Factor 3	0.049	0.819	0.569	**2.16E-11**	**2.20E-16**	0.554	0.16
Factor 4	0.178	0.031	0.869	**6.80E-09**	**2.20E-16**	0.897	0.885
Factor 5	0.949	0.727	0.647	**2.79E-12**	**2.20E-16**	0.291	**2.36E-10**

The p-value is given for each association test (see Methods section). Considering a threshold of 1%, the significant p-values are in bold.

ered as statistical noise, analyzing heterogeneity in gene expression yields a new point of view on transcriptomic data. We show in this study the importance of taking into account these extra sources of variation to be more relevant in the transcriptomic analyses.

Methods

AF expression data set

The data set concerns hepatic transciptome profiles for 11213 genes of 45 half sib male chickens variable for abdominal fatness (AF). The data set was generated to map quantitative trait loci (QTL) for abdominal fatness in chickens and used in a previous study [5]. The sire of this family, generated by successive inter-crossing of two experimental chicken lines divergently selected on abdominal fatness, was known to be heterozygous for an AF QTL on the GGA5 chromosome around 175 cM. Animals, marker genotyping and transcriptome data acquisition and normalization are described in Le Mignon *et al* (2009) [5].

Illustrative examples

For each case, we simulated expression for 1000 genes on 20 arrays divided in two groups using the R programming language. Initially, the expression measurements for each gene were independently drawn from a standard normal distribution. The expression heterogeneity due to simple independent grouping variables was included in the simulated data set by adding a constant value for 7 random individuals for all genes (case 1) or a set of genes (case 2 and 3).

Classical expression analysis

As the variable of interest in the biological study is continuous, we calculated the Pearson correlation coefficient for each gene expression and deduced the number of genes correlated to the trait by considering the P-values under the cutoff 0.05.

Factor-analytic method

Steps and algorithm

The method takes into account the impact of dependence on the multiple testing procedures for high-throughput

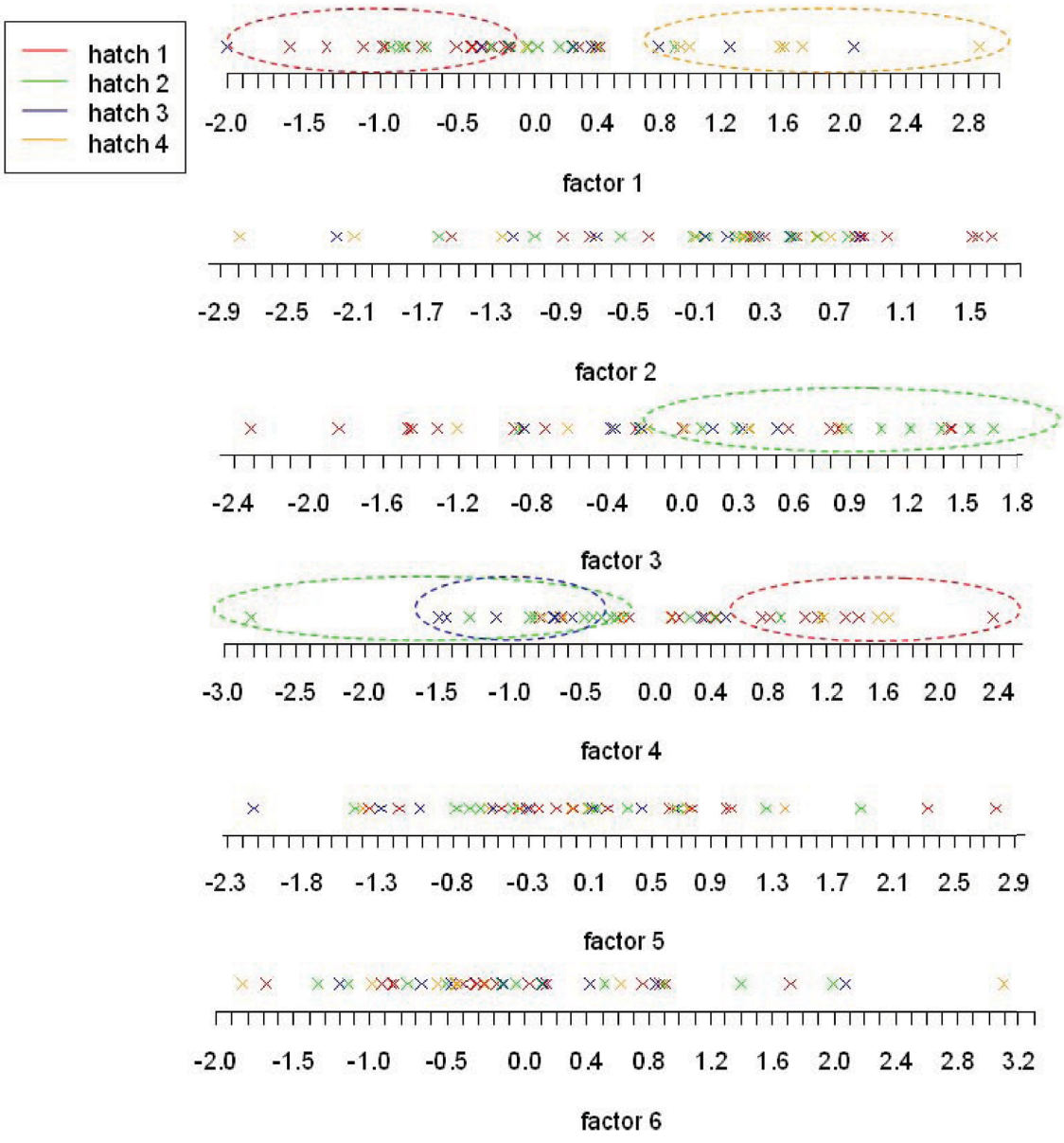


Figure 5 AF data set: individuals representation for each factor. Individuals are represented on the factor using the B matrix of loadings found by FAMT. The individuals are colored according to the variable "hatch" which has 4 levels.

data. The common information shared by all the variables (i.e. gene expressions) is modeled by a factor analysis structure. Let $Y^{(k)} = (Y^{(1)}, Y^{(2)},..., Y^{(m)})'$ be a random m-vector and $x^{(k)} = (x^{(1)},..., x^{(p)})'$ some explanatory variables. The conditional covariance matrix of the responses, given the explanatory variables, is represented by a factor analysis model: $\Sigma = \Psi + BB'$, where Ψ is a diagonal $m \times m$ of uniquenesses and B is a $m \times q$ matrix of factor loadings. In the above decomposition, the diagonal elements in Ψ are referred to as the specific variances of the responses and therefore BB' appears as the shared variance in the common factor structure. This factor analysis representation of the covariance is equivalent to the following mixed effects regression modeling of the data: for $k = 1, ..., m$

$$Y^{(k)} = \beta_0^{(k)} + \mathrm{x}'\beta^{(k)} + \mathrm{b}'_k Z + \epsilon^{(k)}$$

Table 3: Biological terms characterizing factor 1 and 2

FACTOR1				
GO ID	**GO Term**	**Size**	**Count**	**Pvalue**
GO.0007051	spindle organization	6	2	8.20E-03
GO.0050931	pigment cell differentiation	7	2	0.011
GO.0000279	M phase of meiotic cell cycle	79	6	0.012
GO.0000079	regulation of cyclin dependent protein kinase activity	9	2	0.019
GO.0016570	histone modification	13	2	0.038
GO.0015698	inorganic anion transport	53	4	0.039
GO.0007156	homophilic cell adhesion	32	3	0.041
Kegg ID	Kegg pathway	Size	Count	Pvalue
map05216	Thyroid cancer	11	2	0.020
map05130	Pathogenic Escherichia coli infection	12	2	0.024
map04520	Adherens junction	31	3	0.024
FACTOR2				
GO ID	GO Term	Size	Count	Pvalue
GO.0006195	purine nucleotide catabolic process	5	3	2.23E-05
GO.0030168	platelet activation	7	3	7.65E-05
GO.0007051	spindle organization	6	2	2.55E-03
GO.0007596	blood coagulation	24	3	3.76E-03
GO.0030336	negative regulation of cell migration	12	2	0.011
GO.0032879	regulation of localization	103	5	0.012
GO.0001775	cell activation	74	4	0.016
GO.0001890	placenta development	18	2	0.023
GO.0017038	protein import	49	3	0.027
GO.0006403	RNA localization	22	2	0.034
GO.0006816	calcium ion transport	56	3	0.038
Kegg ID	Kegg pathway	Size	Count	Pvalue
map00230	Purine metabolism	64	4	0.013

Enrichment tests were performed on the genes contributing the more to the construction of factor 1 and 2 using the GO BP terms and Kegg pathways. For each enriched term, the identifier (ID), the bioligical term, the size of the whole list of genes related to the term (size), the number of genes in the sub-list related to the term (count), the pvalue of the test.

where b_k is the kth row of B, $Z = (Z^{(1)}, ..., Z^{(q)})$ are latent factors supposed to concentrate the common information in the m-responses and $\epsilon = (\epsilon^{(1)},..., \epsilon^{(m)})'$ is a normally distributed m-vector independent of Z, with mean 0 and variance-covariance Ψ.

An EM algorithm [11] is used to estimate Ψ, B and Z. The number of factors is chosen so that the variance of the number of false discoveries is minimized. A VARIMAX rotation is finally applied on the factors after EM estimation in order to privilege highly dispersed loadings rather than a homogeneous distribution of the loadings. Once the factor model is estimated, factor-adjusted test statistics are obtained by correction of the classical tests from the effect of the common factors. [4] show that the

resulting tests statistics are asymptotically uncorrelated, which improves the overall power of the multiple testing procedure. The algorithm is implemented in the "FAMT" R package available from CRAN. For the subsequent analyses, the raw expression data set is adjusted for the estimated independent factors, which results in the so-called factor-adjusted expression data $\tilde{Y}^{(k)}$:

$$\tilde{Y}^{(k)} = Y^{(k)} - \mathrm{b}_k' Z = \beta_0^{(k)} + \mathrm{x}'\beta^{(k)} + \epsilon^{(k)}.$$

Individual and variable representation

As in PCA, the data set is transformed into a new coordinate system by an orthogonal linear transformation [12]. We can represent the individuals and variables graph through B, the matrix of factor loadings and Z, the matrix of estimated factors. Those two representations are related by a transition formula [12], which enables their simultaneous interpretation. Moreover, each factor can be related to external information which may be available in the experimental design (significance of the relationship is assessed by an analysis of variance test).

QTL and eQTL mapping

QTLMAP software based on an interval mapping method described by Elsen *et al* [13], was used to detect QTL affecting the AF trait and the eQTL affecting the expression of DHRC7. The statistical variable for testing the presence of one QTL (or eQTL) *versus* no QTL (or no eQTL) at one location was an approximate likelihood ratio test (LRT) [14]. Significance thresholds were empirically determined for AF QTL and DHCR7 eQTL from 2000 simulations. For more details, see Le Mignon *et al* (2009) [5].

Gene set enrichment

The enrichment of biological terms among a list of genes was assessed by the probability that an equally high or higher enrichment could be obtained by chance given the frequency of the biological terms among all the genes considered. We first implemented an R program which calculated the P value using a Fisher exact test for over-representation and return the enriched terms. Let $\mathcal{M}_0$ denote the subset of genes related to a given metabolism in a gene set of interest. The Fisher exact test corresponds to the hypergeometric sum as follows: $\sum_{k\in\mathcal{M}_0} P(X=k)$ where $P(X=k) = \binom{B_0}{k}\binom{B-B_0}{m-k}/\binom{B}{m}$, B . B the number of genes contained in the whole population, m the number of genes in the gene set of interest and B_0 the number of genes related to the metabolism. The functional annotations used for this program were generated as indicated in [15] are available on the website: http://www.sigenae.org. They were obtained by a bioinformatics procedure using the Ensembl annotation source [16]. The analysis were done using the Gene ontology (GO) biological processes (BP) terms [17] and the KEGG pathways [18] with a significant threshold of 0.05.

List of abbreviations

AF: Abdominal Fatness; eQTL: Expression Quantitative Trait Loci; GGA5: chromosome 5; FAMT: Factor Analysis for Multiple Testing; SVA: Surrogate Variable Analysis; HCA: Hierarchical Cluster Analysis; PCA: Principal Component Analysis; LRT: Likelihood Ratio Test; GO: Gene Ontology; BP : Biological Process; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Additional material

Additional file 1 Student test for each level of the variable "hatch". Student tests were performed for the levels of the variable "hatch" in order to test their effect on each factor. The crossed out column concern factors for which the global hatch effect were not significant using the Fisher test.

Additional file 2 Real data set: genes representation for each factor. The genes are represented on the factor using the Z matrix of the factors found by FAMT.

Authors' contributions

GLM and SL provided the real expression data set. YB analyzed and interpreted the expression data sets supervised by SL and DC. YB carried out the QTL and eQTL mapping analyses supervised by SL. YB, SL and DC drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

YB is a Ph.D fellow supported by the French Research Ministry and GLM is a Ph.D. fellow supported by the French Technical Institute for Poultry (ITAVI). The research program was supported by grants from a French society for genomics in poultry (AGENAVI), INRA and the Agence Nationale de la Recherche (Grant N°0426). Genotyping was performed at Toulouse-Midi-Pyrénées Genopole (France).

Author Details

[1]Agrocampus Ouest, UMR598, Animal Genetics, 35000 Rennes, France, [2]INRA, UMR598, Animal Genetics, 35000 Rennes, France and [3]Agrocampus Ouest, Applied Mathematics Department, 35000 Rennes, France

Received: 30 March 2010 Accepted: 2 July 2010
Published: 2 July 2010

References

1. Kustra R, Shioda R, Zhu M: **A factor analysis model for functional genomics.** *BMC bioinformatics* 2006, **7**:216.
2. Gordon A, Glazko G, Qiu X, Yakovlev A: **Control of the mean number of false discoveries, Bonferroni, and stability of multiple testing.** *The Annals of Applied Statistics* 2007, **1**:179-190.
3. Leek J, Storey J: **Capturing heterogeneity in gene expression studies by surrogate variable analysis.** *PLoS Genetics* 2007, **3(9)**:.
4. Friguet C, Kloareg M, Causeur D: **A factor model approach to multiple testing under dependence.** *JASA* in press.
5. Le Mignon G, Desert C, Pitel F, Leroux S, Demeure O, Pitel F, Guernec G, Pitel F, Douaire M, P LR, Pitel F, Lagarrigue S: **Using trancriptome profiling to characterize QTL regions on chicken chromosome 5.** *BMC Genomics* in press.
6. Leek J, Storey J: **A general framework for multiple testing dependence.** *Proceedings of the National Academy of Sciences* 2008, **105(48)**:18718.

7. Peltola P, Pihlajamäki J, Koutnikova H, Ruotsalainen E, Salmenniemi U, Vauhkonen I, Kainulainen S, Gylling H, Miettinen T, Auwerx J, *et al.*: **Visceral Obesity is Associated with High Levels of Serum Squalene&ast.** *Obesity* 2006, **14(7):**1155-1163.
8. Miettinen T, Gylling H: **Cholesterol absorption efficiency and sterol metabolism in obesity.** *Atherosclerosis* 2000, **153:**241-248.
9. MIETTINEN T: **Cholesterol production in obesity.** *Circulation* 1971, **44(5):**842.
10. Spearman C: **"General Intelligence," Objectively Determined and Measured.** *The American Journal of Psychology* 1904, **15(2):**201-292.
11. Rubin D, Thayer D: **EM algorithms for ML factor analysis.** *Psychometrika* 1982, **47:**69-76.
12. Lê S, Josse J, Husson F: **FactoMineR: An R package for multivariate analysis.** *Journal of Statistical Software* **25:**1-18.
13. Elsen J, Mangin B, Goffinet B, Boichard D, Le Roy P: **Alternative models for QTL detection in livestock. I. General introduction.** *Genetics Selection Evolution* 1999, **31(3):**213-224.
14. Le Roy P, Elsen J, Boichard D, Mangin B, Bidanel J, Goffinet B: **An algorithm for QTL detection in mixture of full and half sib families.** *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production* 1998, **26:**257-260.
15. Casel P, Moreews F, Lagarrigue S, Klopp C: **sigReannot: an oligo-set re-annotation pipeline based on similarities with the Ensembl transcripts and Unigene clusters.** *BMC proceedings, BioMed Central Ltd* 2009, **3:**S3.
16. **ENSEMBL website** [http://www.ensembl/org]
17. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, *et al.*: **Gene Ontology: tool for the unification of biology.** *Nature genetics* 2000, **25:**25-29.
18. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic acids research* 2006:D354.

doi: 10.1186/1471-2105-11-368
Cite this article as: Blum *et al.*, A factor model to analyze heterogeneity in gene expression *BMC Bioinformatics* 2010, **11**:368

1.2.2 Article 3 : Mach *et al.* JDS, 2012

Apport de l'article

La stratégie développée par Blum et al. (2010) est appliquée dans un autre contexte biologique. Dans cette étude, on s'intéresse à l'effet du polymorphisme du gène *DGAT1* sur la composition du lait et l'expression des gènes de la glande mammaire chez la vache laitière. Au niveau de ce gène, il existe une mutation (substitution de l'acide aminé Alanine (A) par une Lysine (K)) connue pour être associée à une diminution de la production de lait et pour faire varier la composition du lait. Afin d'identifier les gènes et processus biologiques impactés par cette mutation, l'idée est de comparer les données d'expression pour les animaux des 3 génotypes AA, AK et KK. Comme dans Blum et al. (2010), on prend en compte l'hétérogénéité d'expression indépendante du caractère en ajustant les données par les facteurs d'hétérogénéité estimés à l'aide de la méthode FAMT (Friguet et al. (2009)). La méthode FAMT permet de mettre en évidence 30 gènes annotés dont l'expression est différentiellement exprimée selon les génotypes alors que les méthodes classiques ne mettent en évidence aucun gène. Parmi ces gènes, ceux de plus fortes amplitudes d'expression sont testés et validés par RT-qPCR. De plus, comme montré dans Blum et al. (2010), une ACP basée sur les données ajustées par les facteurs d'hétérogénéité permet de capturer sur les premières composantes la variabilité principalement due aux 3 génotypes. Cette approche permet alors d'associer à l'effet des génotypes, des processus biologiques particuliers à l'aide d'une étude fonctionnelle sur les gènes fortement liés aux premières composantes de l'ACP.

J. Dairy Sci. 95:1–12
http://dx.doi.org/10.3168/jds.2012-5348

Pleiotropic effects of polymorphism of the gene diacylglycerol-*O*-transferase 1 (*DGAT1*) in the mammary gland tissue of dairy cows

N. Mach,*[1,2] Y. Blum,†‡ A. Bannink,* D. Causeur,‡ M. Houee-Bigot,‡ S. Lagarrigue,† and M. A. Smits*
*Wageningen UR Livestock Research, Lelystad 8200 AB, the Netherlands
†Agrocampus Ouest, UMR598, Génétique Animale, Rennes 35000, France
‡Agrocampus Ouest, Applied Mathematics, Rennes 35000, France

ABSTRACT

Microarray analysis was used to identify genes whose expression in the mammary gland of Holstein-Friesian dairy cows was affected by the nonconservative Ala to Lys amino acid substitution at position 232 in exon VIII of the diacylglycerol-*O*-transferase 1 (*DGAT1*) gene. Mammary gland biopsies of 9 homozygous Ala cows, 13 heterozygous cows (Ala/Lys), and 4 homozygous Lys cows in midlactation were taken. Microarray ANOVA and factor analysis for multiple testing methods were used as statistical methods to associate the expression level of the genes present on Affymetrix bovine genome arrays (Affymetrix Inc., Santa Clara, CA) with the *DGAT1* gene polymorphism. The data was also analyzed at the level of functional modules by gene set enrichment analysis. In this small-scale experimental setting, *DGAT1* gene polymorphism did not modify milk yield and composition significantly, although expected changes occurred in the yields of C14:0, *cis*-9 C16:1, and long-chain fatty acids. Diacylglycerol-*O*-transferase 1 gene polymorphism affected the expression of 30 annotated genes related to cell growth, proliferation, and development, remodeling of the tissue, cell signaling and immune system response. Furthermore, the main affected functional modules were related to energy metabolism (lipid biosynthesis, oxidative phosphorylation, electron transport chain, citrate cycle, and propanoate metabolism), protein degradation (proteosome-ubiquitin pathways), and the immune system. We hypothesize that the observed differences in transcriptional activity reflect counter mechanisms of mammary gland tissue to respond to changes in milk fatty acid concentration or composition, or both.

Key words: diacylglycerol-*O*-transferase 1 (*DGAT1*) gene, lactating cow, mammary gland, microarray

INTRODUCTION

The diacylglycerol-*O*-transferase 1 (***DGAT1***) gene is mapped to the centromeric end of bovine chromosome 14 (Boichard et al., 2003) and encodes the DGAT1 microsomal enzyme, which catalyzes the last step in triglyceride (**TAG**) synthesis: the esterification of a fatty acyl-CoA to the *sn*-3 position of a diacylglycerol. Evidence exists that the nonconservative Ala (**A**) to Lys (**K**) amino acid substitution at position 232 in exon VIII of the *DGAT1* gene is associated with decreased milk production (Schennink et al., 2007; Banos et al., 2008; Berry et al., 2010), decreased milk protein yield (Thaller et al., 2003; Schennink et al., 2007), increased milk fat yield (Thaller et al., 2003; Schennink et al., 2007), and changes in milk FA composition (Schennink et al., 2007, 2008; Bouwman et al., 2011). The *DGAT1* 232K allele is associated with increased saturated fat and reduced fractions of unsaturated C18 FA and conjugated linoleic acid in milk, which is regarded as unfavorable to human health. Therefore, from the human health perspective, decreasing the frequency of the *DGAT1* 232K allele in dairy cows through breeding programs could be desirable because milk and milk-derived foods are still large contributors to SFA intake in humans (Schennink et al., 2007). It is hypothesized that *DGAT1* gene polymorphism affects the functional properties of the DGAT1 enzyme (e.g., higher activity or alteration in substrate specificity for different FA), which influences milk FA composition (Grisart et al., 2004; Schennink et al., 2007). Although Grisart et al. (2004) did not show quantitative differences in mRNA expression levels of the *DGAT1* gene between alleles, it is not known yet whether *DGAT1* gene polymorphism affects the expression of other genes involved in lipid metabolism in the mammary gland. Functional approaches are required to identify effects of *DGAT1* gene polymorphism on lipid metabolism and other processes in the cow mammary gland tissue. The objective of this study was to determine the effect of *DGAT1* gene polymorphism on the global mRNA expression pattern of genes in the mammary gland tissue of grazing dairy

Received January 16, 2012.
Accepted May 4, 2012.
[1]Corresponding author: nuria.mach@jouy.inra.fr
[2]This author is currently working at INRA UMR 1313 Génétique Animale et Biologie Intégrative, 78352 Jouy-en-Josas, France.

cows to get more insight into the effects of this polymorphism on the physiology of the mammary gland.

MATERIALS AND METHODS

Animals and Diets

Twenty-eight Holstein-Friesian dairy cows in midlactation (DIM; 176 ± 22.8 d) were used. Cows were fed corn silage, grass silage and concentrates as a mixed ration (**MR**). In addition to the MR, each cow received 1 kg of a commercial standard concentrate per day through automatic feeders in the milking parlor. Cows were fed the MR indoors at night, and were grazed on pasture composed of ryegrass (*Lolium perenne* L.), with approximately 20% white clover (*Trifolium repens* L.) during the day. The average paddock size was 5 ha, and the stocking rate was 16 cows/ha. Cows were fed the MR at a level of about 14.5 kg of DM/cow per day, and grazed at a daily herbage intake of approximately 5.5 kg of DM/cow. Ingredients of the concentrates and MR were previously described by Mach et al. (2011). Cows were milked twice per day in the milking parlor at the facility. Two consecutive milk samples (a.m. and p.m. milking) were obtained and pooled. One aliquot was stored at 4°C until analysis of fat, protein, and lactose percentage, and another aliquot was frozen at −20°C until analysis for FA composition by gas chromatography, as described by Mach et al. (2011). Approximately 750 to 1,000 mg of mammary tissue from each cow was obtained by surgical biopsy from the midpoint section of a rear quarter according to the method of Farr (Farr et al., 1996). One part of the tissue was used for extraction of DNA and the other for extraction of RNA.

Genotyping

Prior to genotyping, the quality and quantity of DNA samples isolated from mammary gland tissue were assessed using a NanoDrop spectrophotometer (Isogen Bioscience BV, Maarssen, the Netherlands) and agarose gel electrophoresis. Genotyping of *DGAT1* gene polymorphism was performed using a TaqMan allelic discrimination method in an Applied Biosystems 7,500 quantitative real-time (**qRT**) PCR System (Applied Biosystems, Bleiswijk, the Netherlands), as described by Schennink et al. (2007). The genotype at the *DGAT1* locus was designated homozygous Lys (**KK**), heterozygous Lys/Ala (**KA**), or homozygous Ala (**AA**). A kinship coefficient matrix between animals was calculated by using the Kinship package (version 1.1.0.) in the R programming language (version 2.13; The R Foundation for Statistical Computing, Vienna, Austria), and used to perform 2-way hierarchical cluster analysis (**HCA**) with "1 – cor (x)" as distance and "average" as aggregation criterion. Additionally, the point-biserial correlation method was applied to the dendrogram to identify the optimum number of clusters as described by Odong et al. (2011). The heatmap function was used to generate images.

RNA Isolation, Processing, and Microarray Analysis

Total RNA from mammary gland tissue (50 to 100 mg) was isolated using TRIzol reagent (Invitrogen, Breda, the Netherlands), following the manufacturer's instructions. The RNA purity and concentrations were determined using a NanoDrop ND 1,000 spectrophotometer (Isogen Bioscience BV), and the RNA quality was assessed using a Bioanalyzer 2,100 (Agilent Technologies Netherlands BV, Amsterdam, the Netherlands). The RNA of each biopsy was amplified, biotin labeled, and hybridized to a single-dye Affymetrix GeneChip Bovine Genome Array (#900493; Affymetrix Inc., Santa Clara, CA) by ServiceXS BV (Leiden, the Netherlands), as described in the user's manual (GeneChip Expression Analysis Technical Manual). For further details, see the Supplemental Materials and Methods (available online at http://journalofdairyscience.org/). Arrays were scanned using an Affymetrix GeneChip Scanner 7G and Affymetrix GeneChip operating software version 1.4, following the GeneChip specifications. After scanning, the Affymetrix GeneChip Command Console (AGCC) software automatically acquired and analyzed image data and computed an intensity value for each probe cell. A total of 28 one-color arrays were prepared, one array per RNA sample. All microarray analyses, including preprocessing quality assessment, normalization, and statistical analysis, were carried out using Bioconductor packages (version 2.7) in the R programming language (version 2.13). Details on transcriptome data acquisition, quality control, summarization, and normalization are described in Mach et al. (2011). Briefly, data were quality assessed before and after normalization using several built-in quality control methods implemented in the Bioconductor affycoretools and associated packages to identify eventual irregularities of array hybridization, RNA degradation, and data normalization. Arrays were considered of sufficiently high quality if they showed less than 10% specks in fit probe level model images, if they were not deviating in RNA degradation and density plots, and if they were not significantly deviating in normalized unscaled standard errors (**NUSE**) and relative log expression (RLE) plots. Upon rigorous examination of the built-in quality control methods to identify ir-

regularities of array hybridization, RNA degradation, and data normalization, 2 microarrays related to the AA genotype were discarded. Therefore, we included 26 microarrays in our further analysis: 9, 13, and 4 microarrays for AA, KA, and KK genotypes, respectively. Filters were applied to the data to improve the quality of the normalized data set. Probes that did not have a minimum threshold of 40 raw intensity units in at least 1 array per genotype were omitted from analysis. The threshold of 40 was chosen based on an interquartile range (IQR) offset of 0.25 for at least 1 array per genotype (Caesar et al., 2010). This control filtering criteria reduced the number of probes from 24,128 to 7,486. All microarray experiment data are MIAME compliant and have been deposited in the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/; accession number GSE33720).

Validation of Differential Gene Expression by qRT-PCR

A total of 6 genes differentially expressed between genotypes were validated by qRT-PCR: regulator of G-protein signaling 1 (***RGS1***), ubiquitin D (***UBD***), PHD finger protein 3 (***PHF3***), fibulin 5 (***FBLN5***), pyruvate dehydrogenase (lipoamide) α 1 (***PDHA1***), and UHRF1-binding protein 1-like (***UHRF1BP1L***). These genes exhibited the highest amplitude of variation between the genotypes [at least higher than ±40% between 2 genotypes, absolute value $\log_2$(ratio) >0.485]. Reverse transcription of 1 μg of the isolated total RNA was performed in a 20-μL reaction using Superscript 3 reverse transcriptase (Invitrogen), deoxyribonucleotide triphosphate (dNTP; Roche Diagnostics Nederland BV, Almere, the Netherlands), and random hexamer primers (Roche Diagnostics Nederland BV) for 1 h at 50°C according to the manufacturer protocol (Invitrogen). Templates were amplified after preincubation for 10 min at 95°C, followed by amplification for 40 cycles (10 s at 95°C, 5 s at 60°C, and 5 s at 72°C) on a LightCycler 1.2 qRT-PCR System by using FastStart DNA Master SYBR Green 1 reagents (Roche Diagnostics Nederland BV). The gene-specific primers used are indicated in Supplemental Table S1 (available online at http://journalofdairyscience.org/). All reactions revealed a single product as determined by melting curve analysis. Quantitative measurements were performed by establishing a linear amplification curve from 10-fold serial dilutions of cDNA for each gene, and efficiencies of the used sets of primers were at least 95%. Values were calculated according to the comparative cycle threshold (Ct) method using β-actin (***ACTB***) as the endogenous reference gene.

Statistical Analyses

Milk quality, FA composition, and qRT-PCR data were analyzed using a mixed-effects model (SAS Institute Inc., Cary, NC). The model included genotype as fixed effect and cow as a random effect. Gene expression was analyzed as described above using the microarray ANOVA (**MAANOVA**) package (version 1.1.0.) in the R programming language (version 2.13). The *P*-values were corrected for multiple testing using a false discovery rate (**FDR**) method <10% (*q*-value <0.10), which provides an estimate of the fraction of false discoveries among the significant terms (Bünger et al., 2007). In parallel, the factor analysis for multiple testing (**FAMT**) method (version 2.20.0) was also applied to analyze gene expression in R programming (Causeur et al., 2011). The FAMT method is able to remove expression heterogeneity from the data for the subsequent statistical analysis by adjusting the raw data from the common structure. It is specially designed to select differentially expressed genes in microarray data when the correlation structure among gene expression is strong (Causeur et al., 2011). An FDR <10% was used as a threshold for significance of differential expression. Two-way HCA was performed using the hclust function with "1 − cor (x)" as distance and "average" as aggregation criterion. The heatmap function was used to generate images. Principal component analysis (**PCA**) was performed with the FactoMineR library (http://cran.r-project.org/web/packages/FactoMineR/index.html). The list of significant genes identified was uploaded into Ingenuity Pathways Analysis (IPA version 5.5; Ingenuity Systems Inc., Redwood City, CA).

Moreover, microarray data was analyzed at the level of differentially expressed functional modules by gene set enrichment analysis (**GSEA**). Gene set enrichment analysis discovers a collection of genes that show moderate but coordinated differentiation, even when individual genes are not. Therefore, GSEA can identify more subtle changes in expression than the gene list that results from univariate statistical analysis. To that end, the GSEA method first calculated for each group of genes an enrichment score (**ES**) that reflected the degree of association between the gene expression and the genotypes. Then, it estimated the significance level of the ES by using an empirical phenotype-based permutation test procedure that preserved the complex correlation structure of the gene expression data, and lastly an adjustment for multiple hypothesis testing was done (Subramanian et al., 2005). Gene set size filter considered a minimum of 15 and a maximum of 500 genes, and the number of permutations was set to 1,000. Gene sets were considered significantly enriched at an

Table 1. Effect of diacylglycerol-*O*-transferase 1 (*DGAT1*) gene polymorphism on milk yield and milk composition

	Genotype[1]				
Trait	AA (n = 9)	KA (n = 13)	KK (n = 4)	SEM	*P*-value
DIM	189	171	167	22.8	0.85
Milk production (kg/d)	24.8	20.8	22.6	1.68	0.25
Fat content (%)	4.16	4.34	4.61	0.160	0.50
Fat yield (kg/d)	0.99	0.90	1.11	0.081	0.48
Protein content (%)	3.70	3.81	3.63	0.122	0.72
Protein yield (kg/d)	0.88	0.79	0.88	0.074	0.69
Lactose content (%)	4.28	4.45	4.36	0.090	0.51
Lactose yield (kg/d)	1.01	0.92	1.07	0.071	0.61

[1]The genotype at the *DGAT1* 232 locus was designated KK, KA, or AA for homozygous Lys, heterozygous Lys/Ala, or homozygous Ala, respectively.

FDR <0.05. Normalized ES of significantly enriched pathways between genotypes groups were calculated.

RESULTS

Genotypes and Allele Frequencies

The frequency of the AA, KA, and KK genotype was 0.46, 0.40, and 0.14, respectively. The observed genotype frequencies are consistent with those from Holstein-Friesian populations in the Netherlands (Schennink et al., 2007), the UK (Banos et al., 2008), Ireland (Berry et al., 2010), and France (Gautier et al., 2007). The minor allele frequency (K) was 0.37, similar to that reported by Schennink et al. (2007) in the Netherlands. Additionally, the kinship coefficients were 6.9, 5.5, and 1.7% for AA, KA, and KK genotypes, respectively. Values observed for AA and KA genotypes are consistent with those expected in the Dutch Holstein-Friesian population (5.27%; Danchin-Burge et al., 2011). The discrepancy in the level of kinship coefficient in the KK animals compared with AA and KA genotypes is probably attributed to sample size and differences in the depth of the pedigree available, as 2 of the animals came from Germany and not all the information was available. The kinship coefficient matrix is plotted in Supplemental Figure S1, whereas the resulting heatmap and dendrogram are presented in Supplemental Figures S2 and S3 (available online at http://journalofdairyscience.org/), respectively. The point-biserial correlation was not able to identify a significant number of clusters ($R^2 = 0.10$; $P = 0.98$). The latest result, together with the off-diagonal entries in Supplemental Figure S1 (available online at http://journalofdairyscience.org/), suggests that no global population structure exists within and between genotypes.

Milk Production and Composition

Substituting A for K at the A232K *DGAT1* locus numerically decreased milk yield and increased fat yield, although these effects were not significant (Table 1). With regard to the FA composition (Table 2), the proportion of C14:0 was greater ($P = 0.04$) in AA genotypes as compared with KA and KK genotypes. In

Table 2. Effect of diacylglycerol-*O*-transferase 1 (*DGAT1*) gene polymorphism on milk FA composition (g/100 g of total FA)

	Genotype[1]				
FA	AA (n = 9)	KA (n = 13)	KK (n = 4)	SEM	*P*-value
C4:0	2.74	2.73	2.75	0.110	0.98
C6:0	2.14	2.11	2.15	0.070	0.95
C8:0	1.44	1.37	1.47	0.051	0.66
C10:0	3.25	3.02	3.3	0.142	0.51
C11:0	0.43	0.44	0.50	0.022	0.41
C12:0	4.04	3.73	4.08	0.142	0.32
iso C13:0	0.02	0.02	0.02	0.004	0.79
anteiso C13:0	0.11	0.11	0.14	0.009	0.26
C13:0	0.20	0.21	0.26	0.012	0.16
C14:0	11.82	11.13	11.26	0.173	*

Continued

Table 2 (Continued). Effect of diacylglycerol-*O*-transferase 1 (*DGAT1*) gene polymorphism on milk FA composition (g/100 g of total FA)

	Genotype[1]				
FA	AA (n = 9)	KA (n = 13)	KK (n = 4)	SEM	*P*-value
iso C15:0	0.22	0.21	0.22	0.004	†
anteiso C15:0	0.50	0.49	0.48	0.013	0.87
cis-9 C14:1	1.22	1.34	1.50	0.090	0.42
C15:0	1.02	1.07	1.22	0.030	†
iso C16:0	0.20	0.17	0.17	0.008	0.11
C16:0	28.47	29.47	30.85	0.652	0.25
iso C17:0	0.16	0.16	0.15	0.007	0.88
cis-9 C16:1	1.62	1.96	2.24	0.090	**
anteiso C17:0	0.27	0.21	0.23	0.012	0.13
C17:0	0.61	0.61	0.63	0.012	0.75
cis-9 C17:1	0.16	0.19	0.20	0.012	0.14
C18:0	9.94	8.82	8.16	0.451	0.16
trans-4 C18:1	0.01	0.01	0.01	0.002	0.85
t*rans*-5 C18:1	0.01	0.01	0.01	0.002	0.54
t*rans*-6 + *trans*-8 C18:1	0.29	0.32	0.24	0.010	0.20
trans-9 C18:1	0.16	0.19	0.17	0.011	0.49
trans-10 + *trans*-11 C18:1	1.38	1.80	1.72	0.122	0.14
trans-12 C18:1	0.41	0.46	0.39	0.013	0.11
trans-13 + *trans*-14 C18:1	0.32	0.34	0.40	0.027	0.42
cis-12 C18:1	0.06	0.08	0.07	0.013	0.34
cis-9 C18:1	19.55	19.27	17.64	0.541	0.26
cis-13 C18:1	0.06	0.08	0.07	0.005	0.34
trans-16 + *cis*-14 C18:1	0.55	0.55	0.50	0.022	0.34
cis-15 C18:1	0.24	0.27	0.25	0.013	0.43
cis-11 C18:1	0.30	0.35	0.35	0.033	0.15
trans-9,*trans*-12 C18:2	0.007	0.008	0.016	0.002	0.35
trans-10,*cis*-12 C18:2	0.01	0.01	0.01	0.001	0.58
cis-9, *cis*-12 C18:2	1.48	1.57	1.38	0.040	*
cis-9, *trans*-11 C18:2	0.57	0.67	0.64	0.050	0.40
cis-9, *cis*-12, *cis*-15 C18:3	0.54	0.56	0.53	0.032	0.68
cis-6, *cis*-9, *cis*-12 C18:3	0.02	0.02	0.01	0.002	0.74
C20:0	0.12	0.11	0.11	0.006	0.73
cis-11 C20:1	0.03	0.04	0.02	0.003	*
cis-11, *cis*-14 C20:2	0.02	0.02	0.01	0.002	0.35
cis-11, *cis*-14, *cis*-17 C20:3	0.01	0.01	0.01	0.002	0.53
cis-8,*cis*-11,*cis*-14 C20:3	0.06	0.06	0.06	0.005	0.99
all *cis*-5, -8, -11, -14 C20:4	0.09	0.10	0.11	0.041	0.23
all *cis*-5, -8, -11, -14, -17 C20:5	0.05	0.05	0.04	0.003	0.86
C22:0	0.04	0.05	0.04	0.003	0.36
cis-13 C22:1	0.01	0.01	0.01	0.003	0.31
cis-13, *cis*-16 C22:2	0.04	0.05	0.04	0.002	0.60
all *cis*-7, -10, -13, -16 C22:4	0.01	0.02	0.01	0.002	0.49
C23:0	0.01	0.02	0.01	0.002	0.61
all *cis*-7, -10, -13, -16, -19 C22:5	0.07	0.07	0.09	0.003	0.12
all *cis*-4, -7, -10, -13, -16, -19 C22:6	0.01	0.01	0.01	0.002	0.72
C24:0	0.04	0.03	0.03	0.009	0.56
cis-15 C24:1	0.01	0.01	0.01	0.001	0.74
Unidentified	2.83	3.20	3.04	0.012	0.86
De novo FA[2]	27.4	26.2	27.4	0.56	0.35
C16	30.0	31.4	33.9	0.64	0.26
SFA	67.7	66.2	68.2	0.97	0.49
LCFA[3]	36.5	36.1	33.2	0.76	0.11
MUFA	26.4	27.3	25.8	0.83	0.56
UFA[4]	29.4	30.5	28.8	0.89	0.50
PUFA	2.9	3.2	2.9	0.11	0.18
Total *trans*-FA	3.1	3.7	3.5	0.20	0.19

[1]The genotype at the *DGAT1* 232 locus was designated KK, KA, or AA for homozygous Lys, heterozygous Lys/Ala, or homozygous Ala, respectively.

[2]De novo FA include all FA from C4 to C14.

[3]LCFA = long-chain FA (includes all FA ≥C18).

[4]Unsaturated FA.

†$P < 0.10$; *$P < 0.05$; **$P < 0.01$.

contrast, KK genotypes lead to an increase ($P = 0.01$) in the proportion of *cis*-9 C16:1 FA, whereas they lead to a decrease ($P = 0.03$) in the proportion of *cis*9,*cis*12 C18:2 and *cis*-1 C20:1 compared with AA and KA genotypes. No significant differences could be observed for the other FA.

Differential Expression of Individual Genes in the Mammary Gland

Genes identified by MAANOVA did not meet the threshold for statistical significance after multiple hypotheses testing correction. However, the FAMT method, which adjusts data from heterogeneity components, revealed 30 annotated genes, which differed between genotypes (FDR q-values <0.10). The gene identification, symbol and description of the 30 resulting genes are shown in Table 3. A 2-way HCA was applied to the normalized data set to identify clusters between the different genotypes and the 30 differently expressed genes (Figure 1). The heatmap separated the heterozygous KA animals from the homozygous AA and KK animals. For the gene variables, 2 clusters were identified: cluster 1 (16 genes) and 2 (14 genes). The majority of genes in cluster 1 were significantly downregulated when comparing KK and AA genotypes to KA genotypes. On the contrary, the majority of genes in cluster 2 were downregulated when comparing KA genotypes to KK and AA genotypes. Complementary to these results, the PCA allowed us to identify the genes contributing most to the separation of the different clusters (Figure 2). As highlighted in Figure 2A:1, the first component, which explains the largest contribution to variability in the gene expression data; that is, retains the most information, clearly separates between gene expression patterns of the heterozygous KA animals from the homozygous AA and KK animals. This first component of PCA explained 41.64% of the total variability in gene expression. The second component explained 15.53% and the third 7.91% (Figure 2A:2), which leads to an improved clustering of the KK genotypes. Variable graphs (Figure 2B:1 and 2B:2) have been generated to better understand genes that contribute the most to separate expression patterns of the heterozygous KA animals from the homozygous AA and KK animals. In Figure 2B:1, the quality of representation of a variable on the axis is measured by the squared cosine between the vector issued from the element and its projection on the axis. If this square cosine is close to 1, it means that the element is well projected on the axis (Lê et al., 2008). Projecting the arrows onto the first dimension (Figure 2B:1) we can see that *PDHA1*, *RMB18*, *TMEM9*, and *LOC521568*, and *PHF3*, *ATL2*, and *ZBTB38* are the most important for the first component. For the third principal component (Figure 2B:2), the gene *FBLN5* and *C13H20ORF43* are the most important. Over 85% of the genes showed a high correlation with the first 3 principal components (r >0.6). In this context, most of the genes contributed to split-up genotypes into the 2 main clusters represented in Figure 2A (homozygous vs. heterozygous). Furthermore, the *RGS1*, *UBD*, *PHF3*, *FBLN5*, *PDHA1*, and *UHRF1BP1L* genes not only presented correlation coefficient >0.6 with respect to the principal components represented, but also exhibited the highest amplitude of variation between the genotypes [at least higher than ±40%, absolute value $\log_2$(ratio) >0.485]. These genes are indicated in black in Figure 2B. Therefore, they were used to validate the microarray gene expression by qRT-PCR (Supplemental Figure S4, available online at http://www.journalofdairyscience.org/). Overall, the qRT-PCR data confirmed the microarray data. The Ingenuity Pathways Analysis of the 30 genes revealed that they encode for molecular functions involved in growth, proliferation and development, signaling, molecular transport, cell death modulation, assembly and organization, as well as immune system and inflammatory response (Supplemental Table S2, available online at http://www.journalofdairyscience.org/).

Differential Expression of Functional Modules in the Mammary Gland

Instead of analyzing the differential expression of individual genes, we also analyzed the microarray data at the level of gene sets that together encode for particular functional modules (Figure 3). Notably, the gene sets involved in energy metabolism (oxidative phosphorylation, electron transport chain, citrate cycle, propionate metabolism, and lipid biosynthesis) were enriched in KK and KA compared with AA genotypes. Additionally, genes sets related to immune response were affected when comparing KK with KA and AA genotypes (i.e., IL-7 signaling, and antigen processing and presentation). The cell communication processes were overrepresented when comparing the AA and KK genotypes.

DISCUSSION

Milk Production and Composition

The K allele was associated with numerically reduced milk production and increased milk fat yield, consistent with previous work in dairy cattle (Schennink et al., 2007; Banos et al., 2008; Berry et al., 2010). Absence of statistically significant effects in the present analysis almost certainly relates to the limited sample size.

Table 3. Expression of differentially expressed genes identified by using factor analysis for multiple testing method when comparing the effect of diacylglycerol-*O*-transferase 1 (*DGAT1*) gene polymorphism on the global expression pattern of genes in the mammary gland tissue of dairy cows[1]

			Genotype[2]				
Identity of gene	Gene symbol	Description	AA	AK	KK	SEM	*q*-value
Bt.12614.1.A1_s_at	*EIF3B*	Eukaryotic translation initiation factor 3, subunit B	63.9	103.8	68.6	7.13	†
Bt.13063.1.S1_at	*UBE2A*	Ubiquitin conjugating enzyme E2A	130.3	157.4	129.2	4.34	†
Bt.1370.1.S1_at	*ASS1*	Argininosuccinate synthetase 1	138.3	189.2	146.1	9.81	†
Bt.14993.2.A1_at	*ATL2*	Atlastin guanosine triphosphatase (GTPase) 2	89.6	75.2	97.3	3.43	†
Bt.17703.1.S1_at	*SLC15A2*	Solute carrier family 15 (H^+/peptide transporter), member 2	30.8	54.9	31.3	4.47	†
Bt.19415.1.A1_at	*PDHA1*	Pyruvate dehydrogenase (lipoamide) α 1	35.0	69.8	42.6	4.77	*
Bt.19938.1.A1_at	*SP3*	Predicted: hypothetical LOC540766 (LOC540766)	514.0	431.0	556.1	16.20	*
Bt.20213.2.S1_at	*RBM18*	RNA-binding motif protein 18 (RBM18)	188.8	248.9	169.1	9.65	*
Bt.20400.2.S1_at	*LOC521568*	Similar to ATP-binding cassette, subfamily C, member 4	338.7	440	352	14.9	†
Bt.20639.1.A1_at	*PHF3*	Predicted: similar to KIAA0244 (LOC522928)	43.4	22.2	52.8	3.75	*
Bt.20938.1.S1_at	*ZBTB38*	Zinc finger and BTB domain-containing 38	150.9	116.3	149.1	4.8	†
Bt.21503.1.S1_at	*PPP1CA*	Protein phosphatase 1, catalytic subunit, α isoform	222.3	258.1	225.0	7.82	†
Bt.22594.1.S1_at	*LOC789894*	Similar to PAP-associated domain-containing protein 5 (topoisomerase-related function protein 4-2; TRF4-2)	323.3	264.1	370.7	10.80	*
Bt.23273.1.S1_at	*CIRBP*	Cold-inducible RNA-binding protein	248.2	323.3	276.3	20.12	†
Bt.2374.1.S1_at	*GNB1*	Guanine nucleotide-binding protein (G protein), β polypeptide 1	403.3	471.4	422.3	16.13	†
Bt.24580.1.A1_at	*TMEM9*	Transmembrane protein 9	41.5	56.3	34.4	2.58	†
Bt.27024.1.A1_at	*CCAR1*	Cell division cycle and apoptosis regulator 1	716.3	565.1	858.1	34	*
Bt.28269.1.S1_at	*RHOJ*	Ras homolog gene family, member J	120.0	119.0	87.3	6.6	†
Bt.29411.1.S1_at	*UHRF1BP1L*	Hypothetical LOC520720 (LOC520720)	58.4	32.8	70.19	4.68	*
Bt.3399.1.S1_a_at	*LHPP*	Phospholysine phosphohistidine inorganic pyrophosphate phosphatase	66.1	55.9	70.5	2.79	†
Bt.3539.1.S1_at	*C13H20ORF43*	Chromosome 20 open reading frame 43 ortholog	74.1	72.1	58.3	1.88	†
Bt.425.1.S2_at	*SLTM*	Scaffold attachment factor B (SAFB)-like, transcription modulator	188.9	172.2	225.5	5.56	*
Bt.4310.1.A1_at	*RGS1*	Regulator of G-protein signaling 1	54.6	35.9	161.1	9.41	*
Bt.5709.1.S1_at	*FGFR1OP2*	Fibroblast growth factor receptor 1 (FGFR1) oncogene partner 2	237.4	200.2	203.2	9.36	†
Bt.5897.1.S1_at	*UBD*	Similar to ubiquitin D, transcript variant 2 (LOC512938)	224.0	165.5	385.0	32.40	*
Bt.6449.1.S1_at	*FBLN5*	Fibulin 5	88.0	66.0	45.9	6.97	*
Bt.7413.1.S1_at	*GRN*	Granulin	59.4	76.4	60.6	2.86	†
Bt.7505.1.S1_at	*STK38*	Serine/threonine kinase 38	261.0	217.0	310.1	10.54	†
Bt.8243.1.A1_at	*ZFP91*	Similar to zinc finger protein homologous to mouse Zfp91 (LOC529006)	90.9	66.9	94.9	4.57	†
Bt.9031.1.S1_at	*PTBP1*	Polypyrimidine tract-binding protein 1	42.0	71.0	44.9	4.19	†

[1]A cut-off of false discovery rate *q*-values <0.10 was used.

[2]The genotype at the *DGAT1* 232 locus was designated KK, KA, and AA for homozygous Lys, heterozygous Lys/Ala, and homozygous Ala, respectively.

†*q*-value <0.10; **q*-value <0.05.

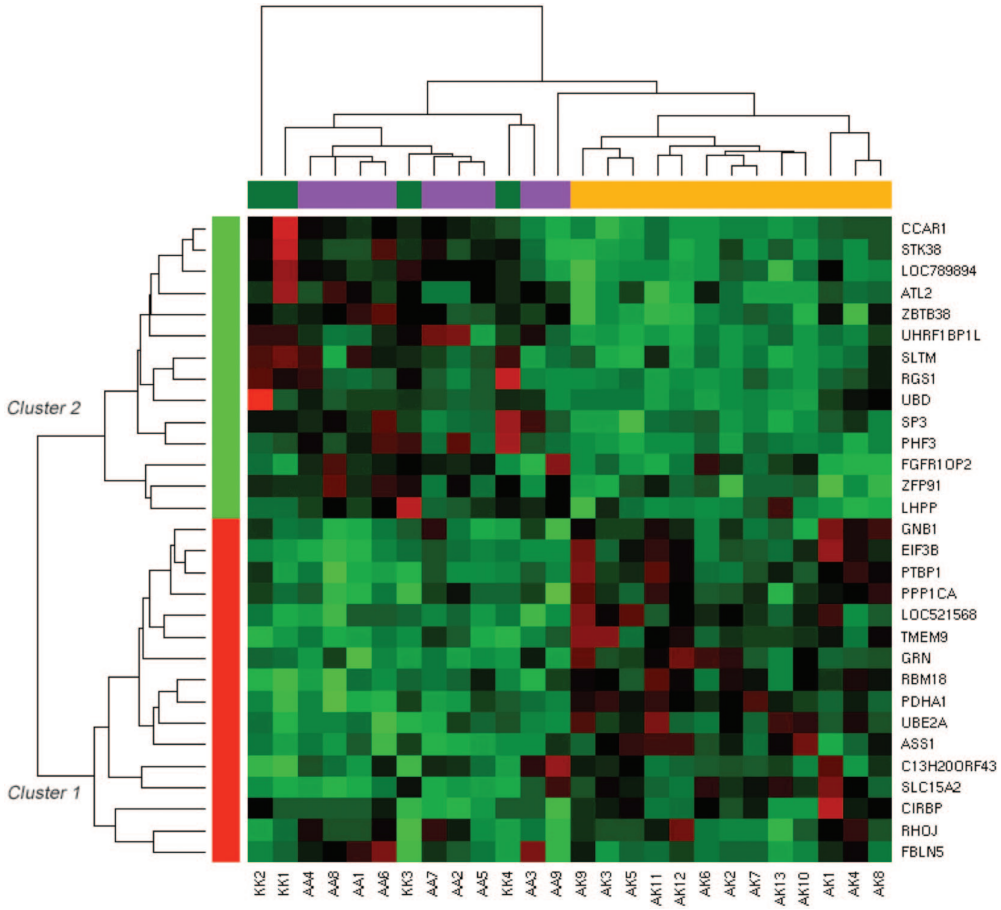


Figure 1. Two-way hierarchical clustering analysis of the genes identified by factor analysis for multiple testing method (q-values <0.10) when comparing the effect of the diacylglycerol-*O*-transferase 1 gene polymorphism on the global expression pattern of genes in the mammary gland tissue of dairy cows. The color-coded gene module is displayed in the color bars to the left of the dendrograms. Each column in the heatmap corresponds to 1 animal (labeled by color: green = homozygous Lys, violet = homozygous Ala, and orange = heterozygous Lys/Ala). In the heatmap, the green color represents low adjacency (negative correlation), whereas red represents high adjacency (positive correlation). Gene clusters 1 and 2 are displayed to the left of the dendrograms.

Nevertheless, the KK genotype was associated with a higher proportion of *cis*-9 C16:1 FA in milk fat, and a lower proportion of *cis*9,*cis*12 C18:2 and *cis*-1 20:1 compared with the AA and KA genotypes. Furthermore, in agreement to Schennink et al. (2007), the KK genotype was associated with lower proportions of C14:0. The observed effects of *DGAT1* gene polymorphism on milk FA composition might be explained by a higher activity of DGAT1 enzyme and alteration of substrate specificity of DGAT1, as suggested previously (Grisart et al., 2004; Schennink et al., 2007). These observations, together with several gene expression effects (see next sections) are the major factor in leading us to believe that *DGAT1* polymorphism was indeed the main factor affecting milk yield and composition, rather than another mutation located in the same chromosome 14 that would be in strong linkage disequilibrium with K232A.

Differential Expression of Genes in the Mammary Gland

The present study is, to our knowledge, the first one in which microarrays were used for analysis of the bovine mammary gland transcriptome in relation to polymorphism in the exon region of the *DGAT1* gene. The analysis with the MAANOVA did not identify individual genes that met the threshold for statistical significance. The differences in gene expression were probably modest relative to the experimental and biological noise, which is inherent to technical features of the experiment (RNA isolation and handling, chip

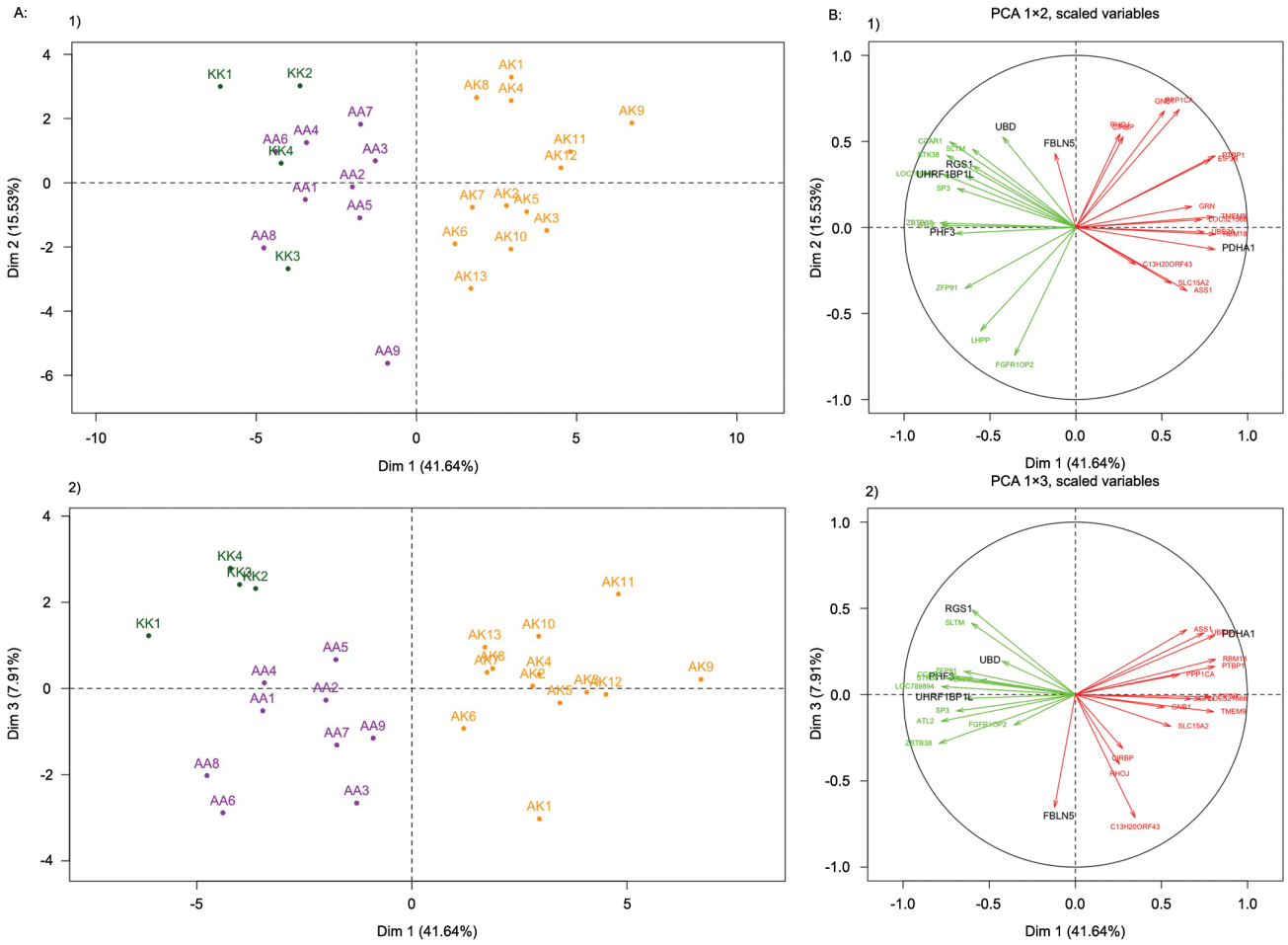


Figure 2. Principal component analysis (PCA) of genes identified by factor analysis for multiple testing method (q-values <0.10) when comparing the effect of the diacylglycerol-*O*-transferase 1 (*DGAT1*) gene polymorphism on the global expression pattern of genes in the mammary gland tissue of dairy cows. A: animal factor map; A:1 is a projection in the X1-X2 plane; :2 is a projection in the X1-X3 plane. Animals are labeled by color: green = homozygous Lys, violet = homozygous Ala, and orange = heterozygous Lys/Ala. The contribution of each animal to the construction of 1 dimension (Dim) allows detection among the genotypes of which ones are extreme and contribute to the construction of the dimension. B: genes factor map; B:1 is a projection in the X1-X2 plane; B:2 is a projection in the X1-X3 plane. The 30 genes identified by factor analysis for the multiple testing method that contribute to the construction of the axes are projected. The quality of representation of a gene on the axis of rank s is measured by the squared cosine between the vector issued from the element and its projection on the axis. If this square cosine is close to 1, it means that the element is well projected on the axis (Lê et al., 2008). The quality of representation of a gene on a plane can be visualized by the distance between the projected variable onto the plane and the correlation circle (circle of radius 1). The genes that appeared to be highly affected by the *DGAT1* gene polymorphism (at least higher than ±40% between 2 genotypes, absolute value $\log_2$(ratio) >0.485; results in Table 3) are indicated in black.

hybridization conditions, and scanner characteristics) and to gene-environment interactions, independent of *DGAT1* gene polymorphism effect. In contrast to MAANOVA, the more advanced FAMT approach allowed us to identify 30 annotated genes whose expression in the mammary gland differed significantly between the *DGAT1* genotypes. The FAMT method is able to capture and remove the hidden dependence structure independent of *DGAT1* gene polymorphism and offers improved high-dimensional multiple testing procedures (Friguet et al., 2009; Blum et al., 2010). The FAMT method removes expression heterogeneity from the data before subsequent analysis of statistical significance by modeling the common information shared by all the genes using a factor analysis structure (Blum et al., 2010; Causeur et al., 2011). As the uncontrolled effects and technological biases are extracted from the statistical noise, FAMT shows an improved capacity to test differential expression of gene expression compared with MAANOVA.

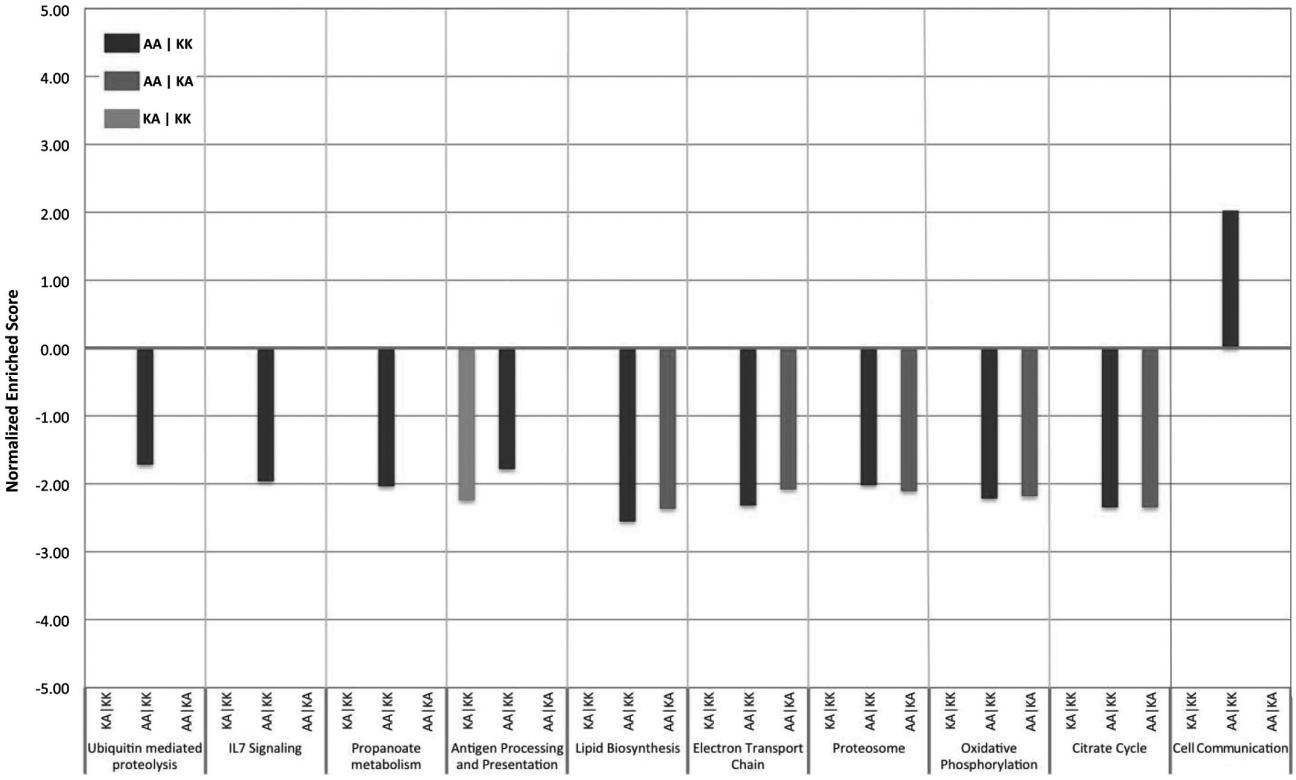


Figure 3. Normalized enriched scores (NES) of enriched gene sets (q-values <0.05) in the mammary gland of dairy cows identified by using the gene set enrichment analysis method when comparing the different diacylglycerol-*O*-transferase 1 (*DGAT1*) genotypes. Gene sets were considered significantly enriched at a false discovery rate q-value <5%. Comparisons between genotypes are labeled by color: green = AA|KA, violet = AA|KK, and orange = KA|KK. The genotype at the *DGAT1* 232 locus was designated KK, KA, or AA for homozygous Lys, heterozygous Lys/Ala, or homozygous Ala, respectively. The value of the NES reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes. A positive NES indicates gene set enrichment at the top of the ranked list; a negative NES indicates gene set enrichment at the bottom of the ranked list. Color version available in the online PDF.

A finding of particular interest was that by using the HCA and PCA approaches, the analyzed groups of animals could be split into homozygous and heterozygous genotypes based on their overall gene-expression profiles. Compared with homozygous animals, heterozygous *DGAT1* animals presented a downregulation of genes that mainly play a role in cellular signaling and immune responses. On the other hand, the heterozygous genotype presented substantial upregulation of genes related to the organization of the mammary gland tissue. Whether this also leads to differences in the cellular and functional dynamics of mammary tissue between homozygous and heterozygous animals is not yet known.

Notably, out of the identified 30 annotated genes whose expression in the mammary gland differed significantly between the *DGAT1* genotypes, *RGS1* and *UBD* genes exhibited the 2 highest expressions in the genotype KK compared with the AA and KA genotypes. Evidence exists that *RGS1* regulates cell signaling, which is accomplished by a myriad of proteins, lipids, ions, and small molecules (McCudden et al., 2005). Donaldson et al. (2005) described that *RGS1* may play a role in the mammary gland of dairy cows in attenuating of the Ras-mediated signal that promotes the activity of peripheral blood lymphocytes. Moreover, Connor et al. (2008) reported that *UBD* is involved in immune responses and inflammation in the mammary gland of dairy cows. The exact molecular function of these genes in mammary tissue of dairy cows remains to be elucidated.

Differential Expression of Functional Modules in the Mammary Gland

In line with the individual gene-based analysis, GSEA showed that *DGAT1* gene polymorphism modified the expression of sets of genes that mainly modulate cell signaling, cell energy metabolism, and immune system response. Remarkably, the gene sets involved in energy

metabolism (oxidative phosphorylation, electron transport chain, citrate cycle, and lipid biosynthesis) were less expressed in AA compared with KK and KA genotypes. The gene sets featured several functional related genes that play a crucial role in TAG synthesis, such as glycerol-3-phosphate acyltransferase mitochondrial (*GPAM*), 1-acylglycerol-3-phosphate *O*-acyltransferase 4 (*AGPAT4*), and insulin-induced gene 1 (*INSIG1*). Also, several genes involved in lipid metabolism [i.e., FA synthase (*FASN*) and stearoyl-CoA desaturase 1 (*SCD1*)] were found to be downregulated in AA compared with KA and KK genotype animals. Although none of those individual genes in the pathway met the threshold for statistical significance after adjustment for multiple testing between genotypes (FDR <0.10), the pathway did. This suggested that pathway approaches can identify genes that show moderated but consistent and coordinated differentiation in expression levels. Furthermore, this pattern of gene expression suggests that compared with the K allele, the A allele is associated with reduced energy metabolism. Although the set of lipid metabolism genes in AA genotyped cows was downregulated, their milk fat content was only numerically lower than in KK genotyped cows, probably as a result of the limited sample size used.

It is noteworthy that we also identified immune system pathways that were sensitive to *DGAT1* polymorphism. Such an effect of *DGAT1* gene polymorphism in the mammary gland in dairy cows has never been reported before. Cows with the KK genotype revealed enrichment in gene sets known to be involved in immune responses compared with the AA and KA genotypes. It featured several genes involved in histocompatibility complex class II. Some of these immune related genes, including genes encoding complement component 5a receptor 1 (*C5AR1*), IFNγ (*IFNG*), interleukins (*IL-1B*, *IL-8*, *IL-6*, and *IL-8RA*), LPS-binding protein (*LBP*) and CD14 antigen (*CD14*), have been found associated with bovine mastitis (Ogorevc et al., 2009). Our observations indicate that *DGAT1* gene polymorphism relates to modifications in the immune response transcripts. The mechanism for these changes in immune response gene expressions in the mammary gland is unknown and requires further investigation. The current findings raise an interesting correspondence with those obtained for lipid mobilization in transition dairy cows. Contreras and Sordillo (2011) suggested that in transition dairy cows, excessive accumulation of lipid components in the hepatocytes and other cells could cause physical damage, including compression and reduction in size and number of organelles, and could induce programmed cell death or apoptosis, intracellular signaling, and alter the expression of proinflammatory mediators through lipid mediator biosynthetic pathways. Based on the observed effects, it is likely that *DGAT1* polymorphism affects the concentration and composition of TAG accumulated in the matrix and lumen of the mammary gland (Koliwad et al., 2010). In turn, this accumulation may not only alter cellular biosynthesis pathways related to lipid metabolism, but also modifies tissue-based processes involved in immune system response or tissue remodeling. This intriguing possibility remains to be demonstrated, although Loor et al. (2005) and Loor et al. (2006) suggested a causal link between hepatic inflammation and liver TAG accumulation in periparturient dairy cows through the activation of genes that have important roles in aspects of cell proliferation and immune response. Additionally, Bionaz et al. (2012) recently reported that also in bovine kidney cells, saturated long-chain FA not only modulate the expression of genes involved in lipid metabolism, but also of genes involved in immune response through peroxisome proliferator-activated receptor α (*PPAR*α) -mediated activation.

CONCLUSIONS

Results of the present study suggest pleiotropic effects of *DGAT1* gene polymorphism on the energy metabolism and immune system in the transcriptome of the mammary gland. These results provide a first overview from which new research strategies can be launched to elucidate the underlying molecular mechanisms for *DGAT1* gene polymorphism and bovine mammary function.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of WUR investment funds (BAS code KB-05-003-040-ASG-V-2), and of "Comissionat per a Universitats i Recerca del DIUE" of the Government of Catalonia. The authors are also grateful to Dirk Anjema, Paul Kroon, Johan Meijer, and Peter Jan ten Haken from Wageningen UR Livestock Research (Lelystad, the Netherlands) for surgical biopsies, and to Karel Houwelingen from Wageningen UR Livestock Research and Jordi Estellé from INRA (Dep. Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France) for critical comments.

REFERENCES

Banos, G., J. A. Woolliams, B. W. Woodward, A. B. Forbes, and M. P. Coffey. 2008. Impact of single nucleotide polymorphisms in leptin, leptin receptor, growth hormone receptor, and diacylglycerol acyltransferase (DGAT1) gene loci on milk production, feed, and body energy traits of UK dairy cows. J. Dairy Sci. 91:3190–3200.

Berry, D. P., D. Howard, S. O'Boyle, S. Waters, J. F. Kearney, and M. McCabe. 2010. Associations between the K232A polymorphism in the diacylglycerol-*O*-transferase 1 (DGAT1) gene and performance in Irish Holstein-Friesian dairy cattle. Ir. J. Agric. Food Res. 49:1–9.

Bionaz, M., B. J. Thering, and J. J. Loor. 2012. Fine metabolic regulation in ruminants via nutrient-gene interactions: Saturated long-chain fatty acids increase expression of genes involved in lipid metabolism and immune response partly through PPAR-α activation. Br. J. Nutr. 107:179–191.

Blum, Y., G. Le Mignon, S. Lagarrigue, and D. Causeur. 2010. A factor model to analyze heterogeneity in gene expression. BMC Bioinformatics 11:368.

Boichard, D., C. Grohs, F. Bourgeois, F. Cerqueira, R. Faugeras, A. Neau, R. Rupp, Y. Amigues, M. Boscher, and H. Levéziel. 2003. Detection of genes influencing economic traits in three French dairy cattle breeds. Genet. Sel. Evol. 35:77–101.

Bouwman, A. C., H. Bovenhuis, M. H. Visker, and J. A. van Arendonk. 2011. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet. 12:43–55.

Bünger, M., H. M. van den Bosch, J. van der Meijde, S. Kersten, G. J. E. J. Hooiveld, and M. Müller. 2007. Genome-wide analysis of PPARα activation in murine small intestine. Physiol. Genomics 30:192–204.

Caesar, R., M. Manieri, T. Kelder, M. Boekschoten, C. Evelo, M. Müller, T. Kooistra, S. Cinti, R. Kleemann, and C. A. Drevon. 2010. A combined transcriptomics and lipidomics analysis of subcutaneous, epididymal and mesenteric adipose tissue reveals marked functional differences. PLoS ONE 5:e11525. http://dx.doi.org/10.1371/journal.pone.0011525.

Causeur, D., C. Friguet, M. Houee-Bigot, and M. Kloareg. 2011. Factor analysis for multiple testing (FAMT): An R package for large-scale significance testing under dependence. J. Stat. Softw. 40:1–19.

Connor, E. E., S. Siferd, T. H. Elsasser, C. M. Evock-Clover, C. P. Van Tassell, T. S. Sonstegard, V. M. Fernandes, and A. V. Capuco. 2008. Effects of increased milking frequency on gene expression in the bovine mammary gland. BMC Genomics 9:362.

Contreras, G. A., and L. M. Sordillo. 2011. Lipid mobilization and inflammatory responses during the transition period of dairy cows. Comp. Immunol. Microbiol. Infect. Dis. 34:281–289.

Danchin-Burge, C., S. J. Hiemstra, and H. Blackburn. 2011. Ex situ conservation of Holstein-Friesian cattle: Comparing the Dutch, French, and US germplasm collections. J. Dairy Sci. 94:4100–4108.

Donaldson, L., T. Vuocolo, C. Gray, Y. Strandberg, A. Reverter, S. McWilliam, Y. Wang, K. Byrne, and R. Tellam. 2005. Construction and validation of a bovine innate immune microarray. BMC Genomics 6:135.

Farr, V. C., K. Stelwagen, L. R. Cate, A. J. Molenaar, T. B. McFadden, and S. R. Davis. 1996. An improved method for the routine biopsy of bovine mammary tissue. J. Dairy Sci. 79:543–549.

Friguet, C., M. Kloareg, and D. Causeur. 2009. A factor model approach to multiple testing under dependence. J. Am. Stat. Assoc. 104:1406–1415.

Gautier, M., A. Capitan, S. Fritz, A. Eggen, D. Boichard, and T. Druet. 2007. Characterization of the DGAT1 K232A and variable number of tandem repeat polymorphisms in French dairy cattle. J. Dairy Sci. 90:2980–2988.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim, A. Kvasz, M. Mni, P. Simon, J. M. Frere, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the *DGAT1* K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. USA 101:2398–2403.

Koliwad, S. K., R. S. Streeper, M. Monetti, I. Cornelissen, L. Chan, K. Terayama, S. Naylor, M. Rao, B. Hubbard, and R. V. Farese Jr.. 2010. *DGAT1*-dependent triacylglycerol storage by macrophages protects mice from diet-induced insulin resistance and inflammation. J. Clin. Invest. 120:756–767.

Lê, S., J. Josse, and F. Husson. 2008. FactoMineR: An R package for multivariate analysis. J. Stat. Softw. 25:1–18.

Loor, J. J., H. M. Dann, R. E. Everts, R. Oliveira, C. A. Green, N. A. Guretzky, S. L. Rodriguez-Zas, H. A. Lewin, and J. K. Drackley. 2005. Temporal gene expression profiling of liver from periparturient dairy cows reveals complex adaptive mechanisms in hepatic function. Physiol. Genomics 23:217–226.

Loor, J. J., H. M. Dann, N. A. Guretzky, R. E. Everts, R. Oliveira, C. A. Green, N. B. Litherland, S. L. Rodriguez-Zas, H. A. Lewin, and J. K. Drackley. 2006. Plane of nutrition prepartum alters hepatic gene expression and function in dairy cows as assessed by longitudinal transcript and metabolic profiling. Physiol. Genomics 27:29–41.

Mach, N., A. A. A. Jacobs, L. Kruijt, J. van Baal, and M. A. Smits. 2011. Alteration of gene expression in mammary gland tissue of dairy cows in response to dietary unsaturated fatty acids. Animal 5:1217–1230.

McCudden, C. R., M. D. Hains, R. J. Kimple, D. P. Siderovski, and F. S. Willard. 2005. G-protein signaling: Back to the future. Cell. Mol. Life Sci. 62:551–577.

Odong, T. L., J. van Heerwaarden, J. H. Jansen, T. J. L. van Hintum, and F. A. van Eeuwijk. 2011. Determination of genetic structure of germplasm collections: Are traditional hierarchical clustering methods appropriate for molecular marker data? Theor. Appl. Genet. 123:195–205.

Ogorevc, J., T. Kunej, A. Razpet, and P. Dovc. 2009. Database of cattle candidate genes and genetic markers for milk production and mastitis. Anim. Genet. 40:832–851.

Schennink, A., J. M. L. Heck, H. Bovenhuis, M. H. P. W. Visker, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2008. Milk fatty acid unsaturation: Genetic parameters and effects of stearoyl-CoA desaturase (*SCD1*) and acyl CoA: Diacylglycerol acyltransferase 1 (*DGAT1*). J. Dairy Sci. 91:2135–2143.

Schennink, A., W. M. Stoop, M. H. P. W. Visker, J. M. L. Heck, H. Bovenhuis, J. J. van der Poel, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2007. *DGAT1* underlies large genetic variation in milk-fat composition of dairy cows. Anim. Genet. 38:467–473.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102:15545–15550.

Thaller, G., W. Kramer, A. Winter, B. Kaupe, G. Erhardt, and R. Fries. 2003. Effects of *DGAT1* variants on milk production traits in German cattle breeds. J. Anim. Sci. 81:1911–1918.

1.3 Décomposition d'un caractère complexe en sous-types d'animaux ayant des profils transcriptomiques homogènes

Cette partie correspond :

– à l'article : **Blum Y**, Le Mignon G, Causeur D, Filangi O, Désert C, Demeure O, Le Roy P, Lagarrigue S. Complex trait subtypes identification using transcriptome profiling reveals an interaction between two QTL affecting adiposity in chicken. BMC Genomics, 2011, 12:567

– au travaux effectués lors de la mobilité à l'étranger de janvier à avril 2012 qui concerne une application de la stratégie de décomposition du caractère sur un croisement de souris F2.

1.3.1 Article 4 : Blum *et al.*, BMC Genomics, 2011

Apport de l'article

Les caractères tels que l'adiposité sont dits complexes étant donné le grand nombre de processus biologiques sous-jacents. Il existe différents chemins biologiques pour un phénotype donné. Cette complexité que l'on peut assimiler à de l'hétérogénéité entre individus peut être oservée grâce aux données transcriptomiques. Les animaux ayant le même phénotype mais des profils d'expression hétérogènes peuvent être classés en sous-types d'animaux plus homogènes en terme de profils d'expression et aussi probablement au niveau de leur génétique. Il a été montré que prendre en compte ces sous-types d'animaux dans les analyses QTL peut permettre de préciser la localisation de certains QTL, voire d'en révéler de nouveaux (Schadt et al. (2003), Le Mignon et al. (2009)). Cette approche peut être vue comme une décomposition du caractère en sous-types phénotypiques. Elle consiste à détecter au préalable une liste de gènes dont l'expression est corrélée au caractère, puis sur la base de ces gènes, d'identifier des sous-types d'animaux pour le caractère d'intérêt ayant des profils transcriptomiques homogènes. Dans cette étude, on utilise la méthode FAMT (Friguet et al. (2009)) pour rechercher les gènes liés au caractère. A partir des expressions ajustées par les facteurs d'hétérogénéité (cf. section 1.2), on peut alors classer les animaux afin de trouver des sous-types pour le caractère d'intérêt. Cette approche est appliquée à une étude concernant une famille de 45 poulets variables pour le gras abdominal (Le Mignon et al. (2009), Blum et al. (2010)). On montre qu'il existe deux sous-types d'animaux parmi les animaux maigres. En enlevant un de ces sous-types dans les analyses QTL, on fait apparaitre un nouveau QTL contrôlant le poids de gras abdominal. De plus, on montre que ce QTL est en interaction avec un autre QTL sur le même chromosome qui est quant à lui détecté en utilisant l'ensemble du dispositif.

Blum et al. BMC Genomics 2011, 12:567
http://www.biomedcentral.com/1471-2164/12/567

BMC Genomics

RESEARCH ARTICLE Open Access

Complex trait subtypes identification using transcriptome profiling reveals an interaction between two QTL affecting adiposity in chicken

Yuna Blum[1,2,3], Guillaume Le Mignon[1,2,4], David Causeur[3], Olivier Filangi[1,2], Colette Désert[1,2], Olivier Demeure[1,2], Pascale Le Roy[1,2] and Sandrine Lagarrigue[1,2*]

Abstract

Background: Integrative genomics approaches that combine genotyping and transcriptome profiling in segregating populations have been developed to dissect complex traits. The most common approach is to identify genes whose eQTL colocalize with QTL of interest, providing new functional hypothesis about the causative mutation. Another approach includes defining subtypes for a complex trait using transcriptome profiles and then performing QTL mapping using some of these subtypes. This approach can refine some QTL and reveal new ones. In this paper we introduce Factor Analysis for Multiple Testing (FAMT) to define subtypes more accurately and reveal interaction between QTL affecting the same trait. The data used concern hepatic transcriptome profiles for 45 half sib male chicken of a sire known to be heterozygous for a QTL affecting abdominal fatness (AF) on chromosome 5 distal region around 168 cM.

Results: Using this methodology which accounts for hidden dependence structure among phenotypes, we identified 688 genes that are significantly correlated to the AF trait and we distinguished 5 subtypes for AF trait, which are not observed with gene lists obtained by classical approaches. After exclusion of one of the two lean bird subtypes, linkage analysis revealed a previously undetected QTL on chromosome 5 around 100 cM. Interestingly, the animals of this subtype presented the same q paternal haplotype at the 168 cM QTL. This result strongly suggests that the two QTL are in interaction. In other words, the "q configuration" at the 168 cM QTL could hide the QTL existence in the proximal region at 100 cM. We further show that the proximal QTL interacts with the previous one detected on the chromosome 5 distal region.

Conclusion: Our results demonstrate that stratifying genetic population by molecular phenotypes followed by QTL analysis on various subtypes can lead to identification of novel and interacting QTL.

Background

In the last decade, integrative genomics approaches that take into account genotypic, molecular profiling and complex traits in segregating populations have been developed to dissect the genetics of complex traits such as human diseases or economically important traits in livestock or plants. Combining QTL mapping and high throughput transcriptome data is proving useful for characterizing QTL regions and elucidating genes and biological pathways that affect complex traits [1-9]. The term "Genetical Genomics" or "Systems Genetics" refers to such a combinatorial approach.

One strategy commonly used by authors working in this context was based on the identification of genes having an eQTL that colocalizes with the QTL responsible for the complex trait of interest. Such a strategy considers the expression level of each gene available on a microarray as a quantitative trait and uses genetic markers to identify genomic regions that regulate gene expression phenotypes; these regions are named eQTL (expression Quantitative Trait Loci). The function of the gene that its mRNA level is controlled by a region can provide new functional information about the candidate gene sought in the eQTL region. Colocalization of

* Correspondence: sandrine.lagarrigue@agrocampus-ouest.fr
[1]INRA, UMR598, Génétique Animale, IFR140 GFAS, 35000 Rennes, France
Full list of author information is available at the end of the article

eQTL with the QTL for complex trait can provide relevant information about the causative gene for the complex trait of interest. This strategy has been widely used in various species (flies [1,10], mice [2-4], rats [5], human [6], eucalyptus [7], Arabidopsis [8], livestock species [9,11] has been reported). When combined with mathematical modeling proposed by Schadt *et al*. [3], this strategy becomes very efficient for distinguishing causal from reactive genes for the complex trait and for identifying the "driver" genes and pathways that are responsible for a complex trait.

Another strategy is based on defining subtypes for a complex trait using gene expression profiles. It is well known that a population measured for a complex trait through one criteria (for example, Body mass index for obesity) may actually have distinct molecular subtypes for this complex phenotype. Use of gene expression profiles may allow the identification of such biologically distinct subtypes. The standard procedure is to identify genes whose expression is correlated to the complex trait and then perform a classification of individuals in order to observe specific subtypes. Applied on a segregating population, the identification of subtypes combined with QTL analysis performed for these subtypes can separately improve sensitivity of QTL detection and reveal new loci. This strategy was first performed by Schadt *et al*. (2003) [4] using a mouse population and then in 2009 by our team using a chicken segregating population [12]. In these two studies, two QTL were observed for the fat mass, one initially observed on the full F2 set and another one only observed when one subtype was removed. As illustrated by these studies, the core of the approach is the determination of subtypes within a segregating population on the basis of the genes correlated to the complex trait. In the present paper, we propose to identify these genes using a method called Factor Analysis for Multiple Testing (FAMT) which takes into account the hidden dependence structure that may result from population structure or/and technical artefact of gene expression profiling experiment, independent of the trait of interest ([11], [13]). We then show the utility of this method to define phenotype subtypes more accurately and to reveal interaction between 2 QTL.

Results and discussion

Identification of animal subtypes for fatness trait using the FAMT method

The first step was to identify which of the 11213 genes expressed in the liver were correlated to the trait of interest, the abdominal fat weight (AF), in the 45 related offspring's. Pearson correlation between hepatic transcript levels and AF trait identified 287 genes significantly associated at the nominal p-value of 0.05 without any correction for multiple tests. To increase the size of this list, Le Mignon *et al*. [12] added to this first list, genes significantly differentially expressed between the 10 leanest and fattest birds in the family. As such, a list of 660 genes was obtained with a significance threshold of 0.05 (Student's t-test p-value and Pearson correlation test p-value) without any correction for multiple tests. It should be noted that applying correction for multiple testing resulted in no gene being differentially expressed. This result might be explained by either a poor genetic variability between animals, which are half sib offsprings, or dependence between genes. Indeed, standard methods to find significant correlation between gene expressions and a variable of interest ignore the correlations among expression profiles [14]. This dependence structure leads to correlation among test statistics, which leads to under representation of the smallest p-values [15]. This can be explained by a number of unmeasured or unmodeled factors independent of the variable of interest (in our study, the AF trait) that may influence the expression of any particular gene ([16], [13]). These factors may induce additional variability in the expression levels and decrease the power to detect the true correlation with the variables of interest. Recently, several studies have introduced models taking into account this gene dependence. In particular, Friguet *et al*. [13] propose to model this sharing of information by a factor analysis structure in a method called Factor Analysis for Multiple Testing (FAMT). The estimated factors in the model capture components of the expression heterogeneity independent of the effects of the variable of interest. We applied this method to our data: 688 transcripts significantly correlated to the AF trait were identified taking into account the existence of six factors containing a common information shared by all genes and independent from the AF trait. The interpretation of these factors was analyzed and discussed in Blum *et al*. [11]. For the further analyses in the present paper, we subtracted the linear dependence kernel defined by the six factors from the 688 raw gene expressions to obtain 688 factor-adjusted expressions as in Blum *et al*. [11].

The second step was to identify the best gene list that distinguishes potential subtypes for the AF trait within the 45 offspring. Separate hierarchical clustering of the birds was performed using either the 287 (Figure 1A) and the 660 genes obtained by classical methods in step one (Figure 1B), or the 688 genes obtained by the FAMT method (Figure 1C). For the latter we used the FAMT adjusted expression values in the clustering algorithm. The results of the three clusterings are shown in Figure 1. The set of 688 genes is clearly more efficient to separate fat and lean birds, and to generate different subtypes for the AF trait. As indicated in Figure 1C, this gene list allows us to clearly distinguish two subtypes

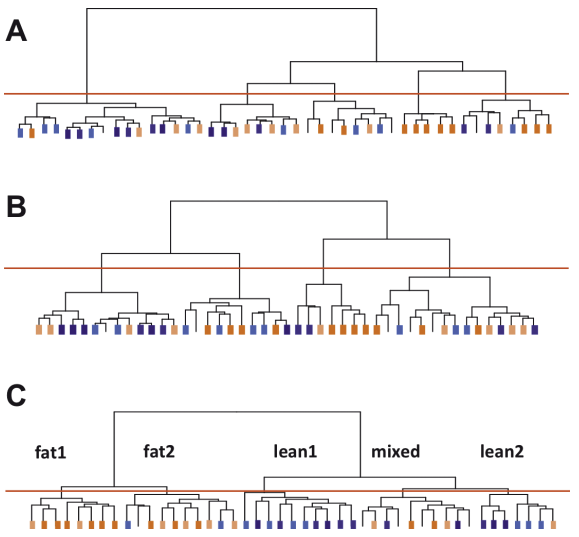


Figure 1 Dendrograms of birds using different gene lists related to fatness and obtained by different statistical methods. **(A)** 287 genes correlated to the fatness trait obtained by a classical approach. **(B)** 660 genes correlated to the fatness trait and/or differentially expressed obtained by classical approaches [12]. **(C)** 688 genes correlated to the fatness trait obtained by FAMT method [11]. Orange/blue bars indicate the 20 fattest/leanest animals respectively. Darkest bars correspond to the 10 extreme animals of the 2 fattest/leanest groups. Colorless bars correspond to the 5 intermediate chickens.

for the fat birds and two other subtypes for the lean birds in addition to one subgroup mixing lean, intermediate and fat birds. This gene list includes almost all of the genes of the 287 genes (93%) but is twice as large. This larger number suggests that correlation between many gene expressions and the variable of interest is underestimated due to the hidden dependence structure. Finally, this gene list is quite different from the 660 gene set with only 69% common genes, suggesting a notable number of false positive genes in the latter due to the absence of correction for multiple testing and for gene dependence.

These results clearly show the importance of taking into account the gene dependence due to additional sources of variation, especially when the expression variation related to the variable of interest may be low and therefore easily impacted by these additional sources.

A new QTL revealed by removing one of the two lean subtypes: genetic characterization of this subtype

Based on the clustering obtained using the FAMT adjusted expression (Figure 2A), we performed linkage analyses for the AF trait on the chromosome 5, either with the whole family or by removing successively one of the five subgroups (Figure 2B). As indicated in Figure 2B, the majority of analyses gave the same LRT curves with the expected AF QTL located around 168 cM on the chromosome 5 [12]. However, after removing the lean subtype called lean2, a new significant QTL (p-value < 0.05) was detected on a proximal chromosome 5 locus around 100 cM with an effect of 1.19 phenotypic standard deviation. Alternative and not necessarily exclusive hypotheses can be drawn to explain the detection of the second QTL after the exclusion of the lean2 group:

1) The first hypothesis is the presence of animals having an AF value in disagreement with the paternal Q/q haplotype in the excluded lean2 group. Removing such birds, especially when their AF values are extreme, can largely increase the power of QTL detection when the design analyzed has a low size. We determined for each offspring the Q/q haplotype corresponding to the proximal QTL (see Methods section). Two out of the 7 birds of the lean2 subtype have the Q paternal haplotype that contributes to a high fat mass (L2 and L7 birds) (Figure 2C). However, we can notice that the lean1 subtype is in the same configuration with 2 extreme lean birds as well (L4 and L5) with the Q haplotype at the new QTL (Table 1) but does not allow to reveal this latter after being removed.

2) The second hypothesis is that the proximal and distal QTL on chromosome 5 interact with each other. In our specific case, this means that the allele configuration of the distal QTL at 168 cM influences the effect of the proximal QTL and therefore masks it when the whole family is used. To investigate this explanation further, we analyzed the paternal haplotype at the QTL around 168 cM for the different birds of the lean2 subtype (Figure 2C). We determined the paternal haplotype for 5 out of the 7 birds belonging to this subtype considering a probability > 99%. Interestingly, all of five birds have the same haplotype q. This observation suggests that the two QTL are interacting: the presence of the Q allele at the distal locus enhances the allelic effect at the proximal QTL and the presence of q allele at the distal locus weakens the allelic effect at the proximal QTL.

Interaction testing between the proximal and distal AF QTL on chromosome 5

Considering a transmission probability greater than 99%, we determined the paternal haplotype for the proximal and distal QTL in 29 birds (40 and 34 birds for the distal QTL (at 168 cM +- 15 cM) and proximal QTL (at 100 cM +- 20 cM)) respectively as shown in Table 1.

Using these 29 birds, we first performed a two-way analysis of variance considering the two QTL as two fixed factors with an interaction between them. As indicated in Figure 3A, the analysis shows clearly a significant interaction between the two QTL (p-value < 0.01). The difference between Q versus q for the proximal

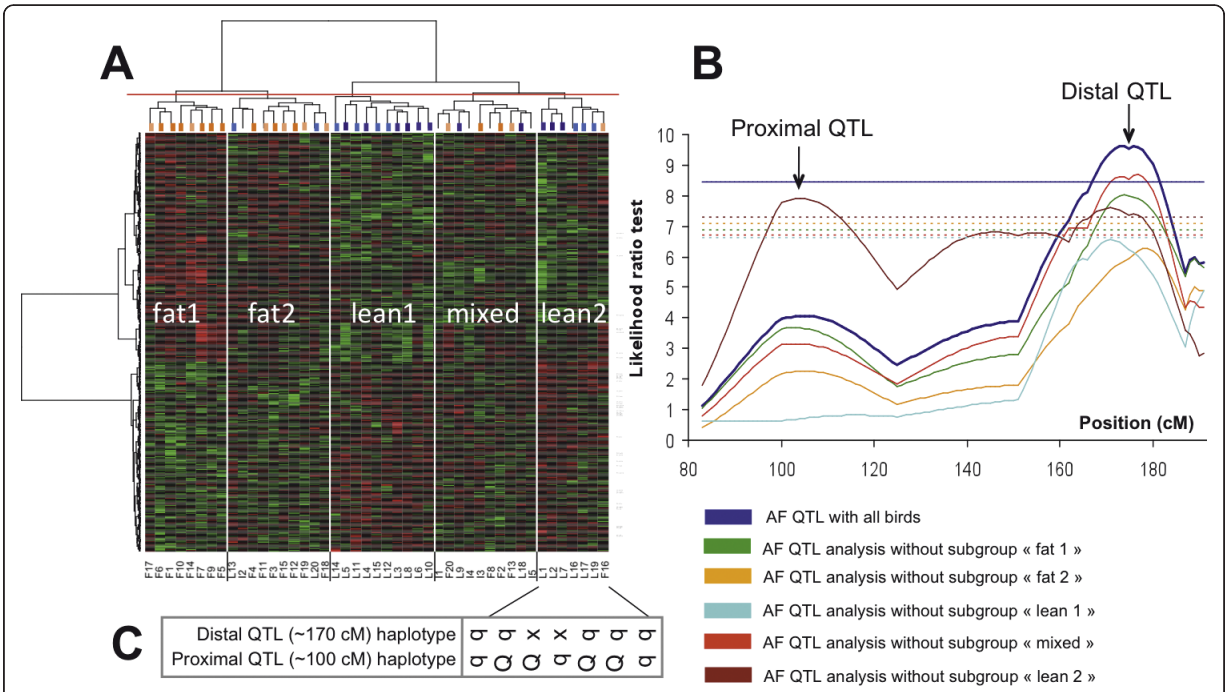


Figure 2 HCA and linkage analysis using subtype combination. **(A)** Heatmap using the 688 genes (Y axis) and the 45 chickens (X axis). **(B)** Interval mapping for the AF trait on chromosome 5, with the whole family (blue) and without one subgroup observed by HCA (other colors). The graph gives the statistical test (Likelihood Ratio Test) related to the test "no QTL" versus "one QTL" (y-axis") for every tested location on the chromosome (centiMorgan, x-axis). The chromosome-wide significance threshold at the 5% level obtained for the analysis including all animals (blue) and for analysis without one of the five subgroups: fat1 (green), fat2 (yellow), lean1 (light blue), mixed (red) or lean2 (brown) are displayed. The genetic distances (cM) and likelihood ratio test (LRT) are shown on the X-axis and Y-axis, respectively. **(C)** The two letters indicate the Q or q haplotype inherited from the sire, with a probability > 99% for the proximal QTL (first letter) or the distal QTL (second letter); × indicates a probability < 99%. The probability of inheriting the paternal Q versus q haplotype was calculated only on the basis of marker information in the region of interest.

QTL is higher when the haplotype is Q at the distal QTL (+15g) than when the haplotype is q (-4g).

We also tested the QTL interaction using the QTLMap software with the "interaction model" ([17], [18]). The procedure tests the model: "No QTL" versus "1 QTL in interaction with another known QTL". We chose to set the location of the distal QTL at 168 cM, corresponding to the maximum LRT. Compared to the analysis of variance, the advantage of this QTL analysis is to set the location for only one of the two QTL presumed in interaction, increasing the number of birds analyzed (40 versus 29 animals) and then allowing to better localize the second QTL. As depicted in Figure 3B, the green curve corresponding to the interaction model analysis shows clearly a significant QTL in the proximal region (p-value < 0.05) in interaction with the fixed QTL at 168 cM. Furthermore, an additive model testing the hypothesis "one QTL" versus "2 QTL" does not highlight the proximal QTL (grey curve, Figure 3B), which is consistent with our expectation that the two QTL are in interaction.

To obtain a better estimate of the proximal QTL location, we developed six novel informative SNP markers in the proximal region at 67, 77, 80, 86, 89 and 95 cM respectively and genotyped the 40 animals accordingly. As indicated in Figure 3B, where the red curve corresponds to the interaction model performed with additional markers, the most probable position of the proximal QTL in interaction with the distal QTL on the chromosome 5 was found at 85 cM (p-value < 0.05) with a Confidence Interval (CI) from 78 to 102 cM.

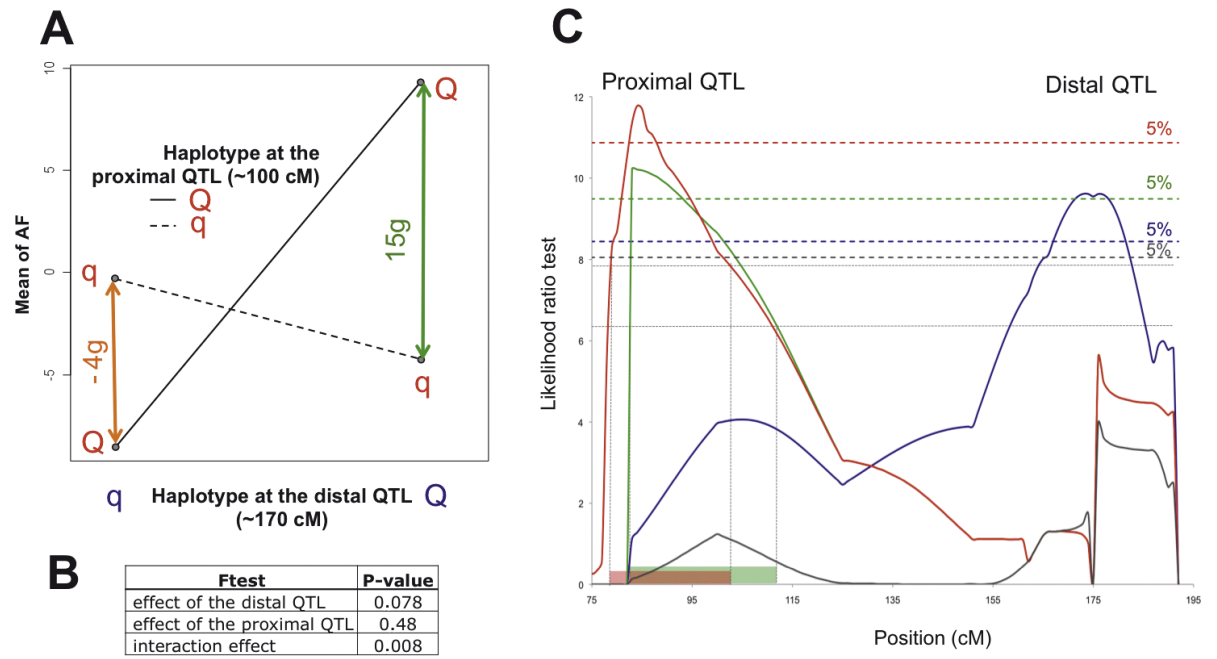
Among the selected 688 genes, we identified 4 genes having a similar QTL profile as the abdominal fatness trait on the chromosome 5 (Table 2). These genes have a distal eQTL colocalizing with the AF distal region (observed by using the classical QTL additive model). They also have a proximal QTL colocalizing with the AF proximal region, with an interaction with the distal eQTL (p-value < 0.1 by using the "interaction model" of QTLMap software and the novel markers). Interestingly, one of these genes has the highest correlation with the AF trait (-0.58 Pearson correlation coefficient).

Table 1 Haplotype determination for the distal and proximal AF QTL on chromosome 5.

Animal ID	Subgroup	Proximal AF QTL haplotype	Distal AF QTL haplotype	Both AF QTL haplotypes
L1	lean2	q	q	q-q
L2	lean2	Q	q	Q-q
L3	lean1	q	Q	q-Q
L4	lean1	Q	q	Q-q
L5	lean1	Q	q	Q-q
L6	lean1	q	x	X
L7	lean2	Q	x	X
L8	lean1	q	Q	q-Q
L9	mixed	q	q	q-q
L10	lean1	x	q	X
L11	lean1	x	q	X
L12	lean1	q	q	q-q
L13	fat2	x	Q	X
L14	lean1	Q	q	Q-q
L15	lean1	x	Q	X
L16	lean2	q	x	X
L17	lean2	Q	q	Q-q
L18	mixed	Q	Q	Q-Q
L19	lean2	Q	q	Q-q
L20	fat2	q	q	q-q
I1	mixed	x	Q	X
I2	fat2	x	Q	X
I3	mixed	q	q	q-q
I4	mixed	Q	Q	Q-Q
I5	mixed	Q	q	Q-q
F1	fat1	Q	Q	Q-Q
F2	mixed	Q	Q	Q-Q
F3	fat2	Q	Q	Q-Q
F4	fat2	Q	Q	Q-Q
F5	fat1	Q	x	X
F6	fat1	q	q	q-q
F7	fat1	Q	Q	Q-Q
F8	mixed	x	Q	X
F9	fat1	Q	x	X
F10	fat1	q	q	q-q
F11	fat2	x	q	X
F12	fat2	Q	Q	Q-Q
F13	mixed	q	q	q-q
F14	fat1	x	Q	X
F15	fat2	x	Q	X
F16	lean2	q	q	q-q
F17	fat1	x	Q	X
F18	fat2	Q	Q	Q-Q
F19	fat2	Q	Q	Q-Q
F20	mixed	q	Q	q-Q
		34 animals	40 animals	29 animals

Animal labels F1 to F20 indicate the 20 fattest chickens, L1 to L20 the 20 leanest chickens and I1 to I5 the 5 intermediate chickens.

The two letters indicate the Q or q haplotype inherited from the sire, with a probability > 99% for the QTL at 84 cM (first column) or the QTL at 168 cM (second column); x indicates a probability < 99%. The probability of inheriting the paternal Q versus q haplotype was calculated by QTLMap software only on the basis of marker information in the region of interest.



Ftest	P-value
effect of the distal QTL	0.078
effect of the proximal QTL	0.48
interaction effect	0.008

Figure 3 Interaction testing. **(A)** Interaction graph is given by a two-way analysis of variance considering the two QTL as two fixed factors with an interaction between them. The "Q-q" QTL effect at the proximal QTL is more important when the animals have the Q allele at the distal QTL (-4g/+15g with the q/Q allele at the distal QTL respectively). **(B)** The table gives the resulting p-value for both factors and the interaction. **(C)** QTL analysis using the "interaction model": the blue curve corresponds to the classical model testing the hypothesis « No QTL versus 1 QTL», the grey curve to the model testing «One QTL versus 2 QTL», the green curve corresponds to the interaction model « No QTL versus 1 QTL in interaction with another one fixed at 168 cM » and the red corresponds to the interaction model after adding the six novel markers. For each case, the significance threshold at the 0.05 level is displayed.

Moreover, all 4 genes were differentially expressed (p-value < 0.1) between the two lean subtypes previously detected (lean1 and lean2). This observation can be interpreted as another illustration of the interaction effect between the proximal and distal AF QTL, but at the gene expression level. Taken together, these observations suggest that at least one of these 4 genes may be a signature of the causative mutation underlying the adiposity trait. These genes produce unknown proteins and/or proteins not particularly related to the adiposity. Further investigations will be necessary to confirm such a signature and clarify the role of these genes in lipid metabolism and adiposity.

Conclusion

In this study, we show the value of determining phenotype subtypes underlying a complex trait by using gene expressions. This subtype identification combined with QTL mapping improves the characterization of QTL responsible for adiposity, by revealing a new QTL in interaction with a previous one. This study also highlights the interest to use FAMT procedure to define

Table 2 Genes for which RNA level is controlled by the two proximal and distal regions in interaction similarly to the adiposity phenotype

			Classical Model		Interaction Model		Test Lean1/lean2	
oligo ID	HGNC	corr	location	maxLRT	location	maxLRT	DE	p-value
RIGG00027	BPNT1	-0.36	182	9.6*	95	9.8^{+}	-/+	+
RIGG05332	NULL	-0.39	178	6.3^{+}	77	11.5*	-/+	*
RIGG07405	P4HA2	-0.38	185	7.3^{+}	73	12.0*	-/+	+
RIGG12578	NULL	-0.58	158	11.7*	86	9.3^{+}	+/-	+

For each gene, are given the oligonucleotide identifier (oligo ID), the HGNC abbreviation (HGNC), the Pearson correlation with the abdominal fatness (corr), the maximum LRT location in cM and the maximum LRT (maxLRT) using either the classical model or the interaction model. The two last columns are related to the statistical test comparing the expression between the lean1 and the lean2 subtypes: the differential expression between the two subtypes (DE = lean1/lean2) and the p-value associated are indicated. * and +: p-value at the chromosome level <0.05 and <0.1 respectively.

more accurately these subtypes for a complex trait compared to classical methods. At the core of the approach proposed here is the phenotype subtype identification, which is still rarely used in the Genetical Genomics field and was reported once a few years ago by Schadt and colleagues [4]. In our report we show the advantage of using such approach in revealing interaction among QTL and discovery of new QTL underlying complex traits.

Methods

Animal design and microarray setup

Animal design, genotyping and microarray setup are previously described by Le Mignon *et al.* [12]. Briefly, the animal design corresponds to 45 male offspring produced by a sire known to be heterozygous for a QTL affecting abdominal fatness (AF) on chromosome 5 with a location confidence interval extended from 156 cM to 187 cM and a significant effect of 1.03 phenotypic standard deviation. This sire are not heterozygous for other AF QTL on GGA1, GGA3 and GGA7 previously detected in a three-generation F0-F1-F2 design performed by intercrossing two experimental chicken lines divergently selected for abdominal fatness from which the sire has been produced. Genotyping for GGA5 chromosome was performed for 10 markers (ADL0292, ADL0023, MCW0238, ADL0233, MCW0026, SEQF0079, SEQF0080, SEQF0082, SEQF0085, ROS330 at 83, 100, 125, 151, 162, 166, 175, 187, 190, 192 cM respectively). Markers were chosen from available markers [19] or developed for this program [12]. The six additional SNP markers were developed from the chicken genome sequence assembly and correspond to rs15678496, rs15683152, rs15685956, rs16689818, rs15691594, rs14531246 at 67, 77, 80, 86, 89 and 95 cM respectively. Gene expression measurements were obtained from the livers of these animals using a 20 K chicken oligo array (Ark-genomics). 11213 genes (55 % of the 20461 genes) were selected as expressed in the liver. The raw and normalized microarray data were deposited in the Gene Expression Omnibus (GEO) public repository [20]. The accession number for the series is GSE12319 and the sample series can be retrieved with accession numbers GSM309564 to GSM309609.

The animal labels were defined as follows: F1 to F20 for the 20 fattest animals, L1 to L20 for the 20 leanest animals and I for the 5 intermediates.

All experiments were conducted under Licence N°; 37-123 from the Veterinary Services, Indre et Loire, France and in accordance with guidelines for care and use of animals in Agricultural Research and Teaching (French Agricultural Agency and Scientific Research Agency).

Classical expression analysis

As the variable of interest in the biological study is continuous, we calculated the Pearson correlation coefficient for each gene expression and deduced the number of genes correlated to the trait by considering the p-values under the cutoff 0.05. To control the False Discovery Proportion (FDR) we performed the Benjamini-Hochberg correction for multiple testing [21].

Factor analysis method

The method takes into account the gene dependence structure and consequently, the impact of dependence on the multiple testing procedures for high-throughput data. Indeed, genes can have similar expression profiles because they are involved in common pathways but independently of the variable of interest (AF in our case). The common information shared by all the variables (i.e. gene expressions) and independent of the variable of interest is modeled by a factor analysis structure. An EM algorithm is used to estimate the model. Once the factor model is estimated, factor-adjusted test statistics are obtained by correction of the classical tests from the effect of the common factors. David Causeur's team showed that the resulting tests statistics are asymptotically uncorrelated, which improves the overall power of the multiple testing procedure ([13], [22]). The algorithm is implemented in the "FAMT" R package available from CRAN. As in Blum *et al.* [11], the raw expression data set is adjusted for the estimated independent factors, which results in the so-called factor-adjusted expression data.

QTL and eQTL mapping

QTL (eQTL) mapping consists in mapping on the genome, regions that control the variation of a complex trait (expression trait). Before QTL analyses, the AF trait values of the sire family (71 birds) were adjusted for hatch and dam effects by two-way variance analysis, including body weight at slaughter as a covariate (SAS GLM procedure). For the eQTL analyses, no adjustment of the gene variables was performed for hatch and dam effects because of the small size of the population studied (45 birds). QTLMap software based on an interval mapping method described by Elsen *et al.* [23], was used to detect QTL (or eQTL) affecting the AF trait (or a gene expression phenotype). The statistical variable for testing the presence of no QTL (or no eQTL) versus one QTL (or one eQTL) at one location and also of one QTL versus two, was an approximate likelihood ratio test (LRT) [24]. Significance thresholds were empirically determined for AF QTL and transcript level eQTL from 2000 simulations performances assuming a polygenic model with a given heritability (h^2 = 0.5). The widely used "one LOD drop-off method" was applied to obtain 95% confidence intervals of the QTL location [25]. QTLMap software was also used to test an interaction between the proximal and distal QTL using the

"interaction model" testing the hypothesis « No QTL versus 1 QTL in interaction with another one fixed in our study at 168 cM » [18]

We considered that a gene has an eQTL colocalizing with an AF QTL if the CI of the eQTL region was overlapping the CI of the QTL region.

List of abbreviations

AF: Abdominal Fatness; FAMT: Factor Analysis for Multiple Testing; GGA: Gallus Gallus; eQTL: Expression Quantitative Trait Locus; CI: Confidence Interval; GEO: Gene Expression Omnibus; HCA: Hierarchical Cluster Analysis; HGNC: HUGO Gene Nomenclature Committee; EM: Expectation Maximization; LRT: Likelihood Ratio Test.

Acknowledgements

YB is a Ph.D fellow supported by the French Research Ministry and GLM was a Ph.D. fellow supported by the French Technical Institute for Poultry (ITAVI). The research program was supported by grants from a French society for genomics in poultry (AGENAVI), INRA and the Agence Nationale de la Recherche (Grant N°0426). Genotyping was performed at Toulouse-Midi-Pyrénées Genopole (France). The authors thanks Frédérique Pitel, Jake A. Lusis and Anatole Ghazalpour for their comments and for editing the English.

Author details

[1]INRA, UMR598, Génétique Animale, IFR140 GFAS, 35000 Rennes, France. [2]Agrocampus Ouest, UMR598, Génétique Animale, IFR140 GFAS, 35000 Rennes, France. [3]Agrocampus Ouest, Applied Mathematics Department, 35000 Rennes, France. [4]ITAVI, F-75008, Paris, France.

Authors' contributions

GLM, CD and SL provided the transcriptomic data set. FP and OD performed genotyping. YB and GLM analyzed the expression data sets supervised by SL and DC. YB and GLM carried out the QTL and the eQTL mapping analyses supervised by SL, PL, OF and OD. SL and YB drafted the manuscript. SL supervised the project. All authors read and approved the final manuscript.

Received: 7 September 2011 Accepted: 21 November 2011
Published: 21 November 2011

References

1. Wayne ML, Pan YJ, Nuzhdin SV, McIntyre LM: **Additivity and trans-acting effects on gene expression in male Drosophila simulans.** *Genetics* 2004, **168(3)**:1413-1420.
2. Ghazalpour A, Wang X, Lusis AJ, Mehrabian M: **Complex inheritance of the 5-lipoxygenase locus influencing atherosclerosis in mice.** *Genetics* 2006, **173(2)**:943-951.
3. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, *et al*: **An integrative genomics approach to infer causal associations between gene expression and disease.** *Nat Genet* 2005, **37(7)**:710-717.
4. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, *et al*: **Genetics of gene expression surveyed in maize, mouse and man.** *Nature* 2003, **422(6929)**:297-302.
5. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, *et al*: **Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease.** *Nat Genet* 2005, **37(3)**:243-253.
6. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, *et al*: **Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics.** *Proc Natl Acad Sci USA* 2003, **100(2)**:605-610.
7. Kirst M, Myburg AA, De Leon JP, Kirst ME, Scott J, Sederoff R: **Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus.** *Plant Physiol* 2004, **135(4)**:2368-2378.
8. DeCook R, Lall S, Nettleton D, Howell SH: **Genetic regulation of gene expression during shoot development in Arabidopsis.** *Genetics* 2006, **172(2)**:1155-1164.
9. Ponsuksili S, Jonas E, Murani E, Phatsara C, Srikanchai T, Walz C, Schwerin M, Schellander K, Wimmers K: **Trait correlated expression combined with expression QTL analysis reveals biological pathways and candidate genes affecting water holding capacity of muscle.** *BMC Genomics* 2008, **9**:367.
10. Wayne ML, McIntyre LM: **Combining mapping and arraying: An approach to candidate gene identification.** *Proc Natl Acad Sci USA* 2002, **99(23)**:14903-14906.
11. Blum Y, Le Mignon G, Lagarrigue S, Causeur D: **A factor model to analyze heterogeneity in gene expression.** *BMC Bioinformatics* 2010, **11**:368.
12. Le Mignon G, Desert C, Pitel F, Leroux S, Demeure O, Guernec G, Abasht B, Douaire M, Le Roy P, Lagarrigue S: **Using transcriptome profiling to characterize QTL regions on chicken chromosome 5.** *BMC Genomics* 2009, **10**:575.
13. Friguet C CD: **A Factor Model Approach to Multiple Testing Under Dependence.** *Journal of the American Statistical Association* 2009, **104(488)**:1406-1415.
14. Kustra R, Shioda R, Zhu M: **A factor analysis model for functional genomics.** *BMC Bioinformatics* 2006, **7**:216.
15. Leek JT, Storey JD: **A general framework for multiple testing dependence.** *Proc Natl Acad Sci USA* 2008, **105(48)**:18718-18723.
16. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis.** *PLoS Genet* 2007, **3(9)**:1724-1735.
17. Elsen JM MB, Goffinet B, Boichard D, Le Roy P: **Alternatives models for QTL detection in livestock.I.General introduction.** *Genetic Selection Evolution* 1999, **31**:213-224.
18. Filangi O MC, Gilbert H, Legara A, Le Roy P, Elsen JM: **QTLMap software in outbred populations.** *9th World Congress of genetics applied to livestock production, German Society for Animal Science* 2010, D787.
19. Groenen MA, Cheng HH, Bumstead N, Benkel BF, Briles WE, Burke T, Burt DW, Crittenden LB, Dodgson J, Hillel J, *et al*: **A consensus linkage map of the chicken genome.** *Genome Res* 2000, **10(1)**:137-147.
20. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles–database and tools update.** *Nucleic Acids Res* 2007, , **35 Database:** D760-765.
21. Benjamini YHY: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society* 1995, **B 57**:289-300.
22. Causeur D FC, Houée-Bigot M, Kloareg M: **Factor Analysis for Multiple Testing (FAMT): An R Package for Large-Scale Significance Testing Under Dependence.** *Journal of Statistical Software* 2011, **40(14)**:1-19.
23. Elsen JM, Mangin B, Goffinet B, Boichard D, Le Roy P: **Alternatives models for QTL detection in livestock.I.General introduction.** *Genetic Selection Evolution* 1999, **31**:213-224.
24. Le Roy P, Elsen JM, Boichard D, Mangin M, Bidanel JP, Goffinet B: **An algorithm for QTL detection in mixture of full and half sib families.** *6th World Congress of Genetic Applied to Livestock Production: 1998; University of Nex England, Armidale* 1998, 257-260.
25. Lander ES, Botstein D: **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps.** *Genetics* 1989, **121(1)**:185-199.

doi:10.1186/1471-2164-12-567
Cite this article as: Blum *et al.*: Complex trait subtypes identification using transcriptome profiling reveals an interaction between two QTL affecting adiposity in chicken. *BMC Genomics* 2011 **12**:567.

1.3.2 Mobilité à UCLA

Revisiting BxD F2 db/db cross data taking into account the heterogeneity in gene expression for QTL analyses: decomposition of the glucose level trait

Résumé des travaux

Une étude à été menée par l'équipe de Jake Lusis sur 435 souris F2 déficientes au récepteur de la leptine, afin d'identifier les facteurs génétiques responsables de la susceptibilité au diabète de type II. Les profils transcriptomiques hépatiques ainsi que plusieurs caractères liés au diabète (taux d'insuline, de triglycéride et de glucose plasmatique) ont été mesurés sur ces souris. Cette étude, en cours de publication à mon arrivée en janvier 2012, a mis en évidence des QTL pour le taux de triglycéride et d'insuline, mais de manière surprenante, aucun QTL significatif n'a été détecté pour le taux de glucose (Davis et al. (2012)).

Un des objectifs de mon séjour était de tester la stratégie de décomposition du caractère sur ce jeu de données ayant l'avantage de porter sur plus d'animaux que l'étude précédente (Blum et al. (2011)). Appliquée au taux de glucose, on montre que la stratégie de décomposition du caractère basée sur les données d'expressions ajustées par les facteurs d'hétérogénéité permet de faire apparaître deux régions QTL pour ce caractère. Une de ces régions co-localise avec un QTL pour le taux d'insuline sur le chromosome 5 et l'autre avec un QTL pour le taux de triglycéride sur le chromosome 4, montrant l'intérêt de notre approche. A noter que le QTL sur le chromosome 4 a également été détecté pour le taux de glucose et de triglycéride sur un dispositif similaire de susceptibilité au diabète de type II (Scherneck et al. (2009)).

Afin d'identifier les facteurs génétiques prédisposant au diabète de type II, on s'intéresse aux souris F2 déficientes au récepteur de la leptine. Il a été montré qu'une déficience de ce récepteur (db/db) entraîne une hyperphagie et une obésité chez les deux lignées de souris C57BL/6 et DBA/2, cependant la lignée DBA/2 engendre un diabète sévère contrairement à la lignée C57BL/6 (Leiter et al. (1981)). L'étude porte sur des souris F2 (db/db) issues du croisement de ces deux lignées (figure 1.1). Les mesures ont été effectuées sur 435 souris F2 comprenant des mâles et des femelles âgés de 5 ou 12 semaines (figure 1.1). Plusieurs caractères liés au diabète ont été mesurés, comme le taux d'insuline, de glucose et de triglycérides plasmatiques (pour plus de détails, voir Davis et al. (2012)). Les profils transcriptomiques on été mesurés dans le foie car plusieurs études ont montré un lien entre la susceptibilité au diabète et l'accumulation lipidique dans le foie (Davis et al. (2010), Lan et al. (2003)). Un pré-traitement des données a permis de conserver 15397 gènes uniques comme étant exprimés dans le foie (Davis et al. (2012)). Une première étude réalisée au

laboratoire de Jake Lusis à UCLA (Davis et al 2011) a mis en évidence des QTL pour le taux de triglycéride et d'insuline, mais aucun QTL significatif n'a été détecté pour le taux de glucose. Un des objectifs de ma mobilité dans le laboratoire de Jake Lusis de janvier à avril 2012, a été d'appliquer la stratégie de décomposition du caractère illustrée plus haut (Blum et al. (2011)) sur ce jeu de données, pour identifier en particulier d'éventuels QTL pour le taux de glucose.

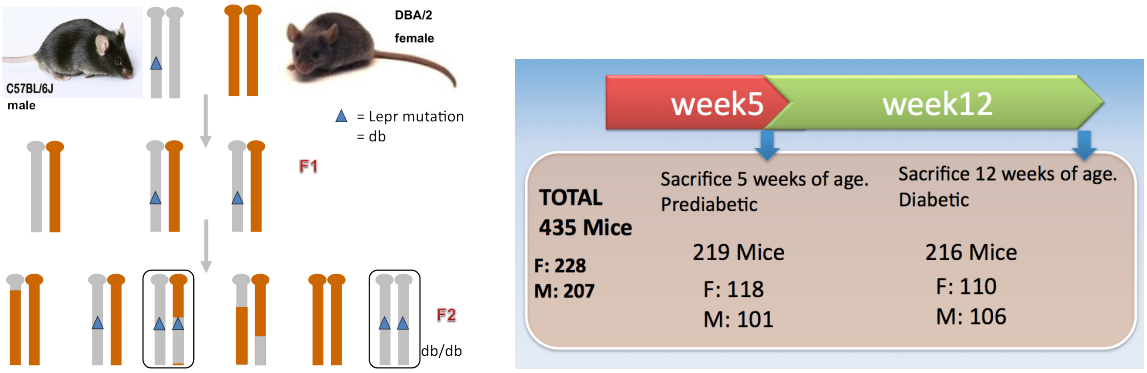


FIGURE 1.1 – Croisement et plan experimental (F : femelle, M : mâle).

La première étape consiste à rechercher les gènes dont l'expression est corrélée au caractère, ici le taux de glucose. Nous utilisons la méthode FAMT de sorte à prendre en compte la variabilité indépendante du caractère d'intérêt. Une liste de 3232 gènes est obtenue avec un seuil de significativité de 5%, alors qu'avec une méthode classique (test de significativité des corrélations de Pearson), 2471 gènes sont trouvés (95 % compris dans la liste des 3232 gènes).
La méthode FAMT identifie 9 facteurs d'hétérogénéité qui peuvent être interprétés en partie par un effet du sexe et de l'âge (figure 1.2).

A

	Age	Sex
Fact 1	0.149	0.122
Fact 2	< 2e-16	9.46e-10
Fact 3	0.973	0.426
Fact 4	6.71e-04	< 2e-16
Fact 5	0.053	0.037
Fact 6	0.509	0.593
Fact 7	0.496	0.955
Fact 8	5.78e-02	1.12e-13
Fact 9	0.375	0.143

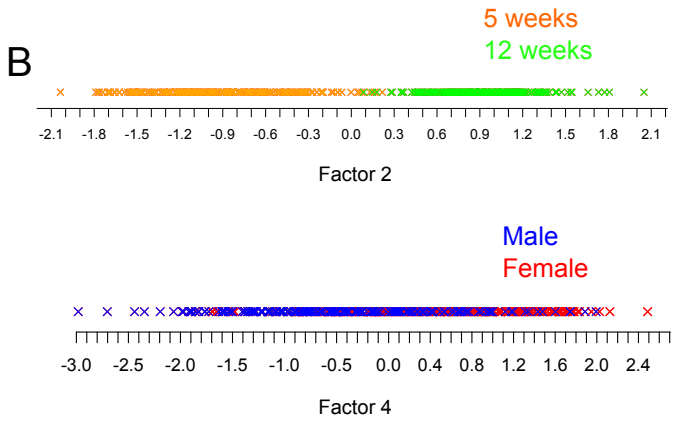


FIGURE 1.2 – Interprétation des facteurs d'hétérogénéité : (A) test de l'association des facteurs d'hétérogénéité avec les facteurs sexe et âge, (B) représentation des individus sur les facteurs d'hétérogénéité 2 et 4.

Une classification basée sur la liste des 3232 gènes dont les expressions sont ajustées pour les facteurs (figure 1.3B) permet d'identifier des sous-types d'animaux pour le taux de glucose : deux sous-types pour un taux élevé en glucose et deux sous-types pour un taux faible. Il est intéressant de noter que comme dans Blum et al. (2011), aucun sous-type pour le caractère n'est observé par une approche classique (figure 1.3A).

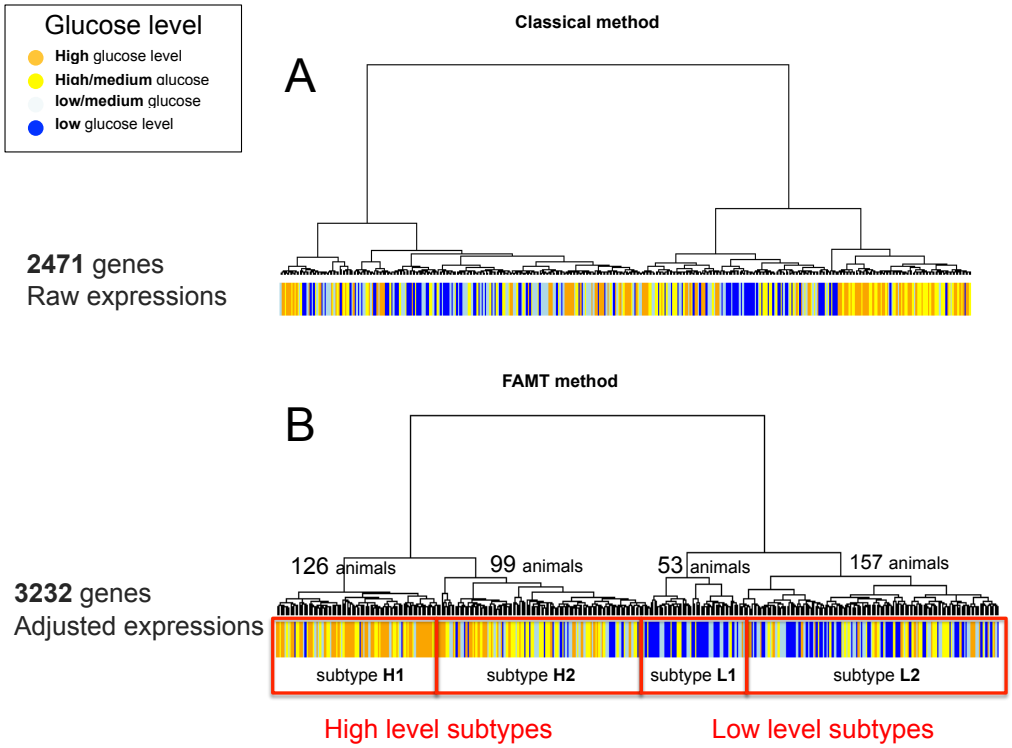


FIGURE 1.3 – Recherche de sous-types d'animaux pour le caractère. Classification des animaux basée sur une liste de gènes dont l'expression est corrélée au caractère, obtenue par : (A) une méthode classique, (B) la méthode FAMT avec ajustement des données d'expression

Une fois les sous-types identifiés, la deuxième étape consiste à réaliser des analyses QTL en enlevant pour chaque analyse un des sous-types phénotypiques déterminés précédemment. Les analyses QTL sont effectuées à l'aide du package "Rqtl" dans le logiciel R (Broman et al. (2003)) et comme dans Davis et al. (2012), les facteurs sexe et âge sont insérés en tant que covariables dans le modèle d'analyse.
Sans le sous-type L1 (figure 1.3B), un QTL est détecté sur le chromosome 4 : le pic est localisé à 127.16Mb et le LOD score (log10 likelihood ratio) est de 5.82 (p-value<10% avec 1000 simulations). En enlevant le deuxième sous-type L2, un QTL est détecté sur le chromosome 5 : le pic est à 131.05Mb et le LOD est de 6.06 (p-value<10% avec 1000 simulations) (figure 1.4). Contrairement à l'étude de Blum et al. (2011) (section 1.3.1), les deux QTL détectés sont suggérés avec l'ensemble du dispositif mais non significatifs à un seuil de 10% (figure 1.4). La méthode de décomposition du caractère permet ici de passer ce niveau de significativité. Il est intéressant de remarquer que le QTL trouvé pour le taux de glucose sur le chromosome 5 co-localise avec un QTL pour le

taux d'insuline (figure 1.4A) et celui sur le chromosome 4 co-localise avec un QTL pour le taux de triglycéride (figure 1.4B). Le QTL sur le chromosome 4 a également été détecté pour le taux de glucose et de triglycéride sur un dispositif similaire de susceptibilité au diabète de type II (Scherneck et al. (2009)), ce qui conforte nos résultats et montre l'intérêt de notre approche. Nous pouvons émettre l'hypothèse qu'il existe des mutations dans les régions évoquées contrôlant à la fois le taux de glucose et de triglycéride d'une part et le taux de glucose et d'insuline d'autre part. Davis et al. (2012) caractérisent les deux régions QTL par une étude fonctionnelle des gènes ayant un cis-eQTL co-localisant avec ces régions (stratégie évoquée dans Le Mignon et al. 2010 §3.2.b). Parmi ces gènes, certains codent pour des protéines à doigts de zinc intervenant comme des facteurs de transcription. En particulier, le facteur de transcription Zfp69, présent dans la région du chromosome 4, serait associé à l'obésité induite par un diabète (Scherneck et al. (2009)) ce qui en fait un bon candidat fonctionnel et positionnel dans notre étude.

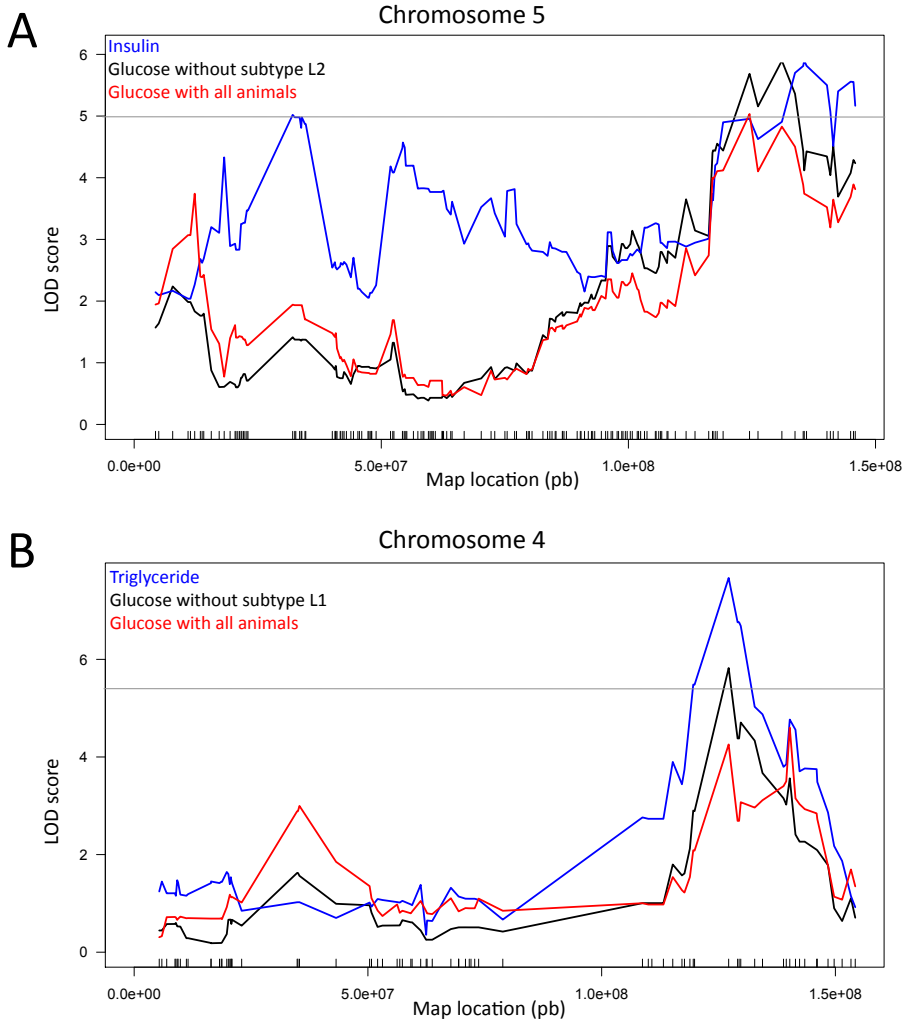


FIGURE 1.4 – Courbes LOD pour le chromosome 5 (A) et le chromosome 4 (B). La ligne horizontale représente le seuil de significativité du LOD score à 10% obtenu avec 1000 simulations

1.4 Conclusion

Dans ce chapitre, on a pu voir l'intérêt de prendre en compte l'hétérogénéité d'expression dans les données transcriptomiques. Au cours des études menées dans différents contextes biologiques, on a montré que l'hétérogénéité du signal ou des profils d'expression masque bien souvent la détection des gènes et des régions du génome liés au caractère d'intérêt. Ce travail contribue ainsi à l'amélioration des stratégies actuelles utilisant des données transcriptomiques pour l'analyse génétique d'un caractère complexe : identification de gènes liés au caractère par analyse différentielle (Blum et al. (2010), Mach et al. (2012)) ; recherche de eQTL co-localisant avec un QTL (Blum et al. (2010)) ; détection de QTL après décomposition du caractère en sous-types d'animaux (Blum et al. (2011), travaux durant la mobilité).

Bien souvent, lors de l'analyse de régions eQTL/QTL, des dizaines voire des centaines de trans-eQTL ainsi que plusieurs gènes candidats peuvent être observés. Or comme nous l'avons évoqué au travers de la figure 2, il existe de multiples relations entre région eQTL/QTL, gènes et caractère d'intérêt. Ainsi, la visualisation d'une structure de dépendance entre les gènes potentiellement impactés par une telle région est une étape clé pour réellement appréhender les mécanismes de régulations sous-jacents. Les travaux du deuxième chapitre sont ainsi centrés sur l'inférence des interactions entre gènes à partir de données transcriptomiques.

Chapitre 2

Modélisation de réseaux de gènes à l'aide de données transcriptomiques : application pour l'analyse génétique d'un caractère complexe

2.1 Introduction

Suite à l'avènement des technologies à haut débit telles que les puces à ADN, l'inférence des réseaux de gènes à partir de données d'expression a suscité un intérêt croissant (D'haeseleer et al. (2000), De Jong (2002), Dobra et al. (2004), Schadt and Lum (2006), Bansal et al. (2007), Liu et al. (2009), De Smet and Marchal (2010)). L'objectif de la reconstruction de réseaux de gènes est de proposer, à partir de données d'expression, des interactions probables entre les gènes et ainsi permettre une meilleure compréhension du fonctionnement global des systèmes biologiques (Wu et al. (2009)). En particulier, la recherche de modules fonctionnels de gènes fortement dépendants ou encore de *hub* gènes (gènes interagissant avec beaucoup d'autres gènes), offre de nouvelles perspectives pour caractériser les processus biologiques sous-jacents à la variabilité d'un caractère d'intérêt et d'autre part, les régulateurs clés de cette variabilité (Gargalovic et al. (2006), Derry et al. (2010)). Pour cela, la visualisation de réseaux géniques sous la forme de graphes peut aider à l'analyse de la structure des réseaux et donc en faciliter l'interprétation biologique (Aittokallio and Schwikowski (2006)). De manière générale, un graphe est un ensemble d'objets distincts représentés par des nœuds, et les interactions entre ces objets sont symbolisées par des flèches ou des arêtes. Dans le cas de réseaux de gènes, un nœud représente un gène et une arête une connexion entre deux gènes (figure 2.1).

Il est important de noter qu'un lien entre deux gènes signifie rarement une interaction directe (physique), mais plutôt que l'activité de l'un causera un changement de l'activité de l'autre, ce changement pouvant être le résultat de différents événements intermédiaires dont les points de départ et d'arrivée sont les protéines codées par ces deux gènes. Les intermédiaires peuvent être

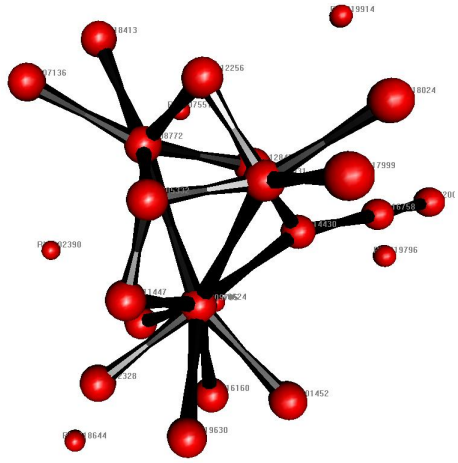


FIGURE 2.1 – Représentation graphique d'un réseau génique : un nœud correspond à un gène et une arête à une dépendance entre deux gènes.

des protéines, des complexe protéiques ou encore des métabolites. Ainsi la représentation d'un réseau génique est une simplification du réseau biochimique global pouvant être vue comme une projection de toutes ces interactions dans l'espace des gènes (figure 2.2) (Brazhnik et al. (2002), Kontos (2009)).

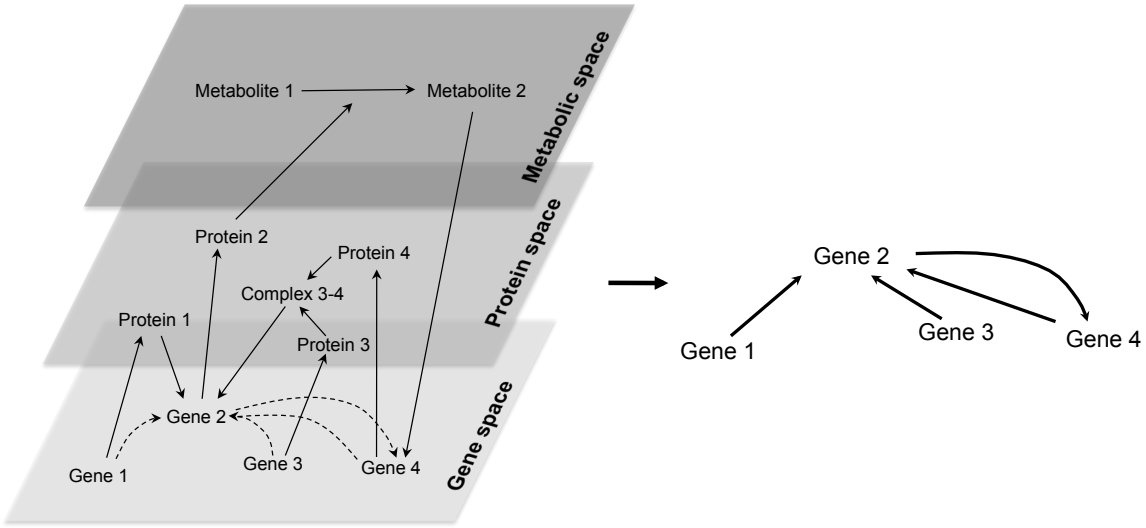


FIGURE 2.2 – Un exemple de réseau biochimique (figure inspirée de Brazhnik et al. (2002)). Les constituants sont organisés en trois niveaux (ou espaces) : ARNm, protéines et métabolites. La projection de l'ensemble des interactions sur l'espace des gènes permet de visualiser les interactions entre gènes uniquement (repris à droite de la figure).

Le principe de la construction de réseaux géniques est donc d'inférer des interactions entre gènes à partir de données d'expression avec ou sans données extérieures (fonctions partagées entre gènes

par exemple). Cette démarche s'avère toutefois ambitieuse du fait de la grande dimension des données : la taille n de l'échantillon (de l'ordre d'une dizaine voire centaine d'individus) est relativement petite comparée au nombre p de variables mesurées (de l'ordre d'une centaine ou milliers de gènes). Ce problème rend les techniques classiques d'apprentissage statistique inappropriées. De plus, il a été montré que beaucoup de réseaux de gènes sont parcimonieux, c'est-à-dire que dans un ensemble de gènes, très peu sont réellement en interaction directe, les régulations opérant plutôt en cascade (Tegner et al. (2003)). Ainsi, l'estimation de ces structures doit à la fois tenir compte de la grande dimension des données et de l'hypothèse d'une faible densité d'interactions dans l'ensemble du réseau.

De nombreuses approches ont été proposées dans la littérature. Dans le cadre des modèles dynamiques qui sont basés sur des données d'expressions temporelles, on peut citer deux approches principales :

- **Réseaux Booléens** (Szallasi and Liang (1998)) : cette approche est une des plus anciennes proposées. Elle considère le réseau comme étant constitué d'éléments logiques : chaque gène est représenté par une variable Booléenne qui exprime un état actif ou inactif. L'idée est alors de trouver des règles logiques permettant de déterminer l'état d'un gène à l'instant t + 1, à partir de l'état de ce même gène et des autres gènes à l'instant t. Ce type d'approche est conçu à l'origine pour l'analyse de systèmes relativement petits.
- **Equations différentielles ordinaires** (De Jong (2002)) : cette approche modélise les régulations géniques par des équations qui expriment la variation de concentration d'un composant x_i du système (par exemple la quantité d'ARNm) comme fonction du temps et de la concentration des autres composants : $\frac{dx_i}{dt} = f_i(t, x_1, ..., x_n)$. Cette approche s'avère très coûteuse en terme de calcul pour des systèmes de taille réaliste.

Dans un cadre plus général, d'autres approches ont été développées basées sur les modèles graphiques qui sont des modèles combinant approches probabilistes et représentation des dépendances à l'aide de graphes : un nœud représente un gène et une arête une dépendance entre deux gènes. Bien que l'aspect dynamique des régulations géniques soit souvent implicite dans ces modèles, ces approches permettent une étude topologique des graphes qui se révèle particulièrement intéressante pour la compréhension des réseaux biologiques de systèmes complexes (Aittokallio and Schwikowski (2006)). On peut distinguer les modèles graphiques non-orientés (les arêtes du graphe ne sont pas directionnelles) des modèles graphiques orientés (les arêtes sont représentées par des flèches). On peut citer deux approches associées aux modèles graphiques non-dirigés : les *relevance networks* et les modèles graphiques Gaussiens (*Gaussian graphical model*). L'approche associée aux modèles graphiques dirigés est connue sous le nom de réseau Bayesien.

- ***Relevance networks*** (Butte et al. (2000)) : cette approche peut être considérée comme le modèle graphique le plus simple par son formalisme. La dépendance entre deux gènes est évaluée à partir du calcul de la corrélation de Pearson entre les mesures d'expression. Si cette corrélation est supérieure en valeur absolue à un seuil fixé au préalable ou considérée comme non nulle après un test de significativité (Transformation de Fisher par exemple), on relie les deux gènes par

une arête dans le graphe. Plus de détails sont donnés dans la section 2.1.1.

- **Modèles graphiques Gaussiens** (Waddell et al. (2000), Schäfer et al. (2005)) : ces modèles font l'hypothèse de la normalité multivariée des données. La dépendance entre deux gènes est dans ce cas mesurée à l'aide de la corrélation partielle, c'est-à-dire la corrélation entre les deux expressions géniques conditionnellement aux autres expressions. A la différence des *relevance networks* où l'on visualise la structure de dépendance marginale, on mesure ici la dépendance conditionnelle permettant de distinguer les liens directs des liens indirects entre les gènes. Les méthodes développées pour l'estimation des corrélations partielles sont détaillées dans la section 2.1.2.
- **Réseaux Bayesiens** (Friedman et al. (2000), Husmeier et al. (2003)) : cette approche peut être vue comme un problème d'optimisation où le but est de trouver la structure qui maximise un score basé sur le calcul des probabilités conditionnelles. Pour chaque profil d'expression X_i, la probabilité conditionnelle est définie par : $\mathrm{P}(X_i|parents(X_i))$, où $parents(X_i)$ sont les régulateurs directs du gène i. Sous l'hypothèse de Markov, la distribution de la probabilité jointe peut s'écrire : $\mathrm{P}(X) = \prod_{i=1}^{n} \mathrm{P}(X_i|parents(X_i))$. Ce score reflète l'aptitude du graphe à "supporter" les données. En d'autres termes, le graphe associé au score le plus élevé est le plus probable sachant les données. Dans le cas de données de grande dimension, différents graphes peuvent avoir le même score. Il est alors parfois judicieux d'introduire des informations biologiques a priori (Mukherjee and Speed (2008)). Cette approche développée à l'origine pour des modèles statiques (données non-temporelles) à été étendue aux modèles dynamiques sous le nom de réseaux Bayesiens dynamiques (Murphy and Mian (1999)). Une étude comparative entre les modèles graphiques Gaussiens et les réseaux bayesiens a montré des performances similaires pour la détection des arêtes dans un graphe (Werhli et al. (2006)). Cependant, contrairement aux modèles graphiques Gaussiens, l'approche des réseaux bayesiens est particulièrement lourde en termes de ressources et de temps de calcul.

Dans cette thèse, nous avons proposé de nouvelles méthodes pour l'inférence de réseaux géniques dans le cadre des *relevance networks* et des modèles graphiques Gaussiens. D'une part, l'approche *relevance network*, très intuitive et simple par son formalisme est extrêmement populaire dans la communauté des biologistes (Steuer (2006)). Récemment une méthode basée sur cette approche nommée WGCNA a donné des résultats particulièrement intéressants en terme d'extraction de modules fonctionnels (Zhang et al. (2005)). D'autre part, les modèles graphiques Gaussiens se sont révélés être un puissant formalisme pour l'inférence des réseaux de gènes, avec l'avantage de pouvoir distinguer les liens directs des liens indirects entre gènes. Ces deux approches sont détaillées dans les sections 2.1.1 et 2.1.2. En section 2.1.3, on présente brièvement l'apport de notre approche basée sur un modèle à facteurs dans le cadre des *relevance networks* et des modèles graphiques Gaussiens. Cette approche est ensuite développée au travers de deux articles présentés en sections 2.2 et 2.3.

2.1.1 *Relevance network*

La construction d'un réseau de gènes peut se décomposer selon les étapes suivantes :

1. choix d'une mesure de lien entre deux gènes
2. calcul du lien pour chaque paire de gènes (figure 2.3 A)
3. définition d'un seuil ou test de significativité du lien (figure 2.3 B)
4. connexion des gènes par une arête si la valeur du lien est supérieure au seuil fixé ou considérée non nulle par le test (figure 2.3 C)

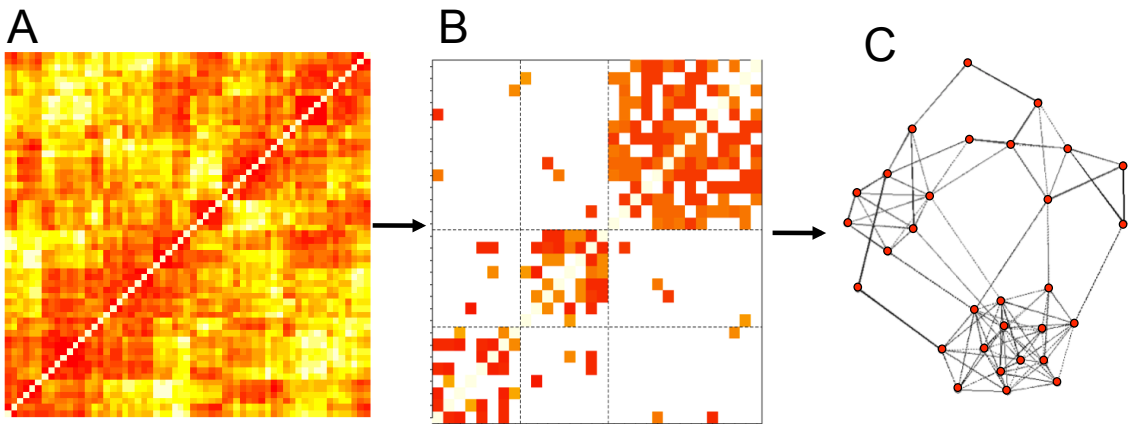


FIGURE 2.3 – Etapes pour la construction d'un réseau de gènes dans le cadre des modèles graphiques non dirigés : (A) calcul de la matrice de similarité S, (B) transformation de S en matrice d'adjacence A, (C) représentation sous forme de graphe des dépendances

Une approche simple et intuitive est de considérer comme mesure de lien entre deux variables y_i et y_j le coefficient de corrélation de Pearson :

$$\text{corr}(y_i, y_j) = \frac{\text{Cov}(y_i, y_j)}{\sqrt{\text{Var}(y_j)\text{Var}(y_j)}}$$

Les méthodes basées sur cette mesure sont rassemblées sous le nom de *relevance network*. Une matrice dite de similarité S est calculée à partir des coefficients de corrélations s_{ij}. Cette matrice peut être transformée en une matrice d'adjacence binaire A contenant des 0 ou des 1 suivant la valeur des coefficients dans S. Ainsi une arête est présente entre les gènes i et j dans le graphe si la valeur a_{ij} correspondante dans A est 1 et inversement absente si la valeur est 0.
Pour tenir compte de la parcimonie du réseau, hypothèse précédemment évoquée (Tegner et al. (2003)), ces méthodes font le choix de règles de décisions plus ou moins strictes désignées sous le nom de *hard thresholding* (Carter et al. (2004), Butte et al. (2000)) ou *soft thresholding* (Zhang et al. (2005), Langfelder and Horvath (2008)).
Une règle de décision de type *hard thresholding* couramment utilisée consiste à considérer un seuil

τ comme suit :

$$a_{ij} = \begin{cases} 1 & \text{si } |s_{ij}| \geq \tau \\ 0 & \text{si } |s_{ij}| < \tau \end{cases}$$

On comprend bien que dans ce cas, le graphe en sortie sera fortement dépendant du choix du seuil τ pouvant entraîner une perte d'information si celui-ci est choisi trop élevé. Certains auteurs testent la significativité du coefficient de corrélation en utilisant une transformation de Fisher (Davidson et al. (2001)) ou en utilisant des permutations (Carter et al. (2004)).
Zhang et al. (2005) proposent une règle de décision de type *soft thresholding* faisant intervenir une fonction puissance :

$$a_{ij} = power(s_{ij}, \beta) = |s_{ij}|^{\beta}$$

Cette transformation à la puissance permet de pondérer les coefficients de la matrice S de sorte à réduire le bruit et favoriser les corrélations les plus fortes. Zhang et al. (2005) appellent ainsi leur méthode WGCNA pour *Weighted Gene Co-expression Network Analysis*. Le coefficient β est choisi de sorte à satisfaire un critère de parcimonie désigné *scale-free topology criterion*. Selon ce critère, la distribution de fréquence de la connectivité $p(k)$, avec $k_i = \sum_{j \neq i} a_{ij}$ la connectivité du gène i, suit approximativement une fonction puissance : $p(k) \sim k^{-\gamma}$ (Jeong et al. (2000), Ravasz et al. (2002)). La valeur de β choisie est la plus petite permettant de satisfaire le critère. En pratique, on considère que le critère est satisfait si R^2 le carré du coefficient de corrélation entre $log_{10}(p(k))$ et $log_{10}(k)$ est supérieur au moins à 0.8.
Afin d'extraire des modules de gènes, c'est-à-dire des groupes de gènes fortement connectés, la méthode WGCNA utilise une mesure de dissimilarité entre les noeuds du graphe appelée *topological overlap dissimilarity measure* (Ravasz et al. (2002)). Intuitivement, plus le nombre de connexions partagées entre deux noeuds est élevé, plus ces noeuds seront considérés comme similaires et inversement. Tout d'abord, on calcule une matrice $\Omega = |\omega_{ij}|$ appelée *TOM* pour *topological overlap matrix* reflétant la similarité entre nœuds :

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}}$$

où $l_{ij} = \sum_{u \neq i,j} a_{iu} a_{uj}$ est d'autant plus fort que le nombre de nœuds connectés à la fois aux nœuds i et j est élevé.
A partir de la matrice TOM la mesure de dissimilarité est ainsi définie :

$$d_{ij}^{\omega} = 1 - \omega_{ij}$$

Une classification hiérarchique sur la matrice de dissimilarité permet de juger du nombre de modules à considérer dans le réseau.
Zhang et al. (2005) montrent qu'il est plus facile d'extraire des modules à partir de la matrice TOM calculée en utilisant les valeurs d'adjacence pondérées plutôt que celles obtenues par un *hard thresholding* (0 ou 1).
La méthode WGCNA, dévelopée par l'équipe de Steve Horvath à UCLA, a été utilisée avec succès

dans plusieurs études biologiques permettant d'identifier des modules géniques fonctionnels en lien avec des caractères complexes (Ghazalpour et al. (2006), Gargalovic et al. (2006), Derry et al. (2010), Davis et al. (2012))

2.1.2 Modèle graphique Gaussien

Le modèle graphique Gaussien introduit par Dempster (1972), est un modèle graphique faisant l'hypothèse de la normalité multivariée des données. Soit Y la matrice des données observées avec n lignes correspondant au nombre d'individus et m le nombre de colonnes correspondant au nombre de gènes. Y est supposée suivre une distribution normale multivariée : $Y \sim \mathcal{N}_m(\mu, \Sigma)$ avec $\mu = (\mu_1, ..., \mu_m)'$ le vecteur des moyennes et $\Sigma = (\sigma_{ij})$ avec $i, j \in [1, m]$ la matrice de variance covariance définie positive.
Ce modèle utilise comme mesure de dépendance entre deux gènes la corrélation partielle que l'on note π :

$$\pi_{i,j} = \text{corr}(y_i, y_j | y_{\setminus i,j})$$

A la différence des *relevance networks*, on mesure ici la dépendance conditionnelle permettant de distinguer les liens directs des liens indirects entre les gènes comme illustré en figure 2.4.

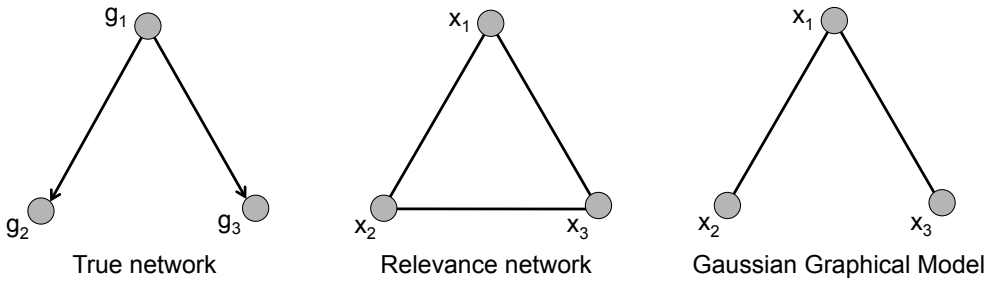


FIGURE 2.4 – Illustration des différents types de dépendances mises en évidence par l'approche *relevance network* et modèle graphique Gaussien. L'approche par un modèle graphique Gaussien permet de détecter uniquement les liens directs alors que l'approche *relevance networks* ne permet pas de distinguer les liens directs (gène 1-2 et 1-3) des liens indirects (gènes 2-3).

Les corrélations partielles peuvent être obtenues à partir de la matrice de variance-covariance inverse Σ^{-1} (Lauritzen (1996)) :

$$\pi_{i,j} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$$

avec $\Sigma^{-1} = (\omega_{i,j})$ pour $i, j \in [1, m]$

Cependant, en grande dimension ($n << m$), la matrice de covariance n'est pas définie-positive et donc non inversible (Dykstra (1970)). Pour contourner ce problème lié à la dimension, plusieurs méthodes ont été développées permettant d'estimer la matrice Σ^{-1} en tenant compte d'une structure parcimonieuse.

Une première approche introduite par Schäfer et al. (2005), consiste à utiliser un estimateur de type *shrinkage* ("estimateur à retrécisseur") de la matrice Σ (Ledoit and Wolf (2004)):

$$\Sigma_{shrink.} = \lambda \mathrm{T} + (1-\lambda)\mathrm{S}$$

où $\lambda \in [0,1]$ est le coefficient de *shrinkage*, S est la matrice de variance-covariance empirique et T un estimateur sous-dimensionné.
Il y a plusieurs façons de choisir T comme par exemple la matrice diagonale constituée des variances empiriques :

$$t_{ij} = \begin{cases} s_{ij} & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

où s_{ij} est le (i,j)-ième élement de S.
Le nombre de paramètres à estimer dans T est petit comparé au nombre de paramètres à estimer dans S (p paramètres au lieu de $p(p+1)/2$). L'estimateur T aura ainsi une variance plus faible que celle de S mais en contrepartie son biais sera plus important. Il s'agit dès lors de choisir λ de sorte à minimiser l'erreur quadratique moyenne (MSE) qui peut s'écrire comme une décomposition biais-variance de l'erreur.
En utilisant comme matrice T celle précédemment décrite, le coefficient de *shrinkage* peut être estimé par :

$$\hat{\lambda} = \frac{\sum_{i \neq j} \widehat{\mathrm{Var}}(s_{ij})}{\sum_{i \neq j} s_{ij}^2}$$

Cette approche permet finalement de construire une matrice $\Sigma_{shrink.}$ facilement inversible en utilisant l'identité de Woodbury (Woodbury (1950)). De plus, pour tenir compte de l'hypothèse de parcimonie, Schäfer et al. (2005) proposent une procédure de test de significativité des corrélations partielles estimées. Pour cela, la distribution des corrélations partielles observées $f(\tilde{\pi})$ est modélisée par un modèle de mélange :

$$f(\tilde{\pi}) = \eta_0 f_0(\tilde{\pi}; \nu) + (1-\eta_0) f_A(\tilde{\pi}),$$

où f_0 est la densité de la distribution sous H_0 ($\pi = 0$) avec ν le degré de liberté, η_0 la proportion (inconnue) de "vraies arêtes" ($\pi \neq 0$), f_A la distribution des corrélations partielles observées pour les vraies arêtes.
La distribution f_0 est donnée par Hotelling (1953). De plus, un algorithme d'estimation de ν, η_0 et f_A est proposé par Efron (2004, 2005). Il est alors possible de calculer pour chaque arête un taux de faux positifs (*FDR : false discovery rate*) local :

$$\mathrm{Prob}(\text{arête nulle}|\tilde{\pi}) = fdr(\tilde{\pi}) = \frac{\hat{\eta_0} f_0(\tilde{\pi}; \hat{\nu})}{\hat{f}(\tilde{\pi})}$$

c'est-à-dire la probabilité a posteriori qu'une arête soit nulle sachant $\tilde{\pi}$. Finalement, une arête est considérée comme présente ou encore significative si son FDR local est inférieur à 0.2 (Efron (2005)).

Cette méthode a été implémentée dans un package R nommé GeneNet (Schäfer et al. (2006)).

Une approche alternative consiste à faire de la régression régularisée. En utilisant les propriétés des variables gaussiennes, on peut exprimer y_i comme combinaison linéaire des $y_{j\neq i}$:

$$y_i = \sum_{j\neq i} \beta_{i,j} y_j + \epsilon_i$$

Dans ce cas, les corrélations partielles sont directement liées aux coefficients de régression :

$$\pi_{i,j} = sign(\beta_{i,j})\sqrt{\beta_{i,j}\beta_{j,i}}$$

Ainsi, l'estimation de la matrice des corrélations partielles peut être ramenée à un problème de régressions linéaires multiples où sélectionner les variables connectées par une arête dans un modèle graphique revient à déterminer un sous-ensemble de variables explicatives. Cette sélection permet de rendre nulles certaines corrélations partielles et donc d'introduire de la parcimonie dans le graphe correspondant. Les méthodes se basant sur cette approche utilisent des techniques de régularisation comme la pénalisation (P) de la norme L_1 (*LASSO : Least Absolute Shrinkage and Selection Operator*) qui permet de réduire certains coefficients et d'en annuler d'autres (contrairement à la pénalisation quadratique *Ridge* qui ne fait que rétrécir les coefficients) :

$$P(\beta) = \lambda\|\beta\|_1 = \lambda\sum_i |\beta_i|,$$

où $\lambda > 0$ est le paramètre de régularisation.
Meinshausen and Bühlmann (2006) ont été les premiers à appliquer cette technique en utilisant une pénalisation *LASSO*. Dans leur méthode, les régressions sont réalisées séparément pour chaque variable ce qui nécessite a posteriori une procédure de symétrisation. Plusieurs améliorations existent dans la littérature (Banerjee et al. (2008), Friedman et al. (2008), Ambroise et al. (2009), Peng et al. (2009)).

2.1.3 Contribution de la thèse : une approche basée sur un modèle à facteurs

Dans le cadre des *relevance networks* et des modèles graphiques Gaussiens, on propose de s'appuyer sur un modèle à facteurs pour modéliser la structure de covariances des profils d'expression géniques. A l'origine, développé dans le domaine de la psychométrie et connu comme une technique de réduction de la dimension (Spearman (1904)), ce modèle est apparu récemment comme une technique d'analyse de la dépendance des données d'expression (Pournara and Wernisch (2007), Carvalho et al. (2008), Friguet et al. (2009), Blum et al. (2010)). Les relations de covariance ou corrélation entre variables sont décrites par l'intermédiaire d'un petit nombre q de variables latentes appelés facteurs communs. La matrice de variance-covariance est décomposée comme suit :

$$\Sigma = BB' + \Psi$$

où B représente la matrice de taille $m\times q$ des coefficients b_k. BB' caractérise l'information commune à l'ensemble des m variables et Ψ est une matrice diagonale caractérisant l'information spécifique

à chacune d'elles.

Ce modèle a des propriétés intéressantes. Tout d'abord, le nombre de paramètres à estimer est beaucoup plus faible que lorsqu'il s'agit d'estimer la matrice de variance-covariance empirique S. Nous verrons dans le cadre des *relevances networks*, que cette caractéristique permet d'améliorer l'estimation des corrélations et présente des avantages pour la prise en compte d'une structure parcimonieuse. De plus, cette structure en facteurs est transposable aux corrélations partielles par un changement de paramétrisation. Nous montrons de même les bonnes performances de notre méthode pour les modèles graphiques Gaussiens par rapport à des méthodes concurrentes au travers d'un exemple simulé.

2.2 Sparse factor model for high-dimensional relevance networks

Cette partie correspond à l'article :

Blum Y, Cadoret M, Houée-Bigot M, Causeur D. Sparse factor models for high dimensional relevance networks.

Cet article sera soumis prochainement au journal *Annals of Applied Statistics*.

Apport de l'article

Une approche très répandue et intuitive pour l'inférence de réseaux de gènes est l'approche *relevance networks* qui se base sur la corrélation empirique pour mesurer la dépendance entre gènes. Dans ce contexte, on se propose d'estimer les corrélations à partir d'un modèle à facteurs. Outre l'avantage d'un plus faible nombre de paramètres à estimer par rapport à la matrice de variance-covariance empirique S, ce modèle s'est avéré particulièrement adapté à la modélisation de la dépendance entre variables. Certains auteurs l'ont utilisé pour identifier des classes de gènes co-exprimés (*clustering*) à partir de données transcriptomiques (Pournara and Wernisch (2007), Carvalho et al. (2008)). En effet, les facteurs du modèle sont associés à une dépendance génique. Ainsi, l'analyse fonctionnelle des gènes fortement associés à un facteur (coefficient b_k élevé) permet d'assimiler parfois une source de dépendance génique à des processus biologiques. Cependant, comme évoqué par Schadt and Lum (2006), la construction de réseau de gènes offre une meilleure compréhension des mécanismes de régulations géniques. Alors que la classification distingue des groupes de gènes co-exprimés, la modélisation de réseaux géniques permet en plus de renseigner sur la connectivité entre les gènes. Il est alors possible d'extraire une structure de dépendance plus épurée permettant de distinguer des modules de gènes fortement connectés ou encore des gènes isolés et donc de détecter plus facilement des régulateurs clés (Schadt and Lum (2006)). Dans notre étude, on propose ainsi de s'appuyer sur un modèle à facteurs pour l'inférence de réseaux géniques. On montre qu'il y a bien un lien entre facteurs et modules géniques. Cette observation suggère de prendre en compte une structure parcimonieuse au travers des coefficients b_k. Différentes procédures sont développées afin d'apporter de la parcimonie au réseau :

- procédures de tests des coefficients : test de l'appartenance d'un gène à un module ou test de significativé des corrélations au travers des coefficients,
- régularisation de type LASSO,
- apport d'informations biologiques : on introduit des connaissances biologiques a priori à l'aide d'annotations fonctionnelles standardisées de type Gene Ontology.

Sur un exemple simulé, on montre les bonnes performances de notre approche par rapport à celles basées sur la corrélation empirique comme la méthode WGCNA (Langfelder and Horvath (2008)), notamment pour identifier les gènes n'appartenant pas à des modules géniques et aussi pour extraire les plus petits modules (non détectés par WGCNA).

Submitted to the Annals of Applied Statistics

SPARSE FACTOR MODELS FOR HIGH-DIMENSIONAL RELEVANCE NETWORKS

By Yuna Blum[*,†] Marine Cadoret[*] Magalie Houée-Bigot[*,†] and David Causeur[*]

Applied Mathematics Department, Agrocampus, Rennes, France[*] and Animal Genetics Department, INRA, Rennes, France[†]

Inference on gene regulatory networks from high-throughput expression data turns out to be one of the main current challenges in systems biology. Such interaction networks are very insightful for the deep understanding of biological behavior and relationships between genes in a system. In particular, a functional characterization of gene modules enables the detection of biological processes underlying complex traits as diseases. Inference on this dependence structure shall account for both the high dimension of the data and the sparsity of the interaction network.

This paper focuses on co-expression networks also known as relevance networks, in which association between genes is completely determined by the correlation matrix of the expression profiles under normality. We propose an EM algorithm to fit a sparse factor model for correlation and demonstrate how it helps extracting modules of genes and more generally improves the gene clustering performance. We also show how this model can be used to take into account a parsimonious structure through testing strategy, LASSO regularization or biological prior. These different estimation methods for sparse factor models are compared to WGCNA method on a simulated example. The good performance of our approach using biological prior is illustrated on a real expression dataset generated to understand adiposity variability in chickens.

1. Introduction. Emergence of high-throughput technologies for large-scale analysis of complex systems, such as brain with functional magnetic resonance infrared (fMRI) or genome with microarrays, has generated new challenges for the statistical methodology. The modern issues of systems biology have especially motivated a new generation of statistical methods for high-dimensional statistics. The False Discovery Rate (FDR) approach of large-scale significance analysis of microarray data is probably the most emblematic example of a renewal of a classical issue, *i.e* multiple testing, to be suited to specific issues of genomic data analysis. Finding out the

AMS 2000 subject classifications: Primary 92C42Relevance network, Factor model, High dimension

genes involved in the variability of complex traits as diseases, *ie.* differentially expressed between conditions, is often a preliminary step for finding key regulators. However, this transcription variability shall also be viewed as resulting from a complex combination of genes individual contributions. In other words, the initial selection of a set of potentially interesting genes and the functional characterization of these genes is still insufficient to explain the transcription system. Exploring the interactions between genes is now expected to give much more insight on the internal structure of expression profiles. Topological analysis of gene interaction networks have therefore received an increased scrutiny lately [4, 21, 5, 25, 1] and graph representation of regulatory networks, where nodes are genes and edges are inhibition or activation connections, have turned out to be intuitive and helpful tools to investigate complex interaction structures [1].

Many mathematical models for interaction graphs have been proposed for gene regulatory networks, with corresponding inference procedures from transcription datasets (see [8]for a survey of methods). Linear modeling of gene regulatory networks often appears as a simple and efficient solution, especially for inference from a large transcription profile observed on a few number of samples. In this context, expression profiles are assumed to be normally distributed and, either the correlations between two gene expressions for the so-called relevance networks or the partial correlations given all the remaining genes for the graphical networks, are viewed as measures of the interaction between those two genes. Graphical models have clearly become very popular because they exclude indirect interactions, namely interactions between two genes induced by their common dependence to other genes, and therefore lead to more sparse networks. However, many biologists prefers dealing with direct correlations, considering they are closer to an intuitive measure of co-expression. This paper proposes a factor modeling approach for inference on relevance networks from high-dimensional transcription datasets.

Extending the idea introduced for differential analysis by Friguet *et al* (2009) [7] and Blum *et al* (2010) [2], we propose to take advantage of a low-dimensional latent linear structure of dependence to improve the stability of correlation estimations. We propose an EM algorithm to fit a sparse factor model for correlations and demonstrate how it helps extracting modules of genes and more generally improves the gene clustering performance. We also show how this model can be used to test for non-module genes or for correlations in order to take into account a parsimonious structure. As an alternative, we propose to use biological knowledge as a prior to infer on the sparsity pattern. Moreover, cyclic coordinate descent algorithm is presented

for a LASSO estimation of the sparse factor model. Finally, these different estimation methods for sparse factor models and leading methods for relevance networks are compared on a simulated example to WGCNA method recently introduced by Langfelder *et al* (2008) [13] and based on empirical correlation estimation. The good performance of our approach using biological prior is illustrated on a real expression dataset generated to understand adiposity variability in chickens [14, 2].

2. Co-expression networks. Gene co-expression networks, or more generally relevance networks, are based on a graph representation of a sparse correlation structure: genes are the vertices of the network and interaction between genes are represented by edges, where an interaction is considered between genes i and j if the correlation ρ_{ij} between the corresponding expressions is nonzero. In other words, the interaction network is entirely defined by the adjacency matrix A of the network, which is a binary version of the correlation matrix: $a_{ij} = 1$ if $\rho_{ij} \neq 0$ and $a_{ij} = 0$ otherwise. Therefore, inference on gene co-expression networks from high-dimensional expression data shares some common objective with large scale significance testing: finding as many true edges as possible with a low proportion of false positives. In many situations, for a more relevant biological interpretation of the network model, it is also insightful to extend the extraction of the network by searching for modules of highly inter-connected genes. As other methods dedicated to the analysis of relevance networks [4, 21, 5], the Weighted Gene Co-expression Network Analysis (WGCNA) method, which is described in [25, 13], provides both tools to estimate the adjacency matrix and to extract a modular structure.

The R package WGCNA, which implements the methods presented in [13], also offers a simulation tool for gene expression data with such a modular interaction structure. Figure 1 displays an image plot of the sample correlation matrix of a dataset simulated using WGCNA with the following parameters: $m = 1000$ genes, $n = 25$ microarrays, 5 modules of genes with respectively $m_1 = 200$ genes, $m_2 = 150$, $m_3 = 80$, $m_4 = 60$ and $m_5 = 40$. The former relative importance of modules is suggested by Langfelder and Horvath (2005) [25] in a tutorial document. The remaining $m - \sum_{i=1}^{5} m_i = 470$ genes are simulated with no mutual interaction and some interaction between the first 2 largest modules, as suggested by figure 1.

A natural hard-thresholding approach to estimate the interaction network from such a dataset consists in testing, for each pair of genes, the significance of the correlation between their expressions and to consider that an interac-

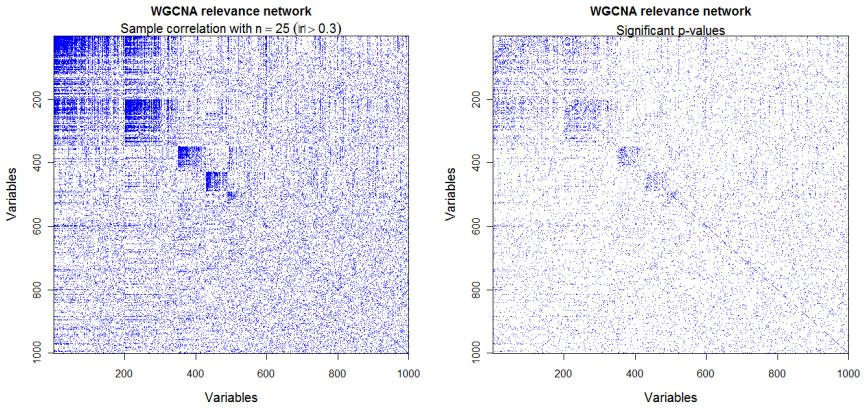


FIG 1. *Image plot of the sample correlation matrix (left plot) and the estimated adjacency matrix (right plot) of a dataset simulated using the R package WGCNA. On the left plot, only correlations with absolute values larger than 0.3 are colored. On the right plot, classification into the two groups of significant and non-significant correlations is based on a hard-thresholding of the estimated local FDR.*

tion exists if the corresponding p-value is lower than a pre-specified threshold. Figure 2 shows the histogram of p-values for the significance tests of correlations in the illustrative dataset with a semi-parametric decomposition of the density into a 2-component mixture of distributions, i.e a non-parametric component for the non-null p-values and a uniform component for the null p-values (see [19] for details). This mixture model for the p-values distribution provides both an estimate for the proportion π_0 of null hypotheses, here $\hat{\pi}_0 = 0.943$, and an estimate, for each correlation coefficient ρ_{ij}, of the so-called local FDR, ℓFDR_{ij}, $i, j = 1, \ldots, m,$. In the present classification issue into the two groups of zero and nonzero correlations, ℓFDR_{ij} can be interpreted as the posterior probability that correlation between genes i and j is nonzero. We propose to consider that genes i and j are not correlated, if ℓFDR_{ij} exceeds a given threshold. In the following, this threshold is chosen so that an approximate proportion $\hat{\pi}_0$ of correlations are not significant. This classification rule applied to the illustrative dataset introduced above results in the adjacency matrix reproduced in figure 1.

[25, 13] propose many variants of a clustering procedure which can be used to extract modules from an adjacency matrix, using a measure of interconnection between genes, called Topological Overlap Measure (TOM) and defined as a normalized number of shared connected genes. In the follow-

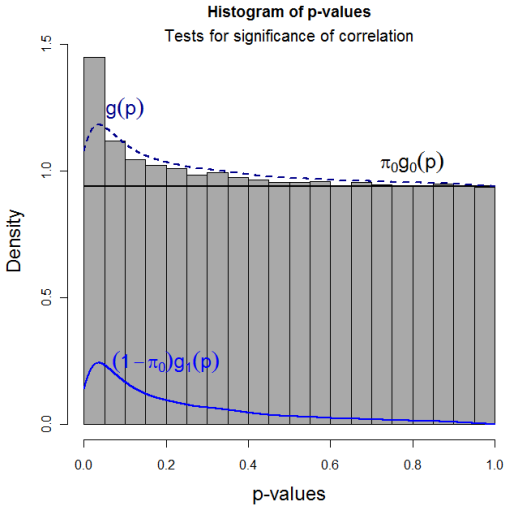


FIG 2. *Histogram of the p-values for the significance tests of correlations in the illustrative dataset. The density function g_1 of the nonnull p-values is unknown and the density function g_0 of the null p-values is uniform: $g_0(p) = 1$ for all p. Both the components of the mixture and the mixing parameter π_0 have been estimated by the semi-parametric method proposed by Robin et al. (2007) [19] and implemented in the R package kerfdr [9].*

TOM classification - Rand index = 0.621									
True modules	grey	turquoise	blue	brown	yellow	green	others	Total	Count
grey	**0.00**	24.50	13.60	5.50	6.60	3.80	46.00	100.00	470.00
turquoise	0.00	**48.00**	12.00	2.00	4.50	1.50	32.00	100.00	200.00
blue	0.00	23.30	**28.70**	0.70	2.70	0.70	44.00	100.10	150.00
brown	0.00	3.80	1.20	**45.00**	1.20	2.50	46.20	99.90	80.00
yellow	0.00	20.00	3.30	1.70	**51.70**	0.00	23.30	100.00	60.00
green	0.00	15.00	2.50	7.50	7.50	**45.00**	22.50	100.00	40.00

TABLE 1

Confusion matrix for clustering of the illustrative 25 × 1000 simulated dataset. Adjacency matrix is based on significance testing of the correlation with classification into the groups of significant and non-significant correlations based on a non-parametric estimate of the local FDR.

ing, the so-called Dynamic Hybrid TOM classification procedure is chosen, because it turns out to be the most performant among the WGCNA classification methods. Table 1 displays the confusion matrix for the adjacency matrix estimated using the hard-thresholding method described above. Although this should be taken as illustrative only, the present hard-thresholding

approach turns out to fail in handling sparsity of the network: it does not find out any unrelated genes and shows a relatively poor performance on the 5 modules.

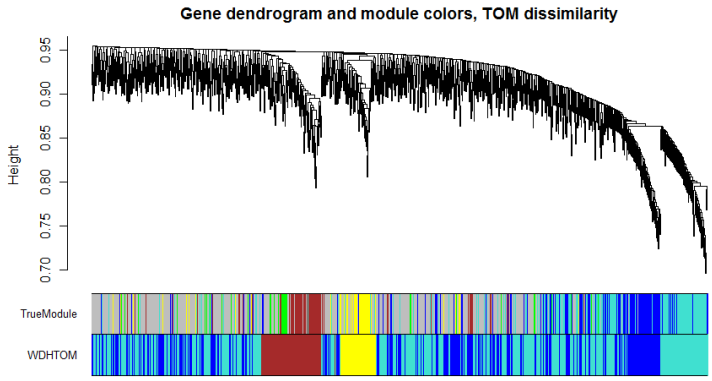


FIG 3. *Gene dendrogram for hierarchical clustering with the WGCNA approach (WDHTOM with $\beta = 2$). Module colors under the dendrogram show the correspondence between true modules and clusters.*

[13] propose to extend the concept of relevance network presented above to weighted co-expression networks in which the weighted adjacency term $0 \leq a^*_{ij} \leq 1$ reflects the level of co-expression between genes i and j as proportional to the correlation on a log-scale: $a^*_{ij} = |\rho_{ij}|^\beta$. This basically transposes the hard-thresholding approach for usual adjacency matrices into a soft-thresholding method which consists in choosing β so that the resulted network satisfies as most as possible the scale-free topology condition (see [25] for details).

For the simulated dataset used above to illustrate the performance of a hard-thresholding approach, the value $\beta = 2$ is obtained by following the suggestion of [13] to choose the lowest value of β ensuring that their R^2-type scale-free topology criterion exceeds 0.90. The gene dendrogram is provided in figure 3 and the results of the clustering procedure is summarized in table 2. Note that the cut-off for the number of clusters is graphically deduced from the dendrogram. The clustering shows a good identification of the 4 largest modules. However, the smallest module is not identified and all the non-module genes are distributed in the 4 largest modules, which reflects the need to better consider the sparse structure of the network. The similarity between true modules and clusters, measured by the Rand index (see [18]), is consequently very similar to the result of the hard-thresholding procedure.

In the following sections, the same simulation settings are used for a more detailed comparison study involving our proposition of a factor analytic decomposition of the correlation structure.

TOM classification								
Weighted adjacency with $\beta = 2$ - Rand index = 0.603								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**0.00**	56.60	33.20	7.20	3.00	0.00	100.00	470.00
turquoise	0.00	**86.00**	12.00	0.50	1.50	0.00	100.00	200.00
blue	0.00	26.70	**70.70**	2.00	0.70	0.00	100.10	150.00
brown	0.00	16.20	22.50	**60.00**	1.20	0.00	99.90	80.00
yellow	0.00	23.30	13.30	0.00	**63.30**	0.00	99.90	60.00
green	0.00	55.00	12.50	32.50	0.00	**0.00**	100.00	40.00

TABLE 2

Confusion matrix for clustering of the illustrative 25 × 1000 dataset with the soft-thresholding approach proposed by Langfelder and Horvath (2005).

3. Sparse factor model for co-expression networks. The failure of the procedure based on significance testing of correlations in the illustrative example introduced in the previous section can be explained by the insufficient power of pairwise significance tests of correlation. In such a situation where the number m of variables is much larger than the sample size n, the number of tests is $m(m-1)/2$ and the BH procedure clearly needs quite a large sample size to be able to detect nonzero correlations. This motivates a modeling approach for the dependence structure to reduce the number of correlation parameters and consequently the number of tests.

As mentioned by Leek and Storey (2008) [16], it can be expected from factor modeling that it increases the overall efficiency of estimation, while keeping quite a large flexibility to model high-dimensional correlation structure. If a common factor structure is obvious, the observations on all the variables are indeed jointly involved in the estimation of each correlation parameter. Moreover, because of this interdependence between the parameters of the correlation model, introduction of sparsity conditions will non simply affect the estimation of nonzero parameters but also increase the efficiency of their estimation by reducing the number of effective parameters.

Sparse factor model

For $1 \leq i \leq n$ and $1 \leq j \leq m$, let Y_{ij} denote the expression of gene j on microarray i. Let $Y_i = (Y_{i1}, \ldots, Y_{im})'$ denote the expression profile for microarray i. It is assumed that the n profiles Y_i are mutually independent and

identically normally distributed with mean μ and variance Σ. Moreover, as suggested in [16, 7], dependence within the expression profile is assumed to be modeled by a latent factor structure: there exists q, with $q \leq \min(n, m)$, unobserved factors $Z_i = (Z_{i1}, \ldots, Z_{iq})'$ independently and identically normally distributed with mean 0 and variance I_q, such that:

$$\mathbb{E}(Y_i \mid Z_i) \quad = \quad \mu + BZ_i \text{ and } \operatorname{Var}(Y_i \mid Z_i) = \Psi$$

where B is the $m \times q$ matrix of loadings with rank q and Ψ is a diagonal matrix which diagonal elements ψ_j^2 are positive. The latent factors can be viewed as sources of dependence across the expression profile in the sense that, conditionally on Z_i, the components of Y_i are independent. It is straightforward checked that the above factor model assumption leads to the following structure for the covariance across the expression profile: $\Sigma = \Psi + BB'$, where Ψ is referred to as the specific variance component and BB' as the common variance component.

Sparsity of the above factor structure is also assumed in the sense that, for some couples (r, s), with $r = 1, \ldots, m$ and $s = 1, \ldots, q$, the corresponding (r, s) loading b_{rs} equals zero. It will be convenient to express this sparsity assumption as a set of linear restrictions on $\operatorname{vec}(B)$ where $\operatorname{vec}(.)$ is the matrix operator transforming a $m \times q$ matrix into the mq-vector obtained by concatenating its row vectors: $b_{rs} = 0$ is equivalent to $R_j' \operatorname{vec}(B) = 0$, where R_j is a $mq-$ vector with 0 entries except for the jth entry which is 1, with $j = q(r-1)+s$. Suppose there are η_0 zero coefficients in B, then the sparsity conditions can be expressed by the linear restrictions $R' \operatorname{vec}(B) = 0$ on B, where R stands for the matrix of sparsity pattern with mq rows and η_0 which all entries are zero except one per column which is 1.

Note that the sparsity pattern can either be obtained from external information such as functional annotations of genes or any preliminary biological knowledge on the interaction network or be extracted from the expression data. The former point will be addressed in the following.

Sparse Maximum-likelihood estimation

The log-likelihood of the above model is given by

$$-\frac{n}{2} \log \det\left(\Psi + BB'\right) - \frac{1}{2} \sum_{i=1}^{n} (Y_i - \mu)' \left(\Psi + BB'\right)^{-1} (Y_i - \mu).$$

It is well-known that this log-likelihood is maximized with respect to μ at $\hat{\mu} = \bar{Y}$, for all Ψ and B. As we focus on the maximum likelihood estimation

(MLE) of the variance parameters, μ will hereafter be replaced by its MLE in the definition of the deviance:

$$\begin{aligned}\mathcal{D}(\Psi, B) &= n \log\det\bigl(\Psi + BB'\bigr) + \sum_{i=1}^{n} (Y_i - \bar{Y})\bigl(\Psi + BB'\bigr)^{-1}(Y_i - \bar{Y}) \\ &= n \log\det\bigl(\Psi + BB'\bigr) + n \operatorname{trace}\bigl[S\bigl(\Psi + BB'\bigr)^{-1}\bigr],\end{aligned}$$

where $S = (1/n)\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})'$. For convenience, and without loss of generality, the profiles will hereafter be considered as centered: the notation Y_i replaces $Y_i - \bar{Y}$.

Direct minimization of $\mathcal{D}(\Psi, B)$ is described by Jøreskog (1969) [11] and is quite popular among users of factor analysis models. However, when the sample size n is small regarding the size m of the expression profile, this algorithm can be numerically unstable and cumbersome. An alternative EM algorithm is proposed by Rubin and Thayer (1982) [20], taking advantage of the factorization of the likelihood based on the conditional distribution of the profile given the latent factors:

$$\begin{aligned}(1) \qquad \mathcal{D}(\Psi, B; Z) &= n \sum_{j=1}^{m} \log \psi_j^2 + \sum_{i=1}^{n} (Y_i - BZ_i)'\Psi^{-1}(Y_i - BZ_i) \\ &\quad - \frac{1}{2}\sum_{i=1}^{n} Z_i' Z_i.\end{aligned}$$

Note that the conditional deviance $\mathcal{D}(\Psi, B; Z)$ can be expressed as follows:

$$\begin{aligned}\mathcal{D}(\Psi, B; Z) &= n \sum_{j=1}^{m} \log \psi_j^2 + \Bigl[\sum_{i=1}^{n} Y_i'\Psi^{-1}Y_i - 2Y_i'\Psi^{-1}BZ_i + Z_i'B'\Psi^{-1}BZ_i\Bigr] \\ &\quad + \sum_{i=1}^{n} Z_i' Z_i, \\ (2) \qquad &= n \sum_{j=1}^{m} \log \psi_j^2 + n\Bigl[\operatorname{trace}\bigl(\Psi^{-1}S\bigr) - 2\operatorname{trace}\bigl(\Psi^{-1}BS_{yz}'\bigr) \\ &\quad + \operatorname{trace}\bigl(B'\Psi^{-1}BS_{zz}\bigr)\Bigr] + n \operatorname{trace}\bigl(S_{zz}\bigr),\end{aligned}$$

where $S_{yz} = (1/n)\sum_{i=1}^{n} Y_i Z_i'$ and $S_{zz} = (1/n)\sum_{i=1}^{n} Z_i Z_i'$. Therefore, the log-likelihood is a linear combination of the sufficient statistics S, S_{yz} and S_{zz}.

The straightforward adaptation of the EM algorithm introduced by Dempster *et al.* (1977) [6] to the present issue is basically obtained by considering the factors Z_i as missing data.

EM algorithm for sparse factor structure

The E-step, which consists in calculating the expectation of $\mathcal{D}(\Psi, B; Z)$ given the profiles Y_i, is not modified by the sparsity restrictions. Conditionally on Y, the expected deviance $\mathbb{E}_y\mathcal{D}(\Psi, B; Z)$ is obtained by replacing the sufficient statistics S, S_{yz} and S_{zz} by their conditional expectation given Y in expression (2):

$$\begin{aligned}\mathbb{E}_y(S) &= S,\\ \mathbb{E}_y(S_{yz}) &= S(\Psi + BB')^{-1}B,\\ \mathbb{E}_y(S_{zz}) &= B'(\Psi + BB')^{-1}S(\Psi + BB')^{-1}B + I_q - B'(\Psi + BB')^{-1}B.\end{aligned}$$

The above expressions are straightforward deduced from the conditional distribution of Z_i: given Y_i, Z_i is normally distributed with mean $B'(\Psi + BB')^{-1}Y_i$ and variance $I_q - B'(\Psi + BB')^{-1}B$. Therefore,

$$\begin{aligned}n^{-1}\mathbb{E}_y\mathcal{D}(\Psi, B; Z) &= \sum_{j=1}^{m}\log\psi_j^2 + \Big[\operatorname{trace}\big(\Psi^{-1}S\big) - 2\operatorname{trace}\big(\Psi^{-1}BC'_{yz}\big)\\ &\quad + \operatorname{trace}\big(B'\Psi^{-1}BC_{zz}\big)\Big] + \operatorname{trace}\big(C_{zz}\big),\end{aligned}$$

where $C_{yz} = \mathbb{E}_y(S_{yz})$ and $C_{zz} = \mathbb{E}_y(S_{zz})$.

Note that this step does not require the direct inversion of the $m \times m$ matrix $\Sigma = \Psi + BB'$. The so-called Woodbury's identity [23], which only involves the inversion of a $p \times p$ matrix, can indeed be used:

$$(\Psi + BB')^{-1} = \Psi^{-1} - \Psi^{-1}B(I_q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}.$$

The M-step now aims at minimizing the expected deviance with respect to B and Ψ under the sparsity restriction $R'\operatorname{vec}(B) = 0$. First, let us focus on the minimization with respect to B using a Lagrange multiplier approach. Noticing that the expected deviance can be expressed as a quadratic form in $\operatorname{vec}(B)$, the following expression has to be minimized with respect to $\operatorname{vec}(B)$ and λ:

$$-2\operatorname{vec}(B)'\operatorname{vec}(\Psi^{-1}C_{yz}) + \operatorname{vec}(B)'\Big[\Psi^{-1}\otimes C_{zz}\Big]\operatorname{vec}(B) + \lambda'R'\operatorname{vec}(B),$$

where λ stands for the η_0- vector of Lagrange multipliers.

It is straightforward deduced that:

$$\text{vec}(\hat{B}) \quad = \quad \Big[\Psi \otimes C_{zz}^{-1}\Big] P_r \,\text{vec}(\Psi^{-1} C_{yz}),$$

where $P_r = I_{mq} - R'(R\Big[\Psi \otimes C_{zz}^{-1}\Big]R')^{-1}R\Big[\Psi \otimes C_{zz}^{-1}\Big]$. The following simpler formulation for the estimate $\hat{b}_r$ of the rth row of B can be deduced from the above expression:

$$\hat{b}_r \quad = \quad \Big[C_{zz}^{-1} - C_{zz}^{-1} C_r^{\star} C_{zz}^{-1}\Big] C_{yz,(r)},$$

where $C_{yz,(r)}$ stands for the rth row of C_{yz} and $C_r^{\star}$ is a $q \times q$ symmetric matrix which entry (i,j) is zero if the corresponding loadings b_{ri} and b_{rj} are nonzero. For (i,j) such that both b_{ri} and b_{rj} are zero, the entry (i,j) is the same as in the inverse matrix of C_{zz}^{-1} restricted to the rows and columns corresponding to zero loadings.

Note that, in the case where all the loadings are nonzero, $C_r^{\star} = 0$ for all $r = 1, \ldots, m$, and the above expression can be formulated in the following equivalent matrix form:

$$\hat{B} \quad = \quad C_{yz} C_{zz}^{-1}.$$

Differentiating $\mathbb{E}_y \mathcal{D}(\Psi, B; Z)$ with respect to the $j$the diagonal term ψ_j^2 of Ψ is straightforward:

$$n^{-1} \frac{\partial}{\partial \psi_j^2} \mathbb{E}_y \mathcal{D}(\Psi, B; Z) \quad = \quad \frac{1}{\psi_j^2} - \Big[\frac{S_{jj}}{\psi_{jj}^4} - 2\frac{(BC'_{yz})_{jj}}{\psi_{jj}^4} + \frac{(BS_{zz}B')_{jj}}{\psi_{jj}^4}\Big].$$

Equating to zero the above derivative leads to:

$$\hat{\psi}_j^2 \quad = \quad S_{jj} - 2(\hat{B}C'_{yz})_{jj} + (\hat{B}C_{zz}\hat{B}')_{jj}.$$

The unrestricted factor model for correlation is now used as an input in the WGCNA soft-thresholding clustering method for the illustrative dataset introduced in the previous section. Before fitting the model, we first need to determine the number of factors. Many methods are proposed in the literature to estimate the dimension of a factor model. Following [15, 16] , we propose to use the parallel analysis method described in [3], which yields $\hat{q} = 3$.

The same rule as used in the previous section for the clustering based on sample correlations is also applied here and provides $\beta = 4$. The results of the clustering are displayed in figure 4 and summarized in Table 3. They

TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.654								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**30.20**	27.20	23.00	14.50	5.10	0.00	100.00	470.00
turquoise	7.00	**71.00**	17.00	2.00	3.00	0.00	100.00	200.00
blue	4.00	22.00	**67.30**	3.30	3.30	0.00	99.90	150.00
brown	10.00	2.50	5.00	**81.20**	1.20	0.00	99.90	80.00
yellow	5.00	13.30	11.70	5.00	**65.00**	0.00	100.00	60.00
green	10.00	35.00	7.50	35.00	12.50	**0.00**	100.00	40.00

TABLE 3

Confusion matrix for clustering of the illustrative 25 × 1000 dataset with the soft-thresholding approach proposed by Langfelder and Horvath (2005) applied with the 3-factor modeling of the correlation.

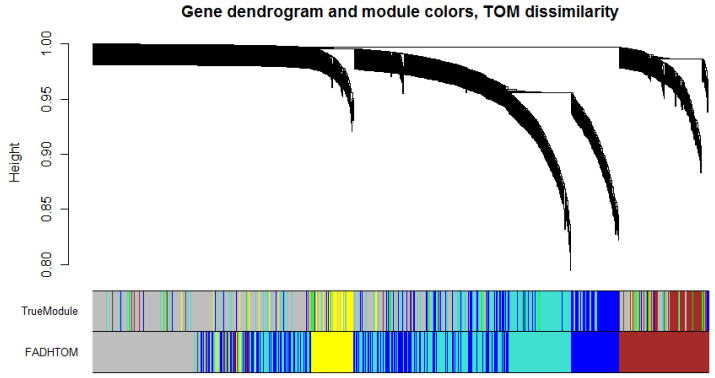


FIG 4. *Gene dendrogram for hierarchical clustering with the WGCNA approach (FADHTOM with $\beta = 4$) based on a 3-factor model for the correlation.*

show that 30 % of non-module genes are identified and that proportions of correct affectations in the 4 largest modules are equivalent to those obtained from the sample correlations (Table 2). Consequently, the Rand index is improved.

The estimated normalized loadings, namely the columns of $\hat{\Psi}^{-1/2}\hat{B}$, and specific variances of the unrestricted 3-factor model are displayed in figure 5. The plot shows a clear but not one-to-one correspondence between factors and modules: the first column of normalized loadings shows larger coefficients for the first two largest modules and the 5th module seems to be harder to identify from the loadings. It also confirms that non-module genes are characterized by corresponding specific variances close to 1.

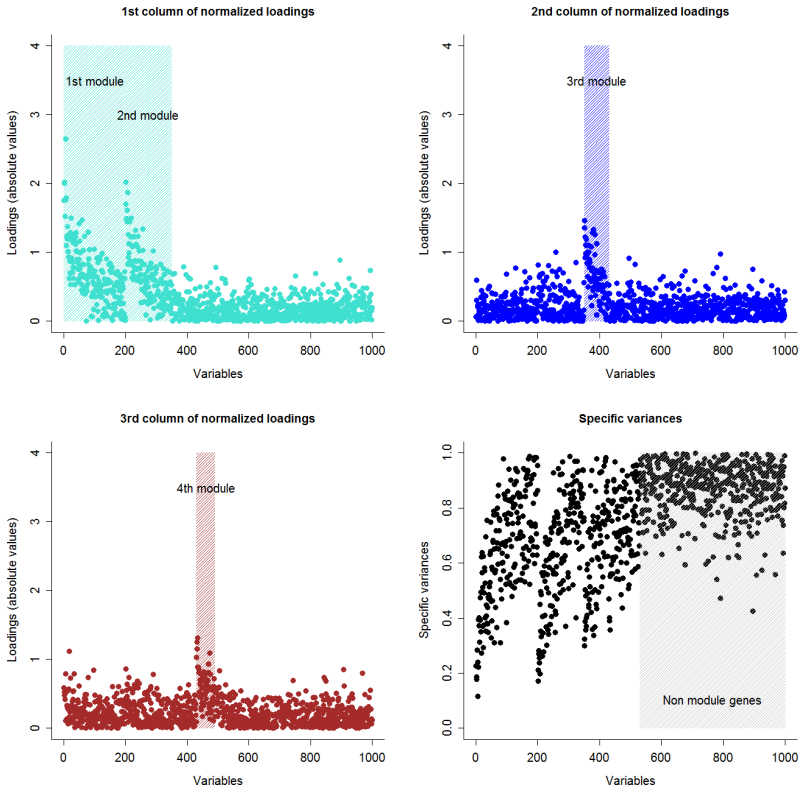


FIG 5. *Normalized loadings and specific variances of the unrestricted 3-factor model for the simulated dataset.*

This suggests that there should be a relationship between the sparsity structure of loadings and the interaction network: genes which all loadings are zero are non-module genes, genes with a large number of nonzero loadings are good candidate to be hubs of the network, pairs of genes with non-zero coefficients on different dimensions are in different modules. As a consequence, the number of factors should depend on the number of modules although there is no reason they are equal. For a network with p modules, at least p factors are indeed necessary to enable the exclusive memberships of some genes to each module.

As 4 modules are identified above, we propose to adapt the modeling strategy using a 4-factor model. The resulting confusion matrix is reproduced in Table 4. Now the 5th module is identified, which suggests to increase

the number of factors to 5, with the same arguments as used above. The confusion matrix given in Table 5 shows a stabilization of the clustering performance.

TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.656								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**23.40**	27.70	20.40	12.60	9.40	6.60	100.10	470.00
turquoise	6.50	**72.00**	16.50	2.00	1.50	1.50	100.00	200.00
blue	2.70	22.00	**66.70**	4.00	4.00	0.70	100.10	150.00
brown	8.80	5.00	5.00	**75.00**	2.50	3.80	100.10	80.00
yellow	3.30	11.70	8.30	0.00	**76.70**	0.00	100.00	60.00
green	7.50	27.50	7.50	15.00	0.00	**42.50**	100.00	40.00

TABLE 4

Confusion matrix for clustering of the illustrative 25 × 1000 dataset with the soft-thresholding approach proposed by Langfelder and Horvath (2005) [25] applied with the 4-factor modeling of the correlation.

TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.649								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**18.10**	28.90	24.30	13.00	10.00	5.70	100.00	470.00
turquoise	5.50	**71.50**	16.00	2.50	2.00	2.50	100.00	200.00
blue	1.30	21.30	**71.30**	4.00	1.30	0.70	99.90	150.00
brown	6.20	6.20	8.80	**76.20**	1.20	1.20	99.80	80.00
yellow	3.30	8.30	11.70	3.30	**73.30**	0.00	99.90	60.00
green	2.50	27.50	7.50	17.50	2.50	**42.50**	100.00	40.00

TABLE 5

Confusion matrix for clustering of the illustrative 25 × 1000 dataset with the soft-thresholding approach proposed by Langfelder and Horvath (2005) [25] applied with the 5-factor modeling of the correlation.

It can also be concluded from the above analysis that identification of the non-module genes while preserving high proportions of correct identifications of the module genes is challenging and needs the explicit introduction of sparsity restrictions on estimation. In the next section, we propose testing strategies to account for sparsity.

4. Testing for sparsity. In the context introduced above of maximum-likelihood estimation in the factor model, asymptotic tests for the significance of the parameters can be used to detect the zero loadings and consequently the edges of the network.

Asymptotic efficiency of the unrestricted EM estimator

Let us consider the following reparameterization of the factor model: for $r = 1, \ldots, m$, $\varphi_r = 1/\psi_r$ and, for $s = 1, \ldots, q$, $\beta_{rs} = b_{rs}/\psi_r$, or, equivalently:

$$\varphi = \Psi^{-1/2},\ \beta = \Psi^{-1/2} B,$$

where φ is the $m \times m$ diagonal matrix which rth diagonal term is φ_r and β is the $m \times q$ matrix which generic (r, s) term is β_{rs}. If ρ_{ij} stands for the correlation coefficient of (Y_i, Y_j), then, under the above assumption of a factor model: ρ_{ij} only depends on β_i and β_j:

$$\rho_{ij} \quad = \quad \frac{\beta_i' \beta_j}{\sqrt{1 + \beta_i' \beta_i}\sqrt{1 + \beta_j' \beta_j}}.$$

It is deduced from the Woodbury's identity that $\Sigma^{-1} = \varphi\big[I_m - \beta(I_q + \beta'\beta)^{-1}\beta'\big]\varphi$. Hence, the log-likelihood is now expressed as follows:

$$\begin{aligned}
\mathcal{L}(\varphi, \beta) &= \frac{n}{2}\log\det\varphi^2 + \frac{n}{2}\log\det\big[I_m - \beta(I_q + \beta'\beta)^{-1}\beta'\big] \\
&\quad -\frac{n}{2}\operatorname{trace}\big[\varphi S \varphi\big[I_m - \beta(I_q + \beta'\beta)^{-1}\beta'\big]\big], \\
&= \frac{n}{2}\log\det\varphi^2 + \frac{n}{2}\log\det K(\beta) - \frac{n}{2}\operatorname{trace}\big[\varphi S \varphi K(\beta)\big],
\end{aligned}$$

where $K(\beta) = I_m - \beta(I_q + \beta'\beta)^{-1}\beta'$.

For convenience, the parameters of the factor model are now considered in the vector form $(\operatorname{diag}(\varphi)', \operatorname{vec}(\beta)')'$, where diag(.) is the matrix operator transforming a $m \times m$ matrix into the m-vector of its diagonal term.

The asymptotic variance of the MLE $(\operatorname{diag}(\hat{\varphi})', \operatorname{vec}(\hat{\beta})')'$ is given in the following proposition.

PROPOSITION 1. *Let $\mathcal{I}$ denote the $m(q+1) \times m(q+1)$ information matrix of the log-likelihood relative to $(\operatorname{diag}(\varphi)', \operatorname{vec}(\beta)')'$ partitioned as follows:*

$$\mathcal{I} \quad = \quad \begin{pmatrix} \mathcal{I}_\varphi & \mathcal{I}_{\beta,\varphi} \\ \mathcal{I}_{\beta,\varphi}' & \mathcal{I}_\beta \end{pmatrix}.$$

Then, the above blocks of $\mathcal{I}$ are given by the following expressions:

$$\begin{aligned}
n^{-1}\mathcal{I}_\varphi &= -\varphi^{-1}\Big[I_m + K^{-1}(\beta) \times K(\beta)\Big]\varphi^{-1}, \\
n^{-1}\mathcal{I}_{\beta,\varphi} &= \big[I_m \otimes (I_q + \beta'\beta)^{-1} + K(\beta) \otimes I_q\big] D_\beta \varphi^{-1}, \\
n^{-1}\mathcal{I}_\beta &= -T(\beta) - K \otimes \big[I_q - (I_q + \beta'\beta)^{-1}\big],
\end{aligned}$$

where $T(\beta)$ is the $mq \times mq$ matrix which (r, r') block $T_{rr'}(\beta)$ is the following $q \times q$ matrix: $T_{rr'}(\beta) = \left[I_q + \beta'\beta\right]^{-1}\beta_{r'}\beta_r'\left[I_q + \beta'\beta\right]^{-1}$, $\times$ stands for the elementwise product operator on matrices with same dimensions, $\otimes$ denotes the Kronecker product of matrices and D_β is the following $mq \times m$ matrix:

$$D_\beta = \begin{pmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \beta_m \end{pmatrix}.$$

It is deduced that the asymptotic distribution of $\sqrt{n}(\mathrm{vec}(\hat{\beta}) - \mathrm{vec}(\beta))$ is normal with mean 0 and variance V_β given by the $mq \times mq$ right lower block of $\mathcal{I}^{-1}$:

$$V_\beta = \mathcal{I}_\beta^{-1} + \mathcal{I}_\beta^{-1}\mathcal{I}_{\beta,\varphi}'(\mathcal{I}_\varphi - \mathcal{I}_{\beta,\varphi}\mathcal{I}_\beta^{-1}\mathcal{I}_{\beta,\varphi}')^{-1}\mathcal{I}_{\beta,\varphi}\mathcal{I}_\beta^{-1}.$$

Proof. see appendix 1. □

Tests for non-module genes

The asymptotic variance is now used to derive tests for the identification of non-module genes. These non-module genes are indeed characterized by their non-correlation with any other genes: if gene i is a non-module gene, then $\beta_i = 0$. This suggests the following χ^2-tests based on the asymptotic normality results shown above: $\hat{\beta}_i'\widehat{V}_{ii}\hat{\beta}_i$, where V_{ii} is the ith $q \times q$ diagonal block of V_β, which follows a χ^2_q distribution under the null hypothesis $H_0^{(i)}$: $\beta_i = 0$.

Figure 6 shows the histogram of p-values for the tests of non-module genes in the illustrative dataset with the similar semi-parametric decomposition of the density into a 2-component mixture of distributions as shown in figure 2. The same classification rule as described in section 2, based on the estimated local FDR $\ell\mathrm{FDR}_i$, $i = 1, \dots, m$, is used to decide which test is significant: gene i is considered as a non-module gene, if $\ell\mathrm{FDR}_i$ exceeds a given threshold, chosen so that an approximate proportion $\hat{\pi}_0 = 0.300$ of genes are non-module genes. This classification rule results in the confusion matrix reproduced in Table 6.

We propose to introduce this preliminary identification of non-module genes as sparsity restrictions on the loadings of the factor model, using the sparse EM algorithm presented in the previous section. A sparse estimation

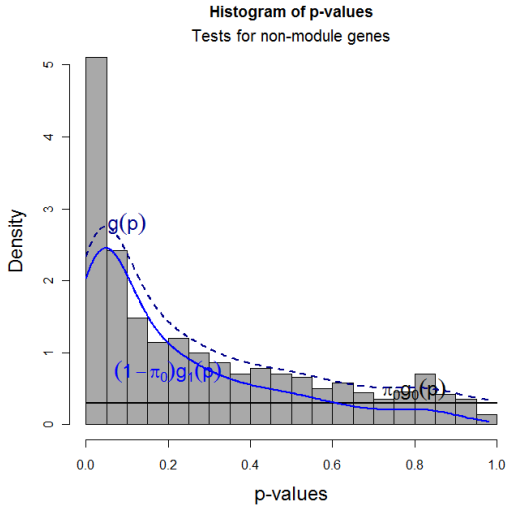


FIG 6. *Histogram of the p-values for the χ^2-tests of non-module genes in the illustrative dataset. The density function g_1 of the nonnull p-values is unknown and the density function g_0 of the null p-values is uniform: $g_0(p) = 1$ for all p. Both the components of the mixture and the mixing parameter π_0 have been estimated by the semi-parametric method proposed by Robin et al. (2007) [19] and implemented in the R package kerfdr [9].*

True modules	non-module	module	Total	Count
non-module	**39.80**	60.20	100.00	470.00
module	11.90	**88.10**	100.00	530.00

TABLE 6
Confusion matrix for clustering of genes into module and non-module genes.

of the factor model is obtained and the TOM classification method is applied to the corresponding correlation matrix. The confusion matrix given in Table 7 shows a better identification of non-module genes with only minor modifications of the classification performance in the 5 modules.

Tests for significance of correlation

We propose here to deduce the sparse structure of the factor model from significance testing of the normalized loadings $\beta_{rs},\ r = 1, \ldots, m,\ s = 1, \ldots, q$, using the t-tests $t_{rs} = \hat{\beta}_{rs}/\sqrt{\hat{v}_{rs}}$, for $r = 1, \ldots, m$ and $s = 1, \ldots, q$, where v_{rs} is the (r, s) term in the matrix V_β. The same classification rule based on the

TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.681								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**40.40**	21.10	14.90	9.60	9.40	4.70	100.10	470.00
turquoise	11.50	**71.00**	12.50	2.00	1.50	1.50	100.00	200.00
blue	10.00	16.00	**68.70**	3.30	1.30	0.70	100.00	150.00
brown	12.50	3.80	8.80	**72.50**	1.20	1.20	100.00	80.00
yellow	10.00	6.70	8.30	3.30	**71.70**	0.00	100.00	60.00
green	22.50	17.50	2.50	15.00	0.00	**42.50**	100.00	40.00

TABLE 7

Confusion matrix for clustering of the illustrative 25 × 1000 dataset with the soft-thresholding approach proposed by Langfelder and Horvath (2005) [25] applied with the 5-factor sparse modeling of the correlation. The sparsity pattern accounts for the non-module genes identified using a χ^2-test.

local FDR is used to obtain the zeros of β. The correlation matrix is now estimated with the sparse EM algorithm, considering that this matrix of zeros defines the sparsity pattern. The resulting confusion matrix is displayed in Table 8 and shows good performance in identifying non-module genes. Nevertheless, the smaller module is no more detected using this procedure.

TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.710								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**62.60**	12.30	3.20	13.60	8.30	0.00	100.00	470.00
turquoise	15.00	**73.50**	4.50	4.50	2.50	0.00	100.00	200.00
blue	12.70	34.70	**42.00**	6.00	4.70	0.00	100.10	150.00
brown	13.80	3.80	1.20	**77.50**	3.80	0.00	100.10	80.00
yellow	15.00	3.30	1.70	3.30	**76.70**	0.00	100.00	60.00
green	32.50	22.50	2.50	32.50	10.00	**0.00**	100.00	40.00

TABLE 8

Confusion matrix for clustering of the illustrative 25 × 1000 dataset with the soft-thresholding approach proposed by Langfelder and Horvath (2005) [25] applied with the 5-factor sparse modeling of the correlation. The sparsity pattern is obtained from significance testing of the normalized loadings.

In the next section, an alternative method is proposed to introduced sparsity in estimation of the factor model, based on a regularization of the EM algorithm using a ℓ_1-penalization of the loadings.

5. LASSO estimation of the factor model. Dealing with sparsity in a high-dimensional situation has received an increased scrutiny for many classical issues such as regression modeling or estimation of correlation and partial correlation matrices. A general class of methods consisting in penal-

izing a loss function, which can be the residual deviance, by a measure of the overall size of the parameters turns out to provide efficient solutions. Among these regularization methods, the Least Absolute Shrinkage and Selection Operator (LASSO) estimation procedure has become very popular, first because it shrinks the estimates while selecting the nonzero parameters and also because efficient algorithms can be used to implement the method.

We propose to modify the EM algorithm described in section 2 by regularization of the M step using a LASSO penalty. In other words, at each M step, the LASSO estimator of the factor model minimizes the criterion $\mathcal{D}(\Psi, B, \lambda)$ defined as follows:

$$\mathcal{D}(\Psi, B, \lambda) \quad = \quad \mathcal{D}(\Psi, B) + \lambda \sum_{r=1}^{m} \sum_{s=1}^{q} |b_{rs}|,$$

where $\lambda \geq 0$ is a tuning parameter controlling the amount of shrinkage and $\mathcal{D}(\Psi, B)$ stands for the unpenalized criterion minimized at each step of the usual EM factor analysis algorithm:

$$\mathcal{D}(\Psi, B) \quad = \quad \log\det\Psi + \operatorname{trace}\big(\Psi^{-1}S\big) - 2\operatorname{trace}\big(\Psi^{-1}BC'_{yz}\big) + \operatorname{trace}\big(B'\Psi^{-1}BC_{zz}\big).$$

The choice of λ is of course crucial since it determines the number of nonzero parameters in the matrix of estimated loadings. In the framework of graphical networks, Banerjee (2008) propose to choose λ so that a BIC criterion is minimized, where the number of parameters in the definition of the criterion is the number of nonzero entries in the estimate. Adapted to our situation, this would yield the following expression:

$$\mathrm{BIC}(\lambda) \quad = \quad \mathcal{D}(\hat{\Psi}_\lambda, \hat{B}_\lambda) + \log(n) \times \#\{(r,s) \in [1,m] \times [1,q], \hat{b}_{rs,\lambda} \neq 0\},$$

where $(\hat{\Psi}_\lambda, \hat{B}_\lambda)$ is the estimate of the factor model parameters obtained with the tuning parameter λ. In our situation, we observe that this strategy often leads to a too sparse solution. Therefore, we propose to choose λ so that the number of zeros in $\hat{B}_\lambda$ equals the estimation resulting from the t-tests introduced in the previous section.

Cyclic Coordinate Descent (CCD) algorithms are especially dedicated to optimization issues involving a large number of parameters and have shown good properties in situations where the objective criterion to be minimized is not continuously differentiable. Therefore, they are very popular for implementing LASSO procedures, in which the ℓ_1-penalty makes the first derivative of the objective criterion with respect to the parameter not continuous. CCD algorithms consist in cyclic minimizations of the marginal coordinate

functions $\mathcal{D}_{rs}(\Psi, B) : \ b_{rs} \mapsto \mathcal{D}(\Psi, B)$ over the parameters b_{rs} until convergence. We propose hereafter a basic CCD algorithm for the LASSO estimation of the factor model, where the cycling follows a natural ordering of the loadings: $b_{11}, b_{12}, \ldots, b_{1q}, b_{21}, \ldots$.

As $\mathcal{D}(\Psi, B)$ is differentiable, the directional derivatives of the penalized criterion are straightforward deduced from the ordinary partial derivative:

$$\begin{aligned}
\frac{\partial \mathcal{D}^+}{\partial b_{rs}}(\Psi, B, \lambda) &= -2\Big[\frac{C_{yz}^{(rs)}}{\psi_r^2} - \frac{\sum_{l=1}^q b_{rl} C_{zz}^{(ls)}}{\psi_r^2}\Big] + \lambda \\
\frac{\partial \mathcal{D}^-}{\partial b_{rs}}(\Psi, B, \lambda) &= 2\Big[\frac{C_{yz}^{(rs)}}{\psi_r^2} - \frac{\sum_{l=1}^q b_{rl} C_{zz}^{(ls)}}{\psi_r^2}\Big] - \lambda.
\end{aligned}$$

Let $\hat{B}$ denote the current update. As described in Wu and Lange (2008), if both directional derivatives evaluated at $\hat{B}$ are nonnegative, the update of $\hat{b}_{rs}$ is skipped. Because the objective function is convex, the directional derivatives cannot be both negative. Now, suppose $\partial \mathcal{D}^- / \partial b_{rs} < 0$, the update is obtained by equating the negative directional derivative to 0:

$$\hat{b}_{rs}^* = \min\left\{0, \hat{b}_{rs} - \frac{\lambda}{2}\frac{\psi_r^2}{C_{zz}^{(ss)}} + \frac{C_{yz}^{(rs)} - \sum_{l=1}^q \hat{b}_{rl} C_{zz}^{(ls)}}{C_{zz}^{(ss)}}\right\}.$$

Similarly, if the directional derivative is negative in the positive direction and positive in the negative direction, then the update $\hat{b}_{rs}^*$ for $\hat{b}_{rs}$ is obtained by equating the directional derivative in the positive direction to zero:

$$\hat{b}_{rs}^* = \max\left\{0, \hat{b}_{rs} + \frac{\lambda}{2}\frac{\psi_r^2}{C_{zz}^{(ss)}} + \frac{C_{yz}^{(rs)} - \sum_{l=1}^q \hat{b}_{rl} C_{zz}^{(ls)}}{C_{zz}^{(ss)}}\right\}.$$

We propose that the above ℓ_1-penalization approach is only used to find the zeros in the matrix of loadings. The correlation matrix is then estimated with the sparse EM algorithm, considering that this matrix of zeros defines the sparsity pattern.

We illustrate the performance of the above method with the same simulated dataset as used in the previous sections. First, it is deduced from the preceding analysis that an appropriate number of factors is 5. As mentioned above, the choice of λ which would result from minimization of the BIC would lead to a too sparse solution here (see Figure 7 which shows how the BIC varies along with λ).

In the present situation, the shrinkage parameter $\lambda = 1.25$ is chosen so that β contains a proportion of 0.25 nonzero entries. This proportion of nonzeros in β is estimated using the 2-component mixture model presented

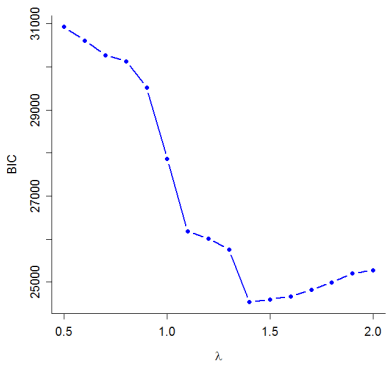


FIG 7. *Values of BIC along with the shrinkage parameter λ.*

above for the p-values of t-tests for the significance of the factor loadings. Figure 8 reproduces the estimated values of the factor model parameters with $\lambda = 1.25$. Note that this choice of λ results in an effective 3-factor model since only 3 factor loadings have nonzero entries.

The correlation matrix is now estimated with the sparse EM algorithm with the sparsity pattern of loadings deduced from the former ℓ_1-penalization approach. Using the same WGCNA clustering rule as in the previous section on the resulting correlation matrix, we obtain the confusion matrix displayed in table 9.

TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.670								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**43.40**	24.30	9.80	14.30	8.30	0.00	100.10	470.00
turquoise	10.00	**71.50**	12.50	3.00	3.00	0.00	100.00	200.00
blue	8.70	38.00	**49.30**	2.00	2.00	0.00	100.00	150.00
brown	11.20	2.50	8.80	**75.00**	2.50	0.00	100.00	80.00
yellow	10.00	6.70	1.70	1.70	**80.00**	0.00	100.10	60.00
green	20.00	22.50	10.00	32.50	15.00	**0.00**	100.00	40.00

TABLE 9

Confusion matrix for clustering of the illustrative 25 × 1000 dataset with the soft-thresholding approach proposed by Langfelder and Horvath (2005) [25] applied with the 5-factor sparse modeling of the correlation. The sparsity pattern is obtained from a ℓ_1 penalization approach with $\lambda = 1.25$.

The results are similar to those obtained from a 5-factor sparse modeling using significance testing procedure of the normalized loadings to define the

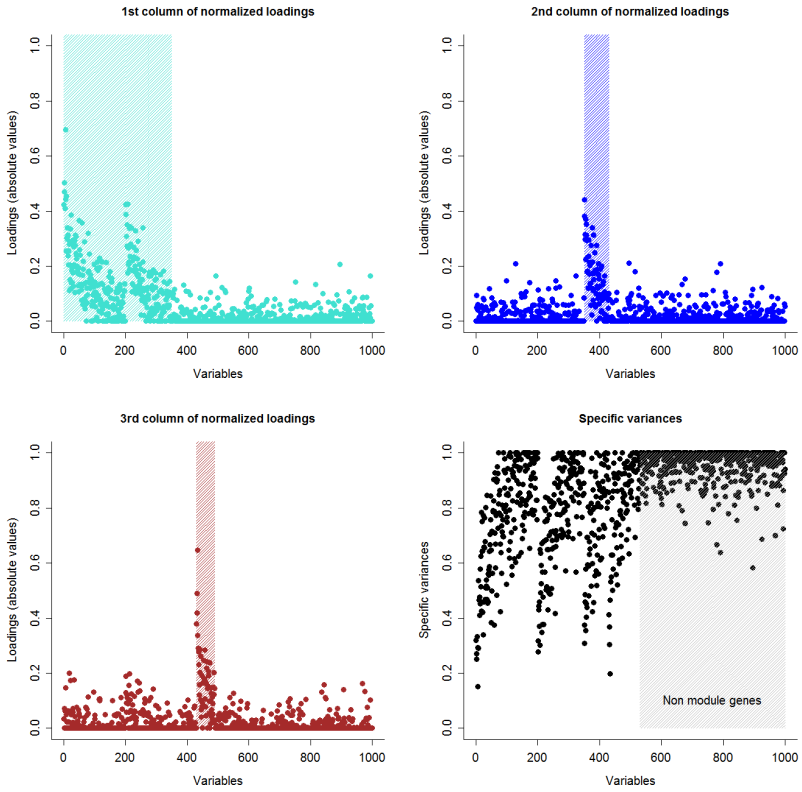


FIG 8. *Normalized loadings and specific variances of the 5-factor model for the simulated dataset, using the ℓ_1-penalization estimation procedure with $\lambda = 1.25$. Note that the 4th and 5th normalized loadings are not represented here since they only have zero entries.*

sparsity pattern (Table 8).

In the next section, another alternative method is proposed to introduce sparsity, based on biological knowledge. The performance of this approach is illustrated on a real expression dataset.

6. Sparsity using biological prior. In this section, we propose to infer sparsity from biological knowledge. We apply our approach to a real expression dataset and use the gene annotation available to define the matrix of sparsity, as suggested in section 3. More precisely, we use the Gene Ontology (GO) [10] biological process (BP) terms, which associate a gene to one or several biological processes. The GO annotation information can be represented by a binary matrix where, genes are in rows and GO terms in columns and a 1 corresponds to an association between a gene and a particular biological process (0 otherwise). A particular characteristic about GO is the hierarchical structure of the terms: some terms concern general processes and comprise a set of specific terms. In order to deal with this redundancy of information and also to get a global view of the common pathways shared among genes, we propose to summarize this biological information. Using a Hierarchical Clustering Analysis on the binary matrix, we define clusters of terms and summarize them into synthetic terms. The ith element of a synthetic term is set to 1 if there is at least one association between the gene i and the GO terms in the considered cluster, and set to 0 otherwise.
We apply our approach to a real expression dataset generated for a study aiming at better understanding adiposity in chickens, which is an economical trait of great interest. The dataset concerns hepatic transcriptome profiles of 45 half sib male chickens variable for their adiposity [14, 2]. In a previous study [2], 328 annotated genes (among a list of 688 genes) where found as being differentially expressed between fat and lean chickens. Inference on interaction between those genes can be particularly useful for a better understanding of the regulatory mechanisms controlling adiposity in this species.
To construct the synthetic GO information matrix, we first select terms which are not associated with too few or too many genes. Indeed, terms associated with too many genes are less informative. For example, very general terms comprise several hundreds of genes. The ones associated with too few genes are too restrictive and not necessary for the synthetic terms construction: because of the hierarchical GO term structure, removing these terms will not lead to loss of information. A selection of terms associated with at least 50 and no more than 200 genes leads to a set of 42 GO terms. Figure 9 shows the heatmap obtained from the corresponding binary matrix. Based

	turquoise module	blue module	Total
brown module	94	27	121
yellow module	6	44	50
green module	28	14	42
grey module (isolated genes)	74	41	115
Total	202	126	

TABLE 10
Contingency table.

on the dendrogram of the GO terms, we choose to summarize the annotation into 4 synthetic variables (see figure 9). The resulting 328 $\times 4$ matrix will be used as sparsity restrictions on the loadings of the factor model.

In order to evaluate the contribution of our biological sparsity based approach, we first apply the unrestricted factor model to the list of 328 genes. The parallel analysis method for determining the number of factor yields to $\hat{q} = 7$. The result of WGCNA soft-thresholding clustering is given in figure 10 and shows the detection of 2 modules without isolated genes (blue and turquoise colors). Interestingly, when applying the biological sparsity restriction approach, 3 smaller modules are highlighted (brown, yellow and green colors) and a proportion of 35% of non-module genes (grey color) is detected.

The contingency table shows that the brown module mainly consists of genes from the turquoise module and yellow module of genes from the blue module (Table 10). The smaller module (green) seems not to be specifically associated with one of the previous module.

These first results reveal that a more precise structure of the network is obtained using biological prior. In order to compare both approaches in terms of biological relevance, we functionally characterize the modules. To complete this task, we perform enrichment test of GO terms using a Fisher's exact test. This procedure tests the over-representation of GO BP terms in a sub-list of genes compared to the whole list of genes (11213 genes in our study [2]). The results obtained using a threshold of 5% are displayed in Table 11. As expected, the brown and turquoise modules on the one hand, and the blue and yellow modules on the other hand, shared common enriched biological processes. Among them, several can be related to the adiposity trait as "steroid biosynthetic process" [17] and "glycerol 3 phosphate metabolic process" [22]. However, the terms "lipid biosynthetic process" and "regulation of lipid metabolic process" enriched in the modules brown and yellow respectively are not found in the blue and turquoise modules. Moreover, the green module shows specific enriched terms concerning the "glycolysis pro-

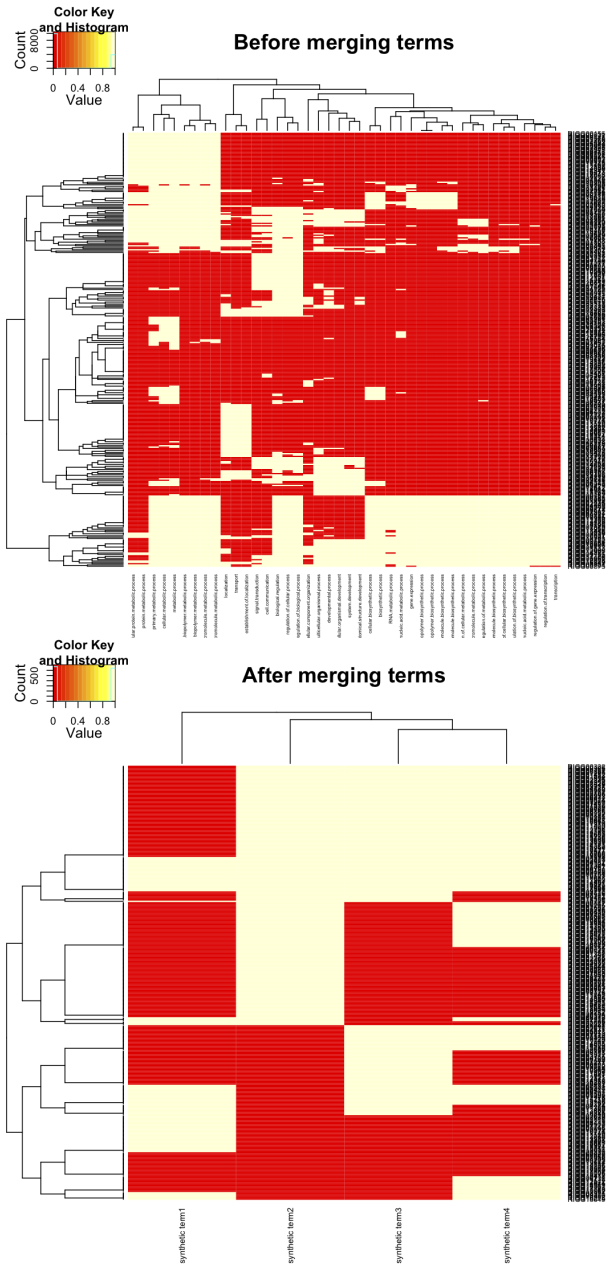


FIG 9. *Heatmap of the binary matrix of annotation consisting in a double hierarchical classification. (top) before merging GO terms, (bottom) after merging terms through synthetic terms*

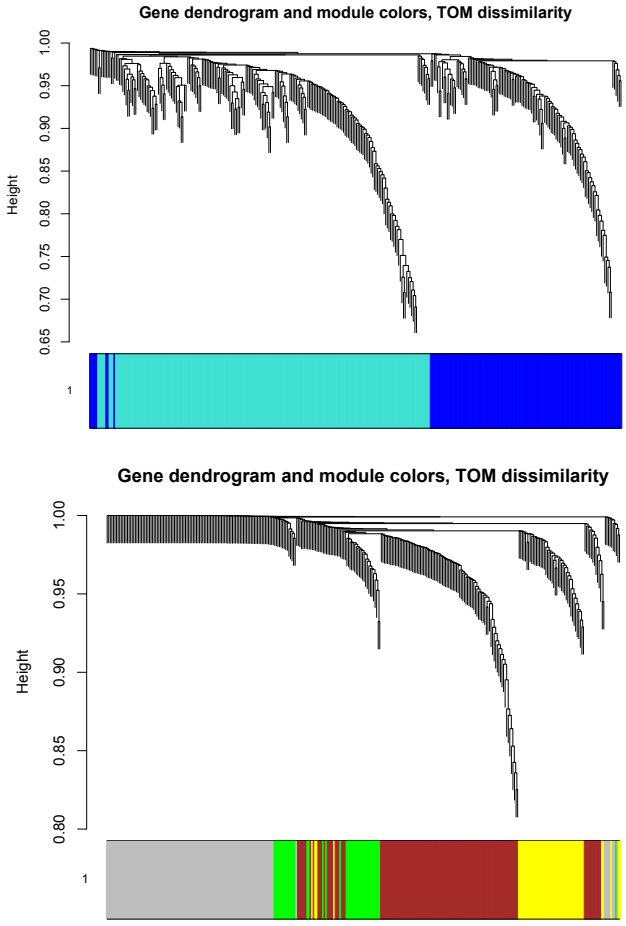


FIG 10. *Gene dendrogram for hierarchical clustering with the WGCNA approach (FAD-HTOM with $\beta = 3$) based on: (top) the unrestricted factor model, (bottom) the sparse factor model using biological prior.*

cess". Interestingly, a recent study showed that enhancing hepatic glycolysis reduces obesity ([24]).

Using the concept of *eigengene* introduced by [12], we evaluate the intensity of correlation between modules and the trait of interest (abdominal fat weight). We show that the green module is indeed negatively correlated to the trait compared to the other modules (figure 11). Finally, we analyze the network in terms of connectivity, which is a measure defined by [13] as the sum of connection strength. WGCNA provide both information about the connectivity of a gene in the entire network and the intra modular connectivity. We focus on the green module specifically highlighted using biological prior in our approach (Table 12). Interestingly, both of the genes involved in the glycolysis metabolism (OGDHL, ENO2), are on the top of the most highly connected genes in this module and have also a high connectivity in the whole network. These two hub genes appear as good key regulator candidates for the adiposity variability in chickens. Deeper investigations will be necessary to clarify the role of these genes in this biological context.

7. Conclusion. In this study, we proposed a novel approach based on a sparse factor model to infer gene co-expression network. We showed a correspondence between factors and modules suggesting to define sparsity through the loadings structure. We introduced several procedures to define a sparse pattern as testing strategies, LASSO regularization or biological prior. Compared to WGCNA on a simulated dataset, all of our approaches seem to improve the clustering performance. Interestingly a sparse factor model using the non-module gene significance testing procedure performed the best whereas the method based on significance testing of normalized loadings or LASSO estimation gave similar results. Furthermore, we showed on a real case study that biological knowledge used as a prior may help in finding a more precise network structure leading to interesting biological interpretations. In this study we used as biological knowledge, Gene Ontology annotation. Nevertheless, our approach offers the possibility to use any other biological information as common eQTL regulations which could be insightful in the context of Genetical Genomics.
Finally, we would like to mention that extensive simulations will be added soon to complete our study.

Module	GO term	Pvalue	Size	Count	HGNC ID
turquoise	regulation.of.muscle.development	1.14e-02	13	3	LEF1;BMP4;NRG1
	steroid.biosynthetic.process	1.74e-02	27	4	CYP17A1;DHCR7; STAR;HMGCS1
	regulation of development	2.19e-02	92	8	FLVCR1;HNF1B;RUNX1; NLE1;PLCG1;MYH9; SALL4;BMP4
	glycerol.3.phosphate.metabolic.process	3.41e-02	8	2	TXNRD3;TIE1
blue	dephosphorylation	6.45e-03	77	6	MTMR9;PPM1E;PTPRF; PPP3CB;PTPN14; PTP4A3
	negative.regulation.of.immune.response	6.73e-03	6	2	COL3A1;PPP3CB
	neuron.migration	5.55e-03	17	3	VAX1;DAB1;PRKG1
	regulation.of.action.potential	1.65e-02	25	3	GAL3ST1;CHRNA4; GNA11
	regulation.of.protein.amino.acid. phosphorylation	2.97e-02	54	4	IBTK;RICTOR;HCLS1; PRKG1
brown	regulation of development	4.05e-06	92	11	FLVCR1;HNF1B;RUNX1; NLE1;PLCG1;MYH9; SALL4;ETNK2; BMP4; LHX1;BCL11B
	regulation.of.muscle.development	2.69e-03	13	3	LEF1;BMP4;NRG1
	glycerol.3.phosphate.metabolic.process	1.28e-02	8	2	TXNRD3;TIE1
	lipid.biosynthetic.process	1.86e-02	94	6	CYP17A1;PDGFA; DHCR7 ;PGDS; HMGCS1;ETNK2
	steroid.biosynthetic.process	2.19e-02	28	4	CYP17A1;DHCR7; HMGCS1; HSD11B1
	pituitary.gland.development	2.86e-02	12	2	LHX3;BMP4
yellow	negative.regulation.of.immune.response	1.04e-03	6	2	COL3A1;PPP3CB
	regulation.of.action.potential	1.16e-03	25	3	CHRNA4;GAL3ST1; GNA11
	synaptic.transmission.cholinergic	1.92e-03	8	2	CHRNA4;LAMA2
	regulation.of.peptidyl.tyrosine phosphorylation	3.05e-03	10	2	RICTOR;HCLS1
	regulation.of.lipid.metabolic.process	1.34e-02	21	2	ACER1;STAR
green	negative.regulation.of.growth	8.64e-03	19	2	BMP10;PTK2
	glycolysis	1.47e-02	25	2	OGDHL;ENO2
	glycolipid.metabolic.process	3.69e-02	5	1	GAL3ST1

TABLE 11

Enrichment tests for each module using GO BP terms. For each term, the size of the whole list of genes related to the term (size), the number of genes in the sub-list related to the term (count), the Pvalue of the test and the HGNC identifier of the associated genes, are given

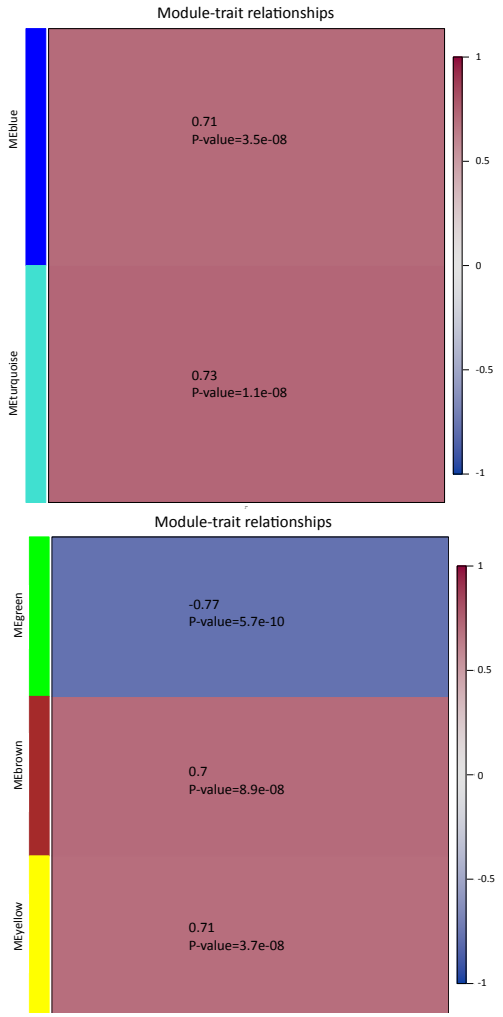


FIG 11. *Image plot of the correlation intensity between a module, summarize into a synthetic variable called eigengene [12], and the trait of interest*

Module	HGNC ID	kTotal	kWithin
green	**OGDHL**	4.36	1.00
(42 genes)	UBLCP1	3.11	0.85
	ENO2	2.63	0.70
	PTPN14	4.25	0.70
	KLF11	0.91	0.45
	SENP2	1.55	0.43
	UBE2R2	1.12	0.40
	NSUN2	0.84	0.36
	PTK2	0.77	0.30
	CPN1	0.77	0.29

TABLE 12

Connectivity in the whole network (kTotal) and intra modular connectivity (kWithin) for the genes in the green module. Top 10 genes with the highest intra modular connectivity are displayed.

References.

[1] T. Aittokallio and B. Schwikowski. Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics*, 7(3):243–255, 2006.

[2] Y. Blum, G. Le Mignon, S. Lagarrigue, and D. Causeur. A factor model to analyze heterogeneity in gene expression. *BMC bioinformatics*, 11(1):368, 2010.

[3] A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509–540, 1992.

[4] A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182, 2000.

[5] S.L. Carter, C.M. Brechbühler, M. Griffin, and A.T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[7] C. Friguet, M. Kloareg, and D. Causeur. A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1415, 2009.

[8] A. Goldenberg, A.X. Zheng, and S.E. Fienberg. *A survey of statistical network models*. Now Publishers, 2010.

[9] M. Guedj, S. Robin, A. Celisse, and G. Nuel. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC bioinformatics*, 10(1):84, 2009.

[10] MA Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(Database issue):D258, 2004.

[11] K.G. Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202, 1969.

[12] P. Langfelder and S. Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology*, 1(1):54, 2007.

[13] P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network

analysis. *BMC bioinformatics*, 9(1):559, 2008.
[14] G. Le Mignon, C. Désert, F. Pitel, S. Leroux, O. Demeure, G. Guernec, B. Abasht, M. Douaire, P. Le Roy, and S. Lagarrigue. Using transcriptome profiling to characterize qtl regions on chicken chromosome 5. *BMC genomics*, 10(1):575, 2009.
[15] J.T. Leek and J.D. Storey. Capturing Heterogeneity In gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9), 2007.
[16] J.T. Leek and J.D. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718, 2008.
[17] T.A. Miettinen and H. Gylling. Cholesterol absorption efficiency and sterol metabolism in obesity. *Atherosclerosis*, 153(1):241–248, 2000.
[18] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, pages 846–850, 1971.
[19] S. Robin, A. Bar-Hen, J.J. Daudin, and L. Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12):5483–5493, 2007.
[20] D.B. Rubin and D.T. Thayer. Em algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76, 1982.
[21] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.
[22] J. Swierczynski, L. Zabrocka, E. Goyke, S. Raczynska, W. Adamonis, and Z. Sledzinski. Enhanced glycerol 3-phosphate dehydrogenase activity in adipose tissue of obese humans. *Molecular and cellular biochemistry*, 254(1):55–59, 2003.
[23] M.A. Woodbury. Inverting modified matrices. *Memorandum report*, 42:106, 1950.
[24] C. Wu, J.E. Kang, L.J. Peng, H. Li, S.A. Khan, C.J. Hillard, D.A. Okar, and A.J. Lange. Enhancing hepatic glycolysis reduces obesity: differential effects on lipogenesis depend on site of glycolytic modulation. *Cell metabolism*, 2(2):131–140, 2005.
[25] B. Zhang, S. Horvath, et al. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.

APPENDIX A: TECHNICAL LEMMAS

LEMMA A.1. *Let β denote a $m \times q$ matrix. Let $K(\beta) = I_m - \beta(I_q + \beta'\beta)^{-1}\beta'$. For $j, r, r' = 1, \ldots, m$ and $s = 1, \ldots, q$*

$$\frac{\partial K_{lj}(\beta)}{\partial \beta_{rs}} = -K_{lr}(\beta)\Big[(I_q + \beta'\beta)^{-1}\Big]'_s \beta_j - K_{rj}(\beta)\Big[(I_q + \beta'\beta)^{-1}\Big]'_s \beta_l.$$

Equivalently, in matrix form:

$$\frac{\partial K(\beta)}{\partial \beta_{rs}} = -K_r(\beta)\big[(I_q + \beta'\beta)^{-1}\big]'_s \beta' - \beta\big[(I_q + \beta'\beta)^{-1}\big]_s K'_r(\beta),$$

where $K_r(\beta)$ stands for the rth column of $K(\beta)$.

Proof. For $s = 1, \ldots, q$, $j = 1, \ldots, m$,

$$\begin{aligned}
\frac{\partial K_{jj}(\beta)}{\partial \beta_{js}} &= \beta_j'(I_q + \beta'\beta)^{-1}\frac{\partial \beta'\beta}{\partial \beta_{js}}(I_q + \beta'\beta)^{-1}b_j - 2\beta_j'(I_q + \beta'\beta)^{-1}\frac{\partial \beta_j}{\partial \beta_{js}}, \\
&= \beta_j'(I_q + \beta'\beta)^{-1}\beta_j\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j + \beta_j'\Big[(I_q + \beta'\beta)^{-1}\Big]_s\beta_j'(I_q + \beta'\beta)^{-1}\beta_j \\
&\quad -2\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j, \\
&= -2\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j K_{jj}(\beta).
\end{aligned}$$

For $r = 1, \ldots, m$ with $r \neq j$,

$$\begin{aligned}
\frac{\partial K_{jj}(\beta)}{\partial \beta_{rs}} &= \beta_j'(I_q + \beta'\beta)^{-1}\frac{\partial \beta'\beta}{\partial \beta_{rs}}(I_q + \beta'\beta)^{-1}\beta_j, \\
&= \beta_j'(I_q + \beta'\beta)^{-1}\beta_r\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j + \beta_j'\Big[(I_q + \beta'\beta)^{-1}\Big]_s\beta_r'(I_q + \beta'\beta)^{-1}\beta_j, \\
&= -2\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j K_{rj}(\beta).
\end{aligned}$$

For $l = 1, \ldots, m$ with $r \neq l$,

$$\begin{aligned}
\frac{\partial K_{lj}(\beta)}{\partial \beta_{rs}} &= \beta_l'(I_q + \beta'\beta)^{-1}\frac{\partial \beta'\beta}{\partial \beta_{rs}}(I_q + \beta'\beta)^{-1}\beta_j, \\
&= \beta_l'(I_q + \beta'\beta)^{-1}\beta_r\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j + \beta_l'\Big[(I_q + \beta'\beta)^{-1}\Big]_s\beta_r'(I_q + \beta'\beta)^{-1}\beta_j, \\
&= -K_{lr}(\beta)\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j - K_{rj}(\beta)\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_l.
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{\partial K_{rj}(\beta)}{\partial \beta_{rs}} &= \beta_r'(I_q + \beta'\beta)^{-1}\frac{\partial \beta'\beta}{\partial \beta_{rs}}(I_q + \beta'\beta)^{-1}b_j - \frac{\partial \beta_r'}{\partial \beta_{rs}}(I_q + \beta'\beta)^{-1}\beta_j, \\
&= \beta_r'(I_q + \beta'\beta)^{-1}\beta_r\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j + \beta_r'\Big[(I_q + \beta'\beta)^{-1}\Big]_s\beta_r'(I_q + \beta'\beta)^{-1}\beta_j \\
&\quad -\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j, \\
&= -K_{rr}(\beta)\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_j - K_{rj}(\beta)\Big[(I_q + \beta'\beta)^{-1}\Big]_s'\beta_r.
\end{aligned}$$

□

APPENDIX B: PROOF OF PROPOSITION 1

The Fisher information matrix $\mathcal{I} = -\mathbb{E}(\mathcal{H})$ of the factor model is now calculated, where $\mathcal{H}$ stands for the $(q+1)m \times (q+1)m$ Hessian matrix of

the log-likelihood with respect to the vector $(\mathrm{diag}(\varphi)', \mathrm{vec}(\beta)')'$ of all the parameters. First, let us recall the expression of the log-likelihood:

$$\begin{aligned}\mathcal{L}(\varphi,\beta) &= \frac{n}{2}\log\det\varphi^2 + \frac{n}{2}\log\det\big[I_m - \beta(I_q+\beta'\beta)^{-1}\beta'\big] \\ &\quad -\frac{n}{2}\,\mathrm{trace}\big[\varphi S\varphi\big[I_m - \beta(I_q+\beta'\beta)^{-1}\beta'\big]\big], \\ &= \frac{n}{2}\log\det\varphi^2 + \frac{n}{2}\log\det K(\beta) - \frac{n}{2}\,\mathrm{trace}\big[\varphi S\varphi K(\beta)\big],\end{aligned}$$

where S is the sample variance matrix and $K(\beta) = I_m - \beta(I_q+\beta'\beta)^{-1}\beta'$.

First, let us differentiate $\mathcal{L}(\varphi,\beta)$ with respect to φ_j, for $j = 1,\ldots,m$:

$$\frac{\partial}{\partial\varphi_j}\mathcal{L}(\varphi,\beta) = \frac{n}{\varphi_j} - nS_j'\varphi K_j(\beta),$$

where $K_j(\beta)$ denotes the jth column of $K(\beta)$.

Now, the first order derivative with respect to β_{rs}, for $r = 1,\ldots,m$ and $s = 1,\ldots,q$ is calculated using lemma A.1:

$$\begin{aligned}\frac{\partial}{\partial\beta_{rs}}\mathcal{L}(\varphi,\beta) &= \frac{n}{2}\frac{\partial}{\partial\beta_{rs}}\log\det\big[K(\beta)\big] - \frac{n}{2}\frac{\partial}{\partial\beta_{rs}}\mathrm{trace}\big[\varphi S\varphi K(\beta)\big], \\ &= \frac{n}{2}\,\mathrm{trace}\,K^{-1}(\beta)\frac{\partial K(\beta)}{\partial\beta_{rs}} - \frac{n}{2}\,\mathrm{trace}\big[\varphi S\varphi\frac{\partial K(\beta)}{\partial\beta_{rs}}\big], \\ &= -n\beta_r'\big[I_q+\beta'\beta\big]_s^{-1} + nK_r'(\beta)\varphi S\varphi\beta\big[I_q+\beta'\beta\big]_s^{-1}, \\ &= -n\big[I_q+\beta'\beta\big]_s^{-1'}\Big[\beta_r - \beta'\varphi S\varphi K_r(\beta)\Big]\end{aligned}$$

Equivalently,

$$\frac{\partial}{\partial\beta_r}\mathcal{L}(\varphi,\beta) = -n(I_q+\beta'\beta)^{-1}\Big[\beta_r - \beta'\varphi S\varphi K_r(\beta)\Big].$$

It is straightforward deduced that,

$$\frac{\partial^2}{\partial\varphi_j^2}\mathcal{L}(\varphi,\beta) = -\frac{n}{\varphi_j^2} - nS_{jj}K_{jj}(\beta).$$

and for $r = 1,\ldots,m$ with $r \neq j$,

$$\frac{\partial^2}{\partial\varphi_r\partial\varphi_j}\mathcal{L}(\varphi,\beta) = -nS_{rj}K_{rj}(\beta).$$

The expectation of S is now needed: $\mathbb{E}(S) = \varphi^{-1}K^{-1}(\beta)\varphi^{-1}$. Therefore, the $m\times m$ Hessian block with respect to $\mathrm{diag}(\varphi)$ is given by the following expression:

$$\mathbb{E}\Big[\frac{\partial^2\mathcal{L}(\varphi,\beta)}{\partial\,\mathrm{diag}(\varphi)\partial\,\mathrm{diag}(\varphi)}\Big] = -n\varphi^{-1}\Big[I_m + K^{-1}(\beta)\times K(\beta)\Big]\varphi^{-1},$$

where $\times$ stands for the elementwise product operator on matrices with the same dimensions.

Now, the second order derivative with respect to φ_j and β_{rs}, for $r = 1,\dots,m$ and $s = 1,\dots,q$ is also deduced from lemma A.1:

$$\begin{aligned}\frac{\partial^2}{\partial\beta_{rs}\partial\varphi_j}\mathcal{L}(\varphi,\beta) &= -nS_j'\varphi\frac{\partial K_j(\beta)}{\partial\beta_{rs}},\\ &= nS_j'\varphi K_r(\beta)\big[(I_q+\beta'\beta)^{-1}\big]_s'\beta_j + nK_{rj}(\beta)S_j'\varphi\beta\big[(I_q+\beta'\beta)^{-1}\big]_s.\end{aligned}$$

Therefore, in matrix form:

$$\frac{\partial^2}{\partial\beta_r\partial\varphi_j}\mathcal{L}(\varphi,\beta) = n(I_q+\beta'\beta)^{-1}\beta_jS_j'\varphi K_r(\beta) + nK_{rj}(\beta)(I_q+\beta'\beta)^{-1}\beta'\varphi S_j.$$

Hence,

$$\begin{aligned}\mathbb{E}\Big[\frac{\partial^2}{\partial\beta_r\partial\varphi_j}\mathcal{L}(\varphi,\beta)\Big] &= \frac{n}{\varphi_j}\delta_{rj}(I_q+\beta'\beta)^{-1}\beta_j + \frac{n}{\varphi_j}K_{rj}(\beta)(I_q+\beta'\beta)^{-1}\beta'K_j^{-1}(\beta),\\ &= \frac{n}{\varphi_j}\delta_{rj}(I_q+\beta'\beta)^{-1}\beta_j + \frac{n}{\varphi_j}K_{rj}(\beta)\beta_j.\end{aligned}$$

Equivalently:

$$\mathbb{E}\Big[\frac{\partial^2\mathcal{L}(\varphi,\beta)}{\partial\,\mathrm{vec}(\beta)\partial\,\mathrm{diag}(\varphi)}\Big] = n\big[I_m\otimes(I_q+\beta'\beta)^{-1} + K(\beta)\otimes I_q\big]D_\beta\varphi^{-1},$$

where D_β is the following $mq\times m$ matrix:

$$D_\beta = \begin{pmatrix}\beta_1 & 0 & \dots & 0\\ 0 & \beta_2 & & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & \beta_m\end{pmatrix}.$$

Finally, the Hessian block with respect to the loadings is calculated: for $r,r' = 1,\dots,m$ and $s,s' = 1,\dots,q$,

$$\begin{aligned}\frac{\partial^2}{\partial\beta_{r's'}\partial\beta_{rs}}\mathcal{L}(\varphi,\beta) &= -n\frac{\partial\big[I_q+\beta'\beta\big]_s^{-1'}}{\partial\beta_{r's'}}\Big[\beta_r-\beta'\varphi S\varphi K_r(\beta)\Big]\\ &\quad -n\big[I_q+\beta'\beta\big]_s^{-1'}\frac{\partial\Big[\beta_r-\beta'\varphi S\varphi K_r(\beta)\Big]}{\partial\beta_{r's'}},\\ &= -n\frac{\partial\big[I_q+\beta'\beta\big]_s^{-1'}}{\partial\beta_{r's'}}\Big[\beta_r-\beta'\varphi S\varphi K_r(\beta)\Big]\\ &\quad -n\big[I_q+\beta'\beta\big]_s^{-1'}\Big[\delta_{rr'}e_{s'}-(\varphi S\varphi)_{r'}'K_r(\beta)e_{s'}-\beta'(\varphi S\varphi)\frac{\partial K_r(\beta)}{\partial\beta_{r's'}}\Big].\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\Big[\frac{\partial^2}{\partial\beta_{r's'}\partial\beta_{rs}}\mathcal{L}(\varphi,\beta)\Big] &= n\big[I_q+\beta'\beta\big]_s^{-1'}\beta' K^{-1}(\beta)\frac{\partial K_r(\beta)}{\partial\beta_{r's'}},\\
&= -n\big[I_q+\beta'\beta\big]_s^{-1'}\beta_{r'}\big[I_q+\beta'\beta\big]_{s'}^{-1'}\beta_r\\
&\quad -nK_{rr'}(\beta)\big[I_q+\beta'\beta\big]_s^{-1'}\beta' K^{-1}(\beta)\beta\big[I_q+\beta'\beta\big]_{s'}^{-1},\\
&= -n\big[I_q+\beta'\beta\big]_s^{-1'}\beta_{r'}\big[I_q+\beta'\beta\big]_{s'}^{-1'}\beta_r\\
&\quad -nK_{rr'}(\beta)\big[I_q-(I_q+\beta'\beta)^{-1}\big]_{ss'}
\end{aligned}$$

Equivalently, the above matrix can be expressed as follows, in matrix form:

$$\mathbb{E}\Big[\frac{\partial^2\mathcal{L}(\varphi,\beta)}{\partial\beta_r\partial\beta_{r'}}\Big] = -n\big[I_q+\beta'\beta\big]^{-1}\beta_{r'}\beta_r'\big[I_q+\beta'\beta\big]^{-1} - nK_{rr'}(\beta)\big[I_q-(I_q+\beta'\beta)^{-1}\big].$$

Standard results on inversion of partitioned matrices are used to obtain the asymptotic variance given in proposition 1. □

Agrocampus Applied Mathematics Department,
65 rue de St-Brieuc 35042 Rennes cedex, France
E-mail: yuna.blum@rennes.inra.fr; marine.cadoret@rennes.inra.fr; magalie.houee@rennes.inra.fr; david.causeur@agrocampus-ouest
Department of Animal Genetics, INRA
65 rue de St-Brieuc 35042 Rennes cedex, France
E-mail: yuna.blum@rennes.inra.fr; magalie.houee@rennes.inra.fr

2.3 Inferring gene networks using a sparse factor Gaussian graphical model

Cette partie correspond à l'article suivant qui est en cours de préparation :

Blum Y, Friguet C, Houée-Bigot M, Lagarrigue S & Causeur D. Inferring gene networks using a sparse factor model approach in the context of Gaussian graphical model.

J'ai présenté ces travaux à l'occasion de plusieurs congrès :
- Statistical Learning and Data Science, Florence, Italy, 7-9 May 2012. (article et communication orale)
- International Biometric Society Channel Network, Bordeaux, 11-13 April 2011. (résumé court et communication orale)
- SFdS Journées de Statistique, Marseille, 25-28 May 2010. (résumé long et communication orale)

Apport de l'article

Les modèles graphiques Gaussiens se sont révélés être un puissant formalisme pour l'inférence des réseaux de gènes, avec l'avantage de pouvoir distinguer les liens directs des liens indirects entre gènes contrairement aux *relevance networks*. Ces modèles s'appuient sur l'hypothèse de normalité des profils d'expression et se basent sur la corrélation partielle comme mesure de dépendance entre les gènes. L'inférence sur les réseaux revient donc ici à estimer les corrélations partielles, ou de manière équivalente, l'inverse de la matrice de variance-covariance des profils, ce qui pose problème en grande dimension. Pour pallier cet inconvénient, de nombreux auteurs ont proposé des approches de régularisation de type *shrinkage* (Schäfer et al. (2005)) ou régressions régularisées (Peng et al. (2009)).

Nous proposons une nouvelle approche s'appuyant sur un modèle à facteurs qui permet de synthétiser la dépendance génique au travers de variables latentes (Pournara and Wernisch (2007), Carvalho et al. (2008), Friguet et al. (2009)). On montre que cette structure est transposable aux corrélations partielles par un changement de paramétrisation. De plus, on introduit plusieurs approches pour prendre en compte une structure parcimonieuse :
- procédure de test des corrélations partielles,
- régularisation de type LASSO.

Les performances de notre méthode sont illustrées dans un premier temps sur un exemple simulé. Les résultats obtenus sont prometteurs par rapport à plusieurs concurrents classiques. Enfin, appliquée à des données réelles dans un contexte de "génétique génomique", notre méthode permet la caractérisation fonctionnelle d'une région eQTL/QTL.

Inferring gene networks using a sparse factor gaussian graphical model

Yuna Blum[1,2,3], Magalie Houée[1,2,3] Chloé Friguet[4], Sandrine Lagarrigue[2,3] & David Causeur[1]

[1] *Laboratoire de Mathématiques Appliquées - Agrocampus Ouest, Rennes*
[2] *INRA, UMR598 Génétique Animale, Rennes*
[3] *Agrocampus Ouest, UMR598 Génétique Animale, Rennes*
[4] *UBS / CNRS - UMR 6205, Laboratoire de Mathématiques de Bretagne Atlantique*

Abstract. The availability of genome-wide expression data to complement the measurements of a phenotypic trait opens new opportunities for identifying biologic processes and genes that are involved in trait expression. Usually differential analysis is a preliminary step to identify the key biological processes implicated in the variability of the trait of interest. However, this variability shall be viewed as resulting from a complex combination of genes individual contributions. In other words, exploring the interactions between genes viewed in a network structure, which vertices are genes and edges stand for inhibition or activation connections gives much more insight on the internal structure of expression profiles. Many currently available solutions for network analysis have been developed, but an efficient estimation of the network from high-dimensional data is still a questioning issue. Extending the idea introduced for differential analysis by Friguet *et al.* (2009) [9] and Blum *et al.* (2010) [4], we propose to take advantage of a factor model structure to infer gene networks. This method shows good inferential properties and also allows an efficient testing strategy for the significance of partial correlations which provides an interesting tool to explore the community structure of the networks. As an alternative, we propose to introduce sparsity through LASSO regularization using a cyclic coordinate descent algorithm. We illustrate the performance of our method comparing it with competitors through a simulated example. Finally, our method is applied to a real dataset in a genetical genomic context, and we show its ability to extract an insightful network structure.

1 Introduction

Inference on gene networks from high throughput expression data is one of the most challenging issues in systems biology. Rather than uncovering single genes for complex traits, a system-based perspective is interesting in elucidating the interactions of genes and environment operating on a complex multicellular biological system [26]. Such a "system" approach involves modeling the relationship among elements of the system such as transcript levels in the form of a network.

Mathematical models for interaction graphs have recently been proposed for gene networks and inference procedures have been derived to estimate networks from transcription datasets [10]. Linear modeling of gene regulatory networks often appears as a simple and efficient solution, especially for inference from a large transcription profile observed on a few number of samples. In this context, partial correlations between two gene expressions given the remaining profile are viewed as measures of the interaction or co-expression between those two genes. Many currently available solutions for network analysis are therefore based on the so-called Gaussian Graphical Model, but an efficient estimation of the network from high-dimensional data is still a questioning issue.
In this framework, a multidimensional Gaussian variable is characterized by a concentration matrix, where conditional independence between pairs of variables is characterized by a zero entry. This matrix may be represented by an undirected graph, where each nodes represents a variable, and an edge connects two nodes if the corresponding pair of random variables are dependent, conditionally on the remaining variables. Thus, partial correlations allow the detection of direct genes interactions only.

2 GGM: general settings

Let Y be the observed data matrix with n rows, corresponding to the number of samples, and m columns, corresponding to the number of genes. Y is assumed to follow a multivariate normal distribution $Y \sim \mathcal{N}_m(\mu, \Sigma)$ with mean vector $\mu = (\mu_1, ..., \mu_m)'$ and positive-definite covariance matrix $\Sigma = (\sigma_{ij})$ with $i, j \in [1, m]$.

Let Π be the partial correlation matrix and $\pi_{i,j}$ the correlation between two variables (gene expressions) i and j conditionally on all the other variables. It can be shown that partial correlation matrix Π is related to the inverse of the covariance matrix Σ as follows [27]:

$$\pi_{i,j} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$$

with $\Sigma^{-1} = (\omega_{i,j})$ for $i, j \in [1, m]$
After a simple rescaling, the matrix Σ^{-1}, also known as the concentration matrix, can be interpreted as the adjacency matrix of an undirected weighted graph $\mathcal{G}$ representing partial correlation structure between variables $Y_1, ..., Y_m$

A challenging issue remains with Σ^{-1} estimation, as in high dimension ($n << m$), the estimated covariance matrix is not positive-definite and thus not invertible.
In order to illustrate this point, let us consider a simple simulated example in which $m = 200$. The R package WGCNA, which implements the methods presented in Langfelder and Horvath (2008) [14], offers a simulation tool for gene expression data with a modular interaction structure. Figure 1 displays an image plot of the sample partial correlation matrix of a dataset simulated using WGCNA with the following parameters: $m = 200$,

$n = 10000$, 5 modules of genes with respectively $m_1 = 80$ genes, $m_2 = 40$, $m_3 = 20$, $m_4 = 10$ and $m_5 = 10$. The former relative importance of modules is suggested by Langfelder and Horvath (2005) [29] in a tutorial document. The remaining $m - \sum_{i=1}^{5} m_i = 40$ genes are simulated with no mutual interaction. Because the partial correlation matrix is estimated on a large number of sampling items, it will be considered that the sample partial correlations matrix displayed in figure 1 is a close estimation of the true partial correlation matrix.

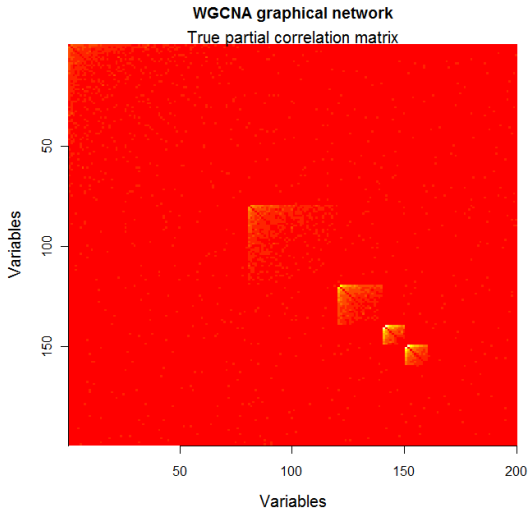


Figure 1: Image plot of the sample partial correlation matrix of a dataset simulated using the R package WGCNA with $m = 200$ variables and $n = 10000$ sampling items (red cells are for absolute values close to zero and yellow for higher absolute values).

Let us now assume that a sample of size $n = 50$ is randomly drawn from the above very large sample. Because the rank of the sample covariance matrix S calculated on this sample is of course much smaller than the number of variables, the calculation of S^{-1} is impossible. However, in a very basic first approach, sample partial correlation matrix can be obtained using the Moore-Penrose g-inverse S^{-} of S. The resulting sample partial correlations are displayed on figure 2. First, the image plot shows that the true modular structure is now very unclear. The obvious lack of association between the true partial correlation structure and the sample partial correlations is confirmed by the right plot of figure 2. Note also that the sample partial correlations are generally more dispersed than the true partial correlations.

This motivates shrinkage estimation procedures ensuring that the estimated partial correlations are smaller than a fixed level. In the next section, such alternative estimation methods are presented to improve the estimation of the true partial correlation matrix.

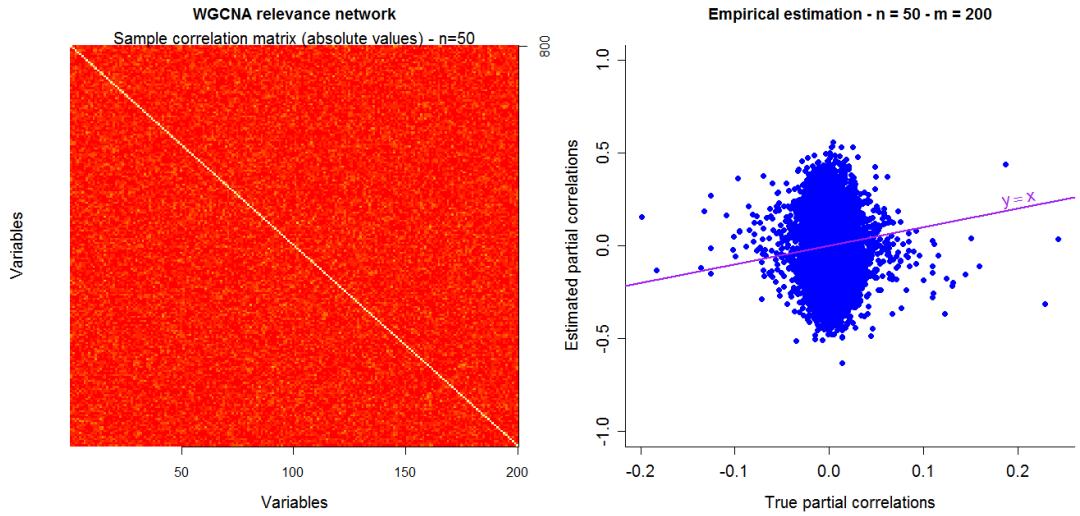


Figure 2: Left plot: image plot of the sample partial correlation matrix of a dataset simulated using the R package WGCNA with $m = 200$ variables and $n = 50$ sampling items (red cells are for absolute values close to zero and yellow for higher absolute values). Right plot: scatterplot showing that the sample partial correlations obviously poorly estimate the true partial correlations.

3 Existing methods based on GGM

Estimating concentration matrix in a high-dimensional context has received an increased scrutiny recently in the statistical literature. Most of the leading methods rely on two main related principles: estimate either the covariance matrix or directly the partial correlation matrix with a shrinkage method and extract the sparsity pattern of the partial correlation structure either directly or using a posterior multiple testing method.

3.1 Shrinkage estimation of the concentration matrix

Applying the principle of James-Stein estimation to address the issue of estimating the covariance matrix with high-dimensional data leads to a wide range of estimating methods often called *ridge* estimation. One of these bias-variance trade-off method was introduced by Schäfer and Strimmer (2005) [25] and is very popular in the biologist community. It relies on a shrinkage estimator of the covariance matrix by a convex linear combination of a target matrix T, which is supposed to be a low dimensional estimate of the covariance matrix, and S:

$$\Sigma_{shrink.} = \lambda T + (1 - \lambda)S$$

where $\lambda \in [0, 1]$ is a tuning hyper-parameter which controls the trade-off between bias and variance.

Usually, without any external biological knowledge suggesting a specific pattern for the target matrix, $T = \text{diag}(S)$, where diag is the matrix operator which sets to zero all the off-diagonal elements. Figure 3 illustrates the impact of shrinkage on the estimated partial correlations on the illustrative dataset introduced above, with $\lambda = 0.1$ and $\lambda = 0.9$.

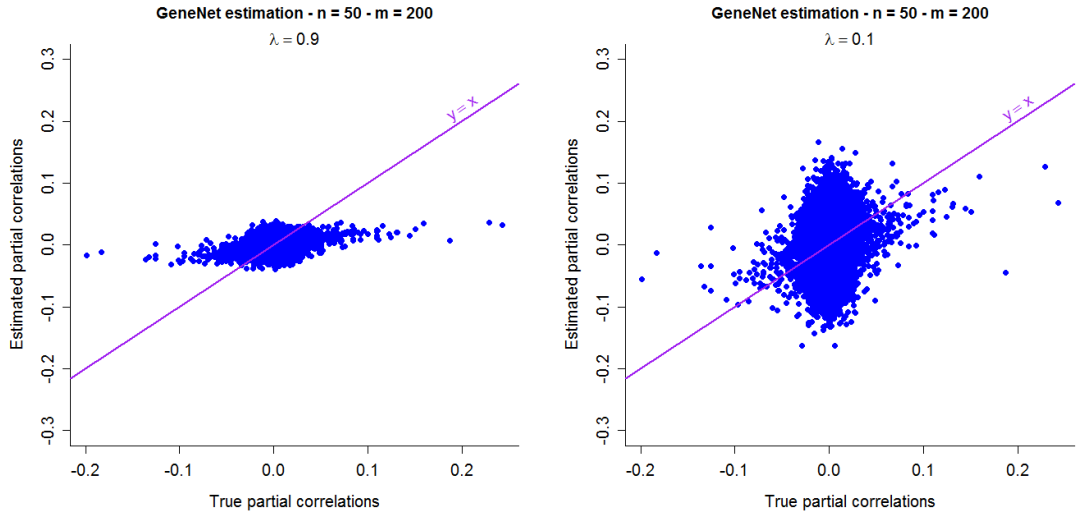


Figure 3: Scatterplots showing the impact of shrinkage in the GeneNet method on the estimation of partial correlations using the illustrative simulated dataset introduced in section 2.

The choice of the tuning parameter λ is of course crucial. Schäfer and Strimmer (2005) [25] propose a strategy which aim is to minimize the mean squared error of estimation of Σ and leads to the following explicit expression of the optimal λ when $T = \text{S}$:

$$\lambda = \frac{\sum_{i \neq j} \text{Var}(s_{ij})}{\sum_{i \neq j} s_{ij}^2}.$$

Figure 4 displays the final result obtained with the optimal λ.

To extract the sparsity pattern of the partial correlation structure, Schäfer and Strimmer (2005) [25] propose a significance testing procedure for the partial correlations based on a 2-component mixture model for the distribution of the estimated partial correlations:

$$f = \eta_0 f_0(\nu) + (1 - \eta_0) f_A,$$

where f is the density function of the mixture distribution, f_0 is the density function of the distribution of the estimated partial correlations under the null hypothesis, depending on

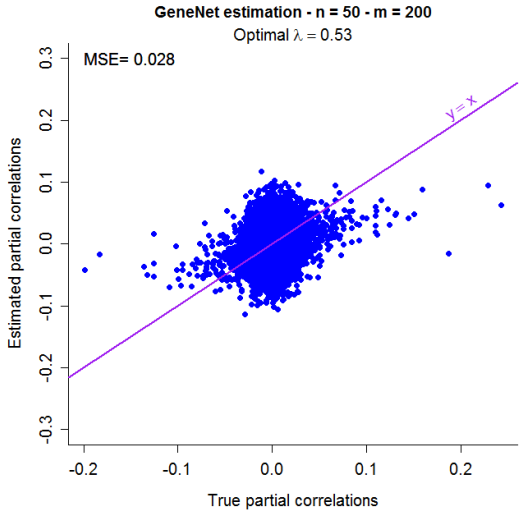


Figure 4: Scatterplot showing the result of the GeneNet method on the estimation of partial correlations using the illustrative simulated dataset introduced in section 2.

the number ν of degrees of freedom, f_A is the density function of the non-null distribution of the estimated partial correlations and η_0 is the mixing coefficient. A non-parametric estimation procedure is used to estimate f_A and η_0 [6, 7], and a posterior probability of no partial correlation is deduced with the local FDR:

$$\mathbb{P}(\pi_{ij} = 0 \mid \hat{\pi}_{ij}) \quad = \quad \ell\text{FDR}_{ij} = \hat{\eta}_0 \frac{f_0((\hat{\pi}_{ij}), \nu)}{\hat{f}_A(\hat{\pi}_{ij})}.$$

A hard-thresholding method is used on the above local FDR values to decide which partial correlations are significant. This method is implemented in the `R` package `GeneNet` [24].

3.2 Regularized estimation

An alternative route is offered by using regularized estimation of the covariance matrix Σ. Up to an additive constant, the log-likelihood of the multivariate normal model is given by:

$$-\frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)'\Sigma^{-1}(Y_i - \mu).$$

It is well-known that this log-likelihood is maximized with respect to μ at $\hat{\mu} = \bar{Y}$, for all Σ. As we focus on the estimation of the concentration matrix $\Omega = \Sigma^{-1}$, μ will hereafter

be replaced by its MLE in the definition of the deviance:

$$\begin{aligned}\mathcal{D}(\Omega) &= -n\log\det\big(\Omega\big) + \sum_{i=1}^{n}(Y_i-\bar{Y})\Omega(Y_i-\bar{Y}) \\ &= -n\log\det\big(\Omega\big) + n\,\mathrm{trace}\big[S\Omega\big],\end{aligned}$$

where $S = (1/n)\sum_{i=1}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})'$.

In order to introduce sparsity into the estimated partial correlation structure and simultaneously shrink the estimate of Ω, Banerjee *et al.* (2008) [3] propose to minimize the ℓ_1-penalized deviance defined as follows:

$$\mathcal{D}(\Omega;\lambda) = \mathcal{D}(\Omega) + \lambda||\Omega||_1,$$

where $\lambda \geq 0$ is a tuning parameter controlling the shrinkage and $||\Omega||_1 = \sum_{i=1}^{m}\sum_{j=1}^{m}|\omega_{ij}|$. Ambroise *el al.* (2009) [2] extends the former method to explicitly account for a latent modular structure of the graphical network.

Another ℓ_1-regularization approach is provided by considering the regression models for each variable against the others:

$$y_i = \beta_{0i} + \beta_{1i}y_1 + \ldots + \beta_{1,i-1}y_{i-1} + \beta_{1,i+1}y_{i+1} + \ldots + \beta_{1m}y_m + \epsilon_i.$$

The regression coefficients can indeed be used to obtain the partial correlations as follow:

$$\pi_{i,j} = sign(\beta_{i,j})\sqrt{\beta_{i,j}\beta_{j,i}}$$

Therefore, partial correlation estimation can be converted into the simultaneous estimation of m regression models. If, moreover, estimates of the regression models are obtained by minimization of a ℓ_1-penalized least-squares criterion, the resulting estimated matrix of slope coefficients is sparse. The so-called Graphical LASSO [8] method is based on the former principle, with additional symmetrization and positiveness restrictions on the matrix of estimated coefficients so that the corresponding matrix Π has the algebraical properties of a partial correlation matrix. [21] propose a modified version of the Graphical LASSO procedure, called Sparse PArtial Correlation Estimation (SPACE), which consists in minimizing a weighted sum over the m regression models of the least-squares criteria with a global ℓ_1-penalization.

The ℓ_1-penalization estimation of Ω is illustrated by using SPACE on the illustrative simulated dataset introduced in section 2. Figure 5 illustrates the impact of shrinkage on the estimated partial correlations on the illustrative dataset introduced above, with $\lambda = 0.1$ and $\lambda = 0.9$.

Following Yuan and Lin (2007) [28], a BIC criterion is also proposed by [21] to determine the optimal value of the shrinkage parameter λ:

$$\mathrm{BIC}(\lambda) = \mathcal{D}(\hat{\Omega}_\lambda) + \log(n)\times\tau(\lambda),$$

where $\hat{\Omega}_\lambda$ stands for the Space estimate of Ω with the value λ of the tuning parameter and $\tau(\lambda)$ is the number of non-zero entries in the upper triangular submatrix of Ω.

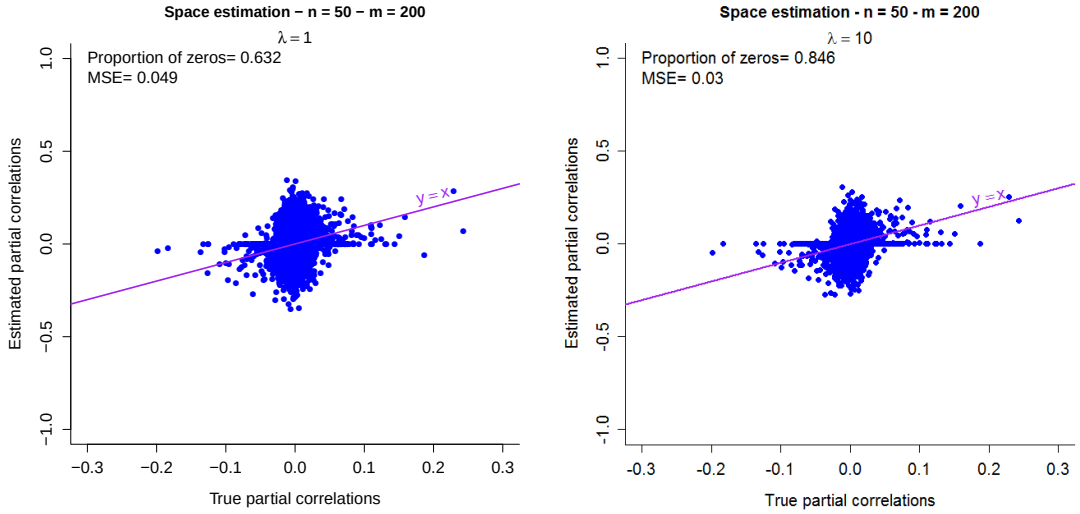


Figure 5: Scatterplots showing the impact of shrinkage on the estimation of partial correlations using the illustrative simulated dataset introduced in section 2.

4 Our proposal: sparse factor model

In the present situation where the number m of variables is much larger than the sample size n, the number of partial correlation parameters is very large, which motivates a modeling approach for the dependence structure.

As mentioned by Leek and Storey (2008) [17], it can be expected from factor modeling that it increases the overall efficiency of estimation, while keeping quite a large flexibility to model high-dimensional correlation structure. If a common factor structure is obvious, the observations on all the variables are indeed jointly involved in the estimation of each correlation parameter. Moreover, because of this interdependence between the parameters of the correlation model, introduction of sparsity conditions will non simply affect the estimation of nonzero parameters but also increase the efficiency of their estimation by reducing the number of effective parameters.

4.1 Factor Gaussian graphical model

The dependence within the expression profile Y_i is assumed to be modeled by a latent factor structure: there exists q, with $q \leq \min(n, m)$, unobserved factors $Z_i = (Z_{i1}, \ldots, Z_{iq})'$ independently and identically normally distributed with mean 0 and variance I_q, such that:

$$\mathbb{E}(Y_i \mid Z_i) \quad = \quad \mu + BZ_i \text{ and } \operatorname{Var}(Y_i \mid Z_i) = \Psi$$

where B is the $m \times q$ matrix of loadings with rank q and Ψ is a diagonal matrix which diagonal elements ψ_j^2 are positive. Note that the latent factors can be viewed as sources of dependence across the expression profile in the sense that, conditionally on Z_i, the components of Y_i are independent. It is straightforward checked that the above factor model assumption leads to the following structure for Σ: $\Sigma = \Psi + BB'$, where Ψ is referred to as the specific variance component and BB' as the common variance component. The term BZ_i is also referred to as the kernel of dependence by Leek and Storey (2008) [17].

Up to an additive constant, the resulting deviance of the above model is given by:

$$\mathcal{D}(\Psi, B) \;=\; n \log\det\big(\Psi + BB'\big) + n\,\mathrm{trace}\big[S\big(\Psi + BB'\big)^{-1}\big].$$

Direct maximization of $\mathcal{D}(\Psi, B)$ is described by Jøreskog (1969) [12] and is quite popular among users of factor analysis models. However, when the sample size n is small regarding the size m of the expression profile, this algorithm can be numerically unstable and cumbersome. An alternative EM algorithm is proposed by Rubin and Thayer (1982) [23], taking advantage of the factorization of the likelihood based on the conditional distribution of the profile given the latent factors (see chapter 2, section 2.1 of the manuscript, for a detailed presentation of the algorithm).

Note that the factor structure on Σ induces an equivalent factor structure on $\Omega = \Sigma^{-1}$, with the same number of factors, and using the following new parameterization:

$$\begin{aligned} \varphi &= \Psi^{-\frac{1}{2}}, \\ \theta &= \Psi^{-\frac{1}{2}} B (I + B'\Psi^{-1}B)^{-\frac{1}{2}}. \end{aligned} \tag{1}$$

The above parameterization is reversible:

$$\begin{aligned} \Psi &= \varphi^{-2}, \\ B &= \varphi^{-1}\theta(I_q - \theta'\theta)^{-\frac{1}{2}}. \end{aligned}$$

It is indeed straightforward deduced from the Woodbury's identity that $\Omega = \varphi(I_m - \theta\theta')\varphi$. In the following, the former decomposition will be referred to as the inverse factor model for Ω. Sparsity of the former factor structure is also assumed in the sense that, for some couples (r, s), with $r = 1, \ldots, m$ and $s = 1, \ldots, q$, the corresponding (r, s) loading θ_{rs} equals zero.

Using the functional invariance property of the maximum likelihood, the ML estimators $\hat{\varphi}$ and $\hat{\theta}$ of φ and θ are now derived by replacing Ψ and B by $\hat{\Psi}$ and $\hat{B}$ respectively using (1).

We now illustrate the gain in the estimation of the partial correlation matrix of a factor model by using the same illustrative simulated dataset introduced in section 2. Before fitting the model, we first need to determine the number of factors. Many methods are proposed in the literature to estimate the dimension of a factor model. Following Leek and Storey (2007, 2008) [18, 17], we propose to use the parallel analysis method described in Buja and Eyuboglu (1993) [5], which yields $\hat{q} = 3$.

Figure 6 illustrates the accuracy of estimation of partial correlations, regarding the performance of the other shrinkage methods introduced above.

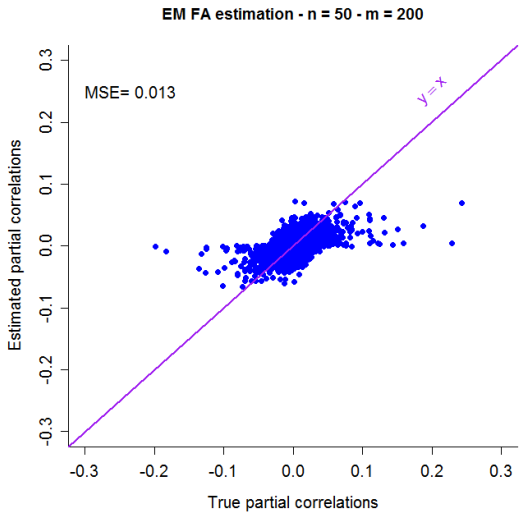


Figure 6: Scatterplots showing the impact of shrinkage on the estimation of partial correlations using the illustrative simulated dataset introduced in section 2.

In the present context of maximum-likelihood estimation, asymptotic tests for the significance of the parameters can be used to detect the zeros in θ and consequently the edges of the graphical network.

4.2 Significance testing in the inverse factor model

The parameters of the factor model are hereafter expressed as $(\text{vec}(\theta)', \text{diag}(\varphi)')'$, where vec(.) is the matrix operator transforming a $m \times q$ matrix into a vector of length mq obtained by the superposition of its row vectors and diag(.) is the operator transforming a diagonal matrix into a vector of length m of its diagonal elements. The asymptotic variance of the ML $(\text{vec}(\theta)', \text{diag}(\varphi)')'$ is given in the following proposition:

Proposition 1 *Let $\mathcal{H}$ denote the $m(q+1) \times m(q+1)$ Fisher's information matrix of the factor model. Then, $\mathcal{H}$ can be expressed as follow:*

$$\mathcal{H} = \begin{pmatrix} \mathcal{H}_\theta & \mathcal{H}_{\theta,\varphi} \\ \mathcal{H}'_{\theta,\varphi} & \mathcal{H}_\varphi \end{pmatrix},$$

where

$$\begin{aligned}
n^{-1}\mathcal{H}_\theta &= \Big[I_m \otimes (I_q - \theta'\theta)^{-1}\Big] \operatorname{vec}(\theta) \operatorname{vec}(\theta)' \Big[I_m \otimes (I_q - \theta'\theta)^{-1}\Big] + (I_m - \theta\theta')^{-1} \otimes (I_q - \theta'\theta)^{-1} \\
&\quad -(I_m - \theta\theta')^{-1} \otimes I_q, \\
n^{-1}\mathcal{H}_{\theta,\varphi} &= -\varphi^{-1} D'_\theta \Big[I_m \otimes (I_q - \theta'\theta)^{-1} + (I_m - \theta\theta')^{-1} \otimes I_q\Big], \\
n^{-1}\mathcal{H}_\varphi &= 2\varphi^{-2} + \varphi^{-1} D'_\theta \Big[I_m \otimes (I_q - \theta'\theta)^{-1}\Big] D_\theta \varphi^{-1} - \varphi^{-1} D'_\theta \Big[(I_m - \theta\theta')^{-1} \otimes I_q\Big] D_\theta \varphi^{-1},
\end{aligned}$$

where $\otimes$ *denotes the Kronecker product of matrices and* D_θ *is the following* $mq \times m$ *matrix:*

$$D_\theta = \begin{pmatrix} \theta_1 & 0 & \dots & 0 \\ 0 & \theta_2 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \theta_m \end{pmatrix}$$

Proof. see appendix B.

The Information matrix can be used to test the significance of partial correlation. Indeed, it is deduced from the standard ML theory that the asymptotic distribution of $\sqrt{n}(\operatorname{vec}(\hat{\theta}) - \operatorname{vec}(\theta))$ is normal with mean 0 and variance V_θ given by the $mq \times mq$ left upper block of $\mathcal{H}^{-1}$. Standard results on inversion of partitioned matrices can be used:

$$V_\theta = \mathcal{H}_\theta^{-1} + \mathcal{H}_\theta^{-1}\mathcal{H}_{\theta,\varphi}(\mathcal{H}_\varphi - \mathcal{H}'_{\theta,\varphi}\mathcal{H}_\theta^{-1}\mathcal{H}_{\theta,\varphi})^{-1}\mathcal{H}'_{\theta,\varphi}\mathcal{H}_\theta^{-1}.$$

We propose to deduce the sparse structure of the factor model from significance testing of the parameters $\theta_{rs},\ r = 1, \dots, m,\ s = 1, \dots, q,$, using the t-tests $t_{rs} = \hat{\theta}_{rs}/\sqrt{\hat{v}_{rs}}$, for $r = 1, \dots, m$ and $s = 1, \dots, q$, where v_{rs} is the (r, s) term in the matrix V_θ.

In the case of the illustrative WGCNA simulation scheme introduced in section 2, it is now arbitrarily assumed that the true graphical network is defined by considering that only the top 10 % of the highest partial correlations are true edges. This gives the adjacency matrix displayed in figure 7.

Figure 8 shows the histogram of p-values for the significance tests of θ in the inverse 3-factor model estimated above with the illustrative dataset. The plot also displays a semi-parametric decomposition of the density into a 2-component mixture of distributions, i.e a non-parametric component for the non-null p-values and a uniform component for the null p-values (see [22] for details). This mixture model for the p-values distribution provides both an estimate for the proportion π_0 of null hypotheses, here $\hat{\pi}_0 = 0.687$, and an estimate, for each θ_{rs}, of the so-called local FDR, $\ell\text{FDR}_{rs},\ r = 1, \dots, m,, s = 1, \dots, q$. In the present classification issue into the two groups of zero and nonzero correlations, ℓFDR_{rs} can be interpreted as the posterior probability that correlation between genes i and j is nonzero. We propose to consider that genes i and j are not correlated, if

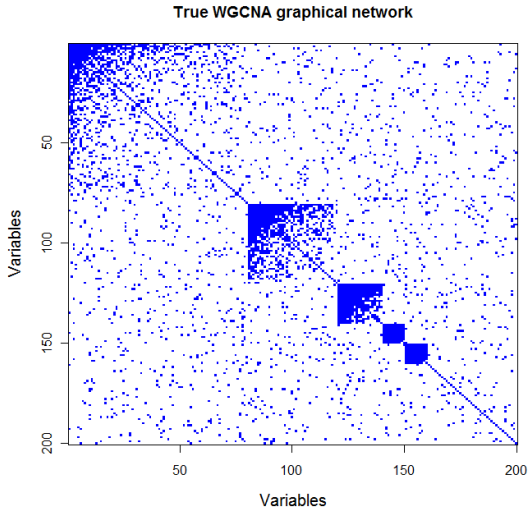


Figure 7: True graphical adjacency matrix obtained by hard-thresholding of the true partial correlation matrix of the WGCNA simulation scheme presented in section 2.

ℓFDR_{rs} exceeds a given threshold. In the following, this threshold is chosen so that an approximate proportion $\hat{\pi}_0$ of parameters are not significant. .

The estimated adjacency matrix of the estimated graphical network is reproduced in Figure 9. It clearly shows a good correspondence with the true graphical network represented in figure 7 (proportion of true negatives=0.91, proportion of true positives=0.32).

As an alternative, we propose in the next section to deduce sparsity using ℓ_1-regularized estimation of the inverse factor model

4.3 ℓ_1-regularized estimation of the inverse factor model

Using the new parameterization, the deviance of the inverse factor model can be expressed as follows:

$$\begin{aligned} n^{-1}\mathcal{D}(\varphi,\theta) &= -\log\det\big[\varphi(I_m-\theta\theta')\varphi\big]+\text{trace}\big[S\varphi(I_m-\theta\theta')\varphi\big], \\ &= -\log\det\big(\varphi^2\big)-\log\det\big(I_m-\theta\theta'\big)+\text{trace}\big[\varphi S\varphi(I_m-\theta\theta')\big], \\ &= -\log\det\big(\varphi^2\big)-\log\det\big(I_q-\theta'\theta\big)+\text{trace}\big(\varphi S\varphi\big)-\text{trace}(\varphi S\varphi\theta\theta')\big]. \end{aligned}$$

Cyclic coordinate algorithms are especially designed for situations where the number of parameters is large and direct minimization of the objective function is not straightforward. They have become very popular in statistics for high-dimensional data for penalized

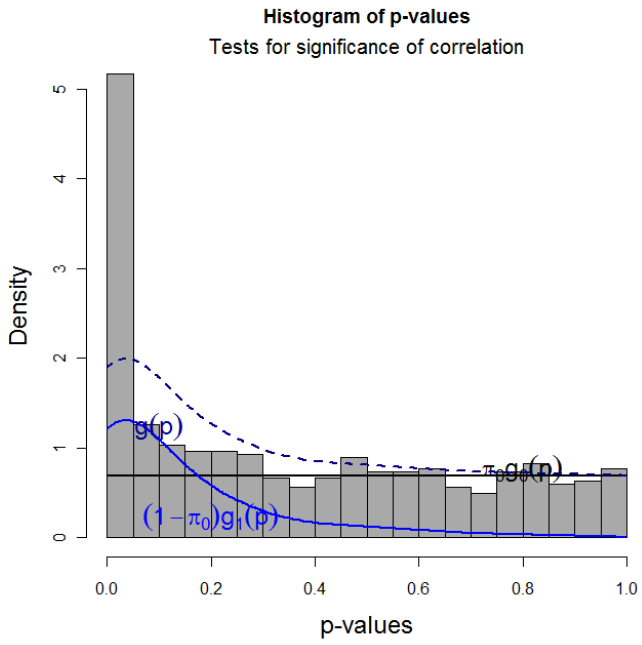


Figure 8: Histogram of the p-values for the significance tests of parameters in θ for the illustrative dataset, using the 3-inverse factor model. The density function g_1 of the nonnull p-values is unknown and the density function g_0 of the null p-values is uniform: $g_0(p) = 1$ for all p. Both the components of the mixture and the mixing parameter π_0 have been estimated by the semi-parametric method proposed by Robin *et al.* (2007) [22] and implemented in the R package kerfdr [11].

regression methods such as LASSO. In the following, a cyclic coordinate descent algorithm is proposed for ML estimation of (φ, θ). First, it is deduced from lemma A.2 that:

$$\begin{aligned} n^{-1}\frac{\partial}{\partial\theta_i}\mathcal{D}(\varphi,\theta) &= -\frac{\partial}{\partial\theta_i}\log\det\big(I_q-\theta'\theta\big)-\frac{\partial}{\partial\theta_i}\operatorname{trace}\big[\varphi S\varphi\theta\theta'\big], \\ &= 2(I_q-\theta'\theta)^{-1}\theta_i-2\varphi_i\theta'\varphi S_i, \end{aligned} \tag{2}$$

where S_i stands for the ith row of S.

Similarly,

$$\begin{aligned} n^{-1}\frac{\partial}{\partial\varphi_i}\mathcal{D}(\varphi,\theta) &= -\frac{\partial}{\partial\varphi_i}\log\det\big(\varphi^2\big)+\operatorname{trace}\Big[(I_m-\theta\theta')\frac{\partial}{\partial\varphi_i}\varphi S\varphi\Big], \\ &= -\frac{2}{\varphi_i}+2S_{ii}\varphi_i-2S_i'\varphi\theta\theta_i. \end{aligned}$$

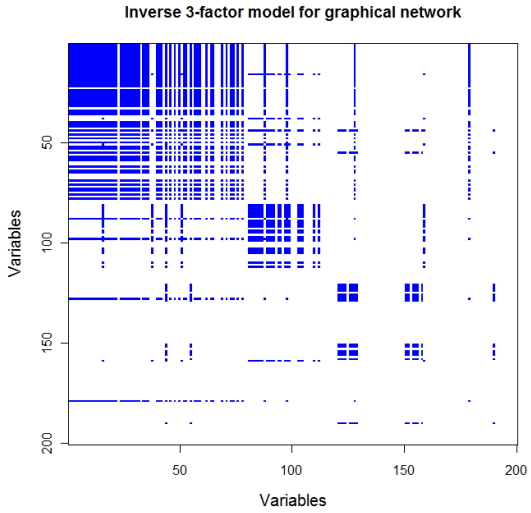


Figure 9: Estimated graphical adjacency matrix obtained by hard-thresholding of the p-values of the t-tests for significance of the parameters in the inverse 3-factor model.

Equating the above derivatives to zero leads to the following explicit expression of the maximum-likelihood estimators of φ_i, given the ML estimators of θ:

$$\hat{\varphi}_i^2 \quad = \quad \frac{1 + \hat{\theta}_i'(I_q - \hat{\theta}'\hat{\theta})^{-1}\hat{\theta}_i}{S_{ii}}.$$

However, there is no explicit solutions for the estimating equations (2). A cyclic coordinate descent algorithm is now proposed to overcome this issue.

First, it is deduced from (2) that the derivative of $\mathcal{D}(\varphi, \theta)$ with respect to its coordinate θ_{rs} is given by:

$$n^{-1}\frac{\partial \mathcal{D}}{\partial \theta_{rs}}(\varphi, \theta) \quad = \quad 2a_s(\theta)'\theta_r - 2\sum_{l=1}^{m}\theta_{ls}S^*_{rl},$$

where $S^* = \varphi S \varphi$, $a_s(\theta)$ stands for the sth column of the $q \times q$ matrix $A = (I_q - \theta'\theta)^{-1}$. At each turn of the cyclic coordinate descent algorithm, the estimate of θ_{rs} is updated by equating the above derivative to 0. It can be shown that:

$$n^{-1}\frac{\partial \mathcal{D}}{\partial \theta_{rs}} \quad = \quad 2\frac{\left[1 - \theta_{rs}u_{rs}\right]u_{rs} + \theta_{rs}v_{rs}}{\left[1 - \theta_{rs}u_{rs}\right]^2 - \theta_{rs}^2 v_{rs}} - 2\theta_{rs}S^*_{rr} - 2w_{rs},$$

where θ^*_{rs} denotes the matrix obtained after replacement of θ_{rs} by 0 in θ, $A^*_{rs} = (I_q - \theta^{*'}_{rs}\theta^{*'}_{rs})^{-1}$, a^*_s is the sth column of A^*_{rs}, $u_{rs} = a^{*'}_s\theta^*_r$, $\tau^*_r = \theta^{'*}_r A^* \theta^*_r$, a^*_{ss} is the sth diagonal

element of A^*, $v_{rs} = a^*_{ss}(1+\tau^*_r)$ and $w_{rs} = \sum_{l=1}^m \theta^*_{ls} S^*_{rl}$. At each step of the cyclic coordinate descent algorithm, $\hat{\theta}_{rs}$ is updated by equating the above derivative to zero, based on the previous estimation of θ obtained at the last iteration:

$$\begin{aligned} -\frac{\big[1-\theta_{rs}u_{rs}\big]u_{rs}+\theta_{rs}v_{rs}}{\big[1-\theta_{rs}u_{rs}\big]^2-\theta^2_{rs}v_{rs}}+\theta_{rs}S^*_{rr}+w_{rs} &= 0, \qquad (3)\\ \frac{-\big[1-\theta_{rs}u_{rs}\big]u_{rs}-\theta_{rs}v_{rs}+(\theta_{rs}S^*_{rr}+w_{rs}\big[(1-\theta_{rs}u_{rs})^2-\theta^2_{rs}v_{rs}\big]}{\big[1-\theta_{rs}u_{rs}\big]^2-\theta^2_{rs}v_{rs}} &= 0, \end{aligned}$$

$$\begin{aligned} &\tfrac{1}{[1-\theta_{rs}u_{rs}]^2-\theta^2_{rs}v_{rs}}\left\{\theta^3_{rs}S^*_{rr}(v_{rs}-u^2_{rs})+\theta^2_{rs}\left[2S^*_{rr}u_{rs}+(v_{rs}-u^2_{rs})w_{rs}\right]\right)\\ &+\theta_{rs}\left[v_{rs}-u^2_{rs}-S^*_{rr}+2u_{rs}w_{rs}\right]+u_{rs}-w_{rs}\big\}=0. \end{aligned}$$

Updating $\hat{\theta}_{rs}$ therefore consists in finding the roots of the cubic polynomial given in the numerator of the above equation. Note that, at each step, $\hat{\theta}$ must be such that $I-\hat{\theta}'\hat{\theta}$ is positive definite, which restricts the solutions of the updating equation to an interval for $\hat{\theta}_{rs}$.

Cyclic coordinate algorithms are usually initialized with zeros for the parameter values: $\hat{\theta}^{(0)}$ and $\hat{\varphi}^{(0)}_r = 1/\sqrt{S_{rr}}$. Using such initial values, the first update of $\hat{\theta}_{11}$ is obtained by solving (3) with $u_{rs}=0$, $v_{rs}=1$ and $w_{rs}=0$.:

$$\theta^3_{rs}S^*_{rr}+\theta_{rs}\Big[1-S^*_{rr}\Big] = 0.$$

Therefore, we propose to initialize the cyclic coordinate descent algorithm with the ML estimate of the inverse factor model.

Figure 10 shows the impact of the choice of λ on the estimated partial correlation matrix using the above algorithm. The same strategy as suggested by Banerjee (2008) [3] is also proposed here, consisting in a minimization of the BIC criterion:

$$\mathrm{BIC}(\lambda) = \mathcal{D}(\hat{\varphi}_\lambda,\hat{\theta}_\lambda)+\log(n)\times\tau(\lambda),$$

where $(\hat{\varphi}_\lambda,\hat{\theta}_\lambda)$ stands for the LASSO estimate of the invere factor model with the value λ of the tuning parameter and $\tau(\lambda)$ is the number of non-zero entries in the $\hat{\theta}_\lambda$.

Interestingly, both methods based either on significance testing or on LASSO regularization of the inverse factor model give equivalent results in terms of MSE.

In the next section, we apply the sparse factor model approach to a real dataset and compare the resulting network structure with the ones obtained with GeneNet and SPACE methods.

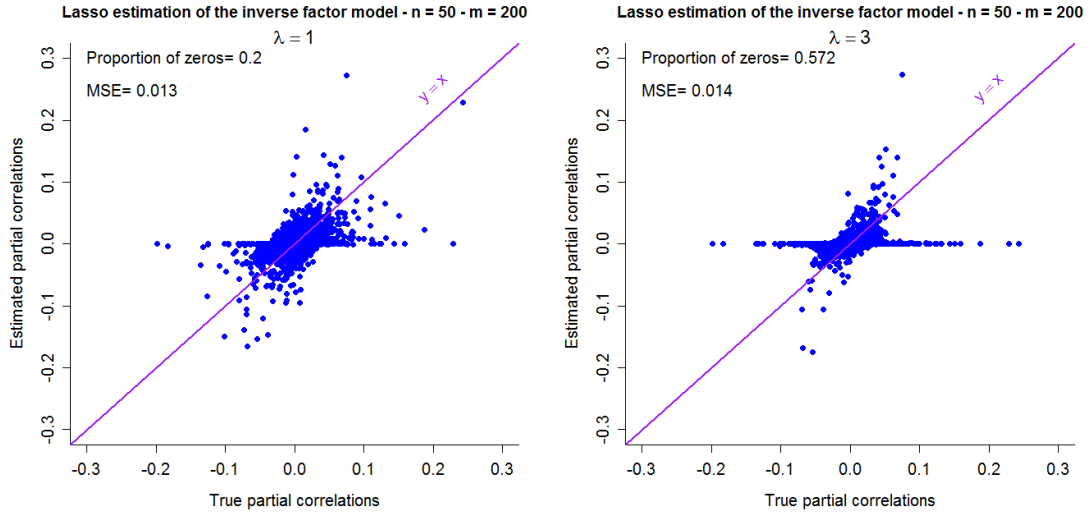


Figure 10: Scatterplots showing the estimated partial correlations using LASSO inverse 3-factor model approach with the illustrative simulated dataset introduced in section 2.

5 Application

We apply our method to a lipid metabolism study that aims at identifying biological processes and genes involved in abdominal fatness variability in chickens [16, 4], which is an economic trait of great interest. The biological purpose is here to better characterize a region of the genome known to control adiposity in chicken. This kind of regions are called QTL regions for Quantative Trait Loci, and can contain several tens of genes. After the detection of a QTL controlling a complex trait, the aim is to identify the causal mutation in the region which is a difficult task [13]. Finding genes which are controlled by this region can provide functional information about the causal mutation [15]. In this context, eQTL analyses were performed by [4] using transcriptome profiles of 45 chickens: 59 genes were found as being regulated by the QTL region (genes having an eQTL co-localizing with the QTL region). In order to better understand the biological mechanisms impacted by the eQTL/QTL region, inferring interactions between those 59 genes and extracting a particular network structure appears as a promising approach.

Our method based on a sparse factor model using significance testing to introduce sparsity is applied to the 59 genes. The parallel analysis methods leads to one factor in the model. We first compare the resulting adjacency matrix with the ones obtained using GeneNet and SPACE method. For fair comparisons, the default parameters are used for both algorithm without additional tuning. Figure 11 shows the heatmap of the adjacency

matrix and the corresponding graph for each methods. Interestingly, very few edges are detected using GeneNet which might result from a too high shrinkage of the extreme partial correlations, as observed in the simulated example (section 3.1). Compared to SPACE, the network obtained using our approach seems to be less sparse but exhibits a modular structure. Surprisingly, only 20% of the edges are in common.

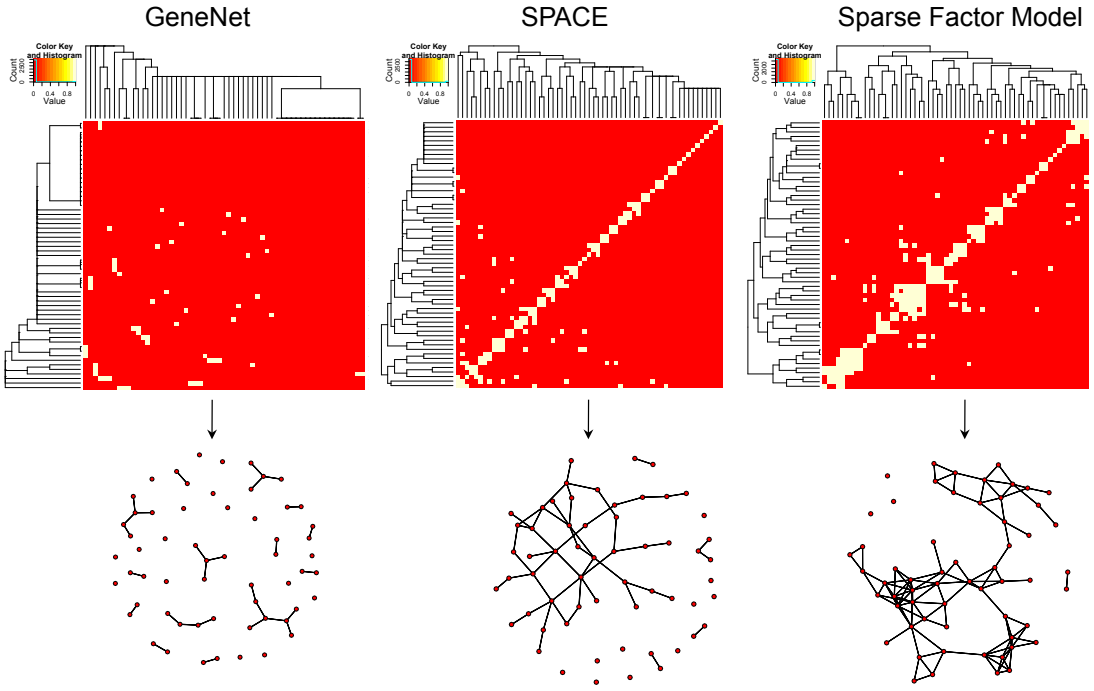


Figure 11: Heatmap of the adjacency matrix and corresponding graph for GeneNet, SPACE and Sparse Factor Model methods.

We focus on the network obtained using our approach which leads to a modular structure. We investigate the biological meaning of this modules using the annotation available. Only 50 % of the genes were annotated among the 59 genes (grey genes in figure 12). Nevertheless, several genes in a same module can be related to common biological processes (figure 12).

Interestingly three genes are implicated in the lipid metabolism and more precisely in the cholesterol metabolism, as DHCR7 encoding for the last enzyme of the cholesterol synthesis process. As several studies show relationships between cholesterol metabolism and obesity [19, 20], these three genes and their connections are maybe the functional signature of the causal mutation in the region. The network analysis provides thus a new functional hypothesis about one of the causal mutations that affect abdominal fatness in

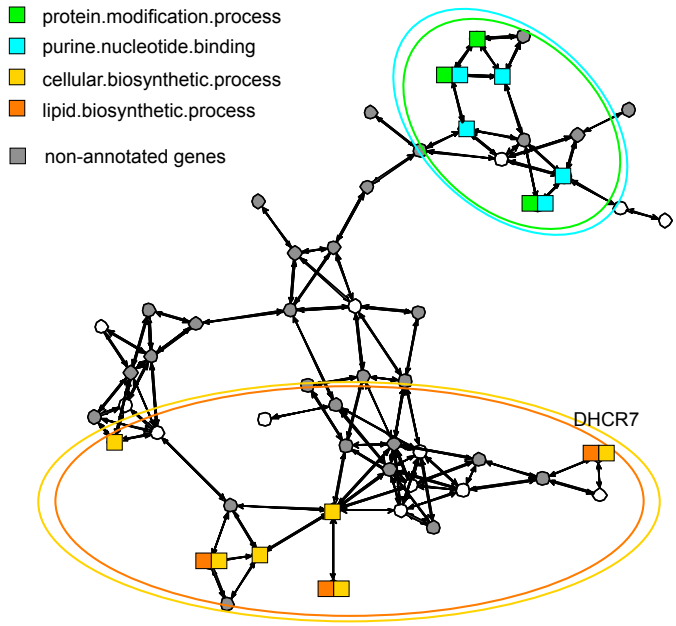


Figure 12: Inferred network of the 59 genes using a sparse Factor Model approach.

chickens. Further investigations will be necessary to confirm such a signature and clarify the role of these genes.

6 Conclusion

In this study, we develop a new approach to infer gene networks based on a factor model, in the context of Gaussian graphical model. Moreover, we propose two procedures to introduce sparsity in the network: one based on significance testing and the other on LASSO estimation. Through a simulated dataset, we show that our method achieves good performance in terms of partial correlation estimation in comparison with two competitors widely used by the biologist community: GeneNet and SPACE. Furthermore, we show that both sparsity procedures in our approach give similar results. Finally, on a real dataset, our method leads to a modular structure compared to the graph obtained either with SPACE or GeneNet, which where very sparse, as already notice in another study [1]. A deeper analysis of the network obtained using the functional annotation available provides interesting hypotheses about the function of a causal mutation and more generally, about the biological processes involved in adiposity in chickens.
Finally, we would like to mention that extensive simulations will be added soon to complete our study.

References

[1] J.D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao. Comparing statistical methods for constructing large scale gene networks. *PloS one*, 7(1):e29348, 2012.

[2] C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009.

[3] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

[4] Y. Blum, G. Le Mignon, S. Lagarrigue, and D. Causeur. A factor model to analyze heterogeneity in gene expression. *BMC bioinformatics*, 11(1):368, 2010.

[5] A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509–540, 1992.

[6] B. Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[7] B. Efron. *Local false discovery rates*. Division of Biostatistics, Stanford University, 2005.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[9] C. Friguet, M. Kloareg, and D. Causeur. A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1415, 2009.

[10] A. Goldenberg, A.X. Zheng, and S.E. Fienberg. *A survey of statistical network models*. Now Publishers, 2010.

[11] M. Guedj, S. Robin, A. Celisse, and G. Nuel. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC bioinformatics*, 10(1):84, 2009.

[12] K.G. Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202, 1969.

[13] S. Lagarrigue and M. Tixier-Boichard. Nouvelles approches de phénotypage pour la sélection animale. *Productions Animales*, 24(4):377, 2011.

[14] P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[15] G. Le Mignon, Y. Blum, O. Demeure, C. Diot, E. Le Bihan-Duval, P. Le Roy, and S. Lagarrigue. Apports de la génomique fonctionnelle à la cartographie fine de qtl. *Productions Animales*, 23(4):343, 2010.

[16] G. Le Mignon, C. Désert, F. Pitel, S. Leroux, O. Demeure, G. Guernec, B. Abasht, M. Douaire, P. Le Roy, and S. Lagarrigue. Using transcriptome profiling to characterize qtl regions on chicken chromosome 5. *BMC genomics*, 10(1):575, 2009.

[17] J.T. Leek and J.D. Storey. A general framework for multiple testing dependence.

[18] J.T. Leek and J.D. Storey. Capturing Heterogeneity In gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9), 2007.

[19] T.A. Miettinen and H. Gylling. Cholesterol absorption efficiency and sterol metabolism in obesity. *Atherosclerosis*, 153(1):241–248, 2000.

[20] P. Peltola, J. Pihlajamäki, H. Koutnikova, E. Ruotsalainen, U. Salmenniemi, I. Vauhkonen, S. Kainulainen, H. Gylling, T.A. Miettinen, J. Auwerx, et al. Visceral obesity is associated with high levels of serum squalene&ast. *Obesity*, 14(7):1155–1163, 2006.

[21] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[22] S. Robin, A. Bar-Hen, J.J. Daudin, and L. Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12):5483–5493, 2007.

[23] D.B. Rubin and D.T. Thayer. Em algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[24] J. Schäfer, R. Opgen-Rhein, and K. Strimmer. Reverse engineering genetic networks using the genenet package. *Journal of the American Statistical Association*, 96:1151–1160, 2006.

[25] J. Schäfer, K. Strimmer, et al. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):32, 2005.

[26] S.K. Sieberts and E.E. Schadt. Moving toward a system genetics view of disease. *Mammalian Genome*, 18(6):389–401, 2007.

[27] J. Whittaker. *Graphical models in applied multivariate statistics*, volume 16. Wiley New York, 1990.

[28] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[29] B. Zhang, S. Horvath, et al. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.

A Technical lemmas

Lemma A.1 *Let B denote a $m \times q$ matrix. For all $q \times m$ matrix A,*

$$\frac{\partial}{\partial B} \operatorname{trace}(AB) \;=\; A'.$$

For all $m \times m$ matrix S_1 and all $q \times q$ matrix S_2:

$$\frac{\partial}{\partial B} \operatorname{trace}(S_1 B S_2 B') \;=\; S_1 B S_2 + S_1' B S_2'$$

Lemma A.2 *Let θ denote any $m \times q$ matrix. If θ_i stands for the ith row of θ, then:*

$$\frac{\partial}{\partial \theta_i} \log\det(I_q - \theta'\theta) \;=\; -2(I_q - \theta'\theta)^{-1}\theta_i.$$

Let A be a symmetric $m \times m$ matrix, then:

$$\frac{\partial}{\partial \theta_i} \operatorname{trace}\big(A\theta\theta'\big) \;=\; 2\theta' a_i,$$

where a_i is the ith row of A.

Proof. According to the matrix cookbook, if θ_{ij} is the jth component of θ_i, then:

$$\begin{aligned}
\frac{\partial}{\partial \theta_{ij}} \log\det(I_q - \theta'\theta) &= \operatorname{trace}\big[(I_q - \theta'\theta)^{-1}\frac{\partial}{\partial \theta_{ij}}(I_q - \theta'\theta)\big], \\
&= -\operatorname{trace}\big[(I_q - \theta'\theta)^{-1}\frac{\partial}{\partial \theta_{ij}}\theta_i\theta_i'\big], \\
&= -2\sum_{k=1}^{q}\big[(I_q - \theta'\theta)^{-1}\big]_{jk}\theta_{ik},
\end{aligned}$$

which shows the first equation in lemma A.2.

Now,

$$\begin{aligned}
\frac{\partial}{\partial \theta_{ij}} \operatorname{trace}\big(A\theta\theta'\big) &= \operatorname{trace}\big(A\frac{\partial}{\partial \theta_{ij}}\theta\theta'\big), \\
&= 2\theta'^{(j)} a_i,
\end{aligned}$$

where $\theta^{(j)}$ is the jth column of θ. □

B Proof of proposition 1

Proof. The Fisher information matrix is given by $-\mathbb{E}(\mathcal{H})$, where $\mathcal{H}$ stands for the $(q+1)m \times (q+1)m$ Hessian matrix of the log-likelihood with respect to the vector $(\mathrm{vec}(\theta)', \mathrm{diag}(\varphi)')'$ of all the parameters. $\mathcal{H}$ can be partitioned as follows:

$$\mathcal{H} \;=\; \begin{pmatrix} \mathcal{H}_\theta & \mathcal{H}_{\theta,\varphi} \\ \mathcal{H}'_{\theta,\varphi} & \mathcal{H}_\varphi \end{pmatrix},$$

where

$$\mathcal{H}_\theta = \frac{\partial^2}{\partial\,\mathrm{vec}(\theta)\partial\,\mathrm{vec}(\theta)}\mathcal{L}(\varphi,\theta),\;\; \mathcal{H}_{\theta,\varphi} = \frac{\partial^2}{\partial\,\mathrm{vec}(\theta)\partial\,\mathrm{diag}(\varphi)}\mathcal{L}(\varphi,\theta)$$

and

$$\mathcal{H}_{\varphi,\varphi} = \frac{\partial^2}{\partial\,\mathrm{diag}(\varphi)\partial\,\mathrm{diag}(\varphi)}\mathcal{L}(\varphi,\theta).$$

Let us now recall the expression of the log-likelihood with the (φ,θ) parameterization:

$$\begin{aligned}
\mathcal{L}(\varphi,\theta) &= \frac{n}{2}\log\det\big[\varphi(I_m-\theta\theta')\varphi\big] - \frac{n}{2}\,\mathrm{trace}\big[S\varphi(I_m-\theta\theta')\varphi\big], \\
&= \frac{n}{2}\log\det\big(\varphi^2\big) + \frac{n}{2}\log\det\big(I_m-\theta\theta'\big) - \frac{n}{2}\,\mathrm{trace}\big[\varphi S\varphi(I_m-\theta\theta')\big], \\
&= \frac{n}{2}\log\det\big(\varphi^2\big) + \frac{n}{2}\log\det\big(I_q-\theta'\theta\big) - \frac{n}{2}\,\mathrm{trace}\big(\varphi S\varphi\big) + \frac{n}{2}\,\mathrm{trace}(\varphi S\varphi\theta\theta')\big],
\end{aligned}$$

Therefore, by lemma A.2:

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\mathcal{L}(\varphi,\theta) &= \frac{n}{2}\frac{\partial}{\partial\theta_i}\log\det\big(I_q-\theta'\theta\big) + \frac{n}{2}\frac{\partial}{\partial\theta_i}\,\mathrm{trace}\big[\varphi S\varphi\theta\theta'\big], \\
&= -n(I_q-\theta'\theta)^{-1}\theta_i + n\varphi_i\theta'\varphi S_i,
\end{aligned}$$

where S_i stands for the ith row of S.

Similarly,

$$\begin{aligned}
\frac{\partial}{\partial\varphi_i}\mathcal{L}(\varphi,\theta) &= \frac{n}{2}\frac{\partial}{\partial\varphi_i}\log\det\big(\varphi^2\big) - \frac{n}{2}\,\mathrm{trace}\big[(I_m-\theta\theta')\frac{\partial}{\partial\varphi_i}\varphi S\varphi\big], \\
&= \frac{n}{\varphi_i} - nS_{ii}\varphi_i + nS'_i\varphi\theta\theta_i.
\end{aligned}$$

The second-order derivatives are deduced: for $j = 1, \ldots, q$

$$\begin{aligned}
\frac{\partial}{\partial\theta_{ij}}\frac{\partial}{\partial\theta_i}\mathcal{L}(\varphi,\theta) &= -n\frac{\partial}{\partial\theta_{ij}}\big[(I_q-\theta'\theta)^{-1}\big]\theta_i - n(I_q-\theta'\theta)^{-1}\frac{\partial}{\partial\theta_{ij}}\theta_i + n\varphi_i\frac{\partial}{\partial\theta_{ij}}\theta'\varphi S_i,\\
&= n(I_q-\theta'\theta)^{-1}\frac{\partial}{\partial\theta_{ij}}\big[I_q-\theta'\theta\big](I_q-\theta'\theta)^{-1}\theta_i - n(I_q-\theta'\theta)^{-1}\frac{\partial}{\partial\theta_{ij}}\theta_i\\
&\quad +n\varphi_i\frac{\partial}{\partial\theta_{ij}}\theta'\varphi S_i,\\
&= -n(I_q-\theta'\theta)^{-1}\theta_i\Big[(I_q-\theta'\theta)^{-1}\Big]_j'\theta_i - n\Big[(I_q-\theta'\theta)^{-1}\Big]_j\theta_i'(I_q-\theta'\theta)^{-1}\theta_i\\
&\quad -n\Big[(I_q-\theta'\theta)^{-1}\Big]_j + n\varphi_i^2 S_{ii}e_j,
\end{aligned}$$

where e_j is the $q\times 1$ vector which jth component is 1 and the other components are 0. Hence:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_i^2}\mathcal{L}(\varphi,\theta) &= -n(I_q-\theta'\theta)^{-1}\theta_i\theta_i'(I_q-\theta'\theta)^{-1} - n\theta_i'(I_q-\theta'\theta)^{-1}\theta_i(I_q-\theta'\theta)^{-1} -\\
&\quad n(I_q-\theta'\theta)^{-1} + n\varphi_i^2 S_{ii}I_q,
\end{aligned}$$

For $k\neq i$,

$$\begin{aligned}
\frac{\partial}{\partial\theta_{kj}}\frac{\partial}{\partial\theta_i}\mathcal{L}(\varphi,\theta) &= -n\frac{\partial}{\partial\theta_{kj}}\big[(I_q-\theta'\theta)^{-1}\big]\theta_i + n\varphi_i\frac{\partial}{\partial\theta_{kj}}\theta'\varphi S_i,\\
&= n(I_q-\theta'\theta)^{-1}\frac{\partial}{\partial\theta_{kj}}\big[I_q-\theta'\theta\big](I_q-\theta'\theta)^{-1}\theta_i + n\varphi_i\frac{\partial}{\partial\theta_{kj}}\theta'\varphi S_i,\\
&= -n(I_q-\theta'\theta)^{-1}\theta_k\Big[(I_q-\theta'\theta)^{-1}\Big]_j'\theta_i - n\Big[(I_q-\theta'\theta)^{-1}\Big]_j\theta_k'(I_q-\theta'\theta)^{-1}\theta_i +\\
&\quad n\varphi_i\varphi_k S_{ik}e_j.
\end{aligned}$$

Hence:

$$\frac{\partial^2}{\partial\theta_i\partial\theta_k}\mathcal{L}(\varphi,\theta) = -n(I_q-\theta'\theta)^{-1}\theta_k\theta_i'(I_q-\theta'\theta)^{-1} - n\theta_k'(I_q-\theta'\theta)^{-1}\theta_i(I_q-\theta'\theta)^{-1} + n\varphi_i\varphi_k S_{ik}I_q.$$

Therefore, in matrix form:

$$\begin{aligned}
\mathcal{H}_\theta &= -n\Big[I_m\odot(I_q-\theta'\theta)^{-1}\Big]\operatorname{vec}(\theta)\operatorname{vec}(\theta)'\Big[I_m\odot(I_q-\theta'\theta)^{-1}\Big] -\\
&\quad n\Big[\theta(I_q-\theta'\theta)^{-1}\theta'\Big]\odot(I_q-\theta'\theta)^{-1} - nI_m\odot(I_q-\theta'\theta)^{-1} + n\Big[\varphi S\varphi\Big]\odot I_q.
\end{aligned}$$

Now, for the second-order derivatives with respect to φ: for $i = 1, \ldots, m$

$$\frac{\partial^2}{\partial\varphi_i\partial\varphi_i}\mathcal{L}(\varphi,\theta) = -\frac{n}{\varphi_i^2} - nS_{ii} + nS_{ii}\theta_i'\theta_i.$$

For $k \neq i$,

$$\frac{\partial^2}{\partial\varphi_k\partial\varphi_i}\mathcal{L}(\varphi,\theta) \;=\; nS_{ik}\theta'_k\theta_i.$$

Therefore, in matrix form:

$$\mathcal{H}_\varphi \;=\; -n\varphi^{-2} - n\,\mathrm{diag}(S) + nD'_\theta\Big[S\odot I_q\Big]D_\theta.$$

Finally, for the cross second-order derivatives: for $i = 1,\dots m$ and $j = 1,\dots,q$

$$\begin{aligned}\frac{\partial^2}{\partial\theta_{ij}\partial\varphi_i}\mathcal{L}(\varphi,\theta) &= nS'_i\varphi\frac{\partial}{\partial\theta_{ij}}\theta\theta_i,\\ &= nS'_i\varphi\theta^{(j)} + nS_{ii}\varphi_i\theta_{ij}.\end{aligned}$$

Hence:

$$\frac{\partial^2}{\partial\theta_i\partial\varphi_i}\mathcal{L}(\varphi,\theta) \;=\; nS'_i\varphi\theta + nS_{ii}\varphi_i\theta'_i.$$

Similarly, for $k \neq i$:

$$\frac{\partial^2}{\partial\theta_k\partial\varphi_i}\mathcal{L}(\varphi,\theta) \;=\; nS_{ik}\varphi_k\theta'_i.$$

Therefore, in matrix form:

$$\mathcal{H}_{\varphi,\theta} \;=\; n\begin{bmatrix} S'_1\varphi\theta & 0 & \dots & 0\\ 0 & S'_2\varphi\theta & & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & & S'_m\varphi\theta\end{bmatrix} + nD'_\theta\Big[(S\varphi)\odot I_q\Big].$$

The expectation of S is now needed: $\mathbb{E}(S) = \varphi^{-1}(I_m - \theta\theta')^{-1}\varphi^{-1} = \varphi^{-2} + \varphi^{-1}\theta(I_q - \theta'\theta)^{-1}\theta'\varphi^{-1}$. Equivalently: for $i,k = 1,\dots,m$,

$$\mathbb{E}(S_{ik}) \;=\; \frac{\delta_{ik} + \theta'_i(I_q - \theta'\theta)^{-1}\theta_k}{\varphi_i\varphi_k},$$

where δ_{ik} is the Kronecker's delta. Therefore,

$$\mathbb{E}\Big[\mathrm{diag}(S)\Big] \;=\; \varphi^{-2} + \varphi^{-1}D'_\theta\Big[I_m\odot(I_q - \theta'\theta)^{-1}\Big]D_\theta\varphi^{-1},$$

and

$$\mathbb{E}\Big[S\varphi\theta\Big] \;=\; \varphi^{-1}\theta(I_q - \theta'\theta)^{-1}.$$

$$\begin{aligned}\mathbb{E}\Big[\frac{\partial^2}{\partial\theta_i\partial\varphi_i}\mathcal{L}(\varphi,\theta)\Big] &= n\mathbb{E}(S_i')\varphi\theta + n\mathbb{E}(S_{ii})\varphi_i\theta_i', \\ &= \frac{n}{\varphi_i}\theta_i'(I_q-\theta'\theta)^{-1} + \frac{n}{\varphi_i}\theta_i'(I_q-\theta'\theta)^{-1}\theta_i.\theta_i'\end{aligned}$$

Similarly, for $k \neq i$,

$$\begin{aligned}\mathbb{E}\Big[\frac{\partial^2}{\partial\theta_k\partial\varphi_i}\mathcal{L}(\varphi,\theta)\Big] &= \mathbb{E}(nS_{ik}\varphi_k\theta_i') \\ &= n\frac{\theta_i'(I_q-\theta'\theta)^{-1}\theta_k}{\varphi_i}\theta_i', \\ &= \frac{n}{\varphi_i}\theta_k'(I_q-\theta'\theta)^{-1}\theta_i\theta_i'\end{aligned}$$

It is deduced from the standard ML theory that the asymptotic distribution of $\sqrt{n}(\text{vec}(\hat{\theta})-\text{vec}(\theta))$ is normal with mean 0 and variance V_θ given by the $mq \times mq$ left upper block of $\mathcal{H}^{-1}$. Standard results on inversion of partitioned matrices are used to obtain the asymptotic variance given in proposition 1. □

2.4 Conclusion

Dans ce travail, nous développons de nouvelles méthodes pour l'inférence de réseaux de gènes s'appuyant sur un modèle à facteurs. Plusieurs procédures sont proposées afin de prendre en compte une structure parcimonieuse :

- tests de significativité des corrélations ou corrélations partielles au travers des coefficients du modèle en facteurs (sections 2.2 et 2.3)
- régularisation de type LASSO (sections 2.2 et 2.3)
- apport d'informations biologiques a priori (section 2.2)

A la fois dans le contexte des *relevance networks* et dans celui des modèles graphiques Gaussiens, les résultats obtenus sur un exemple simulé sont encourageants par rapport à d'autres méthodes existantes. Cependant, avant soumission des deux articles présentés en sections 2.2 et 2.3, nous étofferons les simulations afin de compléter nos travaux.

Dans le cadre des *relevance networks*, notre approche permet d'extraire une structure modulaire plus précise que la méthode WGCNA qui est basée sur les corrélations empiriques : elle détecte mieux les petits modules ainsi que les gènes isolés.

Dans une deuxième étude focalisée sur les modèles graphiques Gaussiens, on montre que notre approche permet de mieux estimer les corrélations partielles par rapport à des méthodes concurrentes classiques (SPACE et GeneNet). Dans un contexte de "génétique génomique", on utilise notre méthode pour inférer le réseau génique sous-jacent à une région eQT/QTL. Le réseau obtenu permet d'identifier plusieurs modules géniques. Une analyse fonctionnelle des gènes peut permettre d'associer un module à un processus biologique. Dans notre étude, seulement 50 % des gènes du réseau ont une annotation fonctionnelle. Il est donc plus difficile d'identifier des processus biologiques communs entre les gènes d'un même module. Cependant, il est intéressant de remarquer que certains auteurs s'appuient sur cette démarche pour faire de la prédiction de fonctions biologiques sur les gènes non annotés (Carvalho et al. (2008)) : l'hypothèse sous-jacente est que si un gène fait partie d'un module génique caractérisé par un processus biologique particulier, alors il est probablement associé à ce processus. Dans notre réseau, un module semble caractérisé par des gènes impliqués dans la synthèse du cholestérol. Or plusieurs études ont montré un lien entre la production de cholestérol et l'obésité (Miettinen and Gylling (2000), Peltola et al. (2006)). Ainsi, cette observation suggère que les gènes de ce module sont les gènes impactés par la mutation causale de la région ce qui apporte une nouvelle hypothèse fonctionnelle sur cette mutation.

Discussion - Perspectives

Ce travail de thèse contribue à l'amélioration et au développement de méthodes utilisant des données transcriptomiques pour l'analyse génétique d'un caractère complexe.

Dans une première partie, on améliore les procédures classiques pour la recherche de gènes ou de régions du génome impliqués dans la variabilité du caractère d'intérêt. L'originalité de notre approche est de prendre en compte l'hétérogénéité d'expression dans les données transcriptomiques. Tout d'abord, on prend en compte l'hétérogénéité du signal du niveau d'expression à l'aide de la méthode FAMT (Friguet et al. (2009)). Cette méthode implémentée dans un package du logiciel R sous le nom de "FAMT" (Kloareg et al. (2009)), permet de capter la variabilité indépendante d'un caractère d'intérêt au travers de facteurs appelées facteurs d'hétérogénéité.
Nos travaux montrent l'intérêt de travailler sur des données d'expression ajustées par ces facteurs à la fois pour les analyses différentielles, les analyses eQTL et la détection de sous-types phénotypiques pour le caractère d'intérêt. En particulier, sur des données d'expression issues de dispositifs familiaux où les profils d'expression entre individus peuvent être peu variables étant donné leur degré d'apparenté, notre approche peut se révéler très fructueuse comparée à des approches classiques (Blum et al. (2010)). Par ailleurs, en utilisant de l'information extérieure au jeu de données concernant les individus (facteur âge par exemple) et les gènes (localisation de l'oligonucléotide correspondant sur la puce à ADN), on interprète partiellement les facteurs d'hétérogénéité. Il est intéressant de remarquer, que bien souvent, les données d'expression sont corrigées pour des facteurs connus par l'expérimentateur. Or on montre que même après correction de ces facteurs connus, la méthode FAMT identifie des facteurs d'hétérogénéité dont certains sont toujours associés à un effet de facteurs connus. Ce résultat confirme bien qu'il existe des facteurs inconnus pouvant avoir un impact sur le niveau d'expression et même pouvant être en interaction avec des facteurs connus ce qui expliquerait la persistance de l'effet de certains facteurs connus même après correction des données pour ces même facteurs.
Suite à ces travaux, une fonctionnalité à été incorporée au package FAMT (Kloareg et al. (2009)) permettant d'interpréter les facteurs d'hétérogénéité détectés à l'aide d'informations extérieures fournies par l'utilisateur.

Dans une deuxième partie, on propose de nouvelles méthodes pour l'inférence de réseaux géniques dans le contexte des *relevance networks* et celui des modèles graphiques Gaussiens. On

peut tout d'abord remarquer que ces deux grandes approches reposent sur des mesures de dépendances linéaires entre gènes. C'est pourquoi certains auteurs utilisent une autre mesure appelée *Mutual Information* permettant de mettre en évidence des dépendance non-linéaires. Cette mesure requiert la discrétisation des profils d'expression et est calculée ainsi entre deux gènes i et j : $MI_{i,j} = H_i + H_j - H_{ij}$, où H, l'entropie, est définie par : $H_i = -\sum_{k=1}^{n} p(x_k)log(p(x_k))$
Butte and Kohane (2000) l'ont utilisé en premier dans le cadre des *relevance networks* en choisissant, tout comme avec les corrélations, un certain seuil pour considérer une dépendance effective. D'autres approches utilisant cette mesure ont été développées par la suite et ont été rassemblées sous le nom de *Information theory based methods* (Margolin et al. (2006), Meyer et al. (2007), Liang and Wang (2008), Hausser and Strimmer (2009)). La méthode ARACNE (Margolin et al. (2006)) par exemple consiste à faire une première sélection des arêtes en choisissant un seuil comme proposé par Butte and Kohane (2000), puis elle considère en plus les triplets de gènes afin d'éliminer des liens indirects avec comme hypothèse que si le gène X_1 interagit avec le gène X_3 via le gène X_2, alors : $MI(X_1; X_3) \leq min(MI(X_1; X_2), MI(X_2; X_3))$. Ainsi, la méthode élimine dans certains cas, l'arête correspondant à la mesure MI la plus faible au sein d'un triplet de gènes qu'elle considère comme une interaction indirecte.

Des études ont comparé des méthodes basées sur des mesures de dépendances linéaires comme les méthodes WGCNA, GeneNet ou SPACE, à d'autres méthodes basées sur la mesure de dépendances non linéaires MI comme la méthode ARACNE. Aussi bien sur données réelles (Steuer et al. (2002)) que sur données simulées (Allen et al. (2012)), ces méthodes donnent globalement des résultats similaires. Dans l'étude comparative récemment réalisée par Allen et al. (2012), les résultats sont même meilleurs pour les méthodes basées sur des dépendances linéaires en termes de spécificité de détection des arêtes ainsi que d'identification des *hub* gènes dans le réseau. Un autre point sujet à discussion est l'hypothèse de normalité des profils d'expression comme soulevé par Giles and Kipling (2003). Cependant, Allen et al. (2012) montrent dans leur étude que sur des données simulées selon une distribution non-normale (mélange de deux distributions normales), la méthode GeneNet, qui fait pourtant l'hypothèse de normalité, donne de meilleurs résultats que la méthode ARACNE.
L'ensemble de ces observations peut s'expliquer finalement par le peu d'information contenu dans les données : l'expression d'un gène est mesurée sur peu d'individus. Ces résultats nous confortent dans l'idée que les hypothèses de linéarité des dépendances et de normalité des profils ne sont pas déraisonnables.

Dans un contexte de "génétique génomique" nous avons pu voir que la modélisation de réseaux géniques permet d'identifier des modules de gènes qui peuvent être associés à des processus biologiques apportant de nouvelles informations quant aux mécanismes de régulations sous-jacents à un caractère complexe. En particulier, nous avons utilisé une des méthodes développées (section 2.3) pour caractériser fonctionnellement une région eQTL/QTL.
Pour faciliter l'analyse des modules géniques, l'équipe de Steve Horvath propose de synthétiser l'information contenue dans chaque module par une variable unique appelée *eigengene* (Langfel-

der and Horvath (2007)). Il est alors possible de repérer rapidement quels sont les modules de gènes fortement corrélés au caractère d'intérêt en mesurant la corrélation entre l'expression du *eigengene* correspondant et les mesures du caractère. De plus, Ghazalpour et al. (2006) utilisent cette variable synthétique pour détecter des "module-QTL" et ainsi identifier des régions *hotspot* (régions contrôlant de nombreux gènes). Le Mignon et al. (2009) ont utilisé une démarche similaire consistant à calculer une variable synthétique pour un ensemble de gènes ayant un eQTL co-localisant avec une région QTL détectée au préalable : leur approche permet alors d'augmenter la puissance de détection du QTL et de gagner en précision sur sa localisation.
Par ailleurs, il est intéressant de remarquer que des études récentes proposent de combiner la construction de réseaux de gènes avec des informations de cartographie (Lum et al. (2006), Chu et al. (2009)). Lum et al. (2006) par exemple introduisent un nouveau concept de connectivité entre gènes basé sur la corrélation de Pearson mais aussi sur les régulations communes entre gènes : deux gènes seront connectés si la corrélation entre leurs expressions est supérieure à un certain seuil et s'ils ont des eQTL en commun. Dans la même idée, il serait intéressant dans l'approche que nous proposons en section 2.2, d'utiliser comme information biologique a priori, les eQTL partagés par les gènes. Le réseau obtenu permettrait d'extraire une structure modulaire tenant compte des régulations eQTL communes entre les gènes.

Enfin, comme évoqué en section 2.1, un réseau de régulation génique est une simplification du réseau biologique global impliquant différents niveaux de régulation (gènes/protéines/métabolites). De plus, une étude récente a montré que le niveau d'ARNm explique seulement 40 % de la variabilité du niveau de protéines ce qui suggère l'existence de nombreux processus de régulation lors de l'étape de la traduction (Schwanhäusser et al. (2011)). Ces résultats incitent à intégrer différentes fenêtres d'observation pour réellement comprendre les mécanismes biologiques sous-jacents à la variabilité de caractères complexes (voir figure 2.5).
Dans cette optique, j'ai participé à la génération de nouvelles données transcriptomiques et métabolites (lipidomiques) sur le même dispositif d'animaux décrit dans les travaux en sections 1.2.1 et 1.3.1, mais sur un échantillon de taille 4 fois plus importante ($\sim$200 individus). Les étapes d'acquisitions expérimentales et de pré-traitements des données sont résumées en Annexe. L'objectif des travaux futurs est d'intégrer les différentes données biologiques à la fois pour les analyses de cartographie et pour la modélisation de réseaux de régulation, afin de mieux comprendre les mécanismes biologiques contrôlant la variabilité d'adiposité chez le poulet de chair (figure 2.6). Comme illustré en figure 2.6, une première stratégie sera d'identifier des régions QTL/eQTL/mQTL (pour métabolite-QTL) contrôlant à la fois le caractère complexe d'intérêt, des gènes et des métabolites (lipides ou acides gras dans notre étude). Ce dispositif plus puissant que le précédent (plus d'individus mesurés) et intégrant différents niveaux d'observation devrait faciliter l'identification des mutations causales sur l'ensemble du génome.
Bien que différentes approches d'intégration génomique ont déjà été proposées dans la littérature (Workman et al. (2006), Ferrara et al. (2008), De Tayrac et al. (2009)), beaucoup reste à faire pour avoir une vision globale des interactions entre les différents niveaux de régulation.

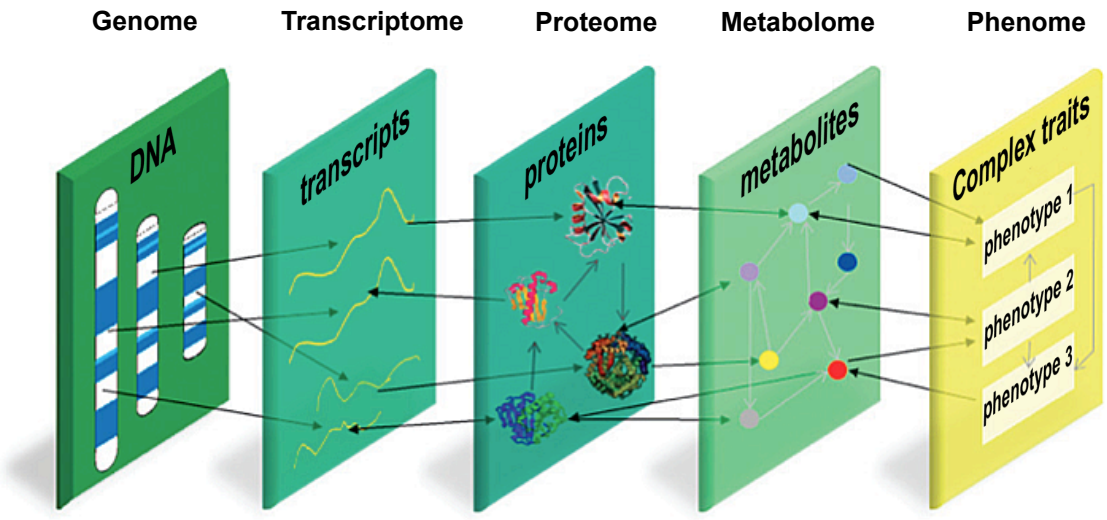


FIGURE 2.5 – Intégration de données biologiques pour l'analyse de caractères complexes (figure empruntée à Wu et al. (2009))

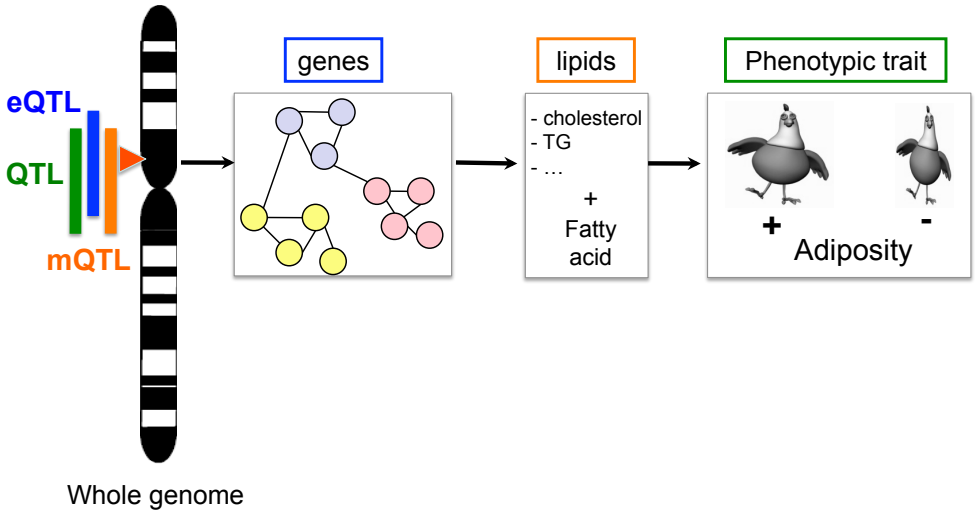


FIGURE 2.6 – Perspective d'intégration de données transcriptomiques, lipidomiques et phénotypiques pour une meilleure compréhension des mécanismes de régulations contrôlant l'adiposité chez le poulet de chair.

Liste des articles et communications

Articles publiés

Mach N, **Blum Y**, Bannink A, Causeur D, Houée M, Lagarrigue S, Smith M.A. Pleiotropic effects of polymorphism of the gene diacylglycerol-o-transferase 1 (DGAT1) in the mammary gland tissue of dairy cows. Journal of Dairy Science, 2012, 95:1-12

Blum Y, Le Mignon G, Causeur D, Filangi O, Désert C, Demeure O, Le Roy P, Lagarrigue S. Complex trait subtypes identification using transcriptome profiling reveals an interaction between two QTL affecting adiposity in chicken. BMC Genomics, 2011, 12:567

Lecerf F, Bretaudeau A, Desert C, **Blum Y**, Lagarrigue S and Demeure O. AnnotQTL: a new tool to gather functional and comparative information of a genomic region. Nucleic Acids Research. 2011

Blum Y, Le Mignon G, Lagarrigue S, and Causeur D. A Factor Model to Analyze Heterogeneity in Gene Expression. BMC Bioinformatics, 2010, 11:368. Highly Accessed

Le Mignon G, **Blum Y**, Demeure O, Diot C, Le Bihan-Duval E, Le Roy P, Lagarrigue S. Contribution of functional genomics to the fine mapping of QTL. Inra Prod.Anim. 2010, 23 (4), 343-358

Articles en préparation

Blum Y, Cadoret M, Houée-Bigot M and Causeur D. Sparse factor models for high dimensional relevance networks.

Blum Y, Friguet C, Houée-Bigot M, Lagarrigue S and Causeur D. Inferring gene networks using a sparse factor Gaussian graphical model.

Communications

Communications orales

Blum Y, Cadoret M, Houée M & Causeur D- Inférence sur réseaux d'interaction parcimonieux par modèles à facteurs, SFdS Journées de Statistique, Bruxelles, 21-25 May 2012.

Blum Y, Houée M, Friguet C, Lagarrigue S & Causeur D- Inferring gene networks using a sparse factor model approach, Statistical Learning and Data Science, Florence, Italy, 7-9 May 2012.

Blum Y, Houée M, Friguet C, Lagarrigue S & Causeur D- Inférence de réseaux de gènes à partir de données transcriptomiques. Séminaire de Biologie Intégrative et Genomique dans le Grand Ouest, Auray, 7-9 November 2011.

Blum Y, Demeure O, Désert C, Guillou H, Bertrand-Michel J, Filangi O, Le Roy P, Causeur D, Lagarrigue S - Integrating QTL controlling fatness, lipid metabolites and gene expressions to genetically dissect the adiposity complex trait in a meat chicken cross. 62nd Annual Meeting EAAP. August 2011, Norway

Mach N, Blum Y, Smits M, Causeur D, Lagarrigue S – Effect of the gene diacylglycerol-O-transferase 1 (DGAT1) polymorphism on the global expression pattern of genes in the mammary gland tissue of dairy cows. 62nd Annual Meeting EAAP. August 2011, Norway

Houée M, Friguet C, Lagarrigue S, Blum Y, Causeur D - Large-scale significance testing of high-thoroughput Data with FAMT. ASMDA 2011 Conference, Rome, 7-10 June 2011

Blum Y, Le Mignon G, Causeur D, Pitel F, Demeure O, Filangi O, Le Roy P, Lagarrigue S - Transcriptome profiling reveals interaction between two QTL for fatness in chicken. The 15th QTL-MAS workshop, Rennes, 19-20 May 2011.

Blum Y, Friguet C, Lagarrigue S & Causeur D - Inferring gene networks using a sparse factor model approach: application in a lipid metabolism study, International Biometric Society Channel Network, Bordeaux, 11-13 April 2011.

Blum Y, Lagarrigue S, Causeur D - Genetic analysis of a complex trait using transcriptomic data: contribution of gene regulatory network modeling. Séminaire des Thésards du Département de Génétique Animale INRA, Pornichet, 5-6 April 2011.

Blum Y, Lagarrigue S & Causeur D - Analyse génétique de caractères complexes, IFR140 Génétique Fonctionnelle, Agronomie et Santé, Rennes, 1srt December, 2010.

Blum Y, Lagarrigue S & Causeur D - A factor model to analyze heterogeneity in gene expression in a context of QTL characterization. International Society for Animal Genetics, Edinburgh, 26-30 July 2010.

Blum Y, Lagarrigue S & Causeur D - Caractérisation fonctionnelle d'un QTL par prise en compte de la dépendance génique, Journée des Doctorants de Rennes & Brest en Biologie-Santé, Rennes, June 2010.

Blum Y, Friguet C, Lagarrigue S & Causeur D - Inférence sur réseaux géniques par Analyse en Facteurs. SFdS Journées de Statistique, Marseille, 25-28 May 2010.

Blum Y, Lagarrigue S & Causeur D - A factor model to analyze heterogeneity in gene expression in a context of QTL mapping. 8th workshop « Statistical Methods for Post-Genomic Data» Luminy, Marseille, 14-15 January 2010.

Posters

Blum Y, Houée M, Friguet C, Lagarrigue S & Causeur D - Genetic analysis of a complex trait using transcriptomic data: contribution of gene regulatory network modeling. Session "Genes and pathways". Plant&Animal Genome XX, San Diego CA USA, 14-18 January 2012.

Blum Y, Modélisation de réseaux de gènes : apport dans le déterminisme génétique de caractères complexes. Séminaire des Thésards du Département de Génétique Animale INRA, Pornichet, 6-7 April 2010.

Blum Y, Lagarrigue S & Causeur D - A factor model to analyze heterogeneity in gene expression in a context of QTL characterization. International Society for Animal Genetics, Edinburgh, 26-30 July 2010.

Blum Y, Le Mignon G, Causeur D, Pitel F, Demeure O, Filangi O, Le Roy P, Lagarrigue S - Transcriptome profiling reveals interaction between two QTL for fatness in chicken. International Society for Animal Genetics, Edinburgh, 26-30 July 2010.

Prix et bourses

INRA-Agrocampus Ouest Grant to attend the PAG XX workshop (Jan. 2012, San Diego, USA) and present the poster "Genetic analysis of a complex trait using transcriptomic data: contribution of gene regulatory network modeling."

Mobility Grant from Rennes Metropole: 4 months in the Jake Lusis Laboratory at UCLA (Los Angeles, USA) from January to April 2012

Yoshi Suzuki Award for the best Abstract and Poster - A factor model to analyze heterogeneity in gene expression in a context of QTL characterization. International Society for Animal Genetics, Edinburgh, 26-30 July 2010.

ISAG bursary to attend the International Society for Animal Genetics workshop (Edinburgh, 26-30 July 2010) for an oral communication and the presentation of two posters : "A factor model to analyze heterogeneity in gene expression in a context of QTL characterization" and "Transcriptome profiling reveals interaction between two QTL for fatness in chicken".

Prize of the best Poster: "Modélisation de réseaux de gènes : apport dans le déterminisme génétique de caractères complexes". Séminaire des Thésards du Département de Génétique Animale INRA, Pornichet, 6-7 April 2010.

Annexe A

AnnotQTL

Au cours de ma thèse, j'ai contribué à la mise en place d'un outil web ayant pour objectif de faciliter l'identification des meilleurs candidats fonctionnels présents dans une région QTL. Ces travaux font l'objet de l'article suivant :

Lecerf F, Bretaudeau A, Desert C, **Blum Y**, Lagarrigue S and Demeure O. AnnotQTL: a new tool to gather functional and comparative information of a genomic region. Nucleic Acids Research. 2011

Apport de l'outil

L'objectif de cet outil est de repérer automatiquement les localisations et annotations fonctionnelles standardisées des gènes d'une région QTL. Les gènes sont localisés à l'aide des bases de données de NCBI et Ensembl et les annotations fonctionnelles standardisées sont extraites à partir d'autres bases (comme Gene Ontology ou HGNC) avec gestion de la redondance. Une option permet de mettre en surbrillance dans la région d'intérêt, les gènes caractérisés par une annotation fonctionnelle en lien avec le caractère choisi par l'utilisateur. Cet outil permet ainsi d'identifier très rapidement les meilleurs candidats dans une région QTL. De plus, il est possible d'identifier des synthénies pour différentes espèces (*Homo sapiens*, *Mus musculus*) et donc de faire des comparaisons inter-espèces des localisations de QTL.

Nucleic Acids Research Advance Access published May 19, 2011

Nucleic Acids Research, 2011, 1–6
doi:10.1093/nar/gkr361

AnnotQTL: a new tool to gather functional and comparative information on a genomic region

F. Lecerf[1,2,*], A. Bretaudeau[3], O. Sallou[3], C. Desert[1,2], Y. Blum[1,2,4], S. Lagarrigue[1,2] and O. Demeure[1,2]

[1]INRA, UMR598 Génétique Animale, F-35000 Rennes, [2]Agrocampus OUEST, UMR598 Génétique Animale, F-35000 Rennes, [3]GenOuest Platform, INRIA/Irisa – Campus de Beaulieu, F-35042 Rennes Cedex and [4]Agrocampus OUEST, Applied Mathematics Department, F-35000, Rennes, France

Received February 9, 2011; Revised April 18, 2011; Accepted April 27, 2011

ABSTRACT

AnnotQTL is a web tool designed to aggregate functional annotations from different prominent web sites by minimizing the redundancy of information. Although thousands of QTL regions have been identified in livestock species, most of them are large and contain many genes. This tool was therefore designed to assist the characterization of genes in a QTL interval region as a step towards selecting the best candidate genes. It localizes the gene to a specific region (using NCBI and Ensembl data) and adds the functional annotations available from other databases (Gene Ontology, Mammalian Phenotype, HGNC and Pubmed). Both human genome and mouse genome can be aligned with the studied region to detect synteny and segment conservation, which is useful for running inter-species comparisons of QTL locations. Finally, custom marker lists can be included in the results display to select the genes that are closest to your most significant markers. We use examples to demonstrate that in just a couple of hours, AnnotQTL is able to identify all the genes located in regions identified by a full genome scan, with some highlighted based on both location and function, thus considerably increasing the chances of finding good candidate genes. AnnotQTL is available at http://annotqtl.genouest.org.

INTRODUCTION

The final steps of genetic mapping research programs require close analysis of several QTL regions to select candidate genes for further studies. Despite several websites (NCBI genome browser, Ensembl Browser, UCSC Genome Browser) or web tools (Biomart, Galaxy) developed to achieve this task, the selection of candidate genes remains a laborious process. The information made available on the more prominent web sites differs slightly in terms of gene prediction and functional annotation, while other websites provide extra information that researchers may want to use (HGNC approved gene symbols, Gene Ontology (GO) Annotation or functional data, conservation of synteny with other species, etc.). It is possible to manually merge and compare this information for one QTL containing few genes, but not for many different QTL regions containing dozens of genes.

Here, we propose a web tool that, for a given region of interest, merges the list of genes available in NCBI and Ensembl, removes redundancy, adds functional annotations from different prominent web sites, and highlights the genes for which functional annotation fits the biological function or diseases of interest. The tool is dedicated to sequenced species of livestock including cattle, pig, chicken and horse as well as dog, i.e. species that have been extensively studied (with over 8000 QTLs detected; see http://www.animalgenome.org/cgi-bin/QTLdb/index). Nevertheless, because of the family designs and the low number of animals used in these species, most of the studies use linkage analysis, and the QTL regions identified remain large (containing dozens of genes). Conversely, in human and model species, most analyses now draw heavily on association studies involving large cohorts, thus providing more power and accuracy, and the web tools already available focus on these species through functional annotation of SNPs in association with the trait (1–8). Most of these tools focus on the SNP annotation itself, describing whether the SNP is located in a gene, or even in a coding sequence, and defining if it could have a functional effect. While these web tools are highly efficient in providing a good annotation for specific SNPs, they

*To whom correspondence should be addressed. Tel: +(33) 2 23 48 59 62; Fax: +(33) 2 23 48 54 70; Email: lecerf@agrocampus-ouest.fr

clearly cannot be used to collect information on the large regions obtained in livestock species.

METHODS

The main objective of AnnotQTL is to minimize redundancy so as to display the maximum amount of information from several sources on the genes in the region of interest. The main AnnotQTL program is implemented in PERL. The data are downloaded from several FTPs or websites (see Figure 1 for details on the data and fields used) and stored on our server for further computation. Location and annotation data from Ensembl are downloaded via BioMart (9) using MartService. AnnotQTL gives several sets of information from comparative mapping of selected species against the human and mouse genomes using a local dump of the data provided by the Narcisse web site (10) and orthologous gene information from the Ensembl comparative database. PERL scripts import the downloaded files into our SQL databases. All PERL scripts and official GO database updates are inserted into a BioMaj (11) workflow to automate the updating process. Updates are performed monthly.

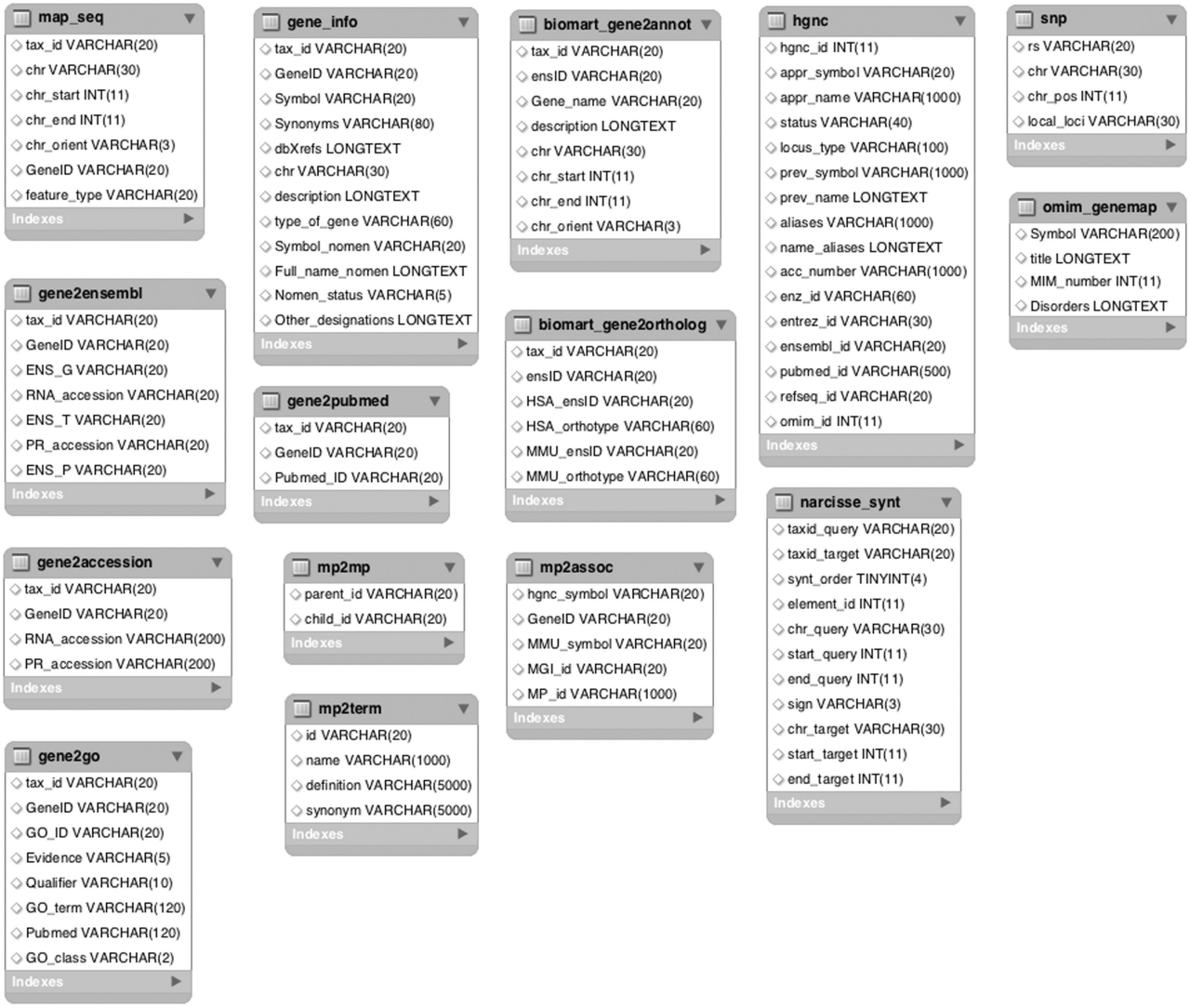


Figure 1. Schematic diagram of the database and source data files. The table map_seq is filled using file xxx_seq_gene.md.gz, where xxx is the species name, located in the NCBI FTP directory: /genomes/xxx/mapview. The tables gene_info, gene2accession, gene2go, gene2ensembl and gene2pubmed are filled using data files stored in the NCBI FTP directory: /gene/DATA. The table biomart_xxx is filled using the BioMart service for the Ensembl databases. For each species, a SQL table is created to store SNP data (here, only one is detailed). The data files are downloaded from NCBI FTP directory: /snp/organisms/xxx/chr_rpts (where xxx is species name). The tables mp2mp, mp2term and mp2assoc are filled using files HMD_HumanPhenotype.rpt and MPheno_OBO.ontology available from the MGI FTP site (ftp.informatics.jax.org/pub/reports). The hgnc table is filled using the data stored at the GeneName website together with its LWP agent (http://www.genenames.org/cgi-bin/hgnc_downloads.cgi). The table omim_genemap is filled using the data file located in the NCBI FTP directory /repository/OMIM. The table narcisse_synt is filled using the comparative data provided by the Narcisse website (http://narcisse.toulouse.inra.fr).

The principle and workflow of AnnotQTL is depicted in Figure 2. Starting with the genome coordinates entered by the user, the program extracts the NCBI GeneID of genes contained in the region, the corresponding annotations (name, description, symbol and so on), plus the associated cross-references (RNA accession number, protein accession number and Ensembl identifier) and the Pubmed identifier. This Pubmed identifier is specific of the requested species and does not list the publications related to this gene in other species. Using the same genome coordinates, the program then extracts Ensembl ID, gene annotation and human and mouse ortholog gene identifiers from the Ensembl database.

The main step now is to remove the redundancy between NCBI and Ensembl data while keeping the specific annotation of both databases. As there is a slight difference in gene location between the two web sites, the filtering process cannot be based on gene location, which leaves two approach options. The first is to use the Ensembl cross-reference provided by NCBI. However, this approach is not exhaustive since few cross-references are missing even for genes annotated in both databases. A second strategy has therefore been developed based on a textual query search in the annotation fields provided by the two sites. Values for the symbol, synonyms, RNA accession and protein accession fields from NCBI are compared against the values in the Ensembl gene name field for the gene of the species of interest. When one or more of these fields match, all the information is combined under one record, thus removing duplicates and enriching the annotation (without losing the annotation specific to both sites). If available, the gene annotations of human and mouse orthologs are also included in this comparison. Each record is also filtered for potential intra-redundancy of annotation between the gene and its orthologs (i.e. the same gene description is found between the requested species and Human or mouse orthologs). This set of genes combining NCBI- and Ensembl-specific information is then compared to the HGNC database. The goal of this procedure is to retrieve the HGNC approved symbol by searching for correspondences between annotation fields and the symbol or aliases fields of the HGNC database. Then, the values found in the HGNC database (symbol and OMIM identifier, if any) are included in the final results output displayed. If the OMIM identifier is still undefined, a search through the OMIM symbol fields is performed using the HGNC symbol or aliases. Where relevant, OMIM identifier, title and related disorders are retrieved from the OMIM database. Finally, the user-specified genome region is cross-compared against the Narcisse database to fetch the human or mouse-orthologous genomic region.

To clarify the output and adapt it to the scientist's query, certain information is only available through menu options. Human and mouse orthology information from Ensembl can be used to more accurately define certain genomic regions left undefined in Narcisse data. Users can also select level of synteny (synt order, see (10) for more details) between studied species and target species (human or mouse). Another option is to upload a set of genetic markers (which can be of any type provided physical location is given) to be inserted in the final results display. User can choose to keep their own marker locations or re-map markers to NCBI genome coordinates (only available with approved marker identifiers). A fourth non-processed column is available for displaying user-defined information. Adding the markers to the results display should ease the identification of the genes that most closely match the most significant markers. Finally, AnnotQTL can highlight genes based on functional annotations provided by GO, Mammalian Phenotype (MP), or OMIM disorders. For GO or MP terms organized in a hierarchical 'parent–children' directory structure, user-inputted keywords provide options for selecting the corresponding terms and associated children. For OMIM, a query is performed against OMIM disorders data retrieved in the previous step with user-input keywords: if the keywords matched, then the OMIM disorders are highlighted in the display. For GO, the genes are highlighted if their GeneID matches with the GO association provided by NCBI. As they do not have a GeneID, the match-up between GO annotation and genes specific to the Ensembl database is based on their HGNC name, where available. Users can improve this 'GOA highlight function' by adding the GOA from human and/or mouse species from orthologs to current genes. For MP, genes are highlighted if their approved symbols match the HGNC approved symbols stored in the MP database. The aim here is to provide functional information and facilitate the identification of genes linked to the trait-of-interest (i.e. functional candidate genes).

APPLICATION

To demonstrate the utility of AnnotQTL and test the efficiency of this web tool, we present different examples using real data aimed at identifying functional and positional candidate genes.

The first example focuses on a bovine mutation controlling muscular hypertrophy. In 1995, the mutation was mapped to the extremity of the BTA2 in a 12 cM interval (12). Using AnnotQTL on this region of the bovine chromosome (0–8 Mb) retrieved the location and functional annotation of 95 genes. We then applied the 'GO highlight function' on this region in two separate queries, using 'muscle' and 'growth' as keywords best describing the observed phenotype. These two terms highlighted two and three genes, respectively, from these 95 genes. Both lists highlighted the MSTN (GDF8) gene, which has been demonstrated as the validated causal gene (myostatin) (13).

A second step analyzed a more extensive set of 21 QTL regions shaping abdominal fatness in chickens (14,15). Average length of these regions was 4.8 Mb. After running AnnotQTL, all the regions were enriched with genes by comparing NCBI and Ensembl information against information provided by either NCBI or Ensembl only (Table 1). For all the genomic regions,

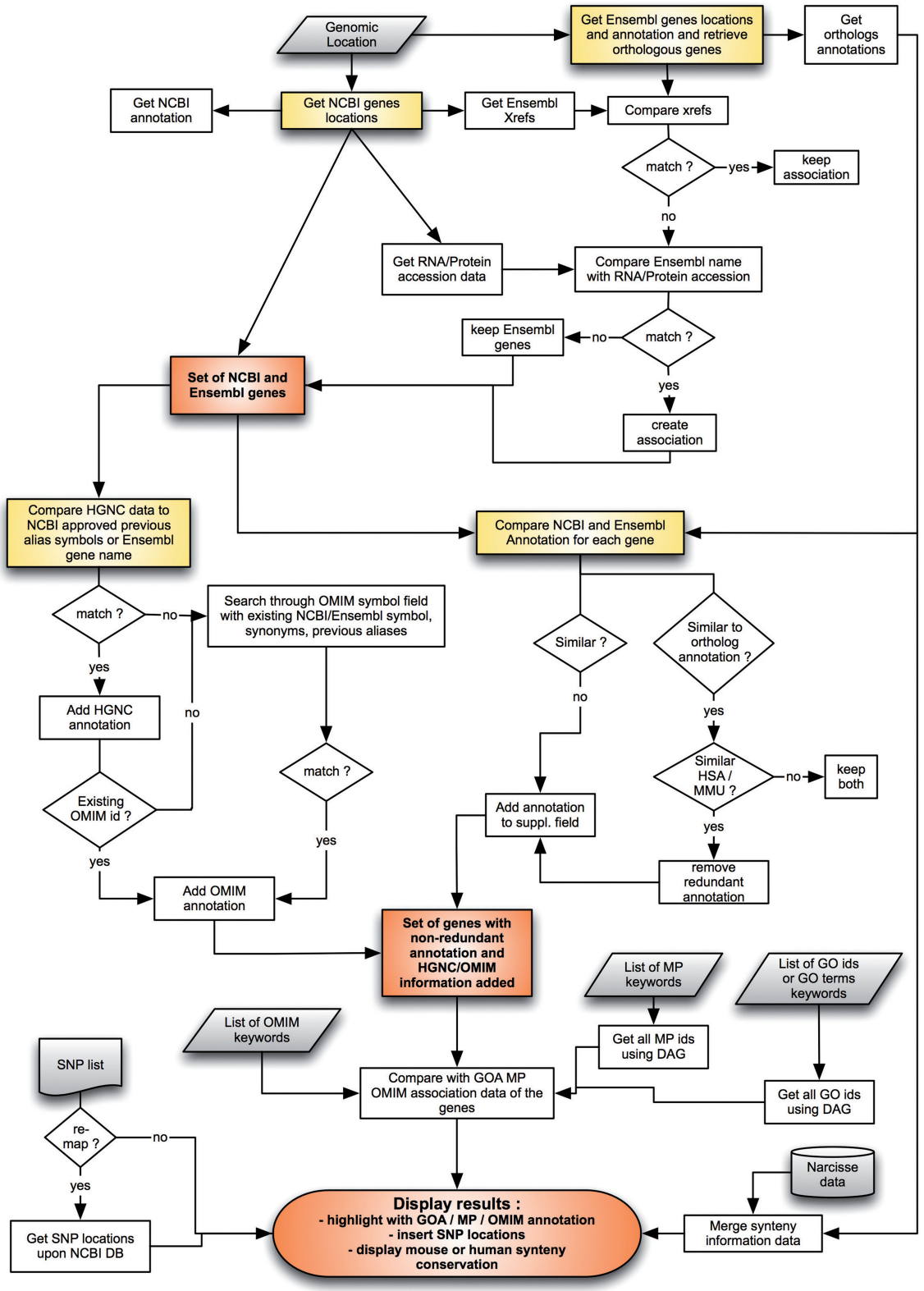


Figure 2. AnnotQTL—principle and workflow. Boxes shaded in gray represent user input or database (i.e. Narcisse) input. Boxes shaded in yellow show the main processes in the AnnotQTL workflow. Boxes shaded in orange represent intermediate results. MP: Mammalian Phenotype.

Table 1. Statistics of the QTL/eQTL regions analyzed using AnnotQTL

	Number of regions	Regions mean size (Mb)	NCBI genes	Ensembl genes	AnnotQTL genes obtained merging NCBI and Ensembl	GO and MP terms screening results	
						Genes found	Average of genes found per region
QTL	21	4.8	1734	1902	2220	127	5.8
eQTL	25	3.4	1198	1283	1506	93	3.7

working from an initial set of 1734 genes from the NCBI database and 1902 genes from the Ensembl database, AnnotQTL retrieved a non-redundant set of 2220 genes. On this large dataset, we applied the 'highlight function' on each region to underline genes whose functional annotation was related to the studied phenotype. Among the 2220 genes located in these 21 QTL regions, 127 were highlighted using the GO term, 'lipid' and the MP term 'adipose' as keywords, with an average 5.4 genes highlighted per region.

Finally, AnnotQTL can also be exploited to look at eQTL regions. Strategies combining transcriptomics and genotyping data have recently been developed to better characterize QTL regions for traits of interest by identifying co-localized eQTLs and QTLs (16–21). Whatever the context, this strategy identifies a much higher number of eQTL regions than in QTL studies, thus creating a need for tools that can efficiently find positional and functional candidate genes. Here, we focus on 25 chicken eQTL regions affecting 70 genes involved in lipid metabolism (i.e. sharing the GO term GO:0006629 'lipid metabolic process'). Average length of these regions is 3.4 Mb. Running AnnotQTL found similar results to those obtained for the QTL regions. All the regions were enriched with genes by comparing NCBI and Ensembl information against information provided by either NCBI or Ensembl only (Table 1): working from an initial set of 1,198 genes from the NCBI database and 1283 genes from the Ensembl database, AnnotQTL retrieved a non-redundant set of 1506 genes. Again, in order to select possible candidate genes, we used the 'highlight function' to pinpoint the genes related to the studied phenotype. Among these 1506 genes, and using the same GO term 'lipid' and MP term 'adipose' as keywords, a total of 93 genes were identified, with an average 3.7 genes highlighted per region.

These examples corresponding to two different contexts (QTL and eQTL analyses) clearly demonstrate how in just a couple of hours, AnnotQTL can accurately analyze the gene content of numerous regions identified by a full genome scan and go on to highlight some of these genes based on both their location and function, whereas in the same time period, a manually run procedure would only have been able to analyze one single region.

CONCLUSION

AnnotQTL is a web tool designed to gather the functional annotation of different prominent web sites while minimizing redundant information. Using all known information substantially accelerates the gene analysis of QTL regions for livestock species traits and improves the selection of candidate genes.

ACKNOWLEDGEMENTS

The authors thank A.T.T. scientific editing services for proofreading the article and all the beta testers for their help on debugging AnnotQTL.

FUNDING

INRA, Agrocampus Ouest, the Regional Council of Brittany; French Ministry in charge of Agriculture (DGER). Funding for open access charge: INRA.

Conflict of interest statement. None declared.

REFERENCES

1. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166–166.
2. Goodswen,S., Gondro,C., Watson-Haigh,N. and Kadarmideen,H. (2010) FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases. *BMC Bioinformatics*, **11**, 311.
3. Wang,P., Dai,M., Xuan,W., McEachin,R.C., Jackson,A.U., Scott,L.J., Athey,B., Watson,S.J. and Meng,F. (2006) SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics*, **22**, e523–e529.
4. Shen,T.H., Carlson,C.S. and Tarczy-Hornoch,P. (2009) SNPit: a federated data integration system for the purpose of functional SNP annotation. *Comp. Methods Prog. Biomedicine*, **95**, 181–189.
5. Ryan,M., Diekhans,M., Lien,S., Liu,Y. and Karchin,R. (2009) LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics*, **25**, 1431–1432.
6. Riva,A. and Kohane,I.S. (2004) A SNP-centric database for the investigation of the human genome. *BMC Bioinformatics*, **5**, 33.
7. Reumers,J., Maurer-Stroh,S., Schymkowitz,J. and Rousseau,F. (2006) SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics*, **22**, 2183–2185.
8. Li,S., Ma,L., Li,H., Vang,S., Hu,Y., Bolund,L. and Wang,J. (2007) Snap: an integrated SNP annotation platform. *Nucleic Acids Res.*, **35**, D707–D710.
9. Haider,S., Ballester,B., Smedley,D., Zhang,J., Rice,P. and Kasprzyk,A. (2009) BioMart Central Portal–unified access to biological data. *Nucleic Acids Res.*, **37**, W23–W27.
10. Courcelle,E., Beausse,Y., Letort,S., Stahl,O., Fremez,R., Ngom-Bru,C., Gouzy,J. and Faraut,T. (2008) Narcisse: a mirror view of conserved syntenies. *Nucleic Acids Res.*, **36**, D485–D490.
11. Filangi,O., Beausse,Y., Assi,A., Legrand,L., Larre,J.M., Martin,V., Collin,O., Caron,C., Leroy,H. and Allouche,D. (2008) BioMAJ: a flexible framework for databanks synchronization and processing. *Bioinformatics*, **24**, 1823–1825.

12. Charlier,C., Coppieters,W., Farnir,F., Grobet,L., Leroy,P.L., Michaux,C., Mni,M., Schwers,A., Vanmanshoven,P., Hanset,R. *et al.* (1995) The mh gene causing double-muscling in cattle maps to bovine Chromosome 2. *Mamm. Genome*, **6**, 788–792.
13. Grobet,L., Martin,L.J., Poncelet,D., Pirottin,D., Brouwers,B., Riquet,J., Schoeberlein,A., Dunner,S., Menissier,F., Massabanda,J. *et al.* (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nat. Genet.*, **17**, 71–74.
14. Lagarrigue,S., Pitel,F., Carre,W., Abasht,B., Le Roy,P., Neau,A., Amigues,Y., Sourdioux,M., Simon,J., Cogburn,L. *et al.* (2006) Mapping quantitative trait loci affecting fatness and breast muscle weight in meat-type chicken lines divergently selected on abdominal fatness. *Genet. Sel. Evol.*, **38**, 85–97.
15. Le Mignon,G., Pitel,F., Gilbert,H., Le Bihan-Duval,E., Vignoles,F., Demeure,O., Lagarrigue,S., Simon,J., Cogburn,L.A., Aggrey,S.E. *et al.* (2009) A comprehensive analysis of QTL for abdominal fat and breast muscle weights on chicken chromosome 5 using a multivariate approach. *Anim. Genet.*, **40**, 157–164.
16. Blum,Y., Le Mignon,G., Lagarrigue,S. and Causeur,D. (2010) A factor model to analyze heterogeneity in gene expression. *BMC Bioinformatics*, **11**, 368.
17. Le Mignon,G., Desert,C., Pitel,F., Leroux,S., Demeure,O., Guernec,G., Abasht,B., Douaire,M., Le Roy,P. and Lagarrigue,S. (2009) Using transcriptome profiling to characterize QTL regions on chicken chromosome 5. *BMC Genomics*, **10**, 575.
18. Ponsuksili,S., Jonas,E., Murani,E., Phatsara,C., Srikanchai,T., Walz,C., Schwerin,M., Schellander,K. and Wimmers,K. (2008) Trait correlated expression combined with expression QTL analysis reveals biological pathways and candidate genes affecting water holding capacity of muscle. *BMC Genomics*, **9**, 367.
19. Babak,T., Garrett-Engele,P., Armour,C.D., Raymond,C.K., Keller,M.P., Chen,R., Rohl,C.A., Johnson,J.M., Attie,A.D., Fraser,H.B. *et al.* (2010) Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. *BMC Genomics*, **11**, 473.
20. Drost,D.R., Benedict,C.I., Berg,A., Novaes,E., Novaes,C.R., Yu,Q., Dervinis,C., Maia,J.M., Yap,J., Miles,B. *et al.* (2010) Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of Populus. *Proc. Natl Acad. Sci. USA*, **107**, 8492–8497.
21. van Nas,A., Ingram-Drake,L., Sinsheimer,J.S., Wang,S.S., Schadt,E.E., Drake,T. and Lusis,A.J. (2010) Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics*, **185**, 1059–1068.

Annexe B

Données générées pendant la thèse

We are conducting a new project on 4 families of chickens corresponding to 45, 53, 42, 37 offsprings originating from crosses between two lines divergently selected for abdominal fatness. The 177 offspring were sacrificed at 9 weeks old, fed ad libitum for a minimum of 4 hours after overnight fasting and measured for their abdominal fatness (AF) and also genotyped for 1536 SNP.
During my thesis, I have participated to the acquisition of transcriptomic and lipidomic profiles on these 177 animals.
First, Colette Désert (engineer at the PEGASE laboratory) and I, hybridized the 177 microarrays (experimental design is summarized in section B.1). After the transcriptomic profile acquisition, I did a pre-treatment of the data (filtering and normalization) using R software (as described in section B.1).
Second, we collaborated with Hervé Guillou (researcher at the INRA of Toulouse) for the lipids and fatty acids extraction. I participated with him to the experiments and I did the analyses of lipids and fatty acids profiles using AZUR software (experimental design is summarized in section B.2). Furthermore, I did a pre-treatment of the data using R software (as described in section B.2).

The eQTL and mQTL analyses are currently performed by a new Ph.D student, Pierre-François Roux. We proposed several approaches to analyze the expression data. One approach consists in adjusting the data for the FAMT method. Another approach consists in considering the experimental factors having a potential effect on the expression, as fixed effects or covariates in the QTL detection model. It will be interesting to compare the results of both strategies and to observe on this powerful experimental design, the contribution of our approach developed in Blum et al. (2010) for eQTL analyses.

B.1 Transcriptomic data acquisition:

Hepatic gene expression acquisition:
Total RNA was isolated with the Nucleospin RNA kit, according to the manufacturer's (Macherey-Nagel, Düren, Germany) instructions. Cyanine-3 labeled cRNA was then prepared from 350ng of

total RNA using the Quick Amp Labeling Kit (Agilent) according to the manufacturer's instructions. For each RNA sample, 1650 ng of labeled cRNA was hybridized onto 4 × 44K custom chicken gene expression 60-mer oligonucleotide microarrays (Agilent Technologies, Palo Alto CA, USA) according to the manufacturer's instructions. Agilent microarray scanner and Feature Extraction software was used to scan, extract and analyze the signals.

Hepatic gene expression filtering and normalization:
The raw data were first filtered to remove probes having more than 60 % of bad quality spots according to the "genepix" flag and the asymmetric spot shape (150 probes were removed). Second, the probes showing a good contrast between spot and background intensities (SNR (Signal to Noise Ratio) defined as greater than 3) were selected and considered as expressed in the liver ($\sim$ 60% of the whole probes). Median Log2 ratio values were then normalized according to the hypothesis that the majority of gene expressions did not differ between two samples. Finally, a signal higher than 3 standard deviation from the mean of a gene profile was considered as an outlier and remove from the data. These different steps led to a total of 28743 genes.
An ANOVA test performed for each gene reveals several experimental factors having a potential effect on the all the gene profiles, as technical batch effect, the hatch of the individuals and also the total weight of the individuals.

B.2 Lipidomic data acquisition:

Hepatic lipid profiles acquisition:
Triglyceride and fatty acid assays were performed as described by Rebouissou et al. (2007). Briefly, following homogenization of tissue samples in methanol/5 mM EGTA (2:1, v/v), lipids corresponding to an equivalent of 2 mg of tissue were extracted according to Bligh and Dyer in chloroform/methanol/water (2.5:2.5:2.1, v/v/v), in the presence of the internal standards: stigmasterol (6 μg), heptadecanoate cholesteryl (6 μg), glyceryl trinonadecanoate (12 μg). Neutral lipids were analyzed by gas-liquid chromatography on a Focus Thermo Electron system using a Zebron-1 Phenomenex fused-silica capillary column (5 m 0.32 mm i.d., 0.50 mm film thickness). Oven temperature was programmed from 200 to 350°C at a rate of 5°C/min, and the carrier gas was hydrogen (0.5 bar). The injector and the detector were at 315 and 345°C, respectively.
To measure total fatty acid methyl ester (FAME) molecular species, lipid amounts corresponding to an equivalent of 2 mg of tissue were extracted in the presence of glyceryl triheptadecanoate (4 μg) as an internal standard. Lipid extract was transmethylated with 1 ml of BF3 in methanol (1/20, v/v) for 150 min at 100°C and evaporated to dryness, and the FAME species were extracted with hexane-water, 1:1. The organic phase was evaporated to dryness and dissolved in 200 μl of ethyl acetate. One μl of FAME was analyzed by gas-liquid chromatography using a 5890 Hewlett Packard system with Restek Famewax fused silica capillary columns (30 m × 0.32 mm ID; 0.25 mm film thickness). Oven temperature changes were programmed from 110°C to 220°C at a rate of 2°C per min, and the carrier gas was helium (0.5 bar). Injector and detector temperatures were

225°C and 245°C, respectively.

Hepatic lipid profiles analysis:
The lipid and fatty acids profiles where analyzed using AZUR software. This software generates automatically the data from the profiles by calculating the integral of each peak corresponding to a specific lipid/fatty acid. Unfortunately, the software does a lot of mistakes in the integration procedure and some lipid or fatty acid names are swapped. Therefore, it is necessary, to check all the integrations manually. Furthermore, peaks with strange shape were not considered in the data. Finally, absolute values higher than 4 standard deviation from the mean for a lipid profile were eliminated. An ANOVA test performed for each variable reveal several experimental factors having a potential effect on the profiles, as the date of extraction, the hatch of the individuals and also the total weight of the individuals. For mQTL analyses, these experimental factors will be added as fixed effects or covariates in the QTL detection model in QTLMap software (Elsen et al. (1999)).

Bibliographie

Aittokallio, T., Schwikowski, B., 2006. Graph-based methods for analysing networks in cell biology. Briefings in bioinformatics 7 (3), 243–255.

Alleman, F., Bordas, A., Caffin, J., Daval, S., Diot, C., Douaire, M., Fraslin, J., Lagarrigue, S., Leclercq, B., 1999. L'engraissement chez le poulet: aspects métaboliques et génétiques. PRODUCTIONS ANIMALES-PARIS-INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE- 12, 257–264.

Allen, J., Xie, Y., Chen, M., Girard, L., Xiao, G., 2012. Comparing statistical methods for constructing large scale gene networks. PloS one 7 (1), e29348.

Ambroise, C., Chiquet, J., Matias, C., 2009. Inferring sparse gaussian graphical models with latent structure. Electronic Journal of Statistics 3, 205–238.

Banerjee, O., El Ghaoui, L., d'Aspremont, A., 2008. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. The Journal of Machine Learning Research 9, 485–516.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., Di Bernardo, D., 2007. How to infer gene networks from expression profiles. Molecular systems biology 3 (1).

Blum, Y., Le Mignon, G., Causeur, D., Filangi, O., Désert, C., Demeure, O., Le Roy, P., Lagarrigue, S., 2011. Complex trait subtypes identification using transcriptome profiling reveals an interaction between two qtl affecting adiposity in chicken. BMC genomics 12 (1), 567.

Blum, Y., Le Mignon, G., Lagarrigue, S., Causeur, D., 2010. A factor model to analyze heterogeneity in gene expression. BMC bioinformatics 11 (1), 368.

Brazhnik, P., de la Fuente, A., Mendes, P., 2002. Gene networks: how to put the function in genomics. TRENDS in Biotechnology 20 (11), 467–472.

Brem, R., Yvert, G., Clinton, R., Kruglyak, L., 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296 (5568), 752–755.

Broman, K., Wu, H., Sen, Ś., Churchill, G., 2003. R/qtl: Qtl mapping in experimental crosses. Bioinformatics 19 (7), 889–890.

Butte, A., Kohane, I., 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Pac Symp Biocomput. Vol. 5. pp. 418–429.

Butte, A., Tamayo, P., Slonim, D., Golub, T., Kohane, I., 2000. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. Proceedings of the National Academy of Sciences 97 (22), 12182.

Carlson, M., Zhang, B., Fang, Z., Mischel, P., Horvath, S., Nelson, S., 2006. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC genomics 7 (1), 40.

Carter, S., Brechbühler, C., Griffin, M., Bond, A., 2004. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20 (14), 2242–2250.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., West, M., 2008. High-dimensional sparse factor modeling: applications in gene expression genomics. Journal of the American Statistical Association 103 (484), 1438–1456.

Chu, J., Weiss, S., Carey, V., Raby, B., 2009. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. BMC systems biology 3 (1), 55.

Davidson, G., Wylie, B., Boyack, K., 2001. Cluster stability and the use of noise in interpretation of clustering. In: Proc. IEEE Information Visualization. Vol. 2001. pp. 23–30.

Davis, R., Castellani, L., Hosseini, M., Ben-Zeev, O., Mao, H., Weinstein, M., Jung, D., Jun, J., Kim, J., Lusis, A., et al., 2010. Early hepatic insulin resistance precedes the onset of diabetes in obese c57blks-db/db mice. Diabetes 59 (7), 1616–1625.

Davis, R., Van Nas, A., Castellani, L., Zhao, Y., Zhou, Z., Wen, P., Yu, S., Qi, H., Rosales, M., Schadt, E., et al., 2012. Systems genetics of susceptibility to obesity-induced diabetes in mice. Physiological genomics 44 (1), 1–13.

De Jong, H., 2002. Modeling and simulation of genetic regulatory systems: a literature review. Journal of computational biology 9 (1), 67–103.

De Smet, R., Marchal, K., 2010. Advantages and limitations of current network inference methods. Nature Reviews Microbiology 8 (10), 717–729.

De Tayrac, M., Lê, S., Aubry, M., Mosser, J., Husson, F., 2009. Simultaneous analysis of distinct omics data sets with integration of biological knowledge: Multiple factor analysis approach. BMC genomics 10 (1), 32.

Dempster, A., 1972. Covariance selection. Biometrics, 157–175.

Derry, J., Zhong, H., Molony, C., MacNeil, D., Guhathakurta, D., Zhang, B., Mudgett, J., Small, K., El Fertak, L., Guimond, A., et al., 2010. Identification of genes and networks driving cardiovascular and metabolic phenotypes in a mouse f2 intercross. PloS one 5 (12), e14319.

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G., West, M., 2004. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis 90 (1), 196–212.

Dykstra, R., 1970. Establishing the positive definiteness of the sample covariance matrix. The Annals of Mathematical Statistics 41 (6), 2153–2154.

D'haeseleer, P., Liang, S., Somogyi, R., 1999. Gene expression data analysis and modeling. In: Pacific symposium on biocomputing. Vol. 99.

D'haeseleer, P., Liang, S., Somogyi, R., 2000. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16 (8), 707–726.

Efron, B., 2004. Large-scale simultaneous hypothesis testing. Journal of the American Statistical Association 99 (465), 96–104.

Efron, B., 2005. Local false discovery rates. Division of Biostatistics, Stanford University.

Elsen, J., Mangin, B., Goffinet, B., Boichard, D., Le Roy, P., 1999. Alternative models for qtl detection in livestock. i. general introduction. Genetics Selection Evolution 31 (3), 1–12.

Ferrara, C., Wang, P., Neto, E., Stevens, R., Bain, J., Wenner, B., Ilkayeva, O., Keller, M., Blasiole, D., Kendziorski, C., et al., 2008. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. PLoS genetics 4 (3), e1000034.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using bayesian networks to analyze expression data. Journal of computational biology 7 (3-4), 601–620.

Friguet, C., Kloareg, M., Causeur, D., 2009. A factor model approach to multiple testing under dependence. Journal of the American Statistical Association 104 (488), 1406–1415.

Gargalovic, P., Imura, M., Zhang, B., Gharavi, N., Clark, M., Pagnon, J., Yang, W., He, A., Truong, A., Patel, S., et al., 2006. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proceedings of the National Academy of Sciences 103 (34), 12741–12746.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E., Drake, T., Lusis, A., et al., 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. PloS Genetics 2 (8), e130.

Giles, P., Kipling, D., 2003. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. Bioinformatics 19 (17), 2254–2262.

Hausser, J., Strimmer, K., 2009. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. The Journal of Machine Learning Research 10, 1469–1484.

Hotelling, H., 1953. New light on the correlation coefficient and its transforms. Journal of the Royal Statistical Society. Series B (Methodological) 15 (2), 193–232.

Husmeier, D., et al., 2003. Reverse engineering of genetic networks with bayesian networks. Biochemical Society Transactions 31 (6), 1516–1518.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z., Barabási, A., 2000. The large-scale organization of metabolic networks. Nature 407 (6804), 651–654.

Kloareg, M., Friguet, C., Causeur, D., et al., 2009. Factor analysis for multiple testing (famt): an r package for simultaneous tests under dependence in high-dimensional data.

Kontos, K., 2009. Gaussian graphical model selection for gene regulatory network reverse engineering and function prediction. Ph.D. thesis, Ph. D. dissertation, Univ. Libre de Bruxelles.

Lagarrigue, S., Tixier-Boichard, M., 2011. Nouvelles approches de phénotypage pour la sélection animale. Productions Animales 24 (4), 377.

Lan, H., Rabaglia, M., Stoehr, J., Nadler, S., Schueler, K., Zou, F., Yandell, B., Attie, A., 2003. Gene expression profiles of nondiabetic and diabetic obese mice suggest a role of hepatic lipogenic capacity in diabetes susceptibility. Diabetes 52 (3), 688–700.

Langfelder, P., Horvath, S., 2007. Eigengene networks for studying the relationships between co-expression modules. BMC systems biology 1 (1), 54.

Langfelder, P., Horvath, S., 2008. Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics 9 (1), 559.

Lauritzen, S., 1996. Graphical models. Vol. 17. Oxford University Press, USA.

Le Bihan-Duval, E., Nadaf, J., Berri, C., Pitel, F., Graulet, B., Godet, E., Leroux, S., Demeure, O., Lagarrigue, S., Duby, C., et al., 2011. Detection of a cis eqtl controlling bmco1 gene expression leads to the identification of a qtg for chicken breast meat color. PloS one 6 (7), e14825.

Le Mignon, G., Blum, Y., Demeure, O., Diot, C., Le Bihan-Duval, E., Le Roy, P., Lagarrigue, S., 2010. Apports de la génomique fonctionnelle à la cartographie fine de qtl. Productions Animales 23 (4), 343.

Le Mignon, G., Désert, C., Pitel, F., Leroux, S., Demeure, O., Guernec, G., Abasht, B., Douaire, M., Le Roy, P., Lagarrigue, S., 2009. Using transcriptome profiling to characterize qtl regions on chicken chromosome 5. BMC genomics 10 (1), 575.

Lecerf, F., Bretaudeau, A., Sallou, O., Désert, C., Blum, Y., Lagarrigue, S., Demeure, O., 2011. Annotqtl: a new tool to gather functional and comparative information on a genomic region. Nucleic acids research 39 (suppl 2), W328–W333.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis 88 (2), 365–411.

Leek, J., Storey, J., 2007. Capturing Heterogeneity In gene expression studies by surrogate variable analysis. PLoS Genetics 3 (9).

Leiter, E., Coleman, D., Hummel, K., 1981. The influence of genetic background on the expression of mutations at the diabetes locus in the mouse. iii. effect of h-2 haplotype and sex. Diabetes 30 (12), 1029–1034.

Liang, K., Wang, X., 2008. Gene regulatory network reconstruction using conditional mutual information. EURASIP Journal on Bioinformatics and Systems Biology 2008.

Listgarten, J., Kadie, C., Schadt, E., Heckerman, D., 2010. Correction for hidden confounders in the genetic analysis of gene expression. Proceedings of the National Academy of Sciences 107 (38), 16465–16470.

Liu, B., Hoeschele, I., de la Fuente, A., 2009. Inferring gene regulatory networks from genetical genomics data. Handbook of Research on Computational Methodologies in Gene Regulatory Networks, 79–107.

Lum, P., Chen, Y., Zhu, J., Lamb, J., Melmed, S., Wang, S., Drake, T., Lusis, A., Schadt, E., 2006. Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. Journal of neurochemistry 97, 50–62.

Mach, N., Blum, Y., Bannink, A., Causeur, D., Houée, M., Lagarrigue, S., Smith, M., 2012. Pleiotropic effects of polymorphism of the gene diacylglycerol-o-transferase 1 (dgat1) in the mammary gland tissue of dairy cows. Journal of Dairy Science.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., Califano, A., 2006. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics 7 (Suppl 1), S7.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34 (3), 1436–1462.

Meyer, P., Kontos, K., Lafitte, F., Bontempi, G., 2007. Information-theoretic inference of large transcriptional regulatory networks. EURASIP Journal on Bioinformatics and Systems Biology 2007, 8–8.

Miettinen, T., Gylling, H., 2000. Cholesterol absorption efficiency and sterol metabolism in obesity. Atherosclerosis 153 (1), 241–248.

Mukherjee, S., Speed, T., 2008. Network inference using informative priors. Proceedings of the National Academy of Sciences 105 (38), 14313.

Murphy, K., Mian, S., 1999. Modelling gene expression data using dynamic bayesian networks. Tech. rep., Technical report, Computer Science Division, University of California, Berkeley, CA.

Nadaf, J., Pitel, F., Gilbert, H., Duclos, M., Vignoles, F., Beaumont, C., Vignal, A., Porter, T., Cogburn, L., Aggrey, S., et al., 2009. Qtl for several metabolic traits map to loci controlling growth and body composition in an f2 intercross between high-and low-growth chicken lines. Physiological genomics 38 (3), 241–249.

Peltola, P., Pihlajamäki, J., Koutnikova, H., Ruotsalainen, E., Salmenniemi, U., Vauhkonen, I., Kainulainen, S., Gylling, H., Miettinen, T., Auwerx, J., et al., 2006. Visceral obesity is associated with high levels of serum squalene&ast. Obesity 14 (7), 1155–1163.

Peng, J., Wang, P., Zhou, N., Zhu, J., 2009. Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association 104 (486), 735–746.

Pournara, I., Wernisch, L., 2007. Factor analysis for gene regulatory networks and transcription factor activity profiles. BMC bioinformatics 8 (1), 61.

Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., Barabási, A., 2002. Hierarchical organization of modularity in metabolic networks. science 297 (5586), 1551–1555.

Rebouissou, S., Imbeaud, S., Balabaud, C., Boulanger, V., Bertrand-Michel, J., Tercé, F., Auffray, C., Bioulac-Sage, P., Zucman-Rossi, J., 2007. Hnf1α inactivation promotes lipogenesis in human hepatocellular adenoma independently of srebp-1 and carbohydrate-response element-binding protein (chrebp) activation. Journal of Biological Chemistry 282 (19), 14437.

Schadt, E., 2005. Exploiting naturally occurring dna variation and molecular profiling data to dissect disease and drug response traits. Current opinion in biotechnology 16 (6), 647–654.

Schadt, E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S., Monks, S., Reitman, M., Zhang, C., et al., 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nature genetics 37 (7), 710–717.

Schadt, E., Lum, P., 2006. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. Journal of lipid research 47 (12), 2601–2613.

Schadt, E., Monks, S., Drake, T., Lusis, A., Che, N., Colinayo, V., Ruff, T., Milligan, S., Lamb, J., Cavet, G., et al., 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422 (6929), 297–302.

Schäfer, J., Opgen-Rhein, R., Strimmer, K., 2006. Reverse engineering genetic networks using the genenet package. Journal of the American Statistical Association 96, 1151–1160.

Schäfer, J., Strimmer, K., et al., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology 4 (1), 32.

Scherneck, S., Nestler, M., Vogel, H., Blüher, M., Block, M., Diaz, M., Herzig, S., Schulz, N., Teichert, M., Tischer, S., et al., 2009. Positional cloning of zinc finger domain transcription factor zfp69, a candidate gene for obesity-associated diabetes contributed by mouse locus nidd/sjl. PLoS genetics 5 (7), e1000541.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M., 2011. Global quantification of mammalian gene expression control. Nature 473 (7347), 337–342.

Spearman, C., 1904. " general intelligence," objectively determined and measured. The American Journal of Psychology 15 (2), 201–292.

Steuer, R., 2006. Review: on the analysis and interpretation of correlations in metabolomic data. Briefings in bioinformatics 7 (2), 151–158.

Steuer, R., Kurths, J., Daub, C., Weise, J., Selbig, J., 2002. The mutual information: detecting and evaluating dependencies between variables. Bioinformatics 18 (suppl 2), S231–S240.

Szallasi, Z., Liang, S., 1998. Modeling the normal and neoplastic cell cycle with 'realistic boolean genetic networks': Their application for understanding carcinogenesis and assessing therapeutic strategies. In: Pacific Symposium on Biocomputing. Vol. 3. pp. 66–76.

Tegner, J., Yeung, M., Hasty, J., Collins, J., 2003. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proceedings of the National Academy of Sciences 100 (10), 5944.

Waddell, P., Kishino, H., et al., 2000. Cluster inference methods and graphical models evaluated on nci60 microarray gene expression data. Genome Informatics Series, 129–140.

Werhli, A., Grzegorczyk, M., Husmeier, D., 2006. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics 22 (20), 2523–2531.

Woodbury, M., 1950. Inverting modified matrices. Memorandum report 42, 106.

Workman, C., Mak, H., McCuine, S., Tagne, J., Agarwal, M., Ozier, O., Begley, T., Samson, L., Ideker, T., 2006. A systems approach to mapping dna damage response pathways. Science's STKE 312 (5776), 1054.

Wu, S., Lusis, A., Drake, T., 2009. A systems-based framework for understanding complex metabolic and cardiovascular disorders. Journal of lipid research 50 (Supplement), S358–S363.

Zhang, B., Horvath, S., et al., 2005. A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology 4 (1), 1128.