Université de Montréal

Études de réseaux d'expression génique: utilité pour l'élucidation des déterminants génétiques des traits complexes

par
Marie Pier Scott-Boyer

Département Biochimie
Faculté de Médecine

Thèse présentée à la Faculté de Médecine
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en Bio-informatique

avril, 2013

Université de Montréal
Faculté de Médecine

Cette thèse intitulée:

Études de réseaux d'expression génique: utilité pour l'élucidation des déterminants génétiques des traits complexes

présentée par:

Marie Pier Scott-Boyer

a été évaluée par un jury composé des personnes suivantes:

Eric Lécuyer,	président-rapporteur
Christian Deschepper,	directeur de recherche
Raphaël Gottardo,	codirecteur
Guillaume Lettre,	membre du jury
Guillaume Bourque,	examinateur externe
Gregor Andelfinger,	représentant du doyen de la Faculté de médecine

Thèse acceptée le: . . . . . .

RÉSUMÉ

Les traits quantitatifs complexes sont des caractéristiques mesurables d'organismes vivants qui résultent de l'interaction entre plusieurs gènes et facteurs environnementaux. Les locus génétiques liés à un caractère complexe sont appelés «locus de traits quantitatifs» (QTL). Récemment, en considérant les niveaux d'expression tissulaire de milliers de gènes comme des traits quantitatifs, il est devenu possible de détecter des «QTLs d'expression» (eQTL). Alors que ces derniers ont été considérés comme des phénotypes intermédiaires permettant de mieux comprendre l'architecture biologique des traits complexes, la majorité des études visent encore à identifier une mutation causale dans un seul gène. Cette approche ne peut remporter du succès que dans les situations où le gène incriminé a un effet majeur sur le trait complexe, et ne permet donc pas d'élucider les situations où les traits complexes résultent d'interactions entre divers gènes.

Cette thèse propose une approche plus globale pour : 1) tenir compte des multiples interactions possibles entre gènes pour la détection de eQTLs et 2) considérer comment des polymorphismes affectant l'expression de plusieurs gènes au sein de groupes de co-expression pourraient contribuer à des caractères quantitatifs complexes. Nos contributions sont les suivantes :

- Nous avons développé un outil informatique utilisant des méthodes d'analyse multivariées pour détecter des eQTLs et avons montré que cet outil augmente la sensibilité de détection d'une classe particulière de eQTLs.

- Sur la base d'analyses de données d'expression de gènes dans des tissus de souris recombinantes consanguines, nous avons montré que certains polymorphismes peuvent affecter l'expression de plusieurs gènes au sein de domaines géniques de co-expression.

- En combinant des études de détection de eQTLs avec des techniques d'analyse de réseaux de co-expression de gènes dans des souches de souris recombinantes consanguines, nous avons montré qu'un locus génétique pouvait être lié à la fois à

l'expression de plusieurs gènes au niveau d'un domaine génique de co-expression et à un trait complexe particulier (c.-à-d. la masse du ventricule cardiaque gauche).

Au total, nos études nous ont permis de détecter plusieurs mécanismes par lesquels des polymorphismes génétiques peuvent être liés à l'expression de plusieurs gènes, ces derniers pouvant eux-mêmes être liés à des traits quantitatifs complexes.

Mots clés: Quantitative trait locus, expression de gènes, réseau de co-expression, masse du ventricule gauche, analyse baysienne multivariée, trait quantitatif complexe, bio-informatique translationnelle, génétique quantitative.

ABSTRACT

Complex quantitative traits are measurable characteristics of living organisms resulting from the interaction between multiple genes and environmental factors. Genetic loci associated with complex trait are called "quantitative trait loci" (QTL). Recently, considering the expression levels of thousands of genes as quantitative traits, it has become possible to detect "expression QTLs " (eQTL). These eQTL are considered intermediate phenotypes and are used to better understand the biological architecture of complex traits. However the majority of studies still try to identify a causal mutation in a single gene. This approach can only meet success in situations where the gene incriminate as a major effect on the complex trait, and therefore can not elucidate the situations where complex traits result from interactions between various genes.

This thesis proposes a more comprehensive approach to: 1) take into account the possible interactions between multiple genes for the detection of eQTLs and 2) consider how polymorphisms affecting the expression of several genes in a module of co-expression may contribute to quantitative complex traits. Our contributions are as follows:

- We have developed a tool using multivariate analysis techniques to detect eQTLs, and have shown that this tool increases the sensitivity of detection of a particular class of eQTLs.

- Based on the data analysis of gene expression in recombinant inbred strains mice tissues, we have shown that some polymorphisms may affect the expression of several genes in domain of co-expression.

- Combining eQTLs detection studies with network of co-expression genes analysis in recombinant inbred strains mice, we showed that a genetic locus could be linked to both the expression of multiple genes at a domain of gene co-expression and a specific complex trait (i.e. left ventricular mass).

Our studies have detected several mechanisms by which genetic polymorphisms may be associated with the expression of several genes, and may themselves be linked to

quantitative complex traits.

Keywords: Quantitative trait locus, gene expression, co-expression network, left ventricular mass, multivariate bayesian analysis, complex quantitative traits, translational bioinformatics, quantitative genetics.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

LISTE DES SIGLES

ANOVA Analysis of variance

ADN Acide desoxyribonucleique

ARN Acide ribonucleique

ARNnc ARN non codant

ARNm ARN messager

Chr Chromosome

CDD chromosome domain-driven

cM CentiMorgan

CTCF CCCTC-binding factor

eQTL expression Quantitative trait locus

EM Expectation-maximisation

ENCODE Encyclopedia of DNA Elements

FDR False discovery rate

H3ac acetylated histone 3

iBMQ integrated Bayesian hierarchical Model for eQTL mapping

iBMQ-cw iBMQ common weigth

IGP invariant genomic probes

indel insertion-deletions

IRCM Institut de recherches cliniques de Montreal

GWAS Genome wide association studies

GO Gene Ontology

kb kilo paires de bases

LINE Long Interspersed Elements

LOD Logarithm of the odds

LVM Left ventricular mass

LTR Long terminal repeat

Mb Méga paires de bases

MVG Masse du ventricule gauche

MCMC Markov Chain Monte Carlo

MOM Mixture Over Marker

mQTL Module QTL

NPV Negative predictive value

NCBI National Center for Biotechnology Information

OMIN Online Mendelian Inheritance in Man

pb Paire de bases

PCA Principal component analysis

PLS Partial Least Square

PPA Posterior probability of association

PPV Positive predictive value

QTL Quantitative trait locus

QTT Quantitative trait transcripts

RI Recombinant Inbred

RIS Recombinant Inbred Strain

RNA-Seq Séquençage à haut débit de l'ARN

ROC Receiver Operating Characteristic curves

SD Standard deviation

SINE Short INterspersed Elements

SNP Single Nucleotide Polymorphism

SPLS Sparse Partial Least Square

SBR Sparse Bayesian Regression

TE Transposable elements

VBQTL Variational Bayes

WGCNA Weighted Gene Co-expression Network Analysis

À mes parents.

REMERCIEMENTS

Je désire premièrement remercier les membres du laboratoire de biologie cardiovasculaire de l'Institut de recherche clinique de Montréal (IRCM). Je voudrais remercier tout spécialement mon directeur de recherche Dr Christian Deschepper pour son aide et sa disponibilité tout au long de ma thèse. Je lui suis également très reconnaissante d'avoir eu un directeur de thèse si extraordinaire, qui m'a communiqué sa passion pour la recherche. Merci beaucoup d'avoir toujours cru en moi et de m'avoir encouragée dans les moments difficiles. Un merci particulier à Sylvie Picard et à Sophie Cardin pour leurs aide au laboratoire.

Je remercie mon codirecteur Raphaël Gottado qui malgré son départ pour le Fred Hutchinson Cancer Research Center à Seattle à continuer à me superviser. Un merci particulier à son étudiant Greg Imholte avec qui j'ai eu la chance de collaborer.

Je remercie Dre Aurélie Labbe de l'Université McGill pour son aide et ses encouragements. Aurélie est pour moi un modèle. Elle est la preuve que l'on peut trouver l'équilibre entre une carrière scientifique et une vie de famille.

Je remercie les membres du jury pour l'intérêt porté à ce travail en acceptant de réviser cette thèse.

Je veux remercier Dr Benjamin Haibe-Kaine et les membres du laboratoire de bio-informatique avec qui j'ai partagé les locaux, spécialement à Joanne Duhaime à qui je souhaite une belle retraite. J'aimerais remercier la communauté de l'Institut de recherche clinique de Montréal (IRCM) et particulièrement Virginie Leduc pour toute son aide avec toute la paperasse administrative. Je voudrais remercier le département de biochimie et le programme de bio-informatique et surtout ma bonne fée Élaine Meunier qui a pris soin de mon dossier étudiant. Je remercie toutes mes collègues du programme de bio-informatique et surtout Julie avec qui j'étudie la bio-informatique depuis les 10 dernières années. Merci pour tous tes conseils !

Je veux aussi remercier mon équipe de triathlon de l'Université de Montréal ; ces moments de défoulement m'ont aidée durant l'écriture de cette thèse. Je veux remercier mes parents de m'avoir laissée développer mon esprit scientifique dès mon plus jeune

âge, malgré les nombreux dégâts dans la cuisine. Merci aussi à ma soeur qui est toujours fière de moi. Un gros merci à mon conjoint Sébastien qui m'a encouragée et qui m'a préparé des bons repas durant la rédaction de cette thèse.

Enfin, j'aimerais remercier Francine sans qui cette thèse ne pourrait être livrée sans fautes d'orthographe.

AVANT-PROPOS

Comme indiqué en couverture, ce doctorat relève du programme de bio-informatique. Comme il s'agit d'une discipline scientifique encore en émergence, j'ai cru opportun d'introduire d'abord quelques concepts préliminaires. La définition formelle de la bio-informatique est celle d'un champ de recherche multidisciplinaire où travaillent de concert biologistes, médecins, informaticiens, mathématiciens et/ou physiciens, dans le but de résoudre de façon quantitative et/ou statistique un problème scientifique posé par la biologie. Cela comprend donc de nombreux domaines possibles d'études, allant de l'analyse du génome à la modélisation de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'images, l'assemblage de génomes et la reconstruction d'arbres phylogénétiques.

Une des sous-disciplines en émergence de la bio-informatique est la bio-informatique translationnelle. C'est un champ qui adresse les problématiques actuelles de l'intégration de grand nombre de données moléculaires et des données cliniques. Cette discipline a pour but de mieux comprendre les bases moléculaires des maladies pour informer la pratique clinique et ultimement améliorer la santé humaine.

Plus concrètement, l'arrivée récente en biologie de nouvelles technologies a généré une quantité incroyable de données. Ainsi, la bio-informatique est devenue nécessaire pour répondre à des besoins spécifiques en matière d'acquisition de données, de stockage, d'analyse et d'intégration. La bio-informatique a permis entre autres d'assembler et d'annoter le génome humain (c'est-à-dire trouver un sens aux séries de bases qui composent les gènes), d'identifier des marqueurs de maladies et des séquences régulatrices et de répertorier les différences entre espèces. Ma thèse est en bio-informatique, car elle consiste à utiliser des méthodes informatiques/statistiques pour intégrer des données biologiques, le but étant de mieux comprendre la biologie des traits complexes.

Comme cette thèse s'adresse à un public très varié, les termes plus techniques en biologie, statistique et informatique ont été définis dans un glossaire à la fin du document (l'annexe I). Ces termes sont suivis du symbole * dans le texte.

CHAPITRE 1

INTRODUCTION

1.1 Introduction

Le but général de ce travail de thèse consiste à développer de nouvelles approches pour analyser de façon intégrée des données génomiques et génétiques et utiliser ces approches pour découvrir de quelle façon des variations génétiques peuvent être liées à des traits quantitatifs complexes. Pour faciliter la compréhension générale de ce travail, ce premier chapitre couvrira les points suivants : une courte introduction sur la génétique, une définition des traits quantitatifs complexes, une description des modèles d'organismes utilisés en génétique, les locus de caractères quantitatifs (QTL), l'expression et la régulation des gènes, les locus de caractères quantitatifs d'expression (eQTL) et finalement les méthodes biostatistiques permettant de détecter les eQTLs et de construire des réseaux permettant d'étudier la co-expression de plusieurs gènes. Les objectifs du projet de recherche seront présentés à la fin de ce chapitre.

1.2 Introduction à la génétique

La génétique est la science qui étudie les mécanismes responsables de la transmission héréditaire de certaines caractéristiques mesurables et/ou observables des organismes vivants. Pour ce faire, elle combine des études portant sur la localisation, la structure et la fonction des gènes à d'autres études dont le but est d'identifier l'étendue des variations dans la composition des génomes de divers individus au sein d'une population. Les sujets d'étude de la génétique incluent (entre autres) la composition et l'évolution des génomes, les signatures de la sélection naturelle, les forces qui influencent la diversité génétique d'une population, et les molécules impliquées dans la formation et le fonctionnement des organismes vivants. Dans cette thèse, nous nous intéresserons surtout à la génétique quantitative*, qui constitue une sous-discipline dédiée à l'étude des facteurs génétiques expliquant la variance des traits quantitatifs complexes (tels que définis plus loin dans le

texte) et les mécanismes de leur héritabilité*.

Dans le paragraphe ci-dessous, des principes de base de la génétique seront introduits et illustrés en prenant l'humain (*homo sapiens*) comme exemple. Chaque humain est constitué d'environ 50 milliards de cellules et chacune de ces cellules réalise des fonctions spécifiques dans le corps. Au coeur de chacune de ces cellules se trouve l'ADN (l'acide désoxyribonucléique), organisé en chromosomes. Le génome de l'humain est organisé en 23 paires de chromosomes : une paire de chromosomes sexuels (XX pour les femelles, XY pour les mâles) et 22 paires de chromosomes non sexuels, c.-à-d. «autosomes». L'ADN est le support de l'information génétique. La structure de l'ADN a été élucidée en 1953 par Watson et Crick [1, 2] ; ils ont montré que les molécules d'ADN sont une double hélice composée de deux brins complémentaires. Chaque brin est composé d'une chaîne de nucléotides. Les 4 types de nucléotides (aussi appelés «bases») sont l'adénine, la cytosine, la thymine et la guanine, communément identifiés par la première lettre de chaque composé (A, C, T et G). L'information génétique est inscrite dans l'ordre de la succession des quatre nucléotides.

En 2003, après une dizaine d'années d'efforts, l'ADN du génome humain a été séquencé au complet. Au total, il contient plus de 3.2 milliards de paires de base. Certaines régions de l'ADN contiennent des gènes (voir Figure 1.1). Les gènes peuvent être activés ou désactivés dans des cellules différentes et à des moments différents. Une des fonctions principales des gènes est de coder pour des protéines, celles-ci représentant les composés participant à la structure des cellules et les instruments leur permettant d'accomplir leurs diverses fonctions. Le dogme central de la biologie moléculaire a été décrit par Francis Crick en 1958 et a été défini comme suit : l'ADN des gènes est d'abord transcrit en une chaîne d'acides ribonucléiques (ARN) et cette dernière est ensuite traduite en protéines. Le type d'ARN responsable de l'encodage des protéines est appelé ARN messager (ARNm). Le dogme central met en évidence le rôle premier accordé à l'ARNm comme support temporaire de l'information génétique. On estime généralement que le génome contient environ 20 000 gènes, ceux-ci représentant seulement 2 à 5 % du génome entier.

La séquence de l'ADN génomique n'est pas identique entre chaque individu d'une

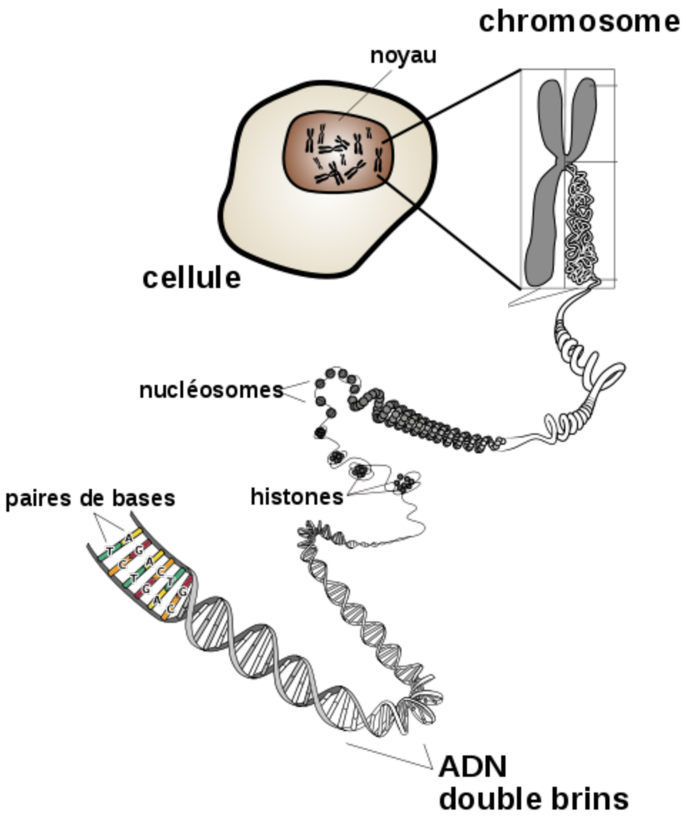


Figure 1.1 – Au coeur de chaque noyau de cellule eucaryote se trouve de l'ADN. La double hélice de l'ADN contient l'information génétique sous forme d'une séquence de nucléotides. L'ADN est organisé en chromosomes. Source : http ://www.genome.gov

population. Les différences de séquences sont appelées «polymorphismes» ou «variations génétiques». Les polymorphismes d'un seul nucléotide (single nucleotide polymorphisms, ou SNP*) représentent un exemple de variation génétique et correspondent à des sites dans le génome où l'ADN diffère d'une simple base entre différents individus d'une population.

Le phénotype* (ou «caractère») est le terme qui décrit une caractéristique mesurable et/ou observable d'un organisme vivant. Il peut ainsi référer à certaines caractéristiques physiques ou comportementales d'un individu, mais peut aussi être observé à l'échelle cellulaire ou moléculaire. Alors que les gènes et l'environnement peuvent tous deux contribuer à l'expression d'un phénotype, la génétique essaye plus particulièrement d'élucider quelles variations génétiques peuvent être responsables de phénotypes particuliers.

1.3 Les caractères quantitatifs complexes

Depuis longtemps, les scientifiques ont tenté de comprendre comment une descendance hérite de certains caractères de leurs parents. Dans cette thèse, lorsque nous parlerons de «trait», nous référerons à une manifestation observable due à une variation génétique. Ces caractéristiques peuvent être des observations anatomiques, morphologiques et/ou moléculaires chez un organisme vivant. Certains traits peuvent avoir un caractère pathologique et/ou représenter une maladie. Dans les cas les plus simples, un trait est dû à la variation d'un seul gène. On parle alors de caractères à «transmission mendélienne», car les principes de leur transmission se conforment aux lois de transmission des gènes tels que décrites par Mendel. Cela ne veut pas dire que le caractère est contrôlé par un seul gène, mais que la variation d'un seul gène est suffisante pour entraîner une variation observable du phénotype. En général, ces traits sont caractérisés par une démarcation claire entre individus sains et affectés. Chez l'humain, les caractères mendéliens les plus étudiés sont des maladies. Généralement, les maladies à transmission mendélienne ont des manifestations cliniques importantes, mais n'affectent qu'un faible pourcentage de la population. Pour étudier ces maladies, il est nécessaire de recruter des familles mul-

tigénérationnelles comportant des individus affectés et non affectés. De telles familles (aussi appelées «pedigrees») sont utilisées pour effectuer des «études de liaison»*, où la transmission de polymorphismes génétiques est comparée à la transmission de maladies de type mendélien. Ces dernières années, de nombreuses études ont ainsi permis d'approfondir notre compréhension des causes et des mécanismes de plusieurs maladies de type mendélien. La base de données OMIM (Online Mendelian Inheritance in Man) [3] a répertorié plus de 12 000 gènes qui présentent des variations liées à des maladies à transmission mendélienne.

Cependant, les maladies les plus fréquentes et affectant le plus d'individus sont d'une autre nature. Bien que des facteurs héréditaires jouent un rôle dans l'incidence de ces maladies, ces facteurs ne peuvent se résumer à une seule variation au sein d'un seul gène. Certaines maladies cardiovasculaires, le diabète de type II, la prédisposition au cancer et les maladies mentales représentent des exemples de ce type d'affections. D'un point de vue génétique, les manifestations de ces maladies sont considérées comme des «caractères quantitatifs complexes». À l'encontre des maladies à transmission mendélienne (qui se manifestent de façon «discontinue», avec une distinction nette entre individus affectés et non affectés), les valeurs de caractères quantitatifs complexes varient entre individus d'une population de façon continue et se distribuent généralement de manière gaussienne (voir Figure 1.2). Cette variation continue résulte des interactions existant à la fois entre de nombreuses variations génétiques (chacune ne contribuant qu'à une petite portion de la variance totale du caractère) et plusieurs facteurs environnementaux. Il est à noter que l'étude des caractères quantitatifs complexes ne s'applique pas uniquement aux maladies humaines, mais concerne de nombreux champs d'étude de la génétique, incluant par exemple les méthodes d'élevage, la sélection végétale appliquée et/ou l'étude des mécanismes moléculaires de l'évolution.

De par la nature différente des caractères quantitatifs complexes, l'élucidation des facteurs génétiques contribuant à leur manifestation nécessite des techniques expérimentales différentes de celles utilisées pour les caractères de type mendélien. Les méthodes actuelles, incluant les études de cartographie de locus de caractères quantitatifs (QTL) et les études d'association* sur le génome complet (GWAS), ont permis d'identifier plu-

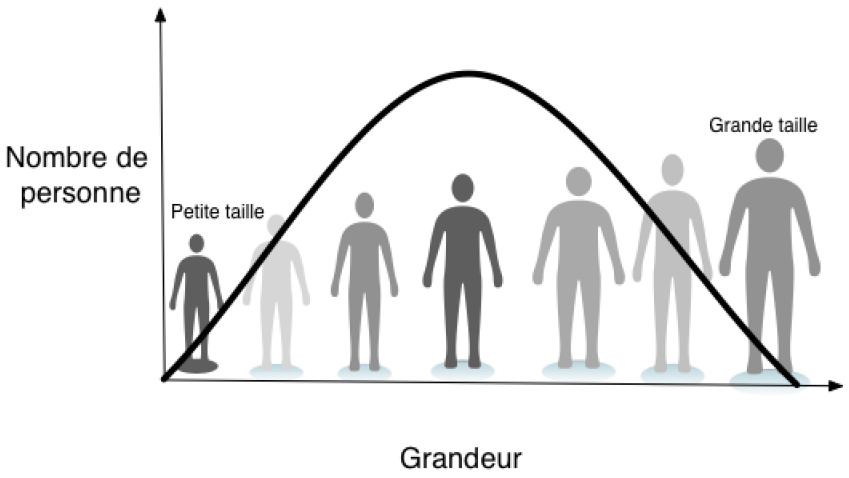


Figure 1.2 – Exemple d'un trait quantitatif complexe. Dans une population, la hauteur des gens suit une courbe normale. Ceci est attribuable à l'interaction entre de nombreuses variations génétiques et des facteurs environnementaux. Bien que les variables environnementales puissent avoir un impact sur la taille adulte, il est clair que la hauteur est principalement déterminée par des allèles spécifiques dont un individu hérite. Cependant, la génétique de la hauteur n'est pas encore complètement comprise. Chez l'humain, plus de 100 gènes ont été associés à la taille [4].

sieurs associations entre des variations géniques et certaines maladies. Cependant, même dans les meilleurs cas, il s'avère que les variations génétiques identifiées ne sont responsables que d'une faible partie de la variance des caractères en question. Par ailleurs, les résultats de ces études ne sont pas toujours reproductibles entre diverses populations. Finalement, même lorsque des locus génétiques sont identifiés, il n'est pas toujours possible d'identifier quel gène au sein du locus contribue au caractère étudié, ou de quelle manière le variant génétique exerce son influence biologique.

1.4 Utilisation d'organismes modèles en génétique

De nombreux mécanismes biologiques fondamentaux, par exemple des voies métaboliques et des voies régulatrices ou développementales, sont conservés entre diverses espèces animales, même entre espèces très éloignées. Ainsi, les mécanismes moléculaires et la transmission de caractères quantitatifs complexes peuvent être disséqués dans

des organismes modèles tels que la mouche drosophile, le ver nématode et la souris. Ces modèles ont de nombreux avantages : 1) ils permettent d'éviter les problèmes éthiques que soulèveraient l'expérimentation chez les humains ; 2) ils constituent une ressource accessible et relativement économique ; 3) ils offrent des avantages de reproductibilité des méthodes expérimentales. Finalement, pour plusieurs de ces organismes, il existe des souches «consanguines», c'est-à-dire des souches où les deux chromosomes sont identiques. Les souches consanguines ont une utilité particulière en génétique, car les croisements entre souches consanguines différentes permettent de créer des individus ou des souches avec des polymorphismes identifiables.

La souris (*mus musculus*) est un modèle animal qui comportent plusieurs avantages : 1) c'est un mammifère (dont le génome est proche de celui de l'homme) ; 2) il est de faible taille et a un cycle de vie court, ce qui permet un élevage relativement rapide et à coût avantageux. La souris comporte 19 paires de chromosomes autosomes et une paire de chromosomes sexuels. Le point de départ de toute étude classique en génétique des caractères complexes est de choisir des souches parentales qui seront croisées pour générer une progéniture où chaque individu aura une répartition différente des variations génétiques présentes dans les souches parentales. Pour le travail de cette thèse, nous avons utilisé un outil génétique particulier, soit des «souches de souris consanguines recombinantes» (RIS, pour Recombinant Inbred Strain). Ce modèle est le fruit d'un travail qui comporte plusieurs étapes : 1) deux souches parentales consanguines originales sont inter-croisées pour générer une progéniture F1 ; 2) des individus F1 sont inter-croisés pour générer des individus F2 ; 3) à partir d'individus F2, plusieurs nouvelles souches consanguines sont créées en répétant des croisements entre frères et soeurs pour plus de 20 générations (voir Figure 1.3). Au final, chacune de ces nouvelles souches consanguines contient des proportions à peu près égales de variations génétiques des deux souches consanguines parentales initiales, mais la répartition de ces variations génétiques est particulière à chaque souche. Comme les nouvelles souches sont consanguines, tous les animaux d'une même lignée sont identiques d'un point de vue génétique et tous les gènes autosomes (et ceux du chromosome X chez les femelles) sont homozygotes. En conséquence, les seules causes de variabilité phénotypique au sein d'individus

d'une même souche et de même sexe sont d'origine non génétique (environnement et/ou méthode utilisée pour mesurer le phénotype).

L'ensemble des diverses souches recombinantes consanguines provenant d'un même croisement initial représente un «panel». Plusieurs panels de lignées recombinantes consanguines de souris ont été développés et peuvent être obtenus auprès d'institutions commerciales et/ou académiques. Pour le travail de cette thèse, nous avons utilisé le panel de souris RIS AXB et BXA. Le panel AXB et BXA est dérivé des souches parentales consanguines C57BL/6J et A/J, le père étant une souris C57BL/6J pour les AXB et une souris A/J pour les BXA. Au total, ce panel comprend 29 souches de souris. Dans notre travail, le nombre de lignées utilisables a été réduit à 24 pour plusieurs raisons : 1) certaines souches sont très semblables génétiquement et sont donc «redondantes» d'un point de vue génétique ; 2) certaines souches ne se reproduisent que très difficilement, ce qui complique l'obtention d'animaux expérimentaux. D'un point de vue expérimental, le panel de souris AXB/BXA offre les avantages suivants :

1. Le génotype de chaque lignée a été bien analysé, et est disponible dans des bases de données publiques [6].

2. La mesure du phénotype peut être répétée sur plusieurs individus génétiquement identiques au sein d'une même lignée. La moyenne résultante augmente la précision de la mesure et permet de minimiser les biais environnementaux et techniques [7].

3. Un phénotype peut être mesuré en fonction du temps sur différents individus d'une même lignée (par exemple à différents âges), même si la mesure nécessite le sacrifice des animaux ou l'utilisation d'une technique invasive [7].

4. Avec les RIS, il est possible de comparer des groupes contrôles et expérimentaux dans une même lignée[7].

Pour le travail de cette thèse, nous avons aussi utilisé des données obtenues par d'autres investigateurs (et disponibles par le biais de bases de données publiques) avec

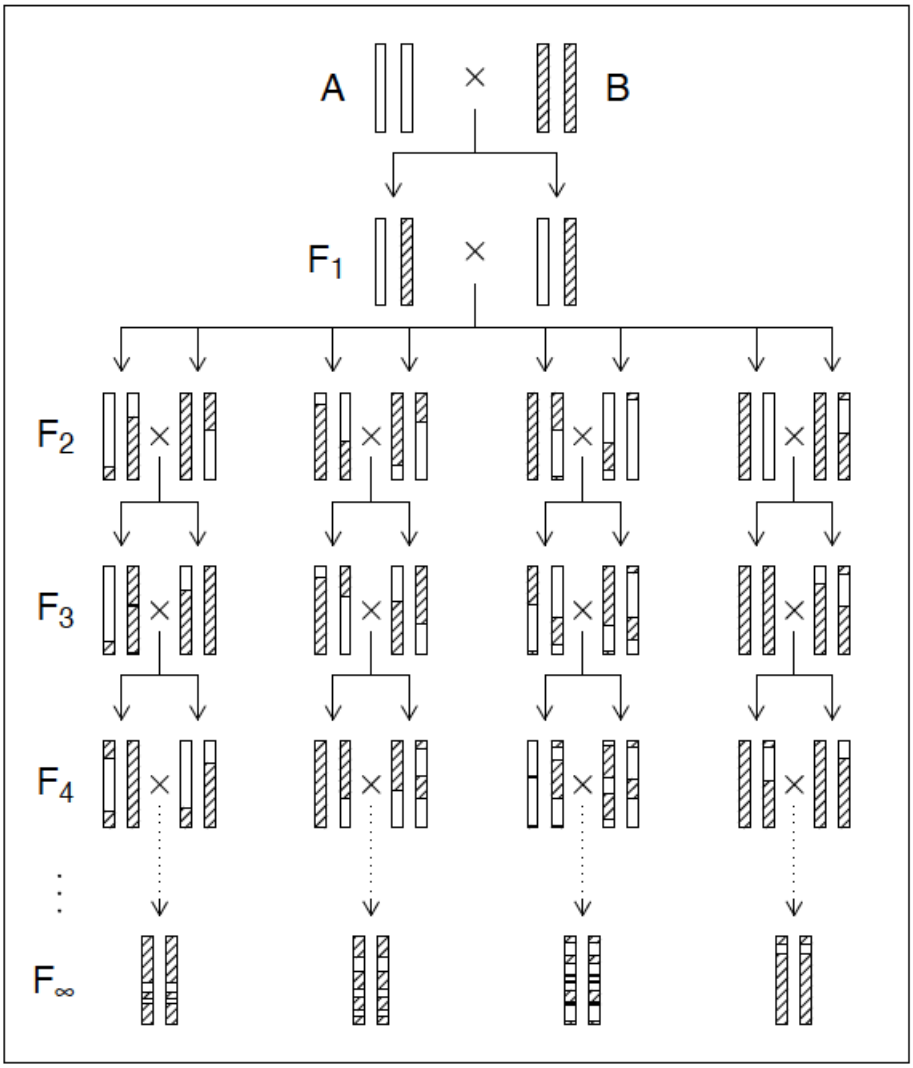


Figure 1.3 – Représentation schématique de la génération de souches recombinantes consanguines. L'élevage des deux premières générations est identique à celle d'un croisement F2. Dans les générations suivantes, les frères et soeurs sont accouplés pour produire une progéniture qui est de moins en moins hétérozygote. Si l'on continue indéfiniment ces croisements, ce processus produira des individus qui sont complètement homozygotes à chaque locus, mais avec des chromosomes qui sont une mosaïque des chromosomes parentaux. Dans la pratique, entre 10 et 20 générations de croisements sont réalisés. Image : [5].

des souches de souris RIS BXD. Ces souris ont été générées à partir des souches parentales C57BL/6J et DBA/2. Théoriquement, le panel de souris BxD offre les mêmes avantages que celui des souris AXB/BXA. De plus, le panel comprend plus de souches que le panel de souris AXB/BXA. Cependant, plusieurs arguments permettent de penser que ces souches comportent encore plusieurs locus hétérozygotes et sont moins homogènes génétiquement que les souris AXB/BXA.

Les croisements génétiques de souris peuvent permettre des avancées importantes dans la compréhension des maladies complexes humaines, incluant l'identification et la validation de gènes candidats contribuant aux manifestations pathologiques des maladies [8]. Cependant, la souris ne reflète pas toujours tous les symptômes cliniques associés aux maladies humaines. De plus, les études en génétique des caractères complexes chez les animaux se limitent souvent à l'analyse de croisements issus de deux souches parentales consanguines particulières. Dans ces conditions, un investigateur ne peut prétendre qu'un modèle animal restreint possède la diversité génétique nécessaire pour saisir la complexité d'un phénotype. Donc, même si un modèle animal peut illustrer certains aspects d'une maladie complexe humaine il ne peut en aucun cas la récapituler complètement [9].

Pour pallier à certains des inconvénients énumérés ci-dessus, comme le manque de diversité génétique, des nouveaux panels de souris tels que le «heterogeneous stock» et le «collaborative cross» ont été développés au cours des dernières années.

Le «heterogeneous stock» [10] sont des animaux dérivés d'un croisement de huit souches de souris consanguines qui ont été élevées pendant 40-50 générations suivant un modèle de croisement qui vise à minimiser la consanguinité. La colonie qui en résulte représente une mosaïque aléatoire des 8 souches fondatrices et dont la distance entre la recombinaison* approche les 2 centimorgans* (cM). Ce croisement a l'avantage d'avoir une grande diversité génétique, mais les souris ne sont pas consanguines donc ce croisement n'est pas idéal pour les analyses intégratives.

Le «collaborative cross» est également un croisement de huit souris consanguines qui sont actuellement en cours de développement [11]. Le «collaborative cross» a été conçu pour remédier à certaines des lacunes des ressources disponibles, comme le petit

nombre de souches disponibles des croisements et le manque de diversité génétique. Le «collaborative cross» est un panel conçu spécifiquement pour l'analyse intégrative de système complexe et donc chacune des souches résultantes sont cosanguines et génétiquement définies.

1.5 Études de locus de traits quantitatifs

Les études de locus de traits quantitatifs sont plus connus sous le nom de QTL, pour «quantitative trait locus». Tel qu'expliqué ci-dessus, la variation de traits quantitatifs complexes est due à l'interaction entre les effets de facteurs environnementaux avec plusieurs locus génétiques. La connaissance du nombre, de la localisation et des effets de ces derniers peuvent mener à de nouvelles découvertes biologiques. Il existe actuellement deux grandes approches en génétique quantitative pour identifier les déterminants génétiques de phénotypes quantitatifs. La première approche, appelée l'analyse de liaison, consiste à cartographier génétiquement des populations bien caractérisées (par exemple les descendants de croisements de souches de référence d'organismes modèles), ce qui permet d'identifier la liaison de phénotypes avec les QTLs qui contiennent des mutations causales. La deuxième approche est représentée par les études d'associations sur le génome complet (GWAS, pour genome-wide association studies). Ces dernières peuvent être effectuées sur des populations moins structurées et permettent d'identifier des marqueurs génétiques communs associés à un phénotype. Plusieurs centaines d'études GWAS et QTL ont été réalisées tant chez l'homme que chez des organismes modèles, conduisant à l'identification de milliers de locus associés à des phénotypes et à des maladies. Pour les besoins de mon projet, je vais restreindre la discussion aux études de QTLs effectuées sur des populations d'organismes modèles. Le point de départ d'une étude de cartographie de QTL est de créer une population génétiquement diverse et de posséder un ensemble de marqueurs génétiques pour chaque individu de la population étudiée. Un marqueur génétique est une région d'ADN variable à travers une population et identifiable par des techniques d'analyse moléculaire. Les SNPs sont un exemple de marqueurs génétiques communément utilisés. La cartographie génétique consiste à ana-

lyser chaque individu d'une population pour identifier l'origine parentale d'un ensemble de marqueurs polymorphiques. L'objectif final est ensuite de trouver le(s) marqueur(s) statistiquement le(s) plus susceptible(s) d'être lié(s) à la valeur quantitative du trait étudié.

Lorsque le trait étudié est un trait quantitatif complexe, un locus peut avoir (en fonction de son origine allélique) des effets quantitatifs sur le phénotype. Dans le cas de traits à transmission mendélienne, un locus peut affecter un phénotype de manière qualitative (effet de type «tout ou rien»). Cependant, la distinction entre un locus de type mendélien et un QTL est souvent artificielle, car les mêmes techniques de détection peuvent être appliquées dans les deux cas. La classification des effets alléliques devrait être considérée comme un continuum, avec les caractères mendéliens à une extrémité et les caractères très polygéniques à l'autre extrême. Une cartographie de QTL sera simplifiée si le caractère considéré est sous le contrôle d'un petit nombre de locus ayant chacun des effets majeurs ou modérément prononcés [12].

1.6 Méthodes de détection de QTLs

Après le phénotypage et la cartographie d'individus d'une population, la détection de QTLs repose sur des méthodes d'analyse statistique [5]. Au fil des années, plusieurs méthodes de complexité croissante ont été développées à cette fin, tel que décrit ci-dessous.

1.6.1 Méthode de régression par marqueur

Les lois de base de la cartographie des QTLs ont été formulées par Soller et al en 1976. La méthode la plus simple pour trouver des QTLs consiste à étudier l'effet des marqueurs sur les valeurs des phénotypes par l'analyse de la variance (ANOVA). Pour chaque marqueur génétique, cette méthode divise les individus de la population en groupes selon leurs génotypes respectifs et compare la valeur des caractères quantitatifs respectifs entre les divers groupes génétiques en faisant un test statistique de significativité (par exemple un test de Student (*t-test*) en cas de 2 variations alléliques possibles, ou

un test de Fisher (F-test) lorsqu'il y a plus de 2 variations alléliques possibles). Cette procédure est répétée pour chaque marqueur. Lorsqu'un marqueur présente un génotype qui co-ségrègue de façon significative avec la distribution phénotypique dans la population, on dit que ce marqueur est «lié» à un QTL (voir Figure 1.4). La méthode de régression par marqueur a l'avantage d'être simple et de prendre en compte des covariables (sexe, traitement, effets environnementaux). Elle peut même être effectuée en l'absence d'une carte génétique complète.

L'approche d'analyse ANOVA pour la cartographie des QTLs a plusieurs faiblesses importantes [13, 14] :

1. On ne peut connaître la localisation précise du QTL.
2. Cette méthode ne permet pas d'avoir une estimation de l'effet du QTL.
3. Il est nécessaire de retirer de l'analyse les individus dont les génotypes sont manquants au niveau du marqueur.
4. Lorsque les marqueurs sont très espacés, les QTLs peuvent être assez loin des marqueurs les plus proches, ce qui diminue la puissance de détection de QTLs.

1.6.2 Méthode de cartographie par intervalles

La méthode de cartographie par intervalles a été proposée par Lander et Botstein en 1989 [13] pour faire face aux faiblesses de la méthode de régression par marqueur et représente actuellement l'approche la plus populaire pour la cartographie des QTLs dans des croisements expérimentaux. Dans cette méthode, chaque localisation dans le génome est considérée comme un emplacement potentiel de QTL et le génotype à cet emplacement est estimé à partir de l'information des marqueurs adjacents. La méthode de cartographie par intervalles se base sur les principes de la recombinaison méiotique et nécessite de calculer la probabilité qu'un QTL ait un certain génotype en fonction de la distance entre les marqueurs flanquants et le génotype de ces derniers (voir Figure 1.5).

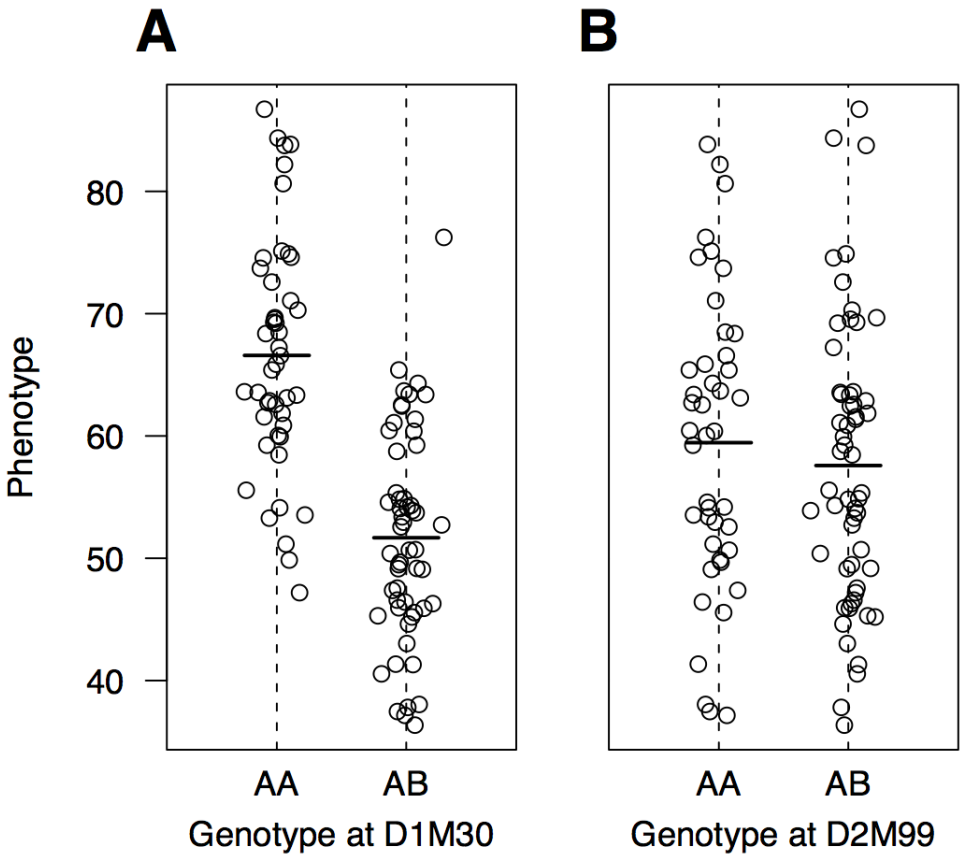


Figure 1.4 – Représentation graphique des valeurs de phénotype séparé selon leur génotype. Exemple pour un QTL significatif pour le marqueur 1 (figure à la gauche) et d'un marqueur non significatif (figure à la droite). Les lignes horizontales montrent la moyenne du groupe. Image : [14].

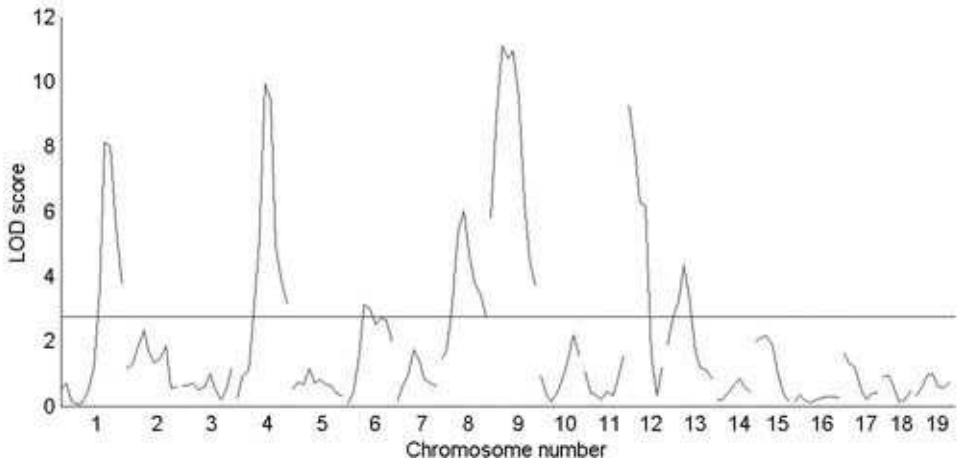


Figure 1.5 – Exemple de cartographie pour analyse génomique de la densité de l'os trabéculaire, d'un croisement entre les souches de souris C57BL/6J et C3H. La ligne horizontale représente un LOD score significatif de 0,05 (tel que déterminé par un test de permutation). Dans cette analyse, il y a plusieurs QTLs significatifs. Image : [14].

La liaison entre le QTL et le phénotype est mesuré avec un score appelé «Logarithm of the odds» (LOD), qui constitue le logarithme du rapport de la probabilité de la présence d'un QTL à la position donnée sur la probabilité de l'absence de QTL dans le génome. Plus le LOD est élevé, plus l'existence d'un QTL est vraisemblable. Ce calcul est effectué pour toutes les positions possibles du génome et les valeurs sont reportées sur un graphique présentant sous forme de courbe comment les probabilités se répartissent sur tous les intervalles de distance sur les chromosomes. Plusieurs algorithmes spécialisés ont été développés afin d'utiliser la méthode de cartographie par intervalles : l'algorithme expectation-maximisation (EM) [13], la méthode de Haley et Knott [15] ou la méthode bayésienne [16].

La méthode par intervalles simples (Interval Mapping) permet de tester l'hypothèse qu'il existe seulement un QTL dans l'intervalle entre deux marqueurs. La limite de cette méthode est que l'on ne peut pas détecter des QTLs proches sur le même chromosome. Pour remédier à cette problématique, la cartographie en intervalles composés (Composite Interval Mapping) consiste à faire dans un premier temps une cartographie d'intervalles simples puis à accorder à chaque marqueur le plus proche des QTLs détectés une valeur en tant que co-facteur lors d'une deuxième analyse.

Une fois la cartographie par intervalles effectuée, il faut établir un seuil de significativité au-delà duquel une valeur de LOD permettra de considérer qu'un locus peut être considéré comme un QTL (c'est-à-dire qu'il est lié de façon significative à la valeur du phénotype en question). Les tests de permutation* représentent une approche bien reconnue pour obtenir des valeurs de seuils ajustées pour un test multiple dans un criblage d'un génome en particulier [17, 18]. De plus, il faut également estimer l'intervalle de confiance pour la localisation des QTLs. La méthode la plus communément acceptée est celle du «saut de LOD» ; elle consiste à considérer les bornes de l'intervalle de confiance comme les régions où les valeurs de LOD sont $\pm$ 1 LOD par rapport à la valeur de LOD observée au pic du QTL. La cartographie par intervalles permet aussi de considérer les interactions entre plusieurs QTLs. Il est donc possible d'identifier des locus qui n'ont aucun effet individuellement, mais qui ont un impact sur le phénotype dans le contexte d'une interaction [16, 19–21].

On travaille encore à améliorer les procédures de cartographie QTL pour les rendre plus faciles et plus efficaces [22]. Par exemple, l'analyse simultanée de plusieurs traits pour augmenter la puissance de détection des QTL.

1.6.3 Méthode bayésienne

En statistique, deux approches différentes ont été développées pour l'étude des probabilités : les statistiques classiques (aussi appelées «fréquentistes») et l'inférence bayésienne. L'approche bayésienne fait le choix de modéliser les résultats attendus en début de processus (quitte à réviser ce premier jugement en donnant des poids de plus en plus faibles aux «a priori» au fur et à mesure des observations). Elle diffère ainsi des statistiques fréquentistes, qui utilisent principalement les propriétés des lois sur les observations, fixent des «a priori» sur une méthode et une hypothèse arbitraire et ne traitent les données qu'ensuite. Malgré leurs différences, les deux approches sont en fait complémentaires. Les statistiques fréquentistes sont en général préférables lorsque les informations sont abondantes et que leur coût de collecte est faible. L'inférence bayésienne s'avère avantageuse dans les cas où le nombre d'échantillons est faible. Comme cette approche exprime toutes les formes d'incertitude en termes de probabilité, elle fournit

un modèle naturel pour modéliser des données complexes et combiner diverses sources d'information.

Une approche bayésienne pour la détection de QTLs, telle que présentée initialement par Sen et Churchill [16], peut se résumer comme suit. Supposons une population de i individus. La valeur du phénotype pour chaque individu est Y_i. Dans une étude classique de QTL, nous pouvons modéliser le phénotype comme une régression linéaire standard sur la valeur du génotype pour chacun des marqueurs génétiques analysés :

$$Y_i = \mu + \sum_{j=1}^{S} X_{ij}\beta_j + \varepsilon_i, \tag{1.1}$$

où

- $i = 1, \cdots, n$ est un individu particulié et $j = 1, \cdots, S$ est un SNP particulier,
- μ représente la moyenne globale du phénotype à travers tous les individus de la population,
- β_j représente l'ampleur de l'effet du marqueur j sur le phénotype,
- X_{ij} représente les éléments de la matrice de SNPs,
- et ε_i représente le terme d'erreur (que nous assumons être distribuée de manière gaussienne).

En pratique, seulement une minorité de marqueurs affectent le phénotype, ce qui fait que la valeur de β dans la majorité des cas devrait être $\beta = 0$. Pour spécifier quels marqueurs doivent être inclus dans le modèle, nous pouvons incorporer un indicateur de variables, γ_j, ce qui affecte le modèle de la manière suivante :

$$Y_i = \mu + \sum_{j=1}^{S} X_{ij}\gamma_j\beta_j + \varepsilon_i, \tag{1.2}$$

où

- γ_j=1 si le marqueur j est associé avec le phénotype, et γ_j=0 dans le cas contraire.

Dans une inférence bayésienne, les distributions a «priori»* sont placées sur les paramètres inconnus, pour permettre de calculer a «posteriori»* les probabilités qu'un SNP soit lié avec le phénotype. Cette valeur est calculée en comptabilisant le nombre de fois que γ prend la valeur de 1. La distribution a «priori» des γ, c'est-à-dire la probabilité a priori que le marqueur j soit associé avec le phénotype, Pr(γ_j=1), est importante, car elle affecte directement la probabilité a «posteriori». Dans une analyse typique, cette valeur est fixée avec une valeur prédéterminée que nous appelons W, et peut être interprétée comme la proportion a «priori» de l'association des marqueurs avec le phénotype. L'estimation des paramètres du modèle se fait à l'aide d'une classe d'algorithmes appelés «Markov Chain Monte Carlo* (MCMC)».

1.7 La masse ventriculaire gauche (MVG) du coeur

Dans notre laboratoire, nous étudions un trait complexe en particulier, soit la masse du ventricule gauche cardiaque (MVG). Le coeur est une pompe composée de quatre cavités : les deux chambres supérieures, appelées oreillettes, servent à collecter le sang circulant vers les chambres cardiaques ; les deux chambres inférieures, appelées ventricules, servent à pomper le sang hors des chambres cardiaques. Les valves séparent les oreillettes des ventricules et permettent au sang de circuler dans le coeur de façon unidirectionnelle. Le septum est la paroi qui sépare le côté droit du côté gauche du coeur. Le coeur est composé de tissu musculaire, ce qui lui permet d'exercer sa fonction de pompe et de faire circuler le sang dans tout l'organisme. Le sang circule dans le système vasculaire, qui comprend les artères, les veines et les capillaires.

Le ventricule gauche est l'une des deux cavités inférieures du coeur. Son rôle est de recevoir le sang oxygéné provenant de l'oreillette gauche, puis de le propulser dans le corps via l'aorte. L'effort fourni par le ventricule gauche pour faire circuler le sang dans le système vasculaire périphérique est plus important que pour le ventricule droit (qui propulse le sang dans la circulation pulmonaire). On peut ainsi observer que le myocarde du ventricule gauche est plus épais que le droit. Ainsi, les variations de la

masse totale du coeur dépendent principalement de variations dans la masse du ventricule gauche (bien que certaines pathologies puissent toucher préférentiellement la masse du ventricule droit).

Les variations de la MVG peuvent résulter de plusieurs causes :

1. **Hypertrophie physiologique suite à une augmentation de l'effort cardiaque.**
   Cette condition est généralement réversible et considérée comme physiologique [23]. Des exemples de ce type d'hypertrophie sont 1) la croissance du coeur pendant la grossesse [24] et 2) la croissance du coeur des athlètes à la suite de l'exercice extrême et / ou répétitive [25]. On peut observer une augmentation de la masse du ventricule gauche chez les sportifs qui suivent un entraînement chronique et soutenu [26]. Ce remodelage est dû à l'adaptation du coeur à la surcharge hémodynamique durant l'effort physique pour mieux oxygéner le corps.

2. **Hypertrophie suite à une augmentation de la charge cardiaque ou la présence d'agents hormonaux hypertrophiants.**
   Des études de populations ont montré qu'une pression artérielle élevée (hypertension) est considérée comme l'indicateur le plus important de l'hypertrophie ventriculaire gauche [27]. Dans l'étude longitudinale de Framingham, les participants souffrant d'hypertension artérielle et ceux qui ont pris des médicaments antihypertenseurs avaient une plus grande augmentation de la masse du ventricule gauche [28]. L'hypertrophie du ventricule gauche peut être induite par des moyens pharmacologiques comme l'angiotensine II [29]. Ce type d'hypertrophie est qualifié de pathologique, car il résulte de remodelage chronique qui mènera à une dilatation cardiomyopathique ou à une défaillance cardiaque. Ce type d'hypertrophie semble être associé avec des maladies cardiovasculaires telles que l'hypertension, les dysfonctions vasculaires, l'athérosclérose et les maladies systémiques telles que l'insuffisance rénale.

3. **Hypertrophie due à des mutations de gènes (les cardiomyopathies hypertrophiques).**

Les cardiomyopathies hypertrophiques (en anglais «hypertrophic cardiomyopathy») [30] est la forme le plus commune de maladies cardiaques monogéniques. Elles affectent 0.2% de la population [31]. Les symptômes cliniques sont très variables ; elles sont une cause fréquente de mort subite chez les jeunes athlètes. Les 20 dernières années, on a identifié plus de 900 mutations dans 23 gènes (principalement des protéines du sarcomère) liés à ces maladies [32]. C'est une hypertrophie pathologique qui peut éventuellement évoluer vers une insuffisance cardiaque.

4. **Variation naturelle de la masse cardiaque chez les individus d'une population.**
   Ces variations ne sont pas pathologiques en soi, et des valeurs élevées de la MVG ne peuvent pas nécessairement être considérées comme des «hypertrophies». Cependant, même lorsqu'elle résulte de variations naturelles, la valeur de la MVG corrèle étroitement avec le risque de mortalité ou de morbidité cardiovasculaires [33]. Pour cette raison, c'est un indice prédictif intéressant.

1.8 La MVG comme trait quantitatif complexe

Les maladies cardiovasculaires sont encore l'une des causes principales de décès dans les pays industrialisés. Une masse du ventricule gauche (MVG) élevée est un facteur prédictif majeur de mortalité et de morbidité cardiovasculaire chez les humains [33–37].

Les variations phénotypiques entre les individus sont dues à des déterminants génétiques, à des facteurs environnementaux et au style de vie. L'héritabilité se réfère à la variation phénotypique d'une population qui est attribuable à la variation génétique entre les individus. L'héritabilité de la MVG a été estimée dans des études de jumeaux [38–42], des familles nucléaires [43–45] et des familles complexes [46–50] et celle-ci variait entre 15 [47] à 84% [41]. Le taux d'héritabilité est important pour la MVG et donc la majeure partie de la variance de la MVG peut être expliquée par des facteurs génétiques.

L'identification de variations génétiques responsables de la variance de la MVG

pourrait ainsi mener à l'identification de facteurs de prédisposition génétique aux maladies cardiaques, à mieux stratifier le risque cardiovasculaire et/ou à mieux comprendre les mécanismes responsables de réguler la MVG. Chez l'humain, peu de progrès ont été faits à ce jour dans l'identification des gènes responsables de traits complexes cardiaques [51]. Des études sur des gènes candidats ont identifié plusieurs SNPs dans des gènes comme *ACE* [52, 53], *PPARA* [54], *GNB3* [55] et *CYP11B2* [54] qui pourraient contribuer à la variabilité de la MVG. Dans le cas de la MVG, une méta-analyse de 5 cohortes comprenant plus de 12 000 individus n'a permis d'identifier que quelques gènes candidats [56]. Ils ont identifié trois locus dans les régions intergéniques près des gènes *NOVA1*, *CALM2* et *MEIS2*.

Une alternative intéressante est de réaliser des études de liaison avec des croisements génétiques animaux. De tels modèles permettent à la fois de : 1) diminuer le nombre d'allèles existant pour chaque marqueur (uniquement deux origines parentales possibles pour la plupart des croisements) ; 2) multiplier les possibilités de recombinaisons génétiques au sein d'une même progénie ; 3) mieux contrôler les variables environnementales (et ainsi, diminuer l'importance de ces dernières sur la variance du phénotype). Au total, ces avantages augmentent la puissance statistique des études et facilitent l'identification de QTLs. Cette approche peut ainsi révéler des gènes candidats qui peuvent être ensuite testés, soit par des manipulations génétiques chez des animaux, soit par association dans des populations humaines.

Le panel de souris RIS AXB/BXA représente un ensemble de souches provenant de croisements réciproques entre les souches consanguines A/J et C57BL/6J (le père original étant C57BL/6J ou A/J dans les souris AXB et BXA, respectivement). En plus de sa disponibilité, ce choix de croisement pour cette étude a été motivé par le fait que les deux souches parentales présentent un phénotype cardiaque bien contrasté, car les souches A/J et C57BL/6J possèdent, respectivement, une MVG faible et élevée (voir Figure 1.6). Sur la base d'une cartographie de QTLs, notre laboratoire a identifié un QTL fortement lié à la MVG sur le chromosome 13 dans l'ensemble de souches AXB/BXA [57]. L'intérêt de cette région est renforcé par le fait que la même région synténique* a été liée à la MVG dans un panel de souches RIS de rats [58]. Notez également que

d'autres QTLs suggestifs pour la MVG ont également été identifiés sur le chromosome 12 et 16 dans l'ensemble de souches AXB/BXA.

1.9 Limitation des études de locus de traits quantitatifs

Les études de QTLs ont permis de trouver des régions liées à des traits complexes. Cependant, trouver un gène candidat dans ces régions n'est pas toujours facile et plusieurs expériences sont nécessaires pour la validation de gènes candidats. Dans les études de type GWAS, il n'est pas rare d'observer que les locus identifiés correspondent à des régions sans gènes et/ou éléments fonctionnels connus (régions parfois appelées «déserts géniques») [59]. Chez l'humain, même dans les quelques cas où on a pu démontrer un effet quantitatif de certaines variations génétiques sur un phénotype particulier, les contributions de ces variations au phénotype ne sont généralement que modestes, et l'ensemble de ces variations n'explique qu'une très petite partie de l'héritabilité globale observée dans la population [60]. Pour augmenter les chances d'identifier les contributions de certains gènes à des phénotypes quantitatifs complexes, une des avenues qui a été explorée au cours des dernières années consiste à combiner des études d'expression de gènes avec des études classiques de génétique.

1.10 La régulation de l'expression des gènes

Le plus souvent, le produit d'un gène est une protéine, mais un gène peut aussi coder d'autres types de molécules qui exerceront des fonctions au niveau des cellules, par exemple les ARNs non codants (ARNnc) [61]. Il y a deux étapes dans la production d'une protéine : 1) la transcription, qui correspond à la production de copies d'ARN à partir de l'ADN génomique ; 2) la translation, qui correspond à l'étape où l'ARNm* est décodé pour produire une chaîne d'acides aminés. Cette dernière entreprend ensuite des changements conformationnels pour finalement produire une protéine fonctionnelle. L'expression de gènes (ou «expression génique») représente ainsi un processus par lequel l'information héréditaire stockée dans un gène est transformée en une molécule qui a un rôle fonctionnel dans une cellule. Le contrôle de l'abondance d'une protéine repré-

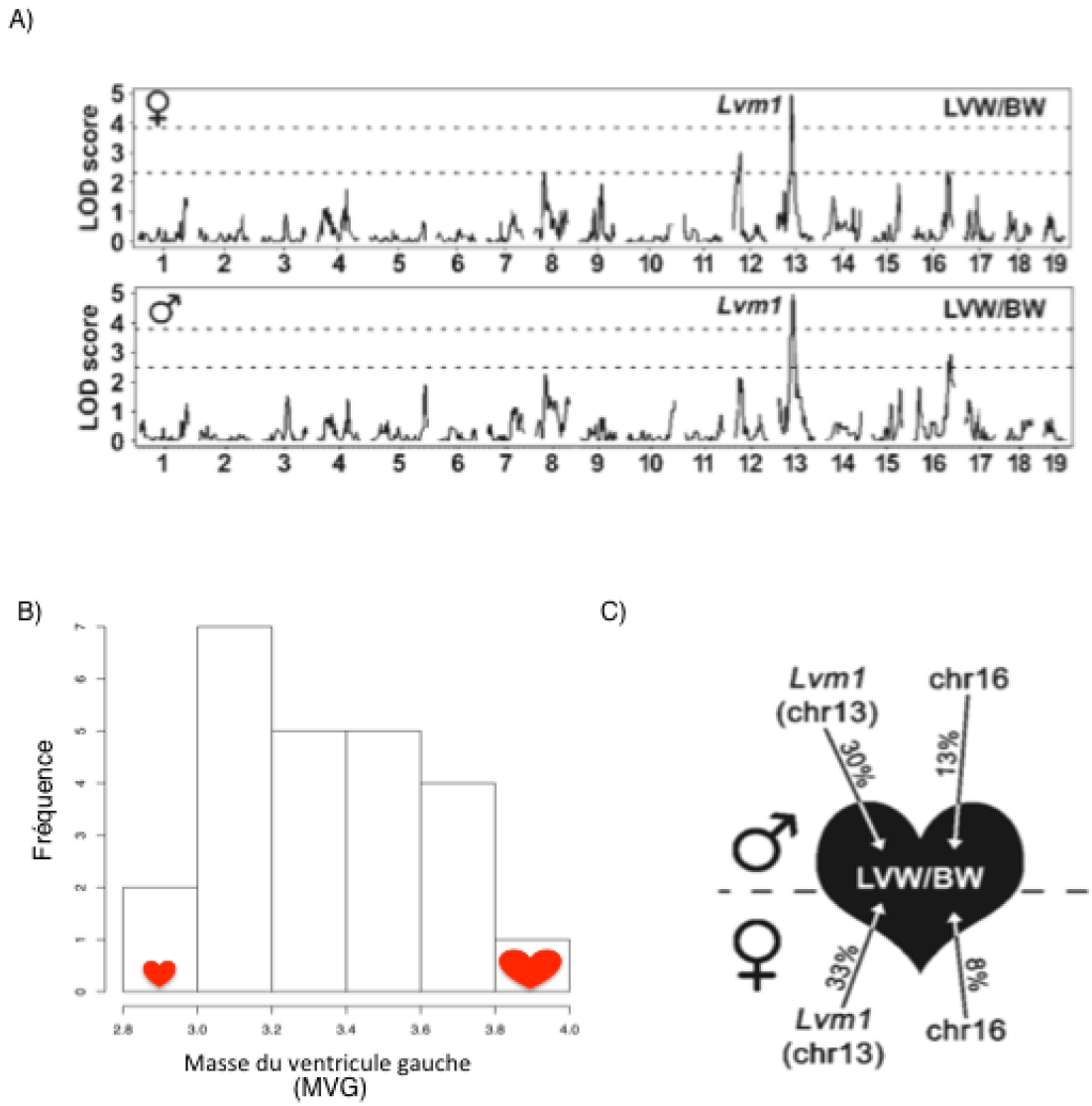


Figure 1.6 – A) La cartographie QTLs pour la masse du ventricule gauche pour les mâles et les femelles. On remarque un locus significatif sur le chromosome 13 et aussi un locus d'intérêt sur le chromosome 16. B) La distribution de la masse du ventricule gauche de la population de souris ABX-BXA. La distribution du phénotype suit une distribution normale, ce qui est caractéristique des traits complexes. C) Résumé des QTLs identifiés pour les mâles et les femelles. Image : [57]

sente donc un des mécanismes par lesquels un gène peut exercer une fonction dans une cellule. Idéalement, la mesure du produit final (c'est-à-dire l'abondance de la protéine) est celle qui permettrait la meilleure compréhension de ces processus. Cependant, les contraintes techniques font en sorte qu'il est plus facile d'inférer le niveau d'expression des gènes en mesurant le messager intermédiaire, soit les niveaux d'ARNm. De plus, les mesures de protéines se prêtent mal aux mesures de criblage à haut débit, alors que les puces à ADN permettent de mesurer dans un échantillon cellulaire ou tissulaire l'abondance de toutes les molécules d'ARNm transcrites par le génome dans cet échantillon. Les puces à ADN mesurent l'abondance de transcrits d'ARNm en utilisant une technologie basée sur les propriétés d'hybridation de l'ARN et de l'ADN. Plus récemment, de nouvelles techniques de séquençage à haut débit ont également été utilisées pour mesurer l'abondance de tous les transcrits d'ARNm au sein d'échantillons (cette technique est communément appelée «RNA-seq») [62].

Dans un organisme, certains gènes ne sont exprimés que dans certaines cellules, sous certaines conditions environnementales, à différents moments et à différents niveaux. La régulation des expressions des gènes comporte l'ensemble des mécanismes de régulation mis en oeuvre pour passer l'information génétique incluse dans une séquence d'ADN à une molécule fonctionnelle (ARN ou protéine). Plusieurs facteurs affectent la régulation de l'expression des gènes : 1) l'ARN polymérase (qui est un complexe enzymatique qui initie et coordonne la synthèse de l'ARN à partir de l'ADN au sein d'un complexe de transcription) ; 2) les promoteurs* (qui sont des séquences d'ADN à proximité d'un gène où se fixe l'ARN polymérase) ; 3) les amplificateurs (qui sont des régions d'ADN où certaines protéines peuvent se fixer pour stimuler la transcription ; 4) les facteurs de transcription (qui sont des protéines nécessaires pour initier et réguler la transcription de gènes) ; ces facteurs se fixent généralement au niveau des promoteurs et/ou des amplificateurs, et peuvent autant activer que désactiver le complexe de transcription ; et 5) la chromatine (qui comprend plusieurs protéines appelées «histones» autour desquelles sont déroulés les filaments d'ADN génomique ; plusieurs types de modifications covalentes au niveau des histones peuvent influencer l'expression des gènes dans les régions correspondantes). Ces dernières années, les nouvelles technologies de puces à ADN* ont

permis d'utiliser les mesures d'expression de gènes comme méthodes complémentaires aux approches purement génétiques, et ainsi faciliter l'identification de gènes contribuant à des caractères complexes. Par exemple, une méthode particulière consiste à corréler l'expression des gènes avec des caractères phénotypiques quantitatifs (voir Figure 1.7). Les gènes dont le niveau d'expression varie avec la valeur de traits quantitatifs complexes sont appelés «quantitative trait transcripts (QTTs)». En sélectionnant les gènes dont les QTTs corrèlent le plus fortement avec des caractères particuliers, il a ainsi été possible d'identifier parmi ceux-ci des gènes qui contribuent de façon causale à certains caractères complexes [63].

L'expression génique permet d'identifier les profils d'expression distincts avec la population. L'intégration de l'expression génique et des données phénotypiques permet d'identifier des gènes corrélant avec un phénotype. L'intégration de l'expression génique et des données génétiques permet l'identification de QTL d'expression.

1.11 Les locus de trait quantitatifs d'expression

Les études de locus de trait quantitatifs d'expression (eQTL) sont un champ d'études aussi appelé «Genetical Genomics». Ce type d'expérience a été proposé pour la première fois en 2001. Le but d'une étude d'eQTLs est d'identifier les régions génomiques qui affectent l'expression de gènes. Dans le cadre de traits complexes, les études sur l'expression des gènes sont intéressantes parce que : 1) l'abondance de transcrits d'ARNm peut être considérée comme un trait intermédiaire entre les variants génétiques et les traits phénotypiques [64, 65] 2) l'effet d'une variation d'ADN sur l'expression des gènes est sans doute plus facile à détecter que son lien avec un trait complexe, en raison des nombreuses étapes intermédiaires supplémentaires menant à la régulation du trait phénotypique. Ceci a mené plusieurs à postuler qu'en intégrant trois types de données (les marqueurs génétiques, l'expression des gènes et les données phénotypiques), il serait possible de mieux disséquer l'architecture des traits quantitatifs complexes. Par exemple, si un SNP est associé à la fois avec un trait complexe et avec l'expression d'un gène au voisinage du SNP, il est possible de prioriser ce gène comme un candidat possible cau-

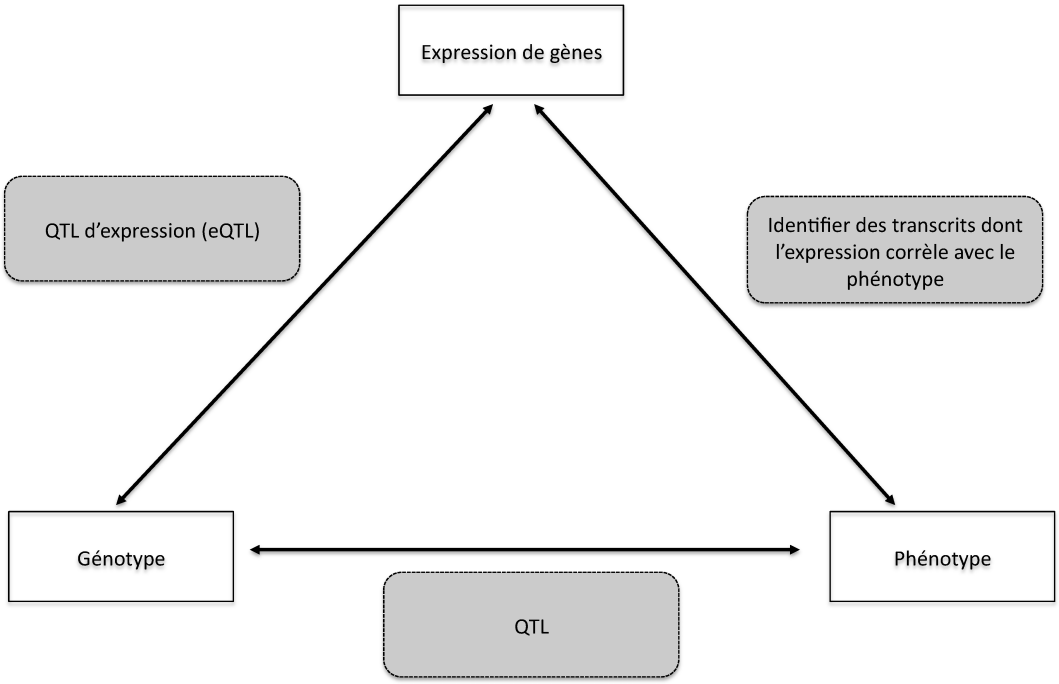


Figure 1.7 – Résultats potentiels de la combinaison de l'expression génique, des données phénotypiques et des données génétiques. L'expression génique permet d'identifier les profils d'expression distincts dans la population. L'intégration de l'expression génique et des données phénotypiques permet d'identifier des gènes corrélant avec un phénotype. L'intégration de l'expression génique et des données génétiques permet l'identification de QTL d'expression.

sant une variation du phénotype.

Le premier résultat obtenu lors d'une analyse de eQTLs est une liste de gènes et de locus qui sont (dépendamment du type d'étude) associés ou liés l'un à l'autre au-delà d'un certain seuil statistique. Les eQTLs qui sont identifiés sont généralement classés en deux grandes catégories : les cis-eQTL* et les trans-eQTL* (voir Figure 1.8).

1.11.1 Les cis-eQTLs

Lorsque l'emplacement d'un eQTL correspond au locus du gène dont l'abondance du transcrit est mesurée, celui-ci est identifié comme un cis-eQTL. La prémisse générale est que les cis-eQTLs correspondent à des situations où une variation génétique dans la région régulatrice des gènes est très probablement responsable de la variation du niveau de transcription (et que la variation régule l'expression du gène «en cis»). Le seuil de distance entre le variant génétique et le gène régulé a été déterminé de façon empirique, et dépend en partie de la densité de la carte génétique [66]. Typiquement, cette distance varie entre 1-5 Mb, selon les études.

Comme cette distance est arbitraire, dans certaines études les cis-eQTL sont appelés «eQTL local» ou «eQTL proximal» et les trans-eQTLs sont appelés «eQTL distal». Pour distinguer les cis-eQTLs des trans-eQTLs, certaines études (RNA-Seq) ont investigué l'effet des allèles hétérozygotes [67]. Pour détecter l'expression allélique, il faut être capable de distinguer quantitativement l'expression de chaque allèle. En principe, l'expression allélique permet un test plus puissant pour détecter les cis-eQTLs [68, 69].

Les cis-eQTLs qui sont également localisés avec le QTL phénotypique et correspondent à des gènes dont l'expression est corrélée avec la variation quantitative du phénotype ont été appelés «c3-eQTLs». Les gènes «c3-eQTLs» ont été utilisés pour hiérarchiser les gènes candidats à considérer dans l'étude des traits complexes.

1.11.2 Les trans-eQTLs

Les trans-eQTLs correspondent à des situations où le QTL régule l'expression de gènes situés dans une région éloignée du QTL et se situent dans la plupart des cas sur des

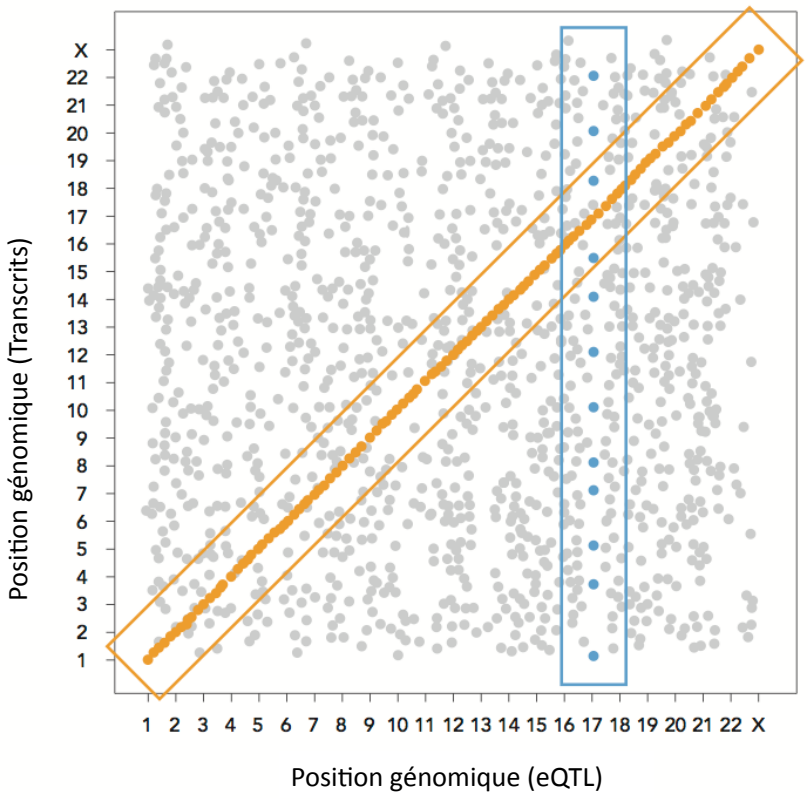


Figure 1.8 – Représentation graphique de l'analyse d' eQTLs. Sur l'axe des X se trouve la position des eQTL et sur l'axe des Y la position des gènes. Les points représentent les eQTLs. Ce graphique illustre les observations de l'analyse eQTL rapportées dans la littérature actuelle. La bande diagonale indique cis-eQTL et la bande horizontale représente un «hotspot» de trans-eQTLs, ce qui suggère que l'expression de plusieurs gènes est associée au même polymorphisme. (Source : http ://www.illumina.com/Documents/products/technotes/ technote_integrating_expression_analysis.pdf)

chromosomes différents. Cependant, les effets génétiques par lesquels les trans-eQTLs sont liés (ou associés) «en trans» à l'expression de gènes sont moins forts que dans les cas de cis-eQTLs. De par ce fait, les trans-eQTLs sont généralement considérés comme plus difficiles à détecter que les cis-eQTLs [66]. Cependant, il existe des situations particulières où un même trans-eQTL peut être lié (ou associé) à l'expression d'un grand nombre de gènes : ce type de locus correspond à ce qui a été appelé un «hotspot» de trans-eQTLs. Ces situations suggèrent qu'une même région polymorphe peut contenir un régulateur important susceptible d'affecter de façon coordonnée l'expression de plusieurs gènes. Si les gènes régulés partagent des fonctions similaires, les «hotspots» de trans-eQTLs peuvent mener à un mécanisme où la variation génétique aura un impact significatif sur certaines fonctions biologiques de l'organisme étudié.

1.12 Outils utilisés pour la détection des eQTLs

Historiquement, les outils informatiques utilisés pour la détection des eQTLs étaient les mêmes que ceux utilisés pour la détection des eQTLs phénotypiques. Ainsi, l'expression de chaque gène était considérée individuellement et l'analyse était répétée pour chaque gène. Cependant, l'analyse d'expression par puces à ADN introduit plusieurs problèmes particuliers : 1) la grande dimensionnalité des données nécessite d'utiliser des procédures de corrections multiples ; 2) ce type d'étude ne comprend souvent qu'un nombre limité d'individus ; 3) en analysant un gène à la fois, l'analyse ne tient pas compte des nombreuses interactions possibles qui peuvent exister entre les gènes. Plusieurs approches sont disponibles pour pallier à ces limitations.

1.12.1 La méthode «Multivariate Sparse Least Square»

Le concept de «Partial Least Square» (PLS) correspond à une technique de réduction de dimension, et peut être vu comme une généralisation d'une «Principal Component Analysis (PCA*)» et d'une régression multiple. La régression PLS cherche des composantes (appelées variables latentes) liées à X (prédicteurs) et à Y (variables à expliquer), servant à exprimer la régression de Y sur ces variables et finalement d'Y sur X. Donc la

«sparse Partial Least Square»(SPLS) est une version «sparse» de la PLS. Ceci indique une situation où le nombre de prédicteurs disponibles (dans ce cas-ci, les marqueurs génétiques) est grand, mais qu'en réalité il n'existe qu'un petit nombre de prédicteurs présentant une corrélation significative avec le trait étudié (c'est-à-dire que les coefficients de corrélation sont différents de la valeur zéro). La méthode «Multivariate Sparse Least Square» est une méthode de régression «sparse» [70] . Cette méthode suit l'algorithme suivant : 1) regroupement de l'expression des gènes(on peut utiliser n'importe lequel algorithme de la littérature) ;2) pour chaque groupe, on voit le profil d'expression des gènes d'un groupe comme une réponse multivariée donc, on fait une régression «Multivariate Sparse Least Square» ; 3) on construit un intervalle de confiance par bootstrap pour les transcriptions sélectionnées.

1.12.2 Les méthodes d'inférence bayésienne

Les méthodes d'inférence bayésienne [71–74] sont basées sur le modèle classique, mais varient dans la façon dont les «a priori» sont définis. Cette spécification des «a priori» permet un contrôle pour la dépendance entre les gènes et pour la dépendance entre les SNPs, et offre la possibilité d'incorporer (ou non) des informations biologiques. La dépendance entre les gènes peut être due à une dépendance génétique (lorsqu'ils sont contrôlés par le même SNP) ou biologique (lorsque les gènes appartiennent à une même famille fonctionnelle ou lorsqu'ils sont contrôlés par une même voie de signalisation). Dans les faits, la dépendance non-génétique est très difficile à modéliser [75] et a été limitée dans des études antérieures à un petit nombre de gènes pour des raisons computationnelles [74].

1.13 Limitations des études de QTL d'expression

La question fondamentale en génétique quantitative est de savoir comment le génotype détermine le phénotype [76]. Une étude de la variation de la transcription n'est qu'une composante de cette question. La détermination de la relation entre la variation biologique n'est pas seulement reflétée au niveau de la transcription, mais dépend aussi

d'autres variables. D'autres limites sont de nature plus technique, et incluent : 1) la précision de la mesure des profils de transcription (bien que l'erreur peut être réduite par l'utilisation d'un grand nombre de réplicats biologiques) ; 2) dans le cas d'utilisation de puces à ADN, le fait que les gènes analysés se limitent à ceux qui sont représentés sur la puce (bien que les nouvelles techniques de RNA-Seq permettent de pallier à cet inconvénient) ; 3) le fait que la mesure d'expression n'est valide que pour un tissu particulier prélevé à un moment particulier.

De plus, une association entre un variant génomique et l'expression d'un gène repose principalement (pour les cis-eQTLs) sur le principe que le variant peut altérer le niveau de transcription du gène. Cependant, de nombreux événements post-transcriptionnels* peuvent jouer des rôles importants dans la modulation de l'activité de produits de gènes. Ces rôles incluent 1) les variations dans la structure de la protéine traduite, 2) la stabilité du transcrit, 3) le transport des ARN traduits vers le cytoplasme, et 4) les évènements d'épissage alternatif*. Sans même compter les erreurs possibles des mesures elles-mêmes, l'expression de gènes peut être influencée par des facteurs non génétiques qui ne sont pas détectés par une puce à ADN, incluant les variations environnementales, les modifications épigénétiques* et les fluctuations aléatoires dans les niveaux d'expression. Néanmoins, il a été montré qu'il existe une forte composante d'héritabilité pour la majorité des gènes.

1.14 Les réseaux de gènes

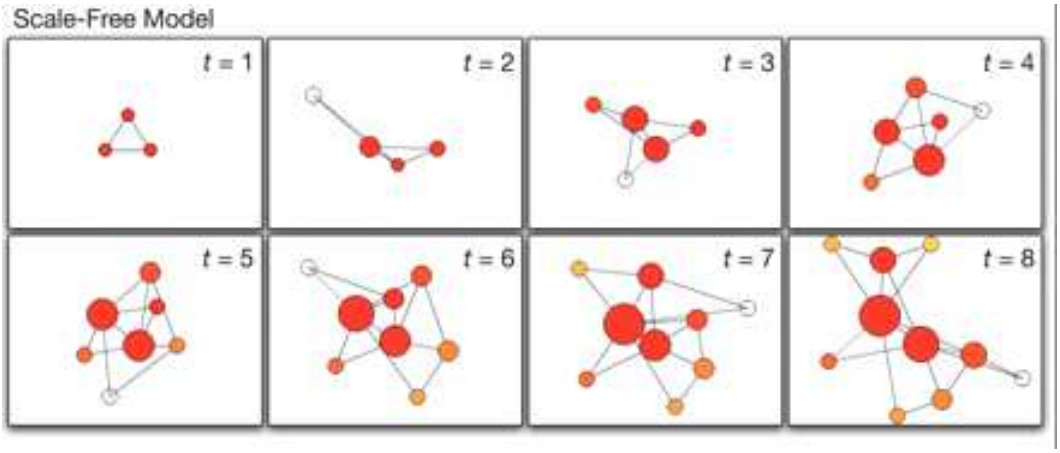
Les réseaux sont une technique d'analyse utilisée dans de nombreux domaines, par exemple pour analyser les interactions entre groupes sociaux, en communication internet ou entre effecteurs biologiques. Les techniques d'analyse des réseaux permettent de mieux comprendre la structure et la dynamique d'interactions complexes. En biologie, certains des réseaux étudiés concernent par exemple les cellules, les protéines ou les gènes.

Dans le contexte des traits complexes, l'approche de réseau de gènes est intéressante, car elle permet de comprendre comment un groupe de gènes peut être régulé dans

le contexte d'un réseau biologique intégré [77, 78]. Bien que ces phénomènes soient encore mal définis, il existe des évidences suggérant l'existence de réseaux modulaires dans lesquels les gènes, les protéines, les métabolites et autres facteurs peuvent opérer en groupes plutôt que de façon isolée. Par exemple, certains groupes de gènes peuvent être régulés de façon coordonnée par des facteurs de transcription, des microARN, des changements dans la méthylation de l'ADN et/ou des phénomènes de remodelage de la chromatine. Ces types de mécanismes correspondent à des situations où des gènes sont organisés au sein de modules de co-expression. Suite au travail de pionnier de Eisen *et al.* [79], un nombre croissant d'études ont utilisé des données d'expression des gènes pour construire des réseaux de régulation fondés sur la co-expression de gènes [80–82]. Ces études ont permis de modéliser des réseaux d'expression de gènes. Ces résultats peuvent, entre autres, être représentés de façon graphique comme un ensemble de noeuds (gènes) reliés par des arêtes (relations entre les gènes). Par rapport aux réseaux de régulation, un réseau de co-expression ne tente pas de distinguer les interactions directes des interactions indirectes entre les gènes. De plus, un réseau de co-expression contient des informations entre gènes voisins qui sont généralement négligées dans des analyses de clustering. [83]. Un certain nombre d'études ont analysé les propriétés topologiques des réseaux de gènes [84, 85] et ont montré que les réseaux de modules co-exprimés ont des propriétés intéressantes. Une de celles-ci correspond au fait que leur architecture est considérée comme étant «scale free». Ceci veut dire que le réseau présente plusieurs connexions parallèles, de sorte que le nombre de connexions de chaque noeud se distribue selon une loi de puissance [86, 87]. La figure 1.9 montre la différence entre un réseau «scale free» par rapport à un réseau aléatoire («random network»).

Une architecture de type «scale-free» a l'avantage fonctionnel suivant : lorsqu'un lien entre deux noeuds du réseau est détruit, il existe une route alternative pour connecter les deux noeuds. Ceci est une propriété importante pour un système biologique qui a besoin d'être robuste, adaptable et efficace, pour survivre à des changements constants[89].

(a)



(b)

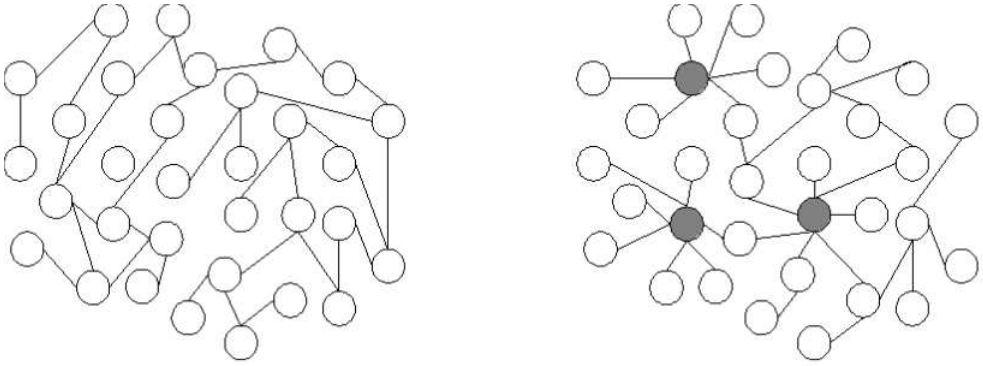


Figure 1.9 – a) Représentation de la naissance d'un réseau «scale free». À partir de trois noeuds connectés (en haut à gauche), un nouveau noeud (représenté par un cercle vide) est ajouté au réseau (en haut). Lors du choix de lier, les nouveaux noeuds préfèrent se fixer aux noeuds plus connectés, un processus connu sous le nom d' «attachement préférentiel». Grâce au processus de croissance et d'attachement préférentiel, un processus «les riches deviennent plus riches» est observé, ce qui signifie que les noeuds fortement connectés acquièrent plus de liens que ceux qui sont moins liés, conduisant à l'émergence naturelle de quelques «hubs» fortement connectés. La taille du noeud est proportionnelle au degré de connexion du noeud. La distribution des degrés de réseau résultant suit la loi de puissance. Image : [86]. b) Représentation graphique de la différence entre un réseau aléatoire (à la gauche) et un réseau «scale free»(à la droite). Image : [88].

1.14.1 Construction d'un réseau de co-expression

La construction d'un réseau de co-expression repose sur un principe simple : si plusieurs gènes appartiennent à un même module, ils auront tendance à être régulés par les mêmes facteurs et leurs niveaux d'expression auront tendance à se ressembler. La première étape nécessite la mesure de l'expression de gènes dans une population génétiquement diverse avec des puces à ADN ou par séquençage à haut débit de l'ARN (RNA-Seq). Pour analyser et structurer les données, il existe plusieurs méthodes analytiques. Une des mieux développées et des plus populaires est la méthode WGCNA [83, 90]. La quantité naturelle de variation dans l'expression des gènes entre individus est utilisée pour analyser comment l'expression d'un gène corrèle avec l'expression des autres gènes dans la population. Le résultat correspond à une matrice contenant toutes les valeurs de corrélation de Pearson entre chaque paire de gènes. Cette matrice est ensuite transformée pour générer «l'adjacence», qui représente une mesure de la force de connexion entre gènes.

Une fois le réseau construit, un point critique est de définir les modules du réseau. La mesure d'adjacence est transformée en une mesure plus robuste appelée «topological overlap»(voir Figure 1.10). Cette mesure prend en compte les connexions entre les gènes voisins de chaque paire de gènes du réseau. La connectivité d'un gène peut être définie par la somme de toutes les valeurs d'adjacence avec les autres gènes. Au sein des modules, les gènes «hub» sont définis comme ceux qui sont les plus connectés au sein de chaque module. Comme chaque module peut contenir entre plusieurs dizaines et plusieurs centaines de gènes, on peut réduire la dimension des données en agrégeant le comportement de ces gènes. Ceci peut être fait par une méthode de PCA, qui résume l'expression de tous les gènes dans le module par celle d'un gène fictif représentatif, appelé «eigengene» *.

Au-delà de l'identification de modules, une question importante concerne leur pertinence et/ou importance biologique. Cet aspect de l'investigation n'a pas encore été établi de façon aussi solide que les méthodes de construction des réseaux. Plusieurs stratégies ont été utilisées par les investigateurs. Par exemple, il est possible de faire des

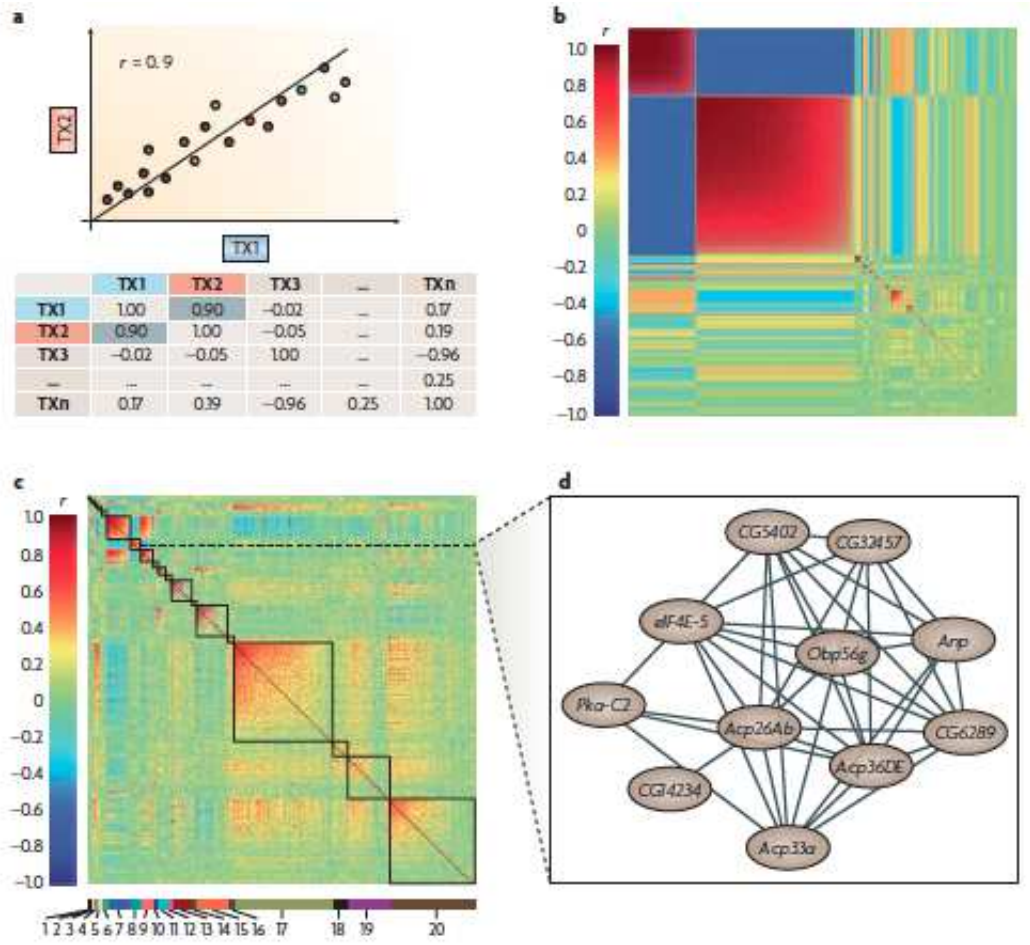


Figure 1.10 – Méthodologie pour générer des réseaux de co-expression à partir de profils d'expression de gènes. Les réseaux de co-expression dépendent de la collection de profils d'expression génique à partir d'une population génétiquement diverse. Dans l'ensemble, les profils de gènes démontrent des similitudes dans les profils d'expression en raison de la corégulation transcriptionnelle. a) La première étape dans la construction de réseaux de co-expression de gènes consiste à calculer une matrice de corrélation entre chaque paire de transcripts. Ici, une représentation d'une telle matrice qui montre la corrélation entre les abondances des deux transcrits, soit les gènes TX1 et TX2. b) Les relations de la co-expression entre les gènes peuvent être quantifiées à l'aide des coefficients de corrélation. c) Les ensembles de gènes corrélés sont groupés en utilisant des algorithmes standards pour identifier les «modules» de gènes co-exprimés. Ici, les modules sont représentés dans les boîtes noires. Les réseaux de co-expression peuvent être visualisés d'un certain nombre de façons, comme un «heatmap» dans lequel les corrélations entre les niveaux de transcription sont indiquées par ombrage plus foncé. d) Les réseaux peuvent aussi être représentés dans un espace multidimensionnel constitué de noeuds (gènes) et d'arêtes (la force de la corrélation). Ici on voit un example d'un module. Image :[91].

tests d'enrichissement pour tester si les gènes de modules identifiés sont enrichis significativement pour des gènes ayant des fonctions particulières (par exemple l'apoptose, le métabolisme, le cycle cellulaire). Ces tests se font à l'aide des annotations de gènes dans la banque de données «Gene Ontology (GO)», où chaque gène est classé de façon hiérarchique dans différentes classes en fonction de leurs termes d'annotation («GO Terms»). Ce test permet de déterminer si les annotations des gènes de modules présentent des enrichissements pour certaines fonctions biologiques. On peut aussi tester si les gènes des modules sont exprimés de façon préférentielle pour certains types cellulaires spécialisés. Certains ont fait des études de simulation pour tenter de valider les modules [92]. Malgré l'utilité de ces tests, il est nécessaire de compléter l'analyse biostatistique par des expériences de validation biologique.

Une autre utilité de l'identification de modules de co-expression est d'intégrer ces données avec des données génétiques et phénotypiques. De façon semblable à une étude de eQTL, on peut tester si les valeurs d'expression des «eigengenes» (qui représentent les valeurs d'expression de tous les gènes d'un module) peuvent être liées ou associées à un phénotype. Ainsi, certaines études ont récemment intégré des analyses de réseaux avec des études plus classiques de eQTLs, ce qui a fourni des informations permettant de mieux comprendre et d'expliquer certains des caractères complexes. Par exemple, une étude a utilisé les réseaux de co-expression pour trouver un «module-QTL» lié au poids des souris [93]. Ces approches consistent à construire un réseau de gènes pour ensuite identifier des eQTLs liés au gène du réseau [94, 95].

Comme toute technologie, les études de réseaux biologiques posent des défis particuliers (par exemple la grande dimension des jeux de données à analyser) et ont des limites techniques, incluant par exemple le bruit inhérent à la collecte de données et l'introduction de biais expérimentaux par la méthode d'échantillonnage.

1.15 Objectifs de la thèse

Tel que décrit ci-dessus, on peut distinguer d'un point de vue génétique deux grandes classes de phénotypes : 1) les traits simples à transmission mendélienne (qui résultent

de mutations très pénétrantes) et 2) les traits quantitatifs complexes, qui sont considérés comme étant le résultat d'interactions entre diverses variantes génétiques et des facteurs environnementaux. Les loci génétiques liés à un caractère complexe sont appelés «loci de caractères quantitatifs» (QTL). Les niveaux d'expression des gènes dans les tissus peuvent également être étudiés comme des caractères quantitatifs complexes. En analysant les niveaux d'expression de milliers de gènes par puces à ADN ou par séquençage haut débit de l'ARN (RNAseq), des études récentes ont montré qu'une part importante de la variabilité de l'expression des gènes dans les populations est due à des variations de séquences d'ADN. Un QTL lié à un niveau d'expression d'un gène dans un tissu est appelé une «expression QTL» (eQTL). Lorsque l'expression d'un gène est associée avec un polymorphisme génétique à proximité du locus génique, l'eQTL correspondant est identifié comme un cis-eQTL. Dans d'autres cas, lorsque le niveau d'expression d'un gène est associé avec un locus clairement distinct de celui du gène, il est défini comme un trans-eQTL. Il est généralement admis que les outils informatiques développés à jusqu'à maintenant ont plus de difficulté à détecter les trans-eQTLs que les cis-eQTLs.

Il y a plusieurs défis pour améliorer la compréhension de caractères quantitatifs complexes. La plupart des succès de la génétique concernent les maladies mendéliennes simples, car elles résultent de mutations fortement pénétrantes qui, souvent, modifient la structure et/ou la fonction des protéines codées. La grande pénétrance de ces mutations fait en sorte qu'il est possible d'identifier les variations génétiques causales dans des familles multi-générationnelles comprenant des individus affectés et non affectés. Cependant, les maladies les plus courantes sont définies comme des caractères quantitatifs complexes, où il est beaucoup plus difficile de détecter les effets de polymorphismes génétiques sur les phénotypes d'intérêt. Ces difficultés sont illustrées par les résultats des nombreuses études d'association pangénomique (GWAS) effectuées récemment. Tout d'abord, les GWAS les plus récentes ont révélé que seulement 11% [96] des associations trouvées pour les maladies les plus courantes concernent de «régions codantes». Ceci indique que des variations génétiques situées en dehors des régions codantes, y compris des éléments de régulation, peuvent constituer d'importants éléments de causalité pour les caractères complexes. Cependant, la nature exacte des éléments fonction-

nels dans ces «déserts géniques» est souvent inconnue. Il existe donc de nombreuses situations où, malgré la détection de signaux positifs dans une étude GWAS, on ne comprend pas les mécanismes par lequel le polymorphisme génétique affecte le trait étudié. Même dans les cas où les signaux correspondent à des gènes bien identifiés, la plupart des variants génétiques détectés à ce jour n'ont que de modestes contributions au trait étudié. Lorsque toutes les variantes génétiques identifiées sont considérées collectivement, dans la plupart des cas, ils n'expliquent collectivement qu'une très petite partie de l'héritabilité globale. Ces problèmes expérimentaux suggèrent que d'autres approches expérimentales pourraient être envisagées.

Il est communément admis que les traits complexes ont des causes complexes polygéniques. Néanmoins, de nombreux efforts ayant pour but d'élucider les déterminants génétiques de caractères complexes visent encore dans la plupart des cas à identifier une mutation causale dans un seul gène. Un des buts de cette thèse a été de développer une approche plus globale pour : 1) tenir compte des multiples interactions possibles entre gènes pour la détection de eQTLs ; 2) considérer comment des polymorphismes affectant l'expression de plusieurs gènes au sein de groupes de co-expression pourraient contribuer à des caractères quantitatifs complexes.

1.15.1 Objectifs spécifiques

1. Développer un outil informatique utilisant des méthodes d'analyse multivariées pour détecter des eQTLs dans des croisements génétiques (chapitres 2 et 3).

2. Analyser des données d'expression de gènes dans des tissus de souris RIS pour déterminer comment des polymorphismes peuvent affecter l'expression de plusieurs gènes au sein de domaines géniques de co-expression (chapitre 4).

3. Combiner des études de détection de eQTLs avec des techniques d'analyse de réseaux de co-expression de gènes pour élucider les déterminants géniques d'un trait complexe particulier, la MVG (chapitre 5).

1.15.2 Organisation des chapitres

Chapitre 1 : Le premier chapitre est une revue de la bibliographie permettant à un lecteur non-spécialiste d'avoir une vue d'ensemble sur les concepts et méthodes utilisées dans cette thèse.

Chapitre 2-chapitre 5 : Les principaux résultats de la thèse présentés sous forme d'articles pour revues scientifiques.

Chapitre 6 : Un chapitre de résultats préliminaires où je présenterai les résultats de l'analyse avec la méthode iBMQ avec les données de l'expression de tissu cardiaque de souris AXB / BXA pour étudier les «hotspots» de trans-eQTLs.

Chapitre 7 : Ce chapitre présentera une discussion sur l'apport de ma recherche à l'étude des traits complexes.

CHAPITRE 2

AN INTEGRATED BAYESIAN HIERARCHICAL MODEL FOR MULTIVARIATE EQTL MAPPING

L'un des objectifs centraux de la génétique/génomique est de mieux comprendre les mécanismes expliquant les variations naturelles de l'expression de gènes. Récemment, en considérant les niveaux d'expression de milliers de gènes comme des traits quantitatifs, il est devenu possible de détecter des «QTLs d'expression» (eQTL). La variation de l'expression génique entre les individus est un déterminant important de la variation phénotypique et de la susceptibilité aux maladies complexes. Présentement, la majorité des analyses d'eQTL sont effectuées avec les mêmes outils que les analyses QTL standard qui étudient seulement quelques phénotypes à la fois. Ainsi, pour analyser un transcriptome qui contient des milliers d'expressions de gènes, on répète l'analyse pour chaque gène indépendamment (l'analyse univariée). Cependant, l'analyse univariée n'intègre pas les interactions entre divers gènes, ce qui ne correspond pas à la réalité. Notre but a été de développer une nouvelle méthode multivariée qui pourra améliorer la détection de eQTLs en empruntant de l'information entre les gènes.

L'article présenté dans ce chapitre présente une méthode bayésienne multivariée pour l'analyse simultanée de plusieurs gènes. En statistique, effectuer une régression multiple sur des milliers de gènes est une opération très complexe. Nous avons développé le modèle suivant : pour chaque combinaison de gènes et de marqueurs génétiques, notre modèle tente d'estimer le paramètre d'indicateur d'inclusion. Le paramètre d'indicateur d'inclusion nous indique si le gène g est associé au SNP j. L'indicateur d'inclusion dépend de la variable de poids. Chaque SNP a une variable de poids et cette variable va prendre une valeur plus grande si plusieurs gènes sont associés à ce SNP pour favoriser la détection de ce SNP. Cette variable de poids représente la différence entre notre modèle et les modèles multivariés précédents.

En informatique, pour valider un modèle, il est nécessaire de faire des simulations. Ainsi on crée artificiellement des données dont on a les résultats et on étudie comment

notre modèle se comporte. Nous avons fait plusieurs simulations et comparaisons avec d'autres méthodes. Les résultats montrent que notre méthode semble mieux détecter les «hotspots» de trans-eQTLs. Biologiquement, ces «hotspots» de trans-eQTLs représentent des régions qui contrôlent l'expression de plusieurs gènes. Le matériel supplémentaire de l'article se trouve à l'annexe II.

Contribution des auteurs à la préparation de l'article :
AT a écrit les équations du modèle. GI a codé l'implémentation du modèle (version parallèle) en langage C. MPSB a fait les analyses de simulations, les comparaisons avec les autres méthodes et l'analyse avec les vraies données. MPSB a écrit sous la supervision de CFD, AL et RG.

Nota bene : L'article suivant a été publié dans le journal *Statistical Applications in Genetics and Molecular Biology*.

MP SCOTT-BOYER, A. TAYEB, G. IMHOLTE, A. LABBE, C.F. DESCHEPPER AND R. GOTTARDO. (2012) An integrated Bayesian hierarchical model for multivariate eQTL mapping. Statistical Applications in Genetics and Molecular Biology, 11(4)

An integrated Bayesian hierarchical model for multivariate eQTL mapping

Marie Pier Scott-Boyer[1], Gregory C Imholte[2], Arafat Tayeb[1], Aurelie Labbe[3], Christian F Deschepper[1], Raphael Gottardo[2]

1. Institut de recherches cliniques de Montréal (IRCM) and Université de Montréal

2. Fred Hutchinson Cancer Research Center

3. University McGill

2.1 Abstract

Recently, expression quantitative loci (eQTL) mapping studies, where expression levels of thousands of genes are viewed as quantitative traits, have been used to provide greater insight into the biology of gene regulation. Originally, eQTLs were detected by applying standard QTL detection tools (using a "one gene at-a-time" approach), but this method ignores many possible interactions between genes. Several other methods have proposed to overcome these limitations, but each of them has some specific disadvantages. In this paper, we present an integrated hierarchical Bayesian model that jointly models all genes and SNPs to detect eQTLs. We propose a model (named iBMQ) that is specifically designed to handle a large number G of gene expressions, a large number S of regressors (genetic markers) and a small number n of individuals in what we call a "large G, large S, small n" paradigm. This method incorporates genotypic and gene expression data into a single model while 1) specifically coping with the high dimensionality of eQTL data (large number of genes), 2) borrowing strength from all gene expression data for the mapping procedures, and 3) controlling the number of false positives to a desirable level. To validate our model, we have performed simulation studies and showed that it outperforms other popular methods for eQTL detection, including QTLBIM, R-QTL, remMap and M-SPLS. Finally, we used our model to analyze a real expression dataset obtained in a panel of mice BXD Recombinant Inbred (RI) strains. Analysis of these data with iBMQ revealed the presence of multiple hotspots showing

significant enrichment in genes belonging to one or more annotation categories.

Keywords: Bayesian multiple regression; eQTL mapping; Markov chain Monte Carlo; multiple testing; sparse modelling; variable selection

2.2 Introduction

”Complex quantitative traits” are typically defined as characteristics that depend in part on inherited factors, but whose magnitude results from interactions between a great number of genes and environmental factors. Originally, investigators studying such traits focused mostly on physical characteristics and/or physiologic responses, and aimed at locating quantitative trait loci (QTL), *i.e.* genomic locations that had an influence on the manifested trait. More recently, since expression levels of genes within tissues can themselves be considered as quantitative traits, several studies have identified so-called “expression quantitative trait loci” (eQTL). The identification of eQTLs has provided greater insights into the biology of gene regulation and/or complex traits [64, 65, 97]. By using DNA microarrays, it has now become feasible to map eQTLs for basically all genes in the genome.

When an eQTL locus corresponds to that of the gene whose transcript abundance is measured, it is identified as a ”cis-acting eQTL” (cis-eQTL), meaning that a genetic variation in the neighborhood of the gene is associated with the differential abundance of its transcript. Equally interesting and abundant are the trans-eQTLs that map to locations distant from the gene region. Many studies have reported strong clustering of trans-eQTLs (*i.e.* multiple genes associated with the same loci) into so-called eQTL hotspots [98, 99], which suggests that these genomic regions harbor polymorphisms that shape the dynamic and global nature of transcriptional regulation.

Since eQTL studies differ from standard QTL studies only in the number of phenotypes, it is not surprising that mostly classical QTL methods have been used to identify eQTLs, one gene at a time. However, this “one gene at-a-time” approach ignores the many important combinatorial effects and interactions between genes. Moreover, the multiplicity problem is such that it is not uncommon to have to perform well over a

million tests, and univariate methods do not deal appropriately with the problem of multiple testing across markers and genes. Over the years, several strategies have emerged in order to address the multiple issues raised by the high dimensionality of the data at both the trait level (thousands of gene expressions) and the genotype level (thousands of SNPs). For instance, [70] proposed using a Sparse Partial Least Square (SPLS) regression technique to account for the high dimension and co-linearity of the genotype data. Dependence among gene expressions is accounted for by clustering the genes according to their expression profile and then applying the SPLS regression at the cluster level. While very appealing, this method has the drawback of identifying markers associated with a "meta-transcript" instead of individual transcripts. Alternatively, [100] proposed a Mixture Over Marker (MOM) modeling technique to facilitate information sharing across both markers and transcripts through an empirical Bayes strategy. Although this method identifies transcripts that map to at least one marker, it has the main disadvantage of identifying at most one eQTL per transcript.

Bayesian models have been widely used to solve the extreme multiplicity problem of eQTL studies. By borrowing information across genes and/or markers, they provide efficient ways to overcome the computational burden imposed by the great number of tests required to analyze one gene/one marker at a time. Several approaches based on Sparse Bayesian Regression (SBR) modeling have been developed specifically for QTL studies. For example, the method of [71], as implemented for eQTL studies in the R-QTLBIM (QTL-Bayesian Interval mapping) package, proceeds by analyzing all SNPs simultaneously but all genes independently. This method was further extended by [72] to handle several traits (genes) simultaneously but is limited in practice to a maximum of five traits, due to computational issues. This approach was generalized to continuous and categorical traits by [74] and implemented in the BAYES software package. As in the work of [72], this implementation also suffers from the same computational downside and cannot be applied to a large number of traits (*i.e.* gene expression profiles) as in a typical eQTL studies. [73] introduced an efficient evolutionary stochastic search algorithms for variable selection and used it to detect eQTLs across multiple tissues, but their approach models each gene separately and no information is explicitly

shared across genes. Recently, [101] proposed a Variational Bayes approach (VBQTL) to jointly model the contribution from genotypes as well as known and hidden confounding factors in a unified Bayesian framework. Even though VBQTL models all genes concurrently, the prior probability of association is assumed to be common across all genes and markers, which is unrealistic for such data. In addition, as in the MOM model [100], the authors constrain each gene to have at most one relevant SNP regulator for computational reasons.

All the Bayesian models described above assume a common prior distribution for the probability of inclusion of a marker in the sparse regression model, across all genes. As we shall see in the simulations studies, this leads to an over-detection of common eQTLs and thus a high number of false positive hotspots. In this paper, we present an integrated Bayesian hierarchical Model for eQTL mapping (iBMQ) that incorporates genotypic and gene expression data (and possibly thousands of SNPs and genes) into a single model while resolving all the issues mentioned above. Specifically, our model is built around flexible prior distributions and is designed to 1) cope with the high dimensionality of eQTL data (large number of genes), 2) borrow strength from all gene expression data for the mapping procedures, and 3) control the number of false positives to a desirable level. Note that the model developed by [102] was developed with similar objectives but for the detection of common eQTLs across tissues. In this slightly different context, the authors had to assume a more restrictive structure on the prior distribution for the probability of inclusion of a marker in the model.

This paper is organized as follows: Section 2.3 introduces our integrated hierarchical Bayesian Model for Multivariate eQTL Mapping (iBMQ) and specifies the different parameters and their priors. In Section 2.4, we evaluate our model using a series of simulation studies and compare its performance to several other previously developed methods: 1) R-QTL [103], 2) QTLBIM [71]; 3) M-SPLS [70]; 4) remMap [104]; and 5) a simplified version of our full iBMQ model that uses a common prior distribution for the probability of inclusion of a marker, *i.e.* iBMQ with common weight (iBMQ-cw). The latter is in fact similar to the BAYES model [74] and VBQTL [101] which both rely on a common prior probability of association. In Section 2.5, we apply our model to

analyze a set of gene expression data obtained in whole eyes from a panel of 68 mice BXD Recombinant Inbred (RI) strains. We conclude in Section 2.6 with a discussion on possible future improvements of the model.

2.3 Model

In this section, we present our integrated hierarchical Bayesian Model used to detect eQTLs (iBMQ) and the full conditional distributions used to perform posterior exploration via Markov chain Monte Carlo (MCMC). In the current application and following examples, individuals are in fact RI strains (where particular combinations of parental alleles have been fixed within strains by extensive inbreeding) and genetic markers are Single Nucleotide Polymorphisms (SNPs).

2.3.1 Model Definition

We model gene expression measurements across individuals as follows,

$$y_{ig} = \mu_g + \sum_{j=1}^{S} x_{ij}\gamma_{jg}\beta_{jg} + \varepsilon_{ig}, \tag{2.1}$$

where

- $g = 1, \cdots, G$ denotes a particular gene or a trait, $i = 1, \cdots, n$ denotes a particular strain or individual and $j = 1, \cdots, S$ denotes a particular SNP;
- y_{ig} is the expression level of gene g for the individual strain i;
- μ_g is the overall mean expression level of gene g (across all strains);
- x_{ij} represents the genotype at locus j for strain i under an additive, dominant or recessive genetic model;
- β_{jg} is the effect size of SNP j on gene g. In practice, only a few markers directly affect the phenotype and thus many of the β's should be exactly zero. In order

to capture the "sparsity" of the model, we need to incorporate indicator variables, γ_{jg}, specifying which marker should be included in the model.

- γ_{jg} is a binary inclusion indicator, i.e $\gamma_{jg} = 1$ if SNP j is included in the model for gene g and $\gamma_{jg} = 0$ otherwise;

- ε_{ig} is an error term assumed to be Gaussian with gene specific variance σ_g^2.

In eQTL studies, we have thousands of gene expression profiles as quantitative phenotypes, and analysis of such data typically requires performing univariate QTL analysis for each gene expression profile.The model we propose is motivated by two key factors: 1) most eQTLs affect more than one expression profile, with some affecting many genes; 2) genes in the same pathway are more likely to be under the influence of common regulators (*i.e.* their expressions are correlated). As a result, there is an opportunity to share information across the hundreds or thousands of gene expression traits in such a way that more informative conclusions can be drawn. This can be done by allowing for a gene/marker-specific probability of QTL, $w_{jg} = \mathbb{P}(\gamma_{jg} = 1)$ a priori, and borrow strength across genes to estimate this probability via flexible genome-wide prior distributions; see Figure 2.1 for a graphical representation. Such a hierarchical structure encourages eQTLs to be associated with more than one gene. The rationale is that true eQTLs are probably associated with more than one transcript, while eQTLs that are associated with a single gene are possibly due to noise and should be down weighted, but not necessarily eliminated.

In the proposed model, we assume that the ε_{ig}'s are independent and identically distributed (*iid*), so that genes are conditionally independent given all model parameters. The gene dependence is introduced via an exchangeable prior on the γ_{gj}'s, thus providing a computationally tractable model with a suitable dependence modeling framework. As we will see in Section 2.4.1, our approach performs well even in the presence of between gene correlations from non-genetic sources.

2.3.2 Prior Distributions

In the following, we describe the different prior distributions of the model (2.1). These priors should reflect our *a priori* knowledge and uncertainty about the model parameters, namely $\theta = (\mu_g, \sigma_g^2, \gamma_{jg}, \beta_{jg}, \omega_{jg}, p_j, a_j, b_j)$. Our priors are defined as follows,

- $\gamma_{jg} \sim \mathscr{B}ernoulli(\omega_{jg})$, where $\mathbb{P}(\gamma_{jg} = 1) = \omega_{jg}$ is an unknown parameter that represents the inclusion probability of SNP j in the model for gene g. In order to reduce the false discovery rate, and since only a small numbers of SNPs act as a determinant of a gene expression, we let the inclusion probability parameters ω_{jg} take the value 0 a priori *most* of the time. When ω_{jg} is not 0, it is assumed to come from a Beta distribution $\mathscr{B}eta(a_j, b_j)$. This can be expressed as a mixture of a Dirac mass at 0 and a Beta distribution with weights p_j and $1 - p_j$ as follow

$$\omega_{jg} \sim p_j \delta_0(\omega_{jg}) + (1 - p_j)\mathscr{B}eta(a_j, b_j)(\omega_{jg}).$$

  The parameter p_j (the probability that ω_{jg} is 0) is identical for all genes. This helps in detecting a stronger signal when a SNP is weakly associated to many gene expressions (Lucas et al. [105]). Furthermore, we use a common conjugate Beta prior for p_j with hyperparameters a_0 and b_0:

$$p_j \sim \mathscr{B}eta(a_0, b_0).$$

  Additionally, a_j and b_j are assumed to follow Exponential distributions with hyperparameters λ_a and λ_b:

$$a_j \sim \mathscr{E}xp(\lambda_a) \text{ and } b_j \sim \mathscr{E}xp(\lambda_b).$$

- $\mu_g \sim \mathscr{N}(m_g, \tau_g^2)$, where m_g and τ_g are the empirical mean and variance of gene expression g
- $\beta_{jg} = 0$ if $\gamma_{jg} = 0$ and $\beta_{jg} \sim \mathscr{N}\left(0, v_{jg}^2\right)$ if $\gamma_{jg} = 1$, with $v_{jg}^2 = c\left(x_j^T x_j\right)^{-1} \sigma_g^2$,

where c is a scaling factor parameter and $\left(x_j^T x_j\right)^{-1} = \left(\sum_{i=1}^{n} x_{ij}^2\right)^{-1}$ mimics the regressor variance, which leads to the well-known g-prior of Zellner and Siow [106]. Here we follow the approach of Yi et al. [71] and consider c to be a constant equal to S the number of SNPs. Bottolo et al. [107] considered an Inverse-Gamma prior $c \sim \mathscr{I}n\mathscr{G}a(\frac{1}{2}, \frac{n}{2})$ based on the Zellner and Siow [106] prior. Recently Petretto et al. [73] considered a common c for all genes with the prior of Liang et al. [108] $c \sim \frac{1}{1+c}$ in the interval $(0, M)$, where the end point M is $M = \max(n, S^2)$. The term σ_g^2, the overall variance of v_{jg}^2, ensures that the parameter σ_g^2 is a nuisance parameter in the model and can be integrated out.

- $\sigma_g^2 \sim \mathscr{I}n\mathscr{G}a(\frac{1}{2}, \frac{1}{2})$ is a vague prior on the error variances.

A graphical representation summarizing our model and its prior specifications is shown in Figure 2.1. Our model has two clear advantages over alternatives. First, it treats a large number of genes at a time, which effectively facilitates the detection of common eQTLs *hotspots* that otherwise could not be detected for genes with weak signals if they were analyzed one at a time. The second advantage is that each gene expression/trait has its own inclusion indicator γ_{jg} at each SNP. In previously published work, the inclusion probability parameters ω_{jg} were either (*i*) considered identical for all SNP positions ($\omega_{jg} = \omega$), with the common ω being considered either given or following a Beta prior distribution [109, 110]; or (*ii*) supposed identical for all genes but depending on the SNP positions ($\omega_{jg} = \omega_j$), with each ω_j following a Beta prior distribution [73, 107]. As we will see in the simulations studies, such assumptions can have a big impact on the performance of the model.

2.3.3 Parameter Estimation

Realizations were generated from the posterior distribution via MCMC algorithms [111]. All updates were done via Gibbs sampling except for a_j and b_j for which no closed form full conditionals are available, and were thus updated via adaptive rejection sampling [112]. All full conditionals are given in Appendix A. We used the method of

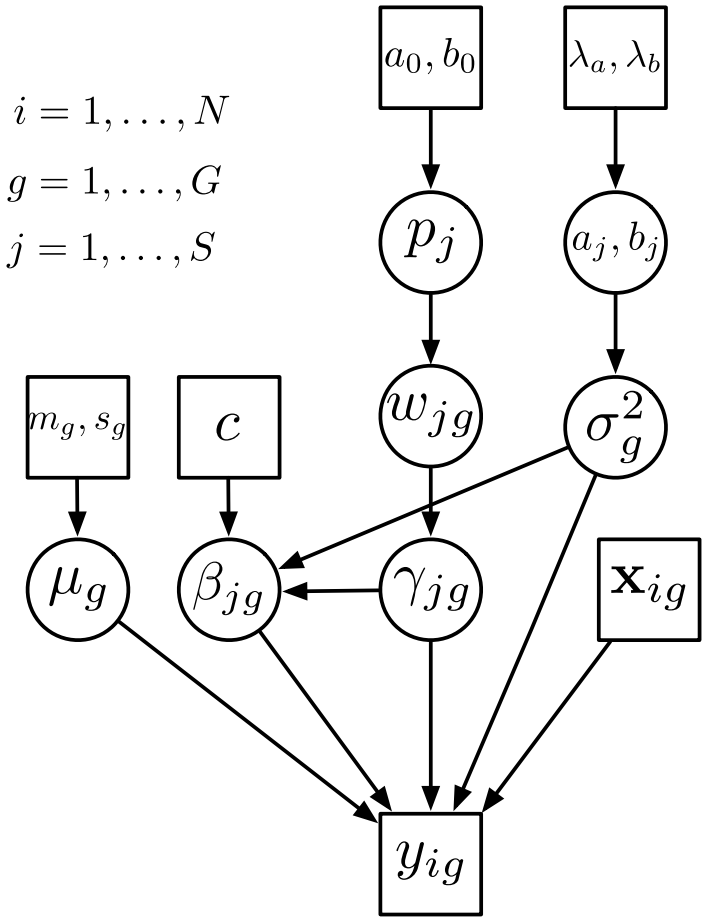


Figure 2.1: Graphical representation of the eQTL model.The rectangles represent either fixed hyperparameters or the data, circles represent unknown (and random) quantities.For each gene, the gene expression phenotype $\mathbf{y}_g$ is expressed as a linear model $y_{ig} = \mu_g + \sum_{j=1}^{S} x_{ij}\gamma_{jg}\beta_{jg} + \varepsilon_{ig}$. The gene/marker specific regression coefficient β_{jg} is assumed to be normally distributed with distribution $\beta_{jg} \sim \mathcal{N}\left(0, c\gamma_{jg}(x_j^T x_j)^{-1}\sigma_g^2\right)$. The prior for μ_g is $\mathcal{N}(m_g, s_g^2)$, and the prior for σ^2 is $\pi(\sigma^2) \propto 1/\sigma^2$. The prior distribution for γ_{jg} is assumed to be Bernoulli with parameter w_{jg}. The w_{jg}'s are given $\omega_{jg} \sim p_j\delta_0(\omega_{jg}) + (1-p_j)\mathscr{B}eta(a_j, b_j)(\omega_{jg})$. Additionally, a_j and b_j are assumed to follow Exponential distributions with hyperparameters λ_a and λ_b and $p_j \sim \mathscr{B}eta(a_0, b_0)$.

Raftery and Lewis [113] to determine the number of iterations, based on a short pilot run of the sampler. For each dataset presented here, this suggested that a sample of no more than about 1,000,000 iterations with 50,000 burn-in iterations was sufficient

to estimate standard posterior quantities. Guided by this, and leaving some margin, we used 2,000,000 iterations after 50,000 burn-ins for each dataset explored here. Results from the diagnostics test and trace plots are presented in Appendix II. Finally, our model depends on four hyperparameters a_0, b_0, λ_a and λ_b that need to be fixed in advance. We can choose these values *a priori* using the expected number of e-QTLs $\mathbb{E}(n_\gamma)$ and its dispersion $\mathbb{V}(n_\gamma)$, as detailed in Appendix B. Using this approach we have chosen $a_0 = \lambda_a = 10$ and $\lambda_b = b_0 = .1$, which favors models with fewer eQTLs.

2.3.4 Inference and Detection of eQTLs

Our ultimate goal is to identify gene/SNP associations, and this can be done using parameter estimates from our model. An eQTL for gene g at SNP j is declared significant if its corresponding marginal posterior probability of association (PPA), *i.e.* $\Pr(\gamma_{jg} = 1|y)$, is greater than a given threshold. In the context of multiple testing and discoveries, a popular approach is to use a common threshold leading to a desired false discovery rate (FDR). In the Bayesian paradigm, derivation of the PPA threshold is trivial and can be calculated using a direct posterior probability calculation as described in [114].

2.4 Simulation Study

We performed two sets of simulation studies: a validation study and a comparison study. The goal of the validation study was to investigate the effects of different factors such as the correlation between SNPs/genes, the effect size and the spatial structure of the true eQTLs with regards to our model's performance. In the comparison study, we compared the performance of our proposed model with other standard methods.

2.4.1 Validation Study

For the first experimental set, we used $n = 100$ individuals, $G = 40$ genes, $S = 1000$ SNPs and $\sigma_g^2 = 0.1$; the latter two values being taken from the experimental data (section 2.5). We also considered two types of correlation structures between SNPs and genes. First, SNPs were considered as either independent or dependent. In the latter case, SNPs

were divided into 100 blocks of 10 SNPs each. We assumed that blocks were independent and imposed a correlation $\rho_x = 0.4$ among the SNPs within each block. Second, genes were also considered as either independent or dependent (*i.e.* correlated, meaning that they show some level of co-expression due to non-genetic causes). In the latter case, genes were divided into 4 independent blocks of 10 consecutive genes each and we chose a correlation $\rho_\varepsilon = 0.5$ within each block. In addition, we simulated two different scenarios for eQTLs positions within gene blocks. These scenarios are illustrated in Figure 2.2 (a-b) and mimic situations where correlations among genes are due to either genetic causes (when they share the same SNP) or non-genetics causes (when genes belong to the same block of genes). In the first scenario, 7 genes share a common eQTL and 8 genes share a common eQTL at another SNP, with 4 genes having both eQTLs in common. In the second scenario, each of the total 40 genes have either 0, 1, 2 or 3 eQTLs. When a gene had more than 1 eQTL, the eQTLs were selected on different SNP blocks. For all scenarios, eQTLs were simulated using two different values of the regression coefficients: $\beta^* = 0.5$ and $\beta^* = 0.2$. These values, based on the values estimated on the experimental data used in section 2.5, allowed us to compare the performance of the model in situations where the magnitude of the effect due to genetic causes varied from small to large. Altogether, the various situations described above amounted to 16 different combinations. In each case, eQTLs were called using an FDR level of 10%.

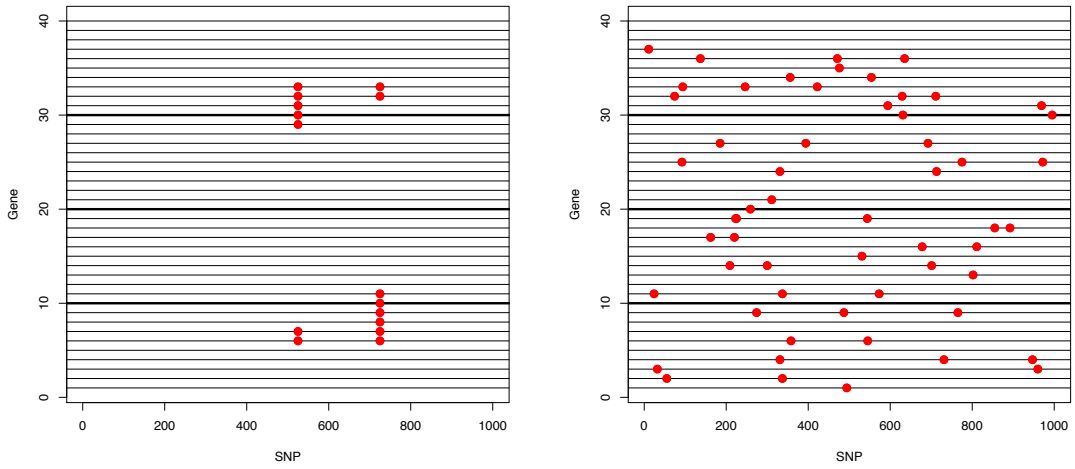
Table 2.I shows the sensitivity, specificity, positive predictive value, and negative predictive value obtained with our model across the different simulation settings we described. These values were computed based on the total number of false positive and false negative across all genes and SNPs (this means that a false positive SNP on 2 different genes would be counted twice). Table 2.I also shows the effects of different parameters on the detection of eQTLs and the capability of our model to perform even in difficult situations. In particular, we observed that correlation between SNPs had a very small impact on eQTL detection, and that the model had difficulty in detecting eQTLs with small effect sizes ($\beta^* = 0.2$). Further investigation showed that this was true except in cases where many genes with weak association values all share one identical SNP (results not shown).

Table 2.I: Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) obtained by our model. The value on the first line is for the first eQTL scenario and the value on the second line is for second eQTL scenario. The standard deviation over the 50 replications are presented in parentheses. We used a FDR threshold value of 10% for calling eQTLs.

	gene	SNP	Sens.	Spec.	PPV	NPV
$\beta^* = 0.5$	indep.	indep.	0.848(0.225)	1(0)	0.17(0.037)	1(0)
			0(0)	0.999(0)	1(0.002)	0.999(0)
		dep.	0.86(0.222)	1(0)	0.182(0.053)	1(0)
			0.001(0.03)	0.999(0)	0.999(0.003)	0.999(0)
	dep.	indep.	0.787(0.227)	1(0)	0.173(0.043)	1(0)
			0(0)	0.99(0)	1(0.002)	0.999(0)
		dep.	0.891(0.198)	1(0)	0.17(0.045)	1(0)
			0(0)	0.99(0)	1(0)	0.999(0)
	gene	SNP	Sens.	Spec.	PPV	NPV
$\beta^* = 0.2$	indep.	indep.	0.12 (0.126)	1(0)	0.226(0.403)	1(0)
			0(0)	1(0)	1(0)	0.99(0)
		dep.	0.104 (0.0114)	1 (0)	0.311 (0.458)	1(0)
			0 (0)	1(0)	0.01(0.02)	0.99(0)
	dep.	indep.	0.16(0.202)	1(0)	0.303(0.4.47)	1(0)
			0(0)	1(0)	1(0)	0.99(0)
		dep.	0.108(0.28)	1(0)	0.28(0.443)	1(0)
			0(0)	1(0)	1(0)	0.99(0)

(a) Validation: scenario 1

(b) Validation: scenario 2



(c) Comparison

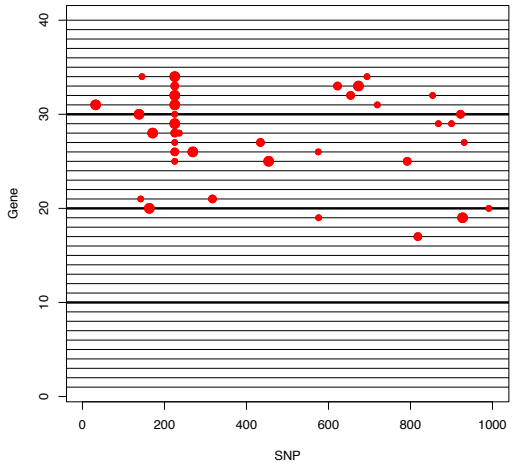


Figure 2.2: Graphical illustration of the scenarios used for the simulation study. Rows represent genes (divided into 4 blocks of 10 correlated genes each), columns represent SNPs and red dots correspond to simulated eQTLs. In the third scenario, the size of red dots is proportional to the strength of the eQTL association and corresponds to $\beta^* = 1$, $\beta^* = 0.5$ and $\beta^* = 0.2$.

2.4.2 Comparison Study

The second set of experiments was based on settings bearing more resemblance to real datasets. The simulated-eQTL distribution is illustrated in Figure 2.2 (c): 10 genes have three eQTLs each (one *cis*-eQTL, one *trans*-eQTL and one *hotspot* common to all 10 genes), one gene has one eQTL, and three genes have two eQTLs each. The regression coefficients were selected (randomly but once for all replications) among the values 0.2, 0.5 and 1. The total number of genes (40) and SNPs (1000) and error variance $\sigma_\varepsilon^2 = 0.1$ were set as in the previous simulation study. The first settings were performed with $n = 75$ and the other settings were performed in order to show the effect of the population size n on the identification and magnitude of detected eQTLs: we repeated the previous settings with $n = 50$ and $n = 25$ individuals all other parameters remaining the same. For each setting, we used 50 replications and results were averaged post-processing over the 50 replications.

In this section, we compare the performance of iBMQ to that of QTLBIM [71], M-SPLS [70], R-QTL [103], remMAP [104], and iBMQ with common inclusion weight (iBMQ-cw). The utility of each tested model and settings used for each are as outlined as follows,

- **QTLBIM**. This Bayesian model is similar to our implementation but was originally designed for classical QTL studies, and thus enables the analysis of only one gene at a time. When applying QTLBIM for comparison, we simply ran it $G = 40$ times, one gene at time. We disabled the options ”genome update” and ”epistasis effect” and we used the same number of iterations, burn-in and recording sweeps as in the other methods we compare. We set the mean prior number of eQTLs to 3 and the maximum number of eQTLs to 8. We noticed that QTLBIM has a tendency to detect accessory signals on SNPs that surround the SNP associated with the main signal. When computing the results, we aggregated main and accessory SNPs as only one signal.

- **M-SPLS**. In this approach, a first step consisted in clustering genes into groups on the basis of their expression similarity. We have used the R package Mclust

[115] for gene clustering and did a sparse partial least-squares (SPLS) regression on the optimal number of clusters. SPLS does not output posterior probabilities, but calculates bootstrapped confidence intervals of SPLS coefficients with default parameters for each simulation.

- **iBMQ-cw**. This simplified version of iBMQ was used (with the same parameters as iBMQ except for w, which is common to all genes/SNPs) and can be representative of other models that make the same assumption, *e.g.* BAYES [74] and VBQTL [101].

- **R-QTL**. This non-Bayesian tool was designed initially for classical phenotypic QTL studies. It corresponds to the simplest method, and is still used in the vast majority of real data studies. When applying R-QTL for comparisons, we simply ran it $G = 40$ times, one gene at time. The interval mapping option was disabled. We have performed a permutation test to get a genome-wide LOD significance threshold per gene.

- **remMap**. We have also performed a comparison using the remMap method , which implements a penalized regression approach , and used the BIC procedure to select tuning parameters.

In each case, eQTLs were called by controlling the false discovery rate (FDR) at 10% except for remMap, which only performs variable selection and does not compute uncertainty measures (*e.g.* p-values or posterior probabilities).

The results of the simulations are shown in Figure 2.2 and Figure 2.3. Figure 2.3 compares the ROC curves of iBMQ to those obtained with the other approaches for the different scenarios. Figure 2.4 shows the PPA plot (for iBMQ, iBMQ-cw and QTLBIM) or the frequency of associations (for M-SPLS and R-QTL) for 20 genes (10 of which sharing the common eQTL "*hotspot*") for the setting with 25 individuals. This figure shows the gain of power of iBMQ compared to QTLBIM and M-SPLS while showing the gain in flexibility compared to iBMQ-cw. Regarding the population size, we observed that all models lose power when the number of individuals decreases. However,

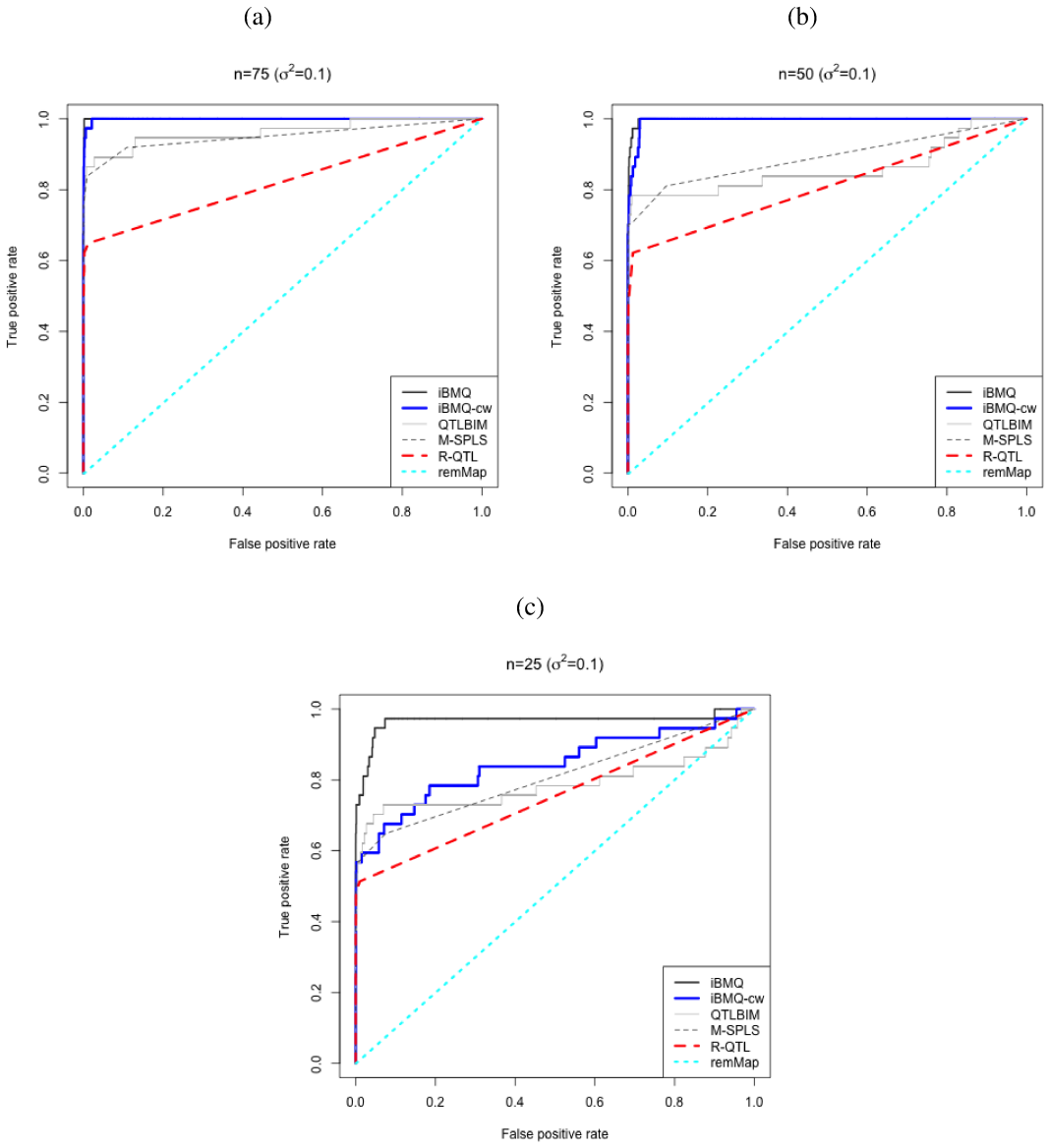


Figure 2.3: The Receiver Operating Characteristic (ROC) curves of iBMQ, iBMQ-cw, QTLBIM, M-SPLS, R-QTL and remMap for the three different simulation scenarios. a) The ROC curves represent results of the $n = 75$, b) the curves present the results of the $n = 50$, c) the curves present the results of the simulation with $n = 25$. Note that remMap does not detect any eQTLs, hence the line $y = x$.

even when simulating a population as small as 25 individuals, iBMQ still detected most eQTLs with β^* coefficient higher than 0.2. Figure 2.3 allows us to understand the behavior of the model in a more visual fashion. A comparison of iBMQ to QTLBIM and

R-QTL shows that iBMQ is better at detecting eQTLs within hotspots. A comparison with iBMQ-cw shows that both models are good at detecting eQTLs within hotspots, but that iBMQ-cw generates noisy signals outside of the hotspots due to the common weight, which has a tendency to include non relevant SNPs into the model. In addition, M-SPLS fails to detect many eQTLs outside of the hotspots, possibly because of the initial clustering, thus showing to what extent the latter can influence the results. iBMQ gains power by sharing information across genes but the model is flexible enough not to create background noise. Overall, the analyses showed that iBMQ increased the power of detecting eQTL hotspots while keeping a low false positive rate, particularly in cases with a small number of individuals. The detection of eQTL hotspots represents an important gain of the multivariate model versus univariate ones in situations where many genes display weak associations with one common shared SNP. Note that for simulations with the present parameters (n $<=$ 75) remMap did not detect any eQTLs.

2.5 Application to Data from Mouse RI Strains

In this section, we have applied our model to the whole eye tissue data generated by Williams and Lu, and available from the Gene Network Website (`genenetwork.com`). This dataset consists of the mRNA profiles of whole eye tissue from $n = 68$ BXD RI mouse strains, as measured using Affymetrix M430 2.0 microarrays [116]. To ease calculation and facilitate comparison with other methods we set $G = 1000$ corresponding to the probes showing the highest variation in expression level, while all 1700 markers (SNPs) were used. Such preselection of high variance genes is often done in eQTL studies to facilitate computation and increase power [102].

After applying the direct posterior probability approach and determining a cutoff corresponding to an FDR of 10% (corresponding PPA= 0.74), iBMQ detected a total of 759 significant eQTLs, in comparison to 182 eQTLs detected by QTLBIM (FDR of 10%, PPA$\geq$ 0.44), 1400 eQTLs detected with M-SPLS (FDR of 10%) and 5727 eQTLs detected with R-QTL (FDR of 10%). The remMap method detected a total of 1365 eQTLs (when considering all results different from zero as eQTLs). The overlap eQTLs

detected by the different methods is presented by in the Table 2.II. The genome-wide distribution of eQTLs found by all 3 methods provides further information about the performance characteristics of each model (see Figure 2.5). Almost all eQTLs detected by QTLBIM were in fact cis-eQTLs (as represented on the diagonal of Figure 2.5b). Our iBMQ method detected (in addition to the cis-eQTLs represented on the diagonal) several "hotspots" of trans-eQTLs (represented by the dots aligning along vertical lines in Figure 2.5a). Finally, M-SPLS did not detect any cis-eQTL, but identified several large groups of trans-eQTLs (Figure 2.5c). Altogether, the number of hotspots containing more than 30 probes amounted to 5 for iBMQ, 3 for remMap, and 16 for M-SPLS. No hotspots were detected by R-QTL and QTLBIM.

To verify whether the hotspots detected by iBMQ (and not by the other 4 methods) showed biologic relevance and coherence, we tested whether corresponding groups of trans-eQTLs showed enrichment in genes belonging to categories within Gene Ontology, using the DAVID Bioinformatics Resources analysis [117] (Table 2.III-2.IV). Five hotspots comprising more than 30 genes were found on 5 different chromosomes. Interestingly, each hotspot showed enrichment for genes related to a GO term dealing with characteristics and properties of epithelial cells (Table 2.III). One hotspot was enriched in genes corresponding to the "Epithelial cell differentiation" GO term. Other genes belonged to 2 other related GO term categories, and comprised almost exclusively gene from the claudin and keratin families, both of which play essential roles in the maintenance of epithelial cell functions. All 3 GO terms thus dealt with the characteristic and properties of epithelial cells, which may be in keeping with the fact that the majority of cells within whole eye tissue (including the eye bulb, the conjunctiva and the cornea) are epithelial in nature. It is interesting to note that some genes were specific to each hotspot while others were found repeatedly in several hotspots on different chromosomes (Table 2.IV). Although the 3 GO categories corresponding to the hotspots were also detected when testing for enrichment in the whole set of genes corresponding to the 1000 probes showing highest level of variance in expression, these 3 GO categories were underrepresented: only 2 out of the 3 above GO categories were represented among the 50 most significant GO categories in the original dataset, and

they ranked only fourteenth and twenty-third in terms of significance of enrichment. In particular, categories corresponding to photoreceptor functions showed most significant enrichments and represented the majority of enriched categories. Thus, the enrichment of the functions related to epithelial cell in the hotspot analysis is not likely to be a mere reflection of category enrichment in the original dataset. Of note, the selection of only a fraction of genes in the dataset was performed to facilitate comparisons across methods. A comprehensive analysis of all eQTL hotspots would require longer calculations using all gene expression data, but could possibly detect other hotspots in addition to the ones reported here, including hotspots of genes related to other functions (such as for instance retinal genes).

Table 2.II: Overlap of eQTL detection between different methods. All numbers originate from tests performed with 5 different methods on the real data set.The total number of eQTLs detected by each method is presented between parentheses. Each column contains the number of eQTLs detected in common for a given method.

	iBMQ	QTLBIM	M-SPLS	R-QTL	remMap
iBMQ (759)					
QTLBIM (182)	66				
M-SPLS (1400)	33	0			
R-QTL (5727)	139	113	2		
remMap (1365)	0	0	0	11	

Table 2.III: Number of probes in each hotspot where iBMQ identified more than 30 trans-eQTL genes, along with information about the GO annotation term for which genes in the hotspot show enrichment for and enrichment statistics. The asterisk (*) identifies annotation terms dealing with characteristics and properties of epithelial cells. The "percent" column represents the percentage of genes found with the GO terms with respect to the total number of genes in the hotspot.

SNP	No of probes	Annotation	Go Term	Percent	P-value
rs4223510 (chr 2)	132	Epithelial cell diff.*	0030855	11.4	4.1e10
rs3687764 (chr 4)	90	Structural molecule act.*	0005198	13.9	0.00012
gnf06.037.785 (chr 8)	32	Structural molecule act.*	0005198	21.7	0.0017
rs13480522 (chr 10)	142	Structural molecule act.*	0005198	11.8	0.000079
rs6376011(chr 12)	33	Intermediate filament*	0005882	16.7	0.00053

Table 2.IV: For each hotspot from Table 2.III, there was significant enrichment for genes belonging to Gene Ontology (GO) categories. All corresponding genes are listed under the GO term they belong to. Genes present on hotspot on different chromosomes are formatted in bold.

Chr 2	Chr 4	Chr 8	Chr 10	Chr 12
GO:0030855	GO:0005198	GO:0005198	GO:0005198	GO:0005882
E74-like factor 3	**claudin 4**	**claudin 23**	**claudin 23**	keratin 1
ets homologous factor	**claudin 7**	**claudin 7**	**claudin 4**	keratin 10
keratin 14	**keratin 13**	**keratin 14**	**claudin7**	**keratin 13**
keratin 17	**keratin 14**	keratin 16	collagen type III	**keratin 14**
keratin 4	**keratin 19**	**keratin 4**	**keratin 13**	
keratin 6A	**keratin 4**		**keratin 14**	
patched homolog 1	**keratin 6A**		keratin 15	
stratifin	**keratin 7**		**keratin 19**	
small proline-rich protein 1A			**keratin 4**	
trans. related protein 63			**keratin 6A**	
			keratin 7	

2.6 Discussion

In this paper, we introduced an integrated Bayesian model for eQTL mapping, iBMQ, that can handle simultaneously thousands of genes and thousands of SNPs. Our methodology is designed to deal with any Bayesian regression problem even when data are available for a limited number of individuals and when the number of measurements (gene expression and/or traits) per individual and the number of regressors (SNPs) are large. The main contribution of our model is that the association binary indicator γ_{jg} (between SNP j and gene g) and the corresponding association probability ω_{jg} of a SNP is specific for each gene and each SNP. In previous studies the association indicators γ_j were common for all genes, and the probability of association was considered as either constant ω over genes and SNPs, or dependent only on SNPs and identical for all genes ω_j. We believe that this is one of the strengths of our modelization as it helps in the detection of *hotspots*, as supported by the results of the simulation studies.

Our model could still be further refined when non-genetic correlations among gene expressions are large compared to the level of genetic correlations. One theoretically "obvious" solution to this issue consists in relaxing the hypothesis of the independence of errors assumed in model (1). In practice, however, this is an unfeasible challenge in terms

of tractability, conjugacy and computability. In fact, if genes are no longer presumed independent, the variances $\sigma_g^2, g = 1, \cdots, G$ need to be replaced by a variance-covariance matrix Σ. To keep the conjugacy of the priors and hence an analytical integration of the posterior distribution for Σ, we have to select an Inverse Wishart (as a generalization to the Inverse Gamma) distribution for Σ as described in Bottolo et al. [107] and Petretto et al. [73], among others. Although this works well in models where the association indicator γ_{jg} is common to all genes, it is not feasible in the model we propose, as it will lead to a loss of the conjugacy property and represent a heavy computational issue. As an alternative, we shall try in future work to incorporate correlations among genes by considering blocks of genes within the model, as this might improve the detection of weak associations.

Other future work will also consider the issues of correlation among SNPs due to linkage disequilibrium. It is unrealistic to totally neglect these issues mainly in highly dense genetic maps: as genetic maps become denser, the correlations between nearby SNPs become higher. A starting point may be the the concept of (left and right) flanking SNPs [71], with the construction of SNP blocks being potentially useful to convert the concept of neighbours in term of probabilities.

In this article, we have compared our model with five alternatives, but there are other methods for analyzing eQTL data. While additional methods are available [101], we chose these five because they are either obvious baseline methods, widely used or have already been compared to other methods [70]. Note that the recent work on Bayesian models for sparse regression analysis of high dimensional data in Richardson et al. [102] also provides a good alternative to our model, as a multiplicative model for the probability structure of the association binary indicator γ_{jg} is presented.

While our model requires more computing than some other methods because it integrates all genes and SNPs jointly via MCMC (section II), we believe that the improved results are worth the additional computing time. In addition, our current C implementation makes use of the openMP API [118] and automatically parallelizes calculations over genes, which can dramatically improve computational time for large datasets. The current implementation of iBMQ in R and C is available from GitHub:

https://github.com/raphg/iBMQ. An R package is currently under preparation to be made available via the Bioconductor project [119].

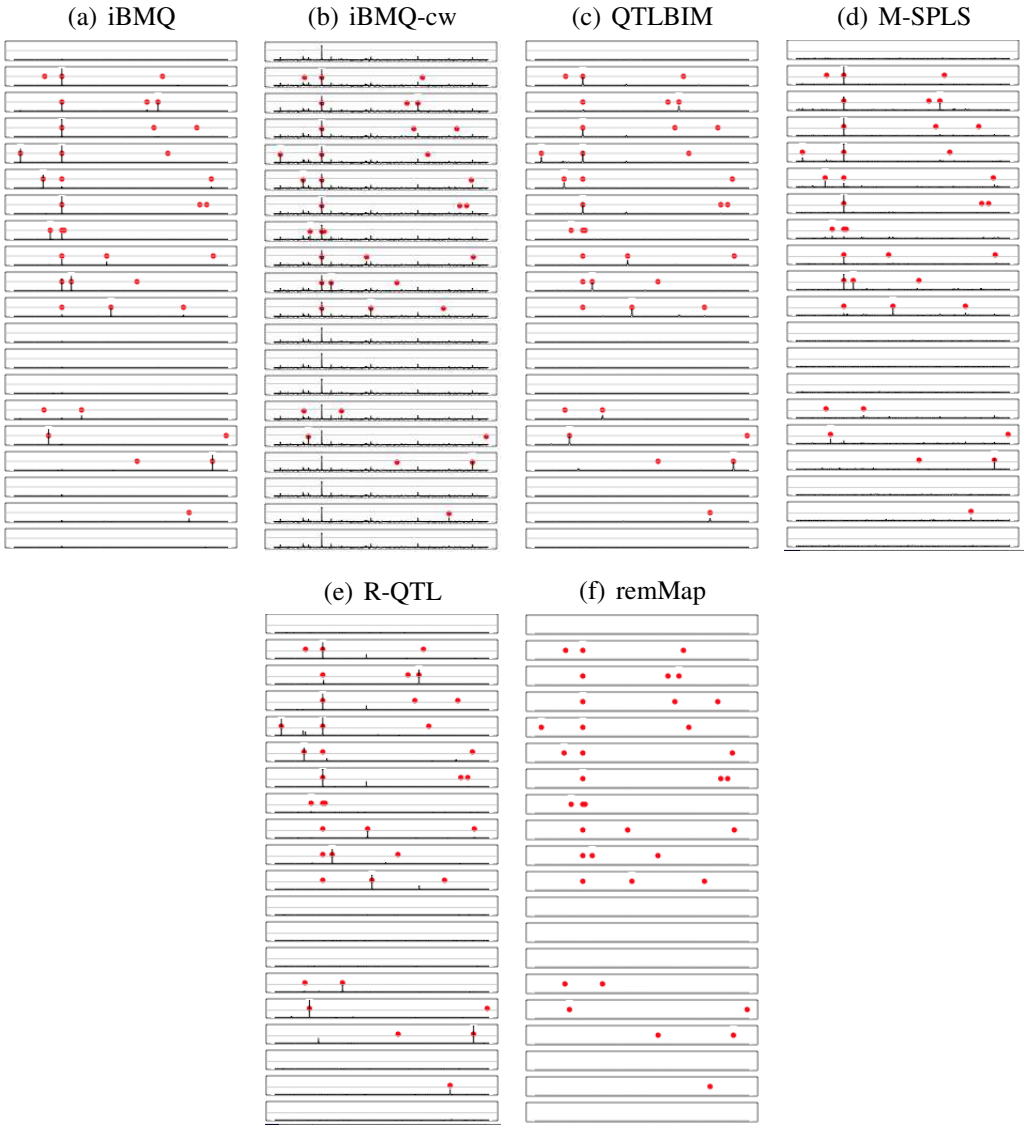
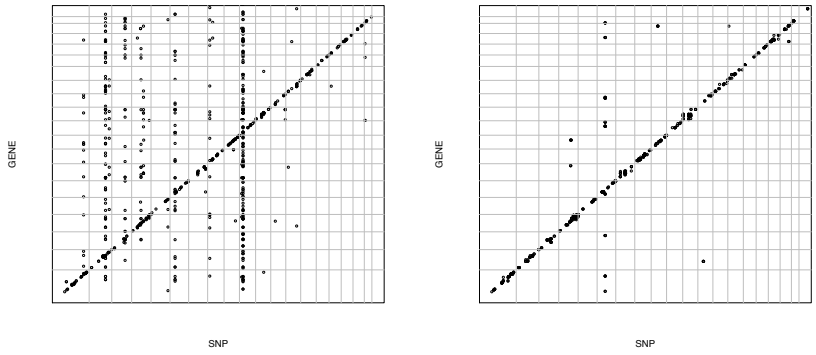


Figure 2.4: Association plots of 20 genes (simulation with 25 individuals): 10 genes share a common eQTL "*hotspot*". The grey horizontal lines correspond to the PPA cut-off used for eQTL detection (corresponding to a False Discovery Rate of 10%), and the red dots represent true eQTLs. a) Posterior Probability plots obtained with iBMQ: the method detects 8/10 genes in the hotspot; b) Posterior Probability plots obtained with iBMQ-cw: the method detects all 10 genes in the hotspot. Although all 10 genes in the hotspot are detected, all other genes (which should not be detected), also display a significant PPA for that SNP; c) Posterior Probability plots obtained with QTLBIM: the method detect only 4/10 genes in the hotspot; d) Frequency of detection of associations with M-SPLS over 50 simulations: the methods detects 7/10 genes in the hotspot; e) Frequency of detection of associations with R-QTL over 50 simulations: the method detects 4/10 genes in the hotspot. f) Frequency of detection of associations with remMap over 50 simulations: this methods did not detect any eQTLs under the present parameters.

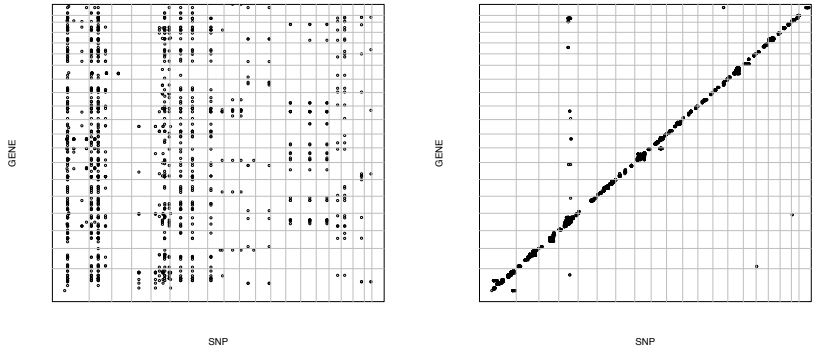
(a) iBMQ

(b) QTLBIM



(c) M-SPLS

(d) R-QTL



(e) remMap

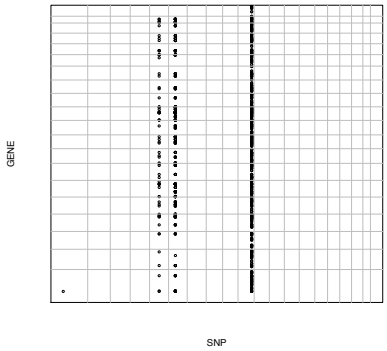


Figure 2.5: Genome-wide distribution of eQTLs found by a) iBMQ, b) QTLBIM , c) M-SPLS, d) R-QTL and e) remMap for the 1000 probes showing most variance of expression in the whole eye tissue from 68 BXD mouse recombinant inbred strains. The x-axis gives the position of each eQTL along the genome; the y-axis gives the position of the probe set target itself. The grey lines mark chromosome boundaries. Cis-QTLs form a diagonal line. Vertical bands represent groups of transcripts linked to a single trans-eQTL. iBMQ detects hotspots of trans-eQTLs (on chromosomes 2,4,8,10 and 12) that are not detected by QTLBIM. No cis-eQTL are detected by M-SPLS. d) R-QTL detect cis-eQTL but virtually no trans-eQTLs. e) No cis-eQTL are detected by remMap and the trans-eQTL hotspot do not overlap those detected by iBQM nor M-SPLS.

CHAPITRE 3

IBMQ : A R/BIOCONDUCTOR PACKAGE FOR INTEGRATED MULTIVARIATE EQTL MAPPING

En bio-informatique, il est important de rendre les méthodes développées sous forme d'outils disponibles aux autres membres de la communauté scientifique. Ainsi ce court chapitre présente une «application note» qui présente le logiciel «iBMQ » et ses fonctionnalités.

Le matériel supplémentaire de l'article se trouve à l'annexe III.

Contribution des auteurs à la préparation de l'article :

MPSB a préparé le package R avec l'aide de GI. MPSB a écrit l'article sous la supervision de CFD, AL et RG. GI a écrit le matériel supplémentaire.

Nota bene : L'article suivant a été soumis à *Bioinformatics*.

iBMQ: a R/Bioconductor package for Integrated Multivariate eQTL Mapping

Gregory C Imholte[1,*], Marie Pier Scott-Boyer[2,*], Aurelie Labbe[3], Christian F Deschepper[2], Raphael Gottardo[1]

1. Fred Hutchinson Cancer Research Center

2. Institut de recherches cliniques de Montréal (IRCM) and Université de Montréal

3. University McGill

*. These authors contributed equally

3.1 Abstract

3.1.1 Motivation

Recently, mapping studies of expression quantitative loci (eQTL) (where the expression levels of thousands of genes are viewed as quantitative traits) have been used to provide greater insight into the biology of gene regulation. To analyze such studies, Bayesian methods have been shown to provide a natural framework for eQTL modeling where information can easily be shared across genes and/or markers to increase the power to detect eQTLs. However, these approaches tend to be computationally demanding and require specialized software. As a result, most eQTL studies are still being analyzed using univariate methods that treat each gene independently, leading to suboptimal results.

3.1.2 Results

We present a powerful, computationally-optimized and free open source R package, iBMQ, for integrated Bayesian Modeling of eQTLs. iBMQ implements a joint hierarchical Bayesian model where all genes and SNPs are modelled concurrently. Model parameters are estimated using an efficient parallel Markov chain Monte Carlo algorithm utilizing the free and widely used openMP parallel library. Using a mice cardiac data, we

show that iBMQ improves the detection of large trans-eQTL bands compared to other state-of-the-art packages for eQTL analysis.

3.1.3 Availability

The R-package iBMQ is available from the Bioconductor web site at http://bioconductor.org and runs on Linux and MAC OS X. iBMQ is distributed under the terms of the Artistic Licence-2.0.

3.2 Introduction

Recently, eQTL mapping studies (where expression levels of thousands of genes are viewed as quantitative traits) have been used to provide greater insight into the biology of gene regulation. It is customary to distinguish two kinds of eQTLs: 1) cis-eQTLs (where the eQTL is on the same locus as the expressed gene); and 2) trans-eQTLs (where the eQTL is on a locus other than that of the expressed gene). Many eQTLs, particularly trans-eQTLs, often form so-called trans-eQTL bands, where one single nucleotide polymorphism (SNP) is linked to the expression of several genes across the genome. Despite this observation, most available tools for identifying eQTLs continue to treat genes as independent during analysis, and as such these methods are underpowered to detect trans-eQTL bands [66]. Here, we present an integrated hierarchical Bayesian model that jointly models all genes and SNPs to detect eQTLs. The iBMQ R/Bioconductor package incorporates genotypic and gene expression data into a single model while 1) specifically coping with the high dimensionality of eQTL data (large number of genes), 2) borrowing strength from all gene expression data for the mapping procedures, and 3) controlling the number of false positives to a desirable level.

3.3 Methods

In **iBMQ**, the source code is written in C for optimal utilization of system resources, and wrapped in more user friendly R code. In addition, **iBMQ** makes use of parallel

computing functionality provided by openMP, which greatly facilitates parallel processing when large datasets are to be processed. **iBMQ** also uses sparse matrix representations for efficient matrix calculations and memory management. More details are given in Supplementary material. **iBMQ** adopts a formal object-oriented programming discipline, making use of existing S4 classes (*e.g. eSet* and *SNPSet*). The main functions used in **iBMQ** are listed below:

1. *eqtlMcmc:* This function generates posterior samples from our hierarchical Bayesian model [120] using a Markov chain Monte Carlo (MCMC) sampler. It takes gene expression values (stored as an *eSet* object) and genomic map data (stored as an *SNPSet* object) as input. The function also has optional arguments related to the MCMC including the number of iterations, number of burn-in iterations, and whether sample parameters should be saved to disk. The output is a matrix of posterior probability of associations (PPAs), which is used for eQTL inference. An eQTL for gene g at SNP j is declared significant if its corresponding marginal posterior probability of association (PPA) is greater than a given threshold.

2. *calculateThreshold:* This function is used to calculate the cutoff that should be used on PPA values to detect significant eQTLs. In the context of multiple testing a popular approach is to use a common threshold leading to a desired false discovery rate (FDR). Our function calculate an FDR threshold based on a direct posterior probability calculation as described in Newton et al. [114].

3. *eqtlFinder:* This function can be used to apply the calculated threshold to our PPAs and identify significant SNPs.

4. *eqtlClassifier:* Given the genomic position of each probe and SNP as input, this function classifies the eQTLs as either cis-eQTLs or trans-eQTL.

5. *hotspotFinder:* This function finds markers where several genes are associated to a single marker and thus identifies trans-eQTL "bands" (also called "hotspots").

3.4 Results

We present some results of iBMQ applied to a dataset generated by our group [120]; available at GeneNetwork (accession number GN421). The dataset consists of a set of G = 8725 genes and 977 markers in cardiac tissue from n = 24 AXB-BXA recombinant inbred strain (RIS) mice, as measured using Illumina microarrays. In this example, we used 1,000,000 iterations with 50,000 burn-in iterations as suggested in previous studies [120]; the total computational time was 9,389 minutes on Mac 2*3.2 Ghz Quad-Core Intel computer using 6 CPUs. After applying the direct posterior probability approach [114] and determining a cutoff corresponding to an FDR of 10% (corresponding PPA=0.693), iBMQ detected a total of 1652 significant eQTLs, among which 278 were cis-eQTLs (where gene start is less that 1 Mb from the peak of eQTL) and 1357 trans-eQTLs. The cis-eQTLs align along a diagonal in the plot (Fig. 3.1). Among trans-eQTLs, iBMQ detected 3 clusters of 50 genes and more which form "trans-eQTL bands" represented by the dots aligning along vertical lines. To verify whether the hotspots detected by iBMQ showed biological relevance and coherence, we tested whether corresponding groups of trans-eQTLs showed enrichment in genes from Gene Ontology (GO) terms categories, using the DAVID Bioinformatics Resources (Table 3.I). In each case: 1) there was significant enrichment for particular GO terms and 2) iBMQ detected more trans-eQTL genes than the univariate R/QTL method [103]. Consequently, the significance of GO term enrichment was higher for trans-eQTL hotspots detected by iBMQ than for the corresponding hotspots detected by R/QTL. One main advantage of iBMQ therefore appears to be its greater sensitivity to detect trans-eQTL bands containing large number of genes. The supplementary material presents the code used in the analysis and additional information about the GO term analysis.

Acknowledgement

Funding:The grant RP-146065 from "Fonds quebecois de recherche Nature et Technologies" (FQRNT) to AL and CFD. RG and GI were funded by NIH grant R01 HG005692.

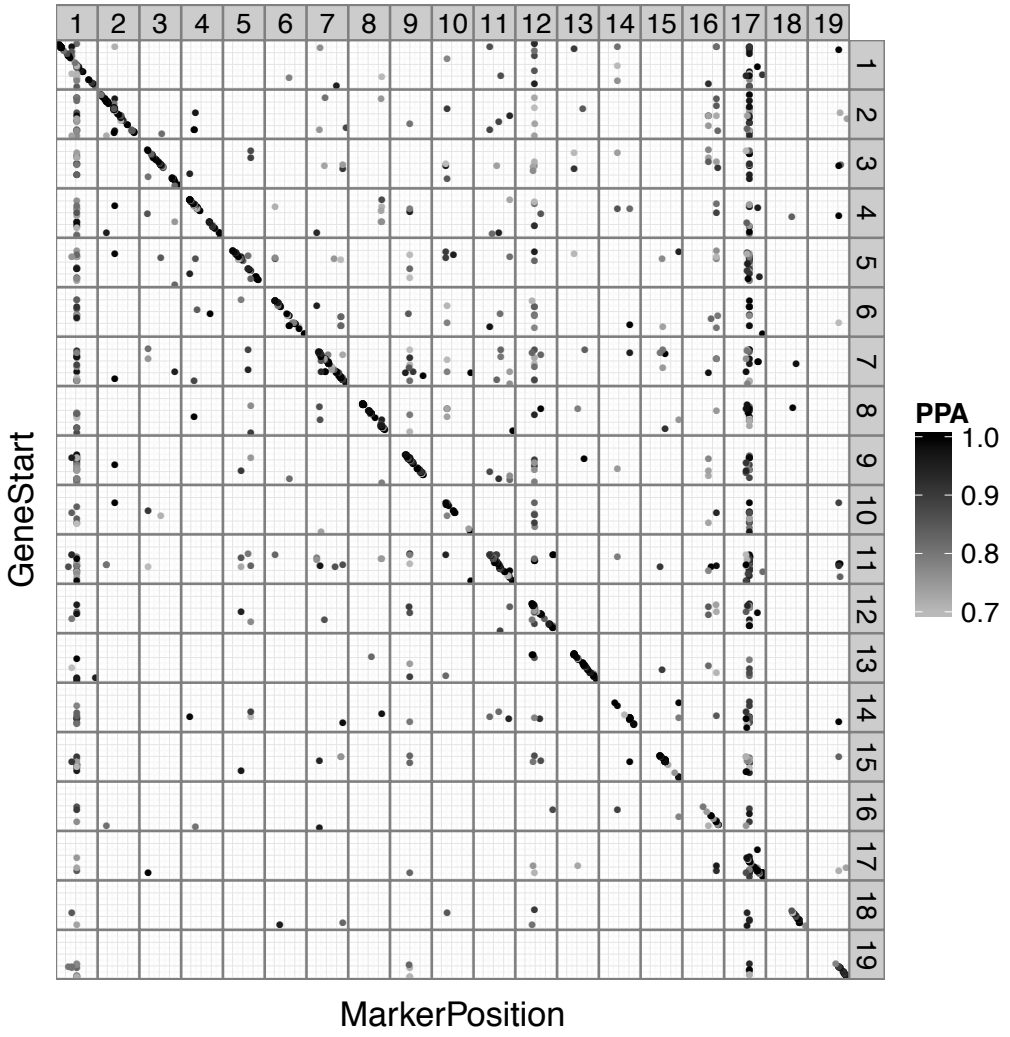


Figure 3.1: Genome-wide distribution of eQTLs found by iBMQ for mice cardiac tissue. The x-axis gives the positions of each eQTL along the genome; the y-axis gives the genomic positions of the probe sets. Chromosome numbers are given in the grey box strips. The cis-QTLs align along the diagonal line. Vertical bands represent groups of transcripts linked to one same trans-eQTL. Each eQTL point is color-coded according to the intensity of its PPA value (see PPA color scale).

SNP	GO term	iBMQ			R/qtl		
		# of genes in trans-eQTL	# of GO term genes	p-val	# of genes in trans-eQTL	# of GO term genes	p-val
1@94.8	0012505	173	14	2.4e-4	27	5	4.0e-3
12@103.5	0007167	53	5	1.5e-3	24	3	6.9e-3
17@72.4	0006955	192	26	2.7e-13	49	6	2.5e-3

Table 3.I: Positions (chr#@pos) of iBMQ-detected trans-eQTL hotspots containing ≥ 50 genes, and comparisons with corresponding hotspots detected by the univariate R/QTL method [103]. The columns list the ID number of the GO terms showing enrichment. For each method, the number of trans-eQTL genes in the band, the number of trans-eQTL genes belonging to the GO term category and the p-value for the term enrichment (as calculated by DAVID, using a modified Fisher exact test) are listed.

CHAPITRE 4

GENOME-WIDE DETECTION OF GENE CO-EXPRESSION DOMAINS SHOWING LINKAGE TO REGIONS ENRICHED WITH POLYMORPHIC RETROTRANSPOSONS IN RECOMBINANT.

Le but de ce chapitre est de comprendre comment un déterminant génétique peut affecter l'expression de plusieurs gènes contigus. Dans ce chapitre, nous avons exploré le potentiel de regroupement de cis-eQTLs, appelés «cis-eQTLs cluster » dans des croisements de souris recombinantes consanguines (RIS).

Jusqu'à présent, il y a très peu d'exemple de domaines de co-expression de gènes chez les mammifères pouvant concerner plus que des doublets ou de triplets de gènes adjacents. Avec l'utilisation de souches recombinantes consanguines de souris, nous avons confirmé l'existence de domaines contenant en moyenne $\pm$ 5 gènes fortement co-exprimés au sein de petits intervalles génomiques. En effectuant des comparaisons avec d'autres régions du génome, nous avons constaté que les régions liées sont caractérisées par un fort enrichissement en rétrotransposons (RT) de type «Short INterspersed Elements »(SINE) polymorphes entre les deux souches parentales du croisement RIS. De plus, les séquences RTs SINE polymorphiques montrent un enrichissement pour le motif de liaison du facteur d'organisation de la chromatine CTCF. Les RTs sont des éléments génétiques qui peuvent s'amplifier eux-mêmes dans un génome. Les plus connus sont les RTs de type «Long terminal repeats »(LTR), les «Short INterspersed Elements »(SINE) et les «Long INterspersed Elements »(LINE). Les RT constituent 50% du génome des mammifères. La majorité des rétrotransposons sont fixés dans le génome, mais les RT plus récents dans l'évolution peuvent être polymorphiques entre les espèces ou les souches. Les RTs peuvent affecter la régulation des gènes par divers mécanismes. Enfin, nous avons effectué des comparaisons multiples avec d'autres tissus de différents croisements RIS et on a montré que nos observations peuvent être reproduites dans d'autres ensembles de données.

Le travail effectué dans cet article est important pour les raisons suivantes : 1) à

l'aide de souches recombinantes consanguines (RIS), nous avons montré l'existence de domaines de co-expression de gènes contenant plus de 2 ou 3 gènes adjacents chez les mammifères. 2) En exploitant le cadre génétique des RIS, nous avons montré que tous les gènes dans ces domaines sont liés à la région de leur lieu propre, formant ainsi des «cis-eQTL clusters» . 3) Ces données suggèrent un mécanisme possible par lequel des polymorphismes non codants peuvent affecter l'expression coordonnée de plusieurs gènes voisins. Le matériel supplémentaire de cet article se trouve à l'annexe IV.

Contribution des auteurs à la préparation de l'article :
MPSB a réalisé toutes les analyses bio-informatiques et a écrit l'article sous la supervision de CFD.

Nota bene : L'article suivant a été publié dans le journal *G3 : Genes, Genomes, Genomics*.

MP SCOTT-BOYER and C.F. DESCHEPPER (2013) Genome-wide detection of gene co-expression domains showing linkage to regions enriched with polymorphic retrotransposons in recombinant inbred mouse strains. G3 (Bethesda), 3(4)

Genome-wide detection of gene co-expression domains showing linkage to regions enriched with polymorphic retrotransposons in recombinant inbred mouse strains

Marie-Pier Scott-Boyer and Christian F Deschepper

Cardiovascular Biology Research Unit, Institut de recherches cliniques de Montréal (IRCM) and Université de Montréal, Montréal, Québec, H2W 1R7, Canada.

4.1 Abstract

Although gene coexpression domains have been reported in most eukaryotic organisms, data available to date suggest that coexpression rarely concerns more than doublets or triplets of adjacent genes in mammals. Using expression data from hearts of mice from the panel of AxB/BxA recombinant inbred mice, we detected (according to window sizes) 42-53 loci linked to the expression levels of clusters of three or more neighboring genes. These loci thus formed "cis-expression quantitative trait loci (eQTL) clusters" because their position matched that of the genes whose expression was linked to the loci. Compared with matching control regions, genes contained within cis-eQTL clusters showed much greater levels of coexpression. Corresponding regions showed 1) a greater abundance of polymorphic elements (mostly short interspersed element retrotransposons), and 2) significant enrichment for the motifs of binding sites for various transcription factors, with binding sites for the chromatin-organizing CCCTC-binding factor showing the greatest levels of enrichment in polymorphic short interspersed elements. Similar cis-eQTL clusters also were detected when we used data obtained with several tissues from BxD recombinant inbred mice. In addition to strengthening the evidence for gene expression domains in mammalian genomes, our data suggest a possible mechanism whereby noncoding polymorphisms could affect the coordinate expression of several neighboring genes.

Keywords: Genetics of gene expression, quantitative trait loci, clustering of co-expressed genes, transposable elements, structural variants

4.2 Introduction

A significant component of gene expression variability in populations is due to variations in the DNA sequence [65]. Accordingly, genetic mapping studies have led to the identification of quantitative trait loci (QTL) linked to the expression levels of particular genes within cells and/or tissues [97]. When the expression of a given gene associates with a genetic polymorphism that maps close to that gene's locus, the corresponding "expression QTL" (eQTL) is identified as a proximal eQTL (also called "cis-eQTL"). In other cases, when the expression level of a gene associates with a locus clearly distinct from that of the gene itself, it is defined as a distal eQTL (also called "trans-eQTL"). In the case of trans-eQTLs, it has been observed that a single genetic locus can show linkage to the abundance of the mRNA transcript of several genes across the genome, forming so-called trans-eQTL "hotspots" or "bands" [121, 122]. One example would, for instance, be that of a cis-eQTL regulating the expression of a transcription factor, which in turns regulates the expression of many other genes belonging to a hotspot of trans-eQTLs.

Contrary to trans-eQTLs, investigators typically associate cis-eQTLs to the expression level of just one gene. The premise is that the cis-acting polymorphism that is located in close proximity to the gene is likely to affect the regulatory machinery of that same gene, and that machinery is unlikely to be shared by genes other than the ones that are immediately adjacent. Nonetheless, if the genome contained features that could influence the expression of several neighboring (and not necessary immediately adjacent) genes, one consequence from a genetic standpoint would be the clustering of several cis-eQTLs within a narrow genetic interval.

Although clustering of cis-eQTLs has not been reported yet, there is evidence that gene co-expression domains exist in several eukaryotic organisms [123, 124]. For instance, about 20% of the genes in Drosophila are arranged into clusters of similarly

expressed genes, with the clusters spanning intervals from 20 to 200 kb and containing 10 to 30 genes each [125]. In mammals, co-expressed genes have been reported to cluster both at either short-range (1Mb) or long-range (> 10 Mb) levels (Woo et al. 2010). One particular case of short-range co-expression clusters concerns that of conserved clusters of paralogous genes arising from tandem duplication (such as for instance Hox, globin and major histocompatibility complex genes). Beyond these paralogous clusters, it was reported in humans that the overall level of co-expression of genes was higher than expected by chance when the genes are located within distances smaller than 1 Mb, although the level of expression did not exceed that of more distant genes by a very large margin [126]. Likewise, in other studies reporting on the clustering of co-expressed genes in mammals, it was found that co-expression rarely concerned more than doublets or triplets of immediately adjacent genes [127, 128]. In such cases, co-expression is generally believed to derive from the sharing of one regulatory element by adjacent genes. Nonetheless, clusters containing in average 2 to 6 coordinately regulated genes within 1 Mb intervals have been observed under special circumstances, such as in fibroblasts during replicative senescence [129].

One limitation of many studies to date is that they have been not been performed within the framework of panels of individuals (or strains) with well-characterized genetic backgrounds. If "short-range" clustering of co-expressed genes could derive from physical elements in the genome, the impact of the latter would be easier to detect in situations where they are polymorphic, such as in animals from genetic crosses. Moreover, since clusters of cis-eQTLs would all map to precise genomic regions, further analysis of these regions might reveal the nature of the polymorphisms associated with coordinate changes in gene expression.

To complement previous genetic studies reporting on QTLs linked to cardiac morphologic characteristics, we have used Illumina microarrays to obtain the profiles of cardiac gene expression in a panel of 24 mouse recombinant inbred strains (RIS). When performing linkage analysis to detect eQTLs for all detected genes, we observed several instances where 3 or more cis-eQTLs clustered within small genomic intervals. Since such clustering of cis-eQTLs had not been reported previously, we used our dataset to

analyze the characteristics of corresponding regions in a more systematic fashion. Likewise, to test to which extent this observation could be generalized to other tissues and/or crosses, we compared our findings to that obtained in other tissues from either the same or other mouse RIS panels. We found that clusters of cis-eQTLs could be detected in all the tissues from all genetic mouse crosses we tested, and that co-expression of cis-eQTLs within these clusters reached very high levels. Further analysis of these regions revealed that they showed enrichment for particular types of structural polymorphisms.

4.3 Material and methods

4.3.1 Detection of eQTLs in hearts from AxB/BxA mouse RIS

The AxB/BxA mouse RIS originate from reciprocal crosses between the two parental C57BL/6 and A/J inbred strains, and were derived from 20 generations of inbreeding of the F2 progeny of these two strains [130]. We had previously used a set of 24 RIS from that panel to detect QTLs linked to cardiac left ventricular mass [57]. Using the same 24 RIS, we extracted total RNA from cardiac left ventricles from 4 male mice for each strain, and used Illumina MouseRef-8 v2.0 BeadChips to obtain the profile of gene expression in the tissues, as described previously [131]. The raw data were obtained by using the BeadStudio software (Illumina) and imported into the R programming environment. The data were processed and normalized using the Limma software [132]. After filtering out genes not detected across the chips (by retaining only genes detected in more than 50% of the biologic replicates for at least one strain), a set of 8725 genes was selected for further analysis. Processed data have been submitted for public access to GeneNetwork (www.genenetwork.org; accession number GN421).

For genomic mapping, genomic DNA was extracted from spleens of all corresponding 24 RIS using the DNeasy tissue kit (Qiagen, Mississauga, ON). All samples were hybridized at The Jackson Laboratory on the Affymetrix Mouse Diversity Array, which contains 623,124 single nucleotide polymorphic (SNP) and 916,269 invariant genomic probes (IGP) [133]. Signal intensities were extracted from CEL files using the MouseDivGeno package [134], and genotyping was performed by comparing the intensity and

contrast of signals in a given line to that in the parental strains [135]. In total, we detected 977 informative SNPs (meaning that they were polymorphic for at least one strain among all 24 strains from the panel) defining intervals averaging 2.59 $\pm$ 2.95 Mb. The average value of the r^2 coefficient (calculated as a descriptor of linkage disequilibrium for all pairs of adjacent informative SNPs) was $r^2 = 0.8$ (where 0 is the value for perfect equilibrium and 1 is the value obtained when two markers have identical information).

To detect and map eQTLs, the data were analyzed with the "R-QTL" tool, all other statistical analyzes being performed with the statistical language R. We used a detection threshold corresponding to a "logarithm-of-the-odds" (LOD) score of 3.3, as suggested previously [136]. For each eQTL, we then determined whether the transcription was regulated in cis or in trans by defining cis-eQTLs as those whose peak eQTL was within 1 Mbp of the physical location of the corresponding gene start. Of note, artifactual detection of eQTLs might theoretically occur when polymorphic SNPs occur within sequences corresponding to the probes used by the microarray. Although the vast majority of SNPs within probes have been shown to have no significant effect on hybridization efficiency for Illumina microarrays [137], we nonetheless used data from the Sanger website to detect all high quality (score > 100) polymorphic SNPs, compared their positions to that of all probes in the microarray (as annotated in GeneNetwork), and verified that the polymorphisms had no impact the eQTL analysis.

4.3.2 Origin of datasets

Lists of transposable elements (TEs) that are polymorphic between the C57BL/6J and the A/J mouse strains were obtained from two different sources. The first one corresponded to a supplementary file from the recent publication of Nellåker et al., which developed a comprehensive catalog of TE variants across 18 mouse strains [138]. The second source corresponded to the MouseIndelDB database (http://variation.osu.edu/mouse_indel/index.html), which reports on structural variants that show polymorphisms between four inbred strains [139, 140] . From both databases, we extracted the locations of elements that showed either "insertion" (i.e. present in C57BL/6J but absent in A/J) or "deletion" (i.e. present in A/J but absent in C57BL/6J)

versus the mm9 reference sequence of the whole genome from C57BL/6J. For simplicity, the above publications used the convention of referring to these two types of structural variants as polymorphic "insertion-deletions" (indels). Although the number of polymorphic TEs reported by MouseIndelDB is lower than that reported by Nellåker et al. (Suppl. Table IV.I), sequences in MouseIndelDB have been characterized in greater detail and contain useful annotations. Of note, the vast majority of TE sequences present in a given species are "fixed", so that in mice, they will be present in both C57BL/6J and A/J, and thus non polymorphic between the 2 strains. To obtain data on the abundance of "fixed" TEs in mice, lists of TEs located within all protein-coding genes were downloaded from the TranspoGene database (http://transpogene.tau.ac.il). Recombination rates across the mouse genomes corresponded to the values calculated in a recent report [141].

Data on genomic binding sites for several factors were obtained from several sources. A list of 33,172 regions associated with CCCTC-binding factor (CTCF) in chromatin from the hearts of adult C57BL/6J mice (as assessed by chromatin immunoprecipitation and massively parallel sequencing, i.e. ChIP-Seq) was obtained from the ENCODE/LICR database of Transcription Factor Binding Sites, using a custom track in the UCSC Genome browser. These data (Release 2, April 2012) correspond to the results of experiments performed by the laboratory of Bing Ren at the Ludwig Institute for Cancer Research. Regions corresponding to binding sites of transcription factors Gata4, Mef2A, Nkx2.5, Srf and Tbx5 in cardiac chromatin (amounting to either 16,753, 1337, 20,573, 23,806 or 55,582 regions, respectively) corresponded to those published by Schlesinger et al. [142]. The 10,486 regions corresponding to the abundance of acetylated histone 3 (H3ac) sites and the 3,596 binding sites for the p300 histone acetylase in cardiac chromatin corresponded to those published by Blow et al. [143] and Scheler et al. [144], respectively.

Lists of other eQTLs obtained in RIS were downloaded from the www.genenetwork.org web site, using the Genograph tool. In short, the tool uses WebQTL to detect all eQTLs associated to the expression levels of genes within a given dataset [145]. Datasets used are listed in Supplementary Table IV.II. In addition to data for whole eyes from

AxB/BxA mice, we used data obtained with five tissues (eye, kidney, hippocampus, hypothalamus and cerebrum) from BxD mouse RIS. The latter originate from crosses between the parental C57BL/6J (B6) and DBA/2J strains. After analysis, eQTLs were selected using a False Discovery Rate threshold of 0.2. As for our own data, we defined cis-eQTLs as those whose peak eQTL was within 1 Mb of the physical location of the corresponding gene start. Using the same parameters as for detection of cis-eQTL clusters (i.e. boxes containing at least 3 cis-eQTL separated by maximum interval of 500 kb), we calculated for each pair of analyzed tissues which proportion of cis-eQTL-containing regions overlapped between the 2 datasets.

4.3.3 Selection and comparative analysis of genomic regions

Clusters of cis-eQTLs were detected by defining regions where cis-eQTLs were separated by maximum distances of either 250, 500 or 750 kb. Control clusters were defined using the same maximum intervals between genes detected by the Illumina array in mouse hearts, and imposing a maximal limit on the overall size of control clusters in order to obtained clusters whose size was not significantly different from that of matching cis-eQTL clusters (Supplementary Table IV.III). To further verify that both types of clusters had similar properties, we calculated the number of "Entrez" genes in each cluster using the biomaRt R package (version 2.10.0) [146] interfaced to BioMart databases. Co-expression levels were quantified by calculating the absolute value of the Pearson correlation coefficient among expression levels of detected genes in the cis-eQTL clusters, and compared to the co-expression levels observed in two other kinds of boxes: 1) control clusters (whose characteristics were similar to cis eQTL clusters in terms of size, number of genes detected in the heart by the Illumina microarray, total number of genes, and overall level of expression of detected genes); and 2) random regions (corresponding to boxes of similar size chosen randomly within the genome).

For comparisons of the abundance of structural variants and/or binding sites of regulatory factors, the regions analyzed were slightly larger than the clusters themselves, and were selected by adding flanking regions of either 250, 500 or 1000 kb to four types of boxes: 1) the same cis-eQTL and control clusters defined above (using maximum

intervals between cis-eQTLS or detected genes of either 250 or 500 kb; and 2) regions with the same size as the previous ones, but either centered around single cis-eQTLs or selected randomly throughout the genome. Data calculated represented the number of features per Mb in each different region (cis-eQTL cluster, control cluster and random region). For easy comparison across different types of features, all data were normalized by dividing them by the mean number of features (i.e. structural variants or binding sites) found in the random group. Accordingly, the mean normalized number of features in random groups was 1 ($\pm$ SD), and the values in other regions corresponded to "fold difference" compared to random regions.

Motif searches were performed in the sequences of polymorphic short interspersed elements (SINE) and long-terminal repeat (LTR)-TEs, using the HOMER bioinformatic package (http://biowhat.ucsd.edu/homer/motif). Sequences analyzed corresponded to those that were present in C57BL/6 but absent in A/J (referred to as C57(+)/A/J(-) in Table 4.I), as these were the only ones where full sequence information was available. To test whether the regions containing polymorphic SINEs had specific characteristics, we use the "annotate_peaks" function provided by the HOMER package.

4.3.4 Statistics

Comparisons between groups were performed by Student's t-test (in the case of 2 groups) or by one-way ANOVA followed be Tukey HSD's post-hoc multiple comparison tests (in the cases of comparisons between more than 2 groups). Differences in the relative abundance of total and polymorphic TEs in different genomic regions were tested by 2 way-ANOVA followed by Tukey HSD's post-hoc tests.

4.4 Results

4.4.1 Detection of cis-eQTL clusters

Using the Illumina MouseRef-8 microarray, we detected a total of 8,725 genes expressed in hearts from the AXB/BXA mice. Further genomic mapping revealed that 777 of these genes were linked to cis-eQTLs, several of them forming clusters of 3 or

more cis-eQTLs within genomic intervals of a few hundreds kb. We thus tested several window sizes to best define cis-eQTL clusters and control clusters (Supplementary Table IV.III). By using maximum intervals of 250 kb, we detected a total of 42 cis-eQTL clusters (containing in average 4.23 $\pm$ 1.9 detected genes, ranging from 3 to 11, within intervals averaging 221.9 $\pm$ 130 kb), and 188 control clusters (containing in average 4.75 $\pm$ 1.85 detected genes, ranging from 3 to 13, within intervals averaging 248 $\pm$ 77.6 kb). By using maximum intervals of 500 kb, we detected a total of 53 cis-eQTL clusters (containing in average 4.9 $\pm$ 3.5 cis-eQTL genes, ranging from 3 to 19, within intervals averaging 467 $\pm$ 486 kb), and 59 control clusters (containing in average 5.2 $\pm$ 2.5 detected genes, ranging from 3 to 17, within intervals averaging 456 $\pm$ 119 kb). There were no significant differences between cis-eQTL and control clusters for any of the aforementioned 12 values (Supplementary Table IV.III). Since genes detected by the Illumina array in heart extracts do not correspond to all genes present in the genome, we also verified the density of all Entrez annotated genes in the same regions to estimate total gene density, and found no significant difference between cis-eQTL and control clusters. When using maximum intervals of 750 kb, we detected a total of 61 cis-eQTL clusters, but detected only 21 matching control clusters (Supplementary Table IV.III). Further analyses were thus restricted to the "250 kb" and "500 kb" clusters. None of the cis-eQTL clusters corresponded to clusters of paralogous genes known to arise from tandem duplication. The coordinates of all 42 "250 kb" cis-eQTL clusters are listed in Supplementary Table IV.IX, along with the symbols of corresponding cis-eQTL genes. All control clusters listed in Supplementary Table IV.X .

Although SNPs within probes are not likely to affect the hybridization efficiency of Illumina microarray probes [137], we nonetheless used data from the Sanger website to detect all high quality (score > 100) SPs that are polymorphic between A/J and C57Bl/6J, and verified that no cis-eQTL within the clusters could represent an artifact resulting from a SNP polymorphism. Accordingly, we found a total of 91 SNPs falling within probe sequences. Among them, 8 corresponded to cis-eQTLs, but none of them corresponded to those found in the cis-eQTL clusters.

Co-expression levels were quantified by calculating the absolute values of the Pear-

son correlation coefficients between each pair of genes within clusters. Within each cis-eQTL cluster, overall co-expression levels were calculated by the averaging all pair-wise co-expression values, and then compared to values obtained in either corresponding control clusters or in 500 "random groups". The latter comprised either 3 to 11 genes (for the "250 kb" clusters) or 3 to 19 genes (for the "500 kb" clusters), all genes being chosen randomly throughout the genome. For cis-eQTL genes within "250 kb" clusters, co-expression level was 0.755 $\pm$ 0.07 (mean $\pm$ SD), this value being significantly higher (P < 10e-16) than that obtained for detected genes in control clusters (0.23 $\pm$ 0.09). Although mean co-expression level in control clusters was about 17% higher than in random groups of genes (0.196 $\pm$ 0.04, P = 7.4e-07), it was also more than 3 times lower than that found in cis-eQTL clusters (Fig. 4.1). Very similar results were obtained in terms of co-expression levels and inter-group differences when using the "500" kb clusters (Fig. 4.1). The increased co-expression of genes within the cis-eQTL clusters was not due to an overall higher level of expression of genes in the cluster: the mean log2 value of expression level of genes within the cis-eQTL clusters was 9.023, this value being not significantly different (p value=0.9) from the level of expression of genes in control clusters (i.e. 9.02).

Given the average size of cis-eQTL clusters (i.e. 248 to 456 kb), the average size of intervals between polymorphic SNPs in the RIS panel (2.59 $\pm$ 2.95 Mb) and the high level of linkage disequilibrium between adjacent informative SNPs (r^2 = 0.8), the vast majority of neighboring and co-expressed cis-eQTLs within cis-eQTL clusters are likely to have the same allelic origin. In contrast, cis-eQTL genes within cis-eQTL clusters did not show homogeneity either in terms of regulation (since consistent up- or down-regulation of cis-eQTL genes was found in only 26% of the cis-eQTL clusters) or in terms of genomic strand origin (since both strands of genomic DNA contributed to the sequences of neighboring and co-expressed cis-eQTL in 77% of the cis-eQTL clusters). There was little evidence to suggest that the cis-eQTL clusters could correspond to either recombinant blocks or to regions with different recombination rates. When defining minimal haplotype blocks as regions flanked by polymorphic markers, we found a total of 930 blocks whose average size (2.59 $\pm$ 2.95 Mb) was considerably larger than that

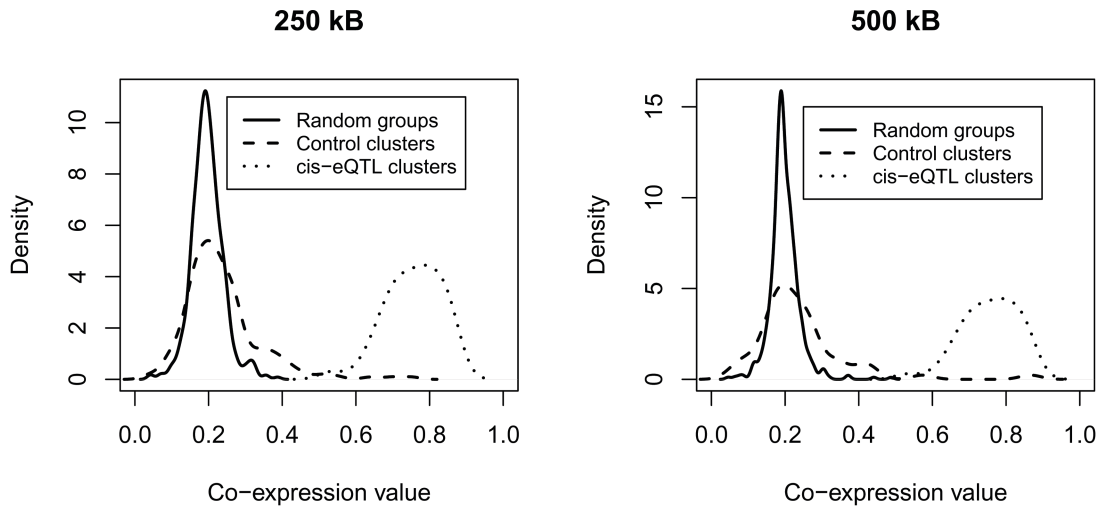


Figure 4.1: Distribution plots of co-expression values (calculated as Pearson's coefficients) of genes in cis-eQTL clusters (dotted line), control clusters (dashed line) and random boxes (plain line). For each cluster, the co-expression values represent the mean of all pairwise coexpression values between genes in the cluster. The absolute co-expression values (mean $\pm$ SD) was 0.76 ± 0.07 in cis-eQTL clusters, the value being significantly higher ($P < 0.001$) than that obtained for genes in either control clusters (0.26 ± 0.12) or random groups of genes (0.19 ± 0.19). Panel A: coexpression levels in "250 kb" clusters; Panel B: coexpression levels in "500 kb" clusters.

of cis-eQTL clusters (248 to 456 kb). Moreover, the average co-expression value of detected genes within these minimal haplotype blocks was 0.25 ± 0.11, and thus much lower than that of cis-eQTLs within cis-eQTL clusters (0.755 ± 0.07). Thus, high co-expression levels were found only for cis-eQTL genes within cis-eQTL clusters, and not for all detected genes throughout the haplotype blocks. Finally, the distribution of recombinant rate values in cis-eQTL clusters did not appear to be different from that of detected genes in control clusters (Supplementary Figure IV.1).

4.4.2 Structural characteristics of regions containing cis-eQTL clusters

The detection of clusters of cis-eQTLs suggested that genetic polymorphisms in some regions could associate with changes in the expression levels of several genes

in the same region. Considering that it would be unlikely that such coordinate changes in the regulation of several neighboring genes could result from SNPs each affecting the expression of corresponding cis-eQTL genes in a proportionate manner, we mined databases to question whether cis-eQTL cluster regions could show enrichment in structural variants (with potential of affecting expression of all cis-eQTLs in the region). Since the majority (i.e. 98%) of mouse structural variants have been reported to correspond to TE variants [139, 147] , we used data from the most recent report that established a catalog of TE variants across mouse strains [138] to test whether cis-eQTL clusters and their surrounding regions would contain more TE variants than either control clusters or regions of similar size centered around single cis-eQTLs. The respective abundances of TEs that were reported as polymorphic between A/J and C57BL/6J are listed in Supplementary Table IV.I. Considering that regulatory regions can be located either upstream or downstream of the genes under consideration, we defined the regions to be analyzed by adding flanking sequences with lengths of either 250 kb, 500 or 1000 kb to the regions corresponding to both types of clusters, thus corresponding to regions of 6 different sizes (Supplementary Table IV.IV and Table IV.V).

For each of the 6 sizes of regions, comparisons were made between the three types of "defined" regions (cis-eQTL clusters, control clusters, and regions centered around single cis-eQTLs) and the fourth type of region consisting of random regions of matching size. Overall, the same inter-regions differences were found regardless of how the regions were defined. For simplicity, the regions corresponding to the "250 kb" clusters augmented by flanking regions of 250 kb were chosen as representative data for presentation (Fig. 4.2). Long interspersed elements (LINEs) and LINE fragments showed some minor (usually non significant) differences in abundance between the four types of regions. Both polymorphic and fixed SINEs and LTRs were more abundant in all three defined regions than in random regions. For SINEs, this is in keeping with the fact that these elements are generally more abundant in gene-rich than in gene-poor regions [148]. However, only polymorphic SINEs were significantly higher ($P < e\text{-}5$) in cis-eQTL clusters compared to both control clusters and regions centered around single eQTLs. Moreover, the fold-enrichment and the significance of these differences were of

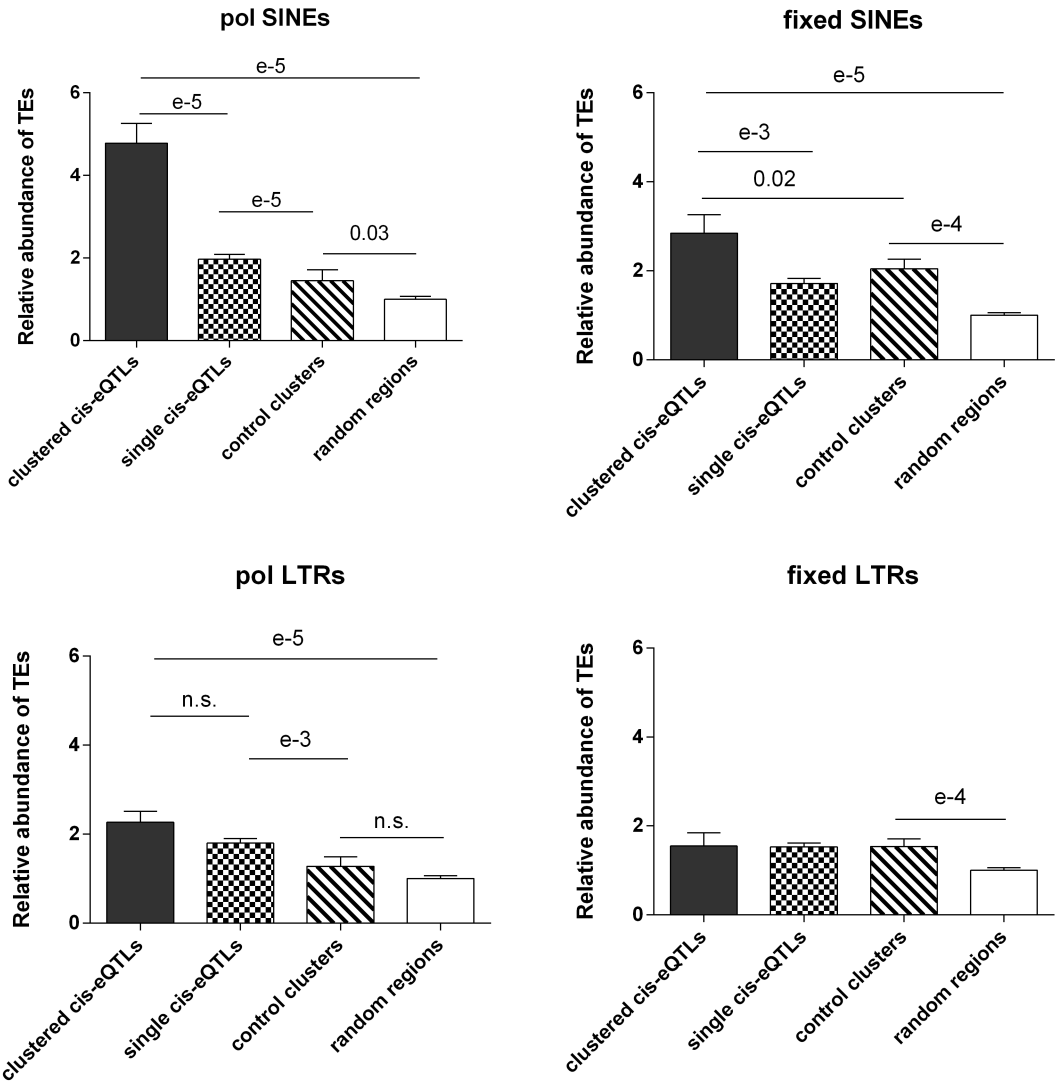


Figure 4.2: Relative abundance of polymorphic and fixed TEs (SINEs and LTR-TEs) in cis-eQTL clusters, control clusters, single cis-eQTL regions and random boxes. Each bar represents mean $\pm$ SEM of the relative abundance of TEs in each genomic region box (absolute number of TEs in each box divided by the mean value of the number of TEs in the corresponding random boxes). P values are as indicated.

greater magnitude for polymorphic SINEs than for fixed SINEs; this indicated that the greater abundance of polymorphic SINEs in cis-eQTL clusters was not a mere conse-

quence of total abundance of TEs, but rather a consequence of the genetic differences between C57BL/6J and A/J mice. These differences were not due to differences in total gene density, because the abundance of total genes in cis-eQTL regions (10.1 $\pm$ 17) was not significantly different than that in control regions (10.18 $\pm$ 3.5). Representative examples comparing two cis-eQTL clusters and control clusters of matching sizes are shown in Fig. 4.3. Altogether, polymorphic SINEs appeared to be a signature characteristic of cis-eQTL regions. We also found that the density of polymorphic SINEs in cis-eQTL cluster regions (taken along with their 250 kb flanking regions) was calculated to be 10.3 $\pm$ 10.7 polymorphic SINEs/Mb, this value being significantly higher ($P < 0.002$) than that found in other regions of corresponding haplotype blocks outside of the cis-eQTL clusters (5.1 $\pm$ 8.1). This provided additional evidence that the cis-eQTL clusters had features that differentiated them from the haplotype blocks that contained them.

To test the possible functional impact of polymorphic SINEs, motif enrichment analyses were performed for TE sequences that are present in C57BL/6 and deleted in A/J mice (because the full sequences of TEs deleted in C57BL/6 and present in A/J are not available yet). The list of most significantly enriched binding sites is shown in Supplementary Table IV.VI. For polymorphic SINEs, the binding site that was most statistically enriched (P = 1e-1283) was that previously reported to be bound by BORIS, a CTCF paralogue that binds a CTCF-like binding site [149]. This site matched the composition of the M1 moiety of the full CTCF binding sited recently described by Schmidt et al. [150]. Most other significantly enriched sites corresponded to binding sites for transcription factors from several families, in addition to another CTCF binding site that in fact matches the composition of the M2 moiety of the full CTCF binding site [150].

According to MouseIndelDB annotation, 19% of SINEs that are polymorphic between C56BL/6J and A/J correspond to B1 SINEs, with the remainder corresponding to B2 SINEs. We also investigated the nature of regions harboring polymorphic TEs (Table 4.I). Notwithstanding a few exceptions, the great majority of polymorphic TEs fell into either intronic or intergenic regions. Not surprisingly, C57(+)/A/J(-) LTR-TEs and SINEs fell into C57 regions previously annotated as having these characteristics. In-

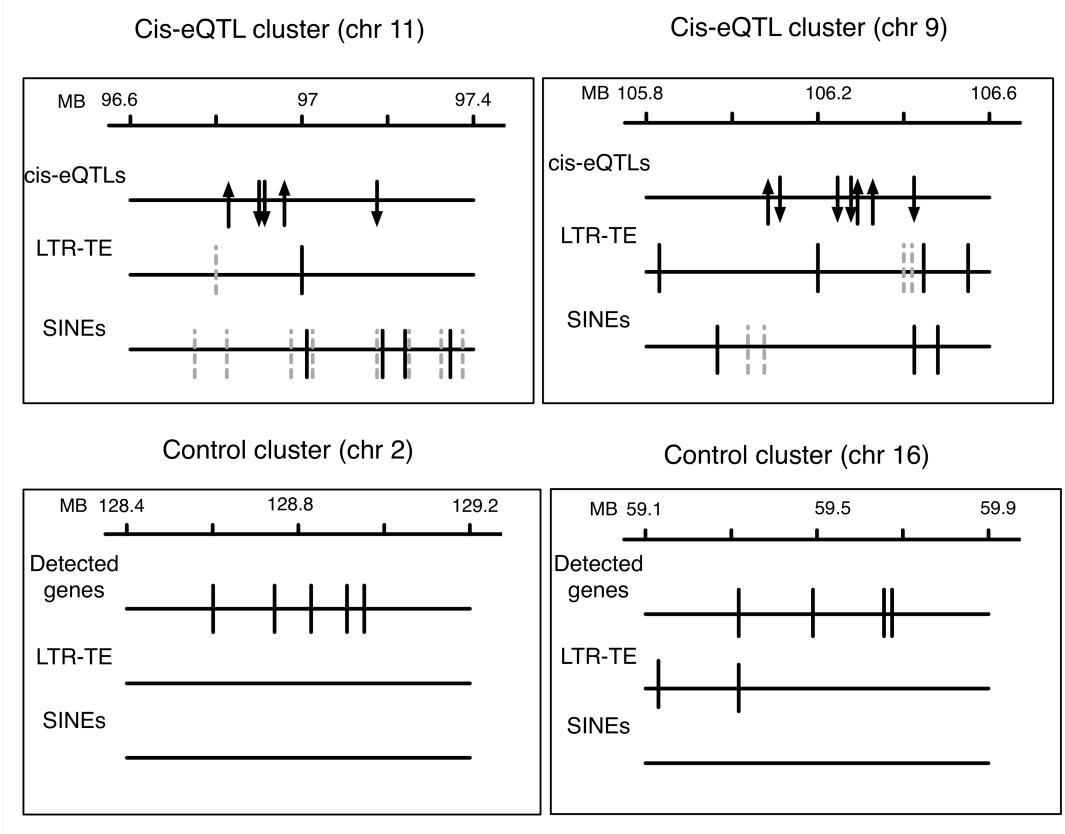


Figure 4.3: Representative examples illustrating two cis-eQTL clusters and control clusters of matching size. For each cluster, the first line (on top) represents respective genomic scales (in Mb); the second line represents the genomic positions of either cis-eQTLs or detected genes; the third and fourth lines represent the genomic positions of polymorphic LTR-TE and SINEs. On the second line, cis-eQTL genes are represented at their respective position by arrows; the arrow for the first gene on the left points upwards; other genes are represented by either upwards or downwards arrows, according to whether they correlated positively or negatively, respectively, with the change in expression of the first gene. On the third and fourth lines, plain vertical lines correspond to TEs that are C57(+)/A/J(-); dashed lines correspond to TEs that are C57(-)/A/J(+).

terestingly, C57(-)/A/J(+) LTR-TEs sometimes fell into regions annotated as containing LINEs or SINEs in C57BL/6J, while C57(-)/A/J(+) SINEs sometimes fell into regions already containing LTR-TEs or LINEs. Beyond polymorphic TEs, we obtained from previous publications (reporting the results of Chip-Seq experiments performed on mouse heart chromatin) lists of all regions corresponding to either binding sites for transcription

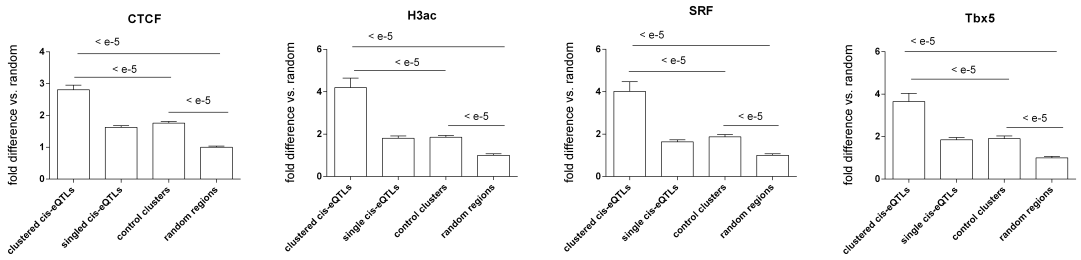


Figure 4.4: Relative abundance of binding sites for CTCF, SRF and Tbx5 and of the H3Ac chromatin marks previously reported in chromatin from mouse hearts cis-eQTL clusters, control clusters, single cis-eQTL regions and random boxes.

factors or chromatin modifications. We then tested whether all corresponding genomic features showed differential abundance between cis-eQTL, control and random regions, using the 6 different types of window sizes described above (Supplementary Table IV.VII and Table IV.VIII). Overall, the same inter-regions differences were found regardless of how the regions were defined, but differences tended to be more pronounced for regions surrounding and comprising the "250 kb" clusters. Accordingly, the regions corresponding to the "250 kb" clusters augmented by flanking regions of 250 kb were chosen as representative examples for presentation (Fig. 4.4). Overall, the abundance of all binding sites was lower in random regions than in all types of defined regions. For some regulatory factors (CTCF, H3Ac, SRF and Tbx5), their abundance was much higher in cis-eQTL regions than in the other tow types of defined regions control regions, and more abundant in the latter than in random regions (Fig. 4.4).

4.4.3 Comparisons with other panels of RIS

To test to which extent cis-eQTL clusters would be conserved across tissues, we questioned whether regions containing cis-eQTL clusters for cardiac genes would overlap with regions containing cis-eQTL clusters for genes expressed in another tissues (with clusters of cis-eQTLs being defined on the basis of maximum intervals of 250 Mb between each cis-eQTL). We first analyzed gene expression data obtained in AxB/BxA

Table 4.I: Properties of regions containing polymorphic TEs. TEs, transposable elements; LTR, long-terminal repeat; SINE, short interspersed element; TSS, transcription start site; UTR, untranslated region; LINE, long 15 interspersed element.

		LTR C57(+)	LTR C57(-)	SINE C57(+)	SINE C57 (-)
Genomic location	intron	10	22	42	42
	Intergenic	25	44	20	29
	Promoter	1	3	1	2
	TSS		1	1	3
	3UTR		1		
	Exon				1
Intergenic/intronic detailed annotations	No specific feature	1	31	1	49
	SINE		7	61	
	LTR	34	9		12
	LINE		9		9
	DNA repeat		3		1
	Low complexity		1		
	Simple repeat		6		

eyes with Illumina microarrays (Supplementary Table IV.II), which allowed us to detect a total of 35 cis-eQTL clusters (while verifying, as explained above, that none of the cis-eQTL genes could represent an artifact due to the presence of a SNP polymorphism within the sequence of the probes). Despite the fact that the latter data were obtained by other investigators, 12 out of the 42 regions containing cis-eQTLs clusters for genes expressed in AxB/BxA hearts also contained cis-eQTL clusters for genes expressed in eyes from the same RIS. We also analyzed gene expression data for other tissues from the BxD RIS mouse panel, where gene expression was analyzed (Supplementary Table IV.II) with either the Affymetrix MoGene 1.0 ST microarray (for hypothalamus) or with the Affymetrix Mouse Genome 430 microarray (for eye, kidney, hippocampus and cerebellum). Of note, consultation of the list of SNP polymorphisms between C57BL/6J and DBA/2 mice revealed a total of 0.005 % and 0.0006 % of probes used by the Affymetrix MoGene 1.0 ST and Affymetrix Mouse Genome 430 microarrays respectively, are affected by such polymorphisms. Corresponding genes were excluded from the cis-eQTL analysis. Analysis of the gene expression data from BxD RIS tissues allowed us to detect 52, 279, 260, 77 and 36 cis-eQTL clusters for kidneys, eyes, hippocampus, hypothalamus and cerebrum, respectively. Again, high proportions of cis-eQTL-containing

regions were shared across tissues. For instance, out of the 52 regions detected for genes expressed in kidneys, the number of regions also detected for genes from other tissues amounted to 44 for eyes, 44 for hippocampus, 15 for hypothalamus and 15 for cerebrum. Finally, to test whether regions containing cis-eQTL clusters in one tissue would overlap with regions containing cis-eQTL clusters in one same tissue from two RIS panels, we compared data obtained for eye genes from AxB/BxA and BxD RIS. Out of the 35 regions containing cis-eQTL clusters for genes from AxB/BxA eyes, 16 corresponded to regions containing cis-eQTL clusters for genes from BxD eyes.

4.4.4 Disucussion

Despite evidence suggesting the existence of gene co-expression domains in several eukaryotic organisms [123] [124], data available to date suggested that co-expression in mammals rarely concerns more than doublets or triplets of immediately adjacent genes [127] [128]. Our data show that within panels of mouse RIS, where gene expression variability is due in part to genetic polymorphisms, it is indeed possible to detect short range clustering of more than 3 neighboring (but not necessarily adjacent) co-expressed genes. Moreover, all co-expressed genes within these domains showed linkage to loci having the same position as that of the domains, thus showing that genetic polymorphisms can associate with the expression levels of several neighboring genes within these domains. Depending on the sizes of windows used to define them, cis-eQTL clusters detected for AxB/BxA hearts contained in average 4.2 to 4.9 cis-eQTL genes within intervals averaging 248 - 456 kb. The number of cis-eQTL clusters detected for hearts from AxB/BxA panel (i.e. 53) was within the same range as the numbers of cis-eQTL clusters detected with three other tissues from the distinct BxD RIS panel (i.e. 60 for cerebrum, 74 for kidneys and 96 for hypothalamus). Moreover, a great proportion of the regions containing cis-eQTL for AxB/BxA hearts were identical to those containing cis-eQTL clusters for eyes from either AxB/BxA or BxD RIS, and many of the regions containing cis-eQTL clusters for BxD kidneys overlapped with those detected for 4 other tissues from the same panel. Altogether, the data indicate that cis-eQTL clusters occur in a fairly robust and consistent manner across different tissues and/or genetic backgrounds in mouse

RIS. Of note, all tissues do not necessarily express the same genes, which may explain in part why some cis-eQTL clusters are distinct between tissues. With some tissues (eyes and hippocampus from BxD RIS), up to 300 cis-eQTL clusters were detected. A comprehensive list of all genomic regions having potential to contain cis-eQTLs and a full understanding of the extent to which they overlap would require data from a more exhaustive list of tissues.

One distinct feature of cis-eQTL clusters was that cis-eQTL genes in these regions showed level of-co-expression that greatly exceeded that found for detected genes in control regions with similar gene density. For instance, in regions selected as controls for the "500 kb" cis-eQTL clusters, only 2 out of 59 regions displayed co-expression levels higher than 0.56 (i.e. the value corresponding to the lowest value for gene co-expression in cis-eQTL clusters). The cis-eQTL clusters thus corresponded to "gene coexpression domain" QTLs. The fact that all coexpressed genes within domains showed linkage to a common locus suggested that the genome contains polymorphisms that can alter coexpression levels. Reasoning that identification of the nature of such polymorphisms could reveal insights as to what drives co-expression of genes within coexpression domains, we mined databases to test whether the cis-eQTL regions harbored any particular types of polymorphic structural variants. Accordingly, we found that polymorphic SINEs ([either C57(+)/A/J(-) or C57(-)/A/J(+)], were significantly more abundant in cis-eQTLs than in any other type of control regions. Given the possibility that polymorphic SINEs could in fact drive these high levels of co-expression in corresponding domains, we tested whether they showed enrichment for particular motifs. In addition to binding sites for various transcription factors belonging to different families, polymorphic SINEs showed enrichment for two sites corresponding to CTCF-binding regions. This is an agreement with the previous report showing that CTCF-binding regions in the mouse genome were preferentially embedded in B2 SINE elements [151]. B2 SINEs constitute types of SINEs that are specific to rodent genomes, where they have undergone waves of amplification [152]. Accordingly, 81% of the SINEs that are polymorphic between C57BL/6J and A/J were in fact B2 SINEs. Moreover, we compared the relative abundance of either chromatin marks or binding sites for either CTCF or cardiac transcription factors in cis-

eQTL clusters vs all 3 other types of regions. Accordingly, we found that binding sites for the transcription factors SRF and TBX5, the chromatin-organizing factors CTCF and p300, and the H3Ac chromatin were all significantly enriched in cis-eQTL clusters vs. all three other regions.

SINEs and CTCF binding sites are of particular interest. The structural and regulatory organization of the mammalian genome is fundamentally dependent on CTCF, which has been dubbed the "master weaver of the genome" [153]. CTCF generally acts as an insulator preventing the spread of inactive heterochromatin, and is often associated with open chromatin [154]. This may be particularly pertinent in the context of our current data further documenting the existence of genetically-controlled gene co-expression domains. As a matter of fact, recent genome-wide studies on chromatin structure have revealed that mammalian genomes are organized into topological domains, the boundaries of which show enrichment for SINEs and CTCF binding, and where the spread of heterochromatin is constrained [155]. Likewise, it was recently shown that in several mammalian species, CTCF-binding events are associated with waves of retrotransposon expansion, thus revealing the mechanism by which these are born [150]. Our data extend these previous reports in several ways: 1) in addition to inter-species differences, polymorphic SINEs can cause differential abundance of CTCF binding sites between strains of one same species; 2) the CTCF-dependent organization of genomic mammalian domains may not be static, as polymorphic SINEs could reshape that organization; and 3) polymorphic SINEs and CTCF-binding sites may constitute a mechanism defining gene co-expression domains, for which there was so far little evidence in mammals beyond doublets of triplets of genes.

Of note, the allelic origin of the cis-eQTL cluster regions did not affect expression of all corresponding cis-eQTL in a consistent manner (as illustrated in Fig. 4.3). However, since cis-eQTL clusters contained an average of 5 polymorphic TEs, changes in corresponding chromatin domains may be more complex than simply corresponding to a chromatin structure that is entirely open or closed for the whole region. Moreover, TEs can affect single gene expression by a variety of different possible mechanisms, as for instance by providing alternative promoters or enhancers, serving as insulators or

transcriptional silencers, disrupting the exon-intron structure and/or causing premature transcriptional termination [156]. TE might also affect gene regulation by mechanisms other than just providing regulatory elements. For instance, we found that some C57(-)/A/J(+) polymorphisms fell themselves within regions containing other TEs, and might thus disrupt in A/J the organization of some TE elements that have regulatory effects within the C57BL/6J strain. In combination with complex changes in chromatin structure, all the above mechanisms might account for the non-uniform effects of the allelic origin of cis-eQTL clusters on gene expression.

Close to half of mammalian genomes is derived from ancient transposable element (primarily retroelements) [157]. Given their great abundance in the genome and the emerging recognition of their role in gene regulation, TEs that show polymorphism between inbred strains are increasingly being recognized as potential players in the genetics of quantitative traits [138]. Accordingly, some classes of polymorphic TEs have been reported to show small but significant enrichment in refined genomic intervals selected on the basis of previous detection of quantitative trait loci for mouse quantitative traits [138]. In terms of gene regulation, recent reports on the effects of TEs have so far concerned mostly single genes [158–160]. However, complex quantitative traits are usually considered to result from the combined regulatory effects of several genes rather than from the highly penetrant effect of single mutations [76]. Gene co-expression domains might thus be of particular interest, as coordinate dysregulation of the expression levels of several genes within cis-eQTL clusters could possibly have greater effects on the phenotypic expression of complex traits than dysregulation of single genes.

Regulatory sequences located outside of coding regions are also interesting in the light of evidence that: 1) up to one third of non-coding sequence variation contributes causally to the traits under investigation in genome-wide association studies (Visel et al. 2009); 2) a great proportion of regulatory variants of gene expression are found at a fairly great distance from transcription start sites [161]; and 3) chromatin structure plays important roles in the organization and regulation of our genes [154]. Genome-wide approaches allowing the discovery and functional characterization of such elements might improve our understanding of their role in human biology and disease susceptibility (42)

[161]. However, if polymorphic TEs turn out to have important consequences on gene regulation, appropriate technologies to detect them genome-wide will need to be developed, as technologies based on detection of SNP polymorphisms are not sufficient in this regard.

FUNDING This work was supported by grant MOP-93583 from the Canadian Institutes for Health Research (CIHR).

CHAPITRE 5

NETWORK ANALYSES REVEAL STRONG CONTRIBUTIONS OF CHROMOSOME DOMAINS TO GENE CO-EXPRESSION MODULES AND A CARDIAC QUANTITATIVE TRAIT IN MICE

Le profilage de l'expression génique peut être utilisé en complément du phénotype pour identifier les gènes liés à des traits complexes. Par conséquent, on peut trouver des gènes associés à des caractères complexes en identifiant des cis-eQTLs au même emplacement qu'un QTL phénotypique, et dont les niveaux d'expression corrèlent avec le phénotype d'intérêt. Les cis-eQTLs qui sont également localisés avec le QTL phénotypique et correspondent à des gènes dont l'expression est corrélée avec la variation quantitative du phénotype ont été appelés «c3-eQTLs». Les gènes «c3-eQTLs» ont été utilisés pour hiérarchiser les gènes candidats à considérer dans l'étude des traits complexes. Toutefois, étant donné que les traits complexes résultent de l'interaction entre plusieurs gènes et l'environnement, nous avons cherché à étudier si des groupes eQTLs au sein de réseau de gènes pourraient être liés à des traits complexes.

Ce chapitre s'intéresse à trouver des gènes liés à la masse du ventricule gauche cardiaque. La masse du ventricule gauche est un trait complexe avec un intérêt médical, car une masse élevée du ventricule gauche corrèle avec la mortalité cardiaque.

Nous avons, pour étudier ce trait, utilisé une population de 24 souris AXB-BXA. Des études précédentes avaient déjà détecté, sur le chromosome 13, un QTL lié à la masse ventriculaire gauche. Pour prolonger cette étude, nous avons donc utilisé des puces à ADN pour obtenir le profil d'expression des gènes dans le coeur de chaque souche. Nous avons constaté par analyse de cartographie que l'expression d'un groupe de huit gènes adjacents est liée à la même région génique que la masse du ventricule gauche. De plus, nous avons identifié des modules de co-expression par des analyses de réseaux de gènes. Dans le module qui corrèle le plus fortement avec la masse du ventricule gauche, les huit gènes regroupés sont parmi les plus connectés et sont également connectés à d'autres gènes liés à un QTL présentant le même profil que le QTL lié à la MVG. La

répartition des variations structurales et des rétrotransposons polymorphes suggère que les changements dans l'expression des gènes regroupés et associés pourraient résulter de modifications au niveau de domaines chromosomiques, possiblement via des modifications de la chromatine. Cet article indique que la connectivité entre les gènes étend considérablement le nombre de gènes candidats à envisager au sein des locus de caractères quantitatifs. Le matériel supplémentaire de cet article se trouve à l'annexe V.

Contribution des auteurs à la préparation de l'article :
MPSB a réalisé les analyses de eQTL, les analyses de réseau de gènes, les analyses bioinformatiques et a rédigé l'article sous la supervision de CFD. SP a généré les ARNs pour les analyses de transcriptomiques.

Nota bene : L'article suivant a été soumis au journal *Genome Research*.

Network analyses reveal strong contributions of chromosome domains to gene co-expression modules and a cardiac quantitative trait in mice.

Marie-Pier Scott-Boyer, Sylvie Picard and Christian F Deschepper

Cardiovascular Biology Research Unit, Institut de recherches cliniques de Montréal (IRCM) and Université de Montréal, Montréal, Québec, H2W 1R7, Canada.

5.1 Abstract

Using Weighted Gene Co-expression Network Analysis, we analyzed cardiac gene expression data to complement earlier mapping studies where we had identified on mouse chromosome 13 (chr13) *Lvm1*, a quantitative trait locus (QTL) showing linkage to heart size in a panel of mouse recombinant strains (RIS). Accordingly, we detected 49 modules of highly connected genes, where the one showing the highest correlation to heart size: 1) contained a higher than expected proportion of genes from chr13 (with many of them having expression levels linked in cis to single nucleotide polymorphisms on chr13), and 2) was linked as a whole to a "module QTL" (mQTL). A total of 21 out of the 49 detected modules showed similar evidence of being genetically-driven: one chromosome contributed in average to 41% of their genes, with the latter: 1) showing higher connectivity than module genes from other chromosomes, and 2) clustering in regions averaging less than 20 Mb. The latter regions also showed enrichment in polymorphic structural variants and transposable elements, altogether indicating that chromosome domains were one of the main organizing principles of these genetically-driven modules. The mQTL of the module correlating with heart size had a profile matching closely that of *Lvm1* on chr13: although many genes in that module correlated with heart size individually, they correlated mostly in a manner that was directly proportional to their inter-connectivity. This greatly extends the number of candidate genes to consider as contributors to the phenotype within the QTL, beyond just those closest to the QTL peak.

5.2 Introduction

Contrary to genetic disorders with mendelian inheritance (which result from highly penetrant mutations of single genes), complex quantitative traits are generally considered to result from the interactions between the allelic variants of multiple genes (each having presumably small effects) and environmental factors [162]. The identification of quantitative trait loci (QTLs) represents one of the strategies used to elucidate the genetic determinants of complex quantitative traits [163], particularly in genetic crosses between animal inbred strains, where experimental conditions can be better controlled and genetic complexity is reduced compared to human populations [164]. However, even in such populations, the size of most detected QTLs is usually such that they typically contain a great number of genes, which complicates the identification of causing mutations. The more recent availability of high-density maps of single-nucleotide polymorphisms (SNPs) has spurred an ever-increasing number of genome-wide associations studies (GWAS). Despite the higher resolution provided by such maps and the fact that hundreds robustly replicated loci have been identified, only a very low number of causal gene allelic variants have been identified [76]. One difficulty stems from the fact that less than 12% of associated loci lie close to the protein-coding regions of genes, the majority of them being in either intergenic or intronic regions [76].

Functional genomic studies, which study the functional consequences of genetic variations on intermediate molecular traits, have been proposed as a means to improve the detection of gene variants causally linked to phenotypic traits [165, 166]. Gene expression constitutes the most commonly studied type of intermediate molecular phenotypes, with QTLs linked to gene expression being called "expression QTLs" (eQTLs) [166]. When the expression of a given gene associates with a genetic polymorphism that maps close to that gene's locus, the corresponding eQTL is referred to as a "cis-eQTL", with the presumption that a cis-acting polymorphism within the regulatory machinery of that gene affects its expression. Cis-eQTLs that both colocalize with the phenotypic QTL and correspond to genes whose expression correlates with quantitative variation of the phenotype have been called "c3-eQTLs", and have been used to prioritize genes to be

considered as candidates harboring causal mutations [167]. However, there are several limitations to this strategy: 1) only a minority of complex quantitative traits are believed to result from the dysregulation of only one gene [162]; 2) the abundance of eQTLs and the strong correlation structure in the genome make it likely that some of their overlaps with phenotypic QTLs are coincidental and not driven by the same functional variants [165]; and 3) instead of representing the sum of the individual actions of several independent biomolecules, biological systems are more typically organized as modular networks [95, 168]. Since functionally related genes are likely to show mutual dependence in their expression network, one alternative to the identification of c3-QTLs has been to construct gene co-expression networks, to then define highly inter-connected gene modules and identify which ones correlate best with variations in complex traits [95, 168, 169].

Cardiac left ventricular mass (LVM) is a quantitative complex trait that constitutes an important and independent predictor of cardiovascular mortality and mortality [27, 51]. Since this trait is also highly heritable [51], the identification of genetic determinants of LVM might lead to a better stratification of cardiovascular risk, and possibly improve our understanding of the mechanisms governing LVM. Using a panel of 24 AxB/BxA mouse recombinant inbred strains (RIS), we have previously identified on chromosome 13 (chr13) one major QTL linked to LVM (identified as "*Lvm1*") [57]. To extend these studies, we have used Illumina microarrays to profile gene expression in the hearts of four male individuals from each strain. We used these data to perform both c3-eQTL and weighted gene co-expression network analyses. By combining both approaches, we found that some modules contained a disproportionately high abundance of cis-eQTLs originating from restricted domains on single chromosomes. One of such modules showed linkage to a "module QTL" (mQTL) that had the same profile as QTL *Lvm1*. Genes contained in that module correlated with LVM not only by their level of expression, but also in a manner that was directly proportional to their level of interconnectivity with other genes in the module.

5.3 Material and methods:

5.3.1 Gene expression and mapping analyses

The AxB/BxA mouse RIS originate from reciprocal crosses between the two parental C57BL/6J and A/J inbred strains, and were derived from 20 generations of inbreeding of the F2 progeny of these two strains [130]. We have previously used a set of 24 strains from that panel to detect QTLs linked to normalized cardiac LVM (defined as LV weight corrected for whole body weight, and simply referred hereafter as "LVM") [57]. Using 4 male 12 week-old individuals from each of the same strains, we extracted total RNA from the cardiac LVs of mice, and used them to profile gene expression using Illumina MouseRef-8 v2.0 BeadChip. Raw data have been deposited into the GeneNetwork database (www.genenetwork.org) under accession number GN421. Precautions were taken to randomize all 96 samples across all lanes in all 12 microarray slides, as described previously [170]. However, since two separate batches of hybridizations were needed to hybridize the slides, possible batch effects were normalized using the ComBat software [171]. Further analyses, involving genotyping of all 24 RIS and mapping of eQTLs, were performed as described previously [172]. In short, gene expression data (corresponding for each strain to the average of gene profile values obtained in 4 individuals per strain) were analyzed (along with genomic maps) with the "R-QTL" tool [103], using a detection threshold corresponding to a "logarithm-of-the-odds" (LOD) score of 3.3, as suggested previously [136]. For each eQTL, we then determined whether the transcription was regulated in cis or in trans by defining cis-eQTLs as those whose peak eQTL was within 1 Mb of the physical location of the corresponding gene start. Confidence intervals were determined by calculating the 1.5-LOD support interval [173]. For of each cis-eQTL, we calculated the Pearsons correlation coefficient of the expression level of its corresponding gene with the value of LVM in corresponding strains. To determine which cis-eQTL genes had expression values that correlated significantly with LVM, Westfall-Young adjusted p-values were calculated on the basis of 1,000,000 permutations, using R. To find a threshold corresponding to a "false discovery rate" (FDR) = 0.1, adjusted p-values were then transformed into q-values, using the "qvalue" R pack-

age.

5.3.2 Gene co-expression networks and modules

On the basis of the expression data of all 8725 genes detected with the Illumina microarray in the LVs of male individuals from all 24 strains, we used the "Weighted Gene Co-expression Network Analysis" (WGCNA) R package [90] to construct a gene co-expression network. Within a network, each gene represents a node, and the connections between nodes are defined as edges. Network analyses were performed on the basis of the following calculations: 1) estimation of a particular β power value was performed by using the scale-free topology criterion described previously [174], which led us to the power β=6 value for all groups; 2) measures of topological overlap between nodes were calculated on the basis of the number of shared neighbors; and 3) a hierarchical clustering of the above values was performed to produce dendrograms. To define modules (i.e. clusters of highly interconnected genes), branches of the hierarchical clustering tree were cut using the dynamic tree cut algorithm implemented in the dynamicTreeCut R package. "Eigengene" values (defined as the first principal component of the expression data) were then calculated for each module. Since eigengenes can be considered as representatives of the gene expression profiles in corresponding modules, eigengene values can be used to either detect modules correlating with a given phenotype, or to detect "module-QTLs" (mQTLs), i.e. QTLs showing linkage to entire gene co-expression modules. Mapping of mQTLs was performed with the "R-QTL" tool [103], using a detection threshold corresponding to a "logarithm-of-the-odds" (LOD) score of 3.3, as suggested for RIS crosses [136]. Modules whose eigengene value correlated strongly with LVM ($P < 0.01$) were visualized graphically with the Cytoscape software [175], using the values of connection strength for the edges and that of connectivity (defined as the sum of connection strengths of each node with all other network genes) for each node, as calculated by WGCNA. In some cases, comparisons were performed between several groups of genes within modules, using ANOVA tests followed by Tukey's post-hoc multiple comparison tests.

5.3.3 Analysis of structural variants

A list of mouse genomic structural variants (including deletions, insertions and copy number variants) was obtained from the Sanger database (http://www.sanger.ac.uk/cgi-bin/modelorgs/mousegenomes/snps.pl). Structural variants defined as polymorphic between the parental A/J and C57BL/6J strains were those showing either "insertion" (i.e. present in C57BL/6J but absent in A/J) or "deletion" (i.e. present in A/J but absent in C57BL/6J) versus the mm9 reference sequence of the whole genome from C57BL/6J. A list of short interspersed nuclear elements (SINEs) was obtained from the recent study of Nellåker et al., which developed a comprehensive catalog of transposable element variants across 18 mouse strains [138]. We extracted from that report the list of SINEs that are polymorphic between the parental A/J and C57BL/6J strains. For certain mQTLs (see below), we examined the abundance of both insertion-deletions (indels) and polymorphic SINEs in consecutive 2 Mb regions extending on both sides of the mQTL peaks (up to total distances of 18 Mb). The profiles of abundance of these elements were compared to those found in regions of similar size surrounding a total of 500 polymorphic SNPs randomly selected in the entire genome. Comparisons between groups were performed by ANOVA followed by Tukey's post-hoc multiple comparison tests.

5.4 Results

5.4.1 Identification of c3-eQTLs in hearts from AxB/BxA mouse RIS

On the basis of gene expression profiling performed with Illumina microarrays in extracts of cardiac LVs from mouse AxB/BxA RIS, we found (after filtering out genes not detected in more than 50% of the biologic replicates for at least one strain) a total of 8725 genes whose expression was detectable. Genetic mapping of these gene expression values revealed a total of 10,530 eQTLs above the 3.3 LOD threshold. Among those, a total of 777 loci had a peak that was located within less than 1 Mb from the transcription start site of the gene whose expression was measured, and thus could be defined

as cis-eQTLs (Fig. 5.1). Out of those, only 33 corresponded to genes whose expression level correlated significantly (FDR = 0.1) with the values of LVM in the panel of mouse RIS. In this dataset, the threshold correspond in fact to r^2 values > 0.54. These 33 cis-eQTLs thus corresponded to c3-QTLs (Fig. 5.1). Strikingly, 8/33 c3-eQTL were clustered between positions 59.74 and 64.53 Mb within an interval of 5.8 Mb on chr13 (i.e. the same chromosome as *Lvm1*, the QTL we have identified previously for LVM/BW in the same strains). The eQTLs of these eight c3-QTLs on chr13 all had confidence intervals that overlapped with that of *Lvm1* (whose peak was located at position 57.8 Mb on chr13) (Suppl. Figure V.1). Within that cluster, 6 cis-eQTL genes corresponded in fact to 6 contiguous genes all contained with a 250 kb interval. Their identity (and corresponding LOD scores) are as follows: *Zfp367* (LOD=8.9), *Habp4* (LOD=5.9), *Cdc14b* (LOD=7.7), *1110018J18RIK* (LOD=14.5), *Ctsl* (LOD=13.6) and *Cdk20/Ccrk*(LOD=10.8). There was one additional cis-eQTL for the *Fastkd3* gene that also correlated significantly with LVM. Its transcription start site is at position 68.7 Mb on chr13, and the confidence interval of it's QTL bordered (but did not overlap with) the confidence interval of *Lvm1*. On chr17, we detected another cluster of 5 cis-eQTLs, all corresponding to genes contained within an interval of 375 kB and correlating significantly with LVM. Their identity (and corresponding LOD scores) are as follows: *Rnps1* (LOD= 13.6), *Gfer* (LOD= 17.1), *Ndufb10* (LOD= 11.5), *Fadh1* (LOD= 7.91), and *Cacna1h* (LOD = 6.9). Their confidence interval also overlapped with that of a minor LVM QTL on chr17. Although the latter QTL had a weak and non-significant linkage score for LVM (LOD=1.5), its profile was closely matched by that of these 5 cis-eQTLs on chr17 (Suppl. Figure V.2).

5.4.2 Weighted gene co-expression network analyses

Network analysis and selection of gene co-expression networks were performed using WGCNA package. This allowed us to detect a total of 49 modules, each containing at least 40 genes and being identified by a color name. By correlating the values of the eigengene of each module with that of the LVM values, two distinct modules were found to correlate significantly ($P < 0.01$) with LVM: 1) the module "thistle2" contained

a total of 48 well-annotated genes, and its correlation coefficient with LVM was 0.66 (p-value = 0.0004); 2) the module "plum2" contained 49 well-annotated genes, and its correlation coefficient with LVM was -0.57 (p-value = 0.004). QTL mapping analyses were performed for the eigengene of these two modules to detect corresponding mQTLs (Fig. 5.2). The mQTL of thistle2 module had a strong peak on chr13 (LOD = 12.2) and a profile that matched closely that of the *Lvm1* QTL on chr13. The mQTL of plum2 module had a strong peak on chr17 (LOD = 16) and a profile that matched closely that of a minor (and non-significant) LVM QTL on chr17.

5.4.3 Properties of co-expression modules correlating with LVM

A graphic representation of the thistle2 module, where the size of each node/gene and the thickness of each edge is proportional to their connectivity and strength, respectively, is shown in Fig. 5.3. We separated genes in the module into 3 distinct groups according to a combination of criteria that included their connectivity, the physical position of their locus and/or their genetic linkage with LVM. The first group comprised to a cluster of 11 eQTL genes all comprised within a 8 Mb interval on chr 13 (from positions 60.8 to 68.7 Mb), among which 10 genes corresponded to the most connected genes in the module. Within this 8Mb interval, 6 out of the 11 cis-eQTLs corresponded to the 6 contiguous c3-QTL genes within a 250 kb interval on chr13 that we identified on the basis of eQTL analysis (see above). This tighter cluster of 6 genes also corresponded to one of the "cis-eQTL cluster/gene co-expression domains" we reported previously in this RIS panel [172]. The second group in the module comprised 5 genes that were not physically located on chr13, but are all trans-eQTL genes whose eQTL profiles match that of *Lvm1* on chr13 (Supp. Figure V.3). The third group comprised all other genes from the module, three of which also belonged to chr13. Altogether, as much as 14 (out of the total 48) genes originated from chr13. This number of genes is much higher than the number of genes expected to originate from chr13 if the latter contributed to a module of 48 genes in a random fashion: given that the ENTREZ database reports that 808 out of all 20,369 genes originate from chr 13, one expects only 1.9 genes to originate from chr13. In addition to the 14 genes originating physically from chr13, the contribution of

this chromosome to the module is even higher if one considers that it links genetically to 5 other trans-eQTLs. A second important observation was that the correlation coefficient of each gene with LVM was directly proportional to its connectivity value (Fig. 5.3). Both connectivity and correlation decreased progressively across groups in the following order: 1) the physical cluster of 11 genes on chr13; 2) the group of 5 trans-eQTLs on chr13; and 3) all other network genes.

A similar graphic representation is shown for the plum2 module in Fig. 5.4, with groups of genes being defined in the following fashion. The first group corresponded to a cluster of 22 eQTL genes all contained within a 6 Mb interval on chr 17 (from positions 21 to 26.5 Mb). These 22 genes were all comprised within the group of the 31 most connected genes in the module. Within this 6Mb interval, there was a tighter cluster of 5 cis-eQTLs, all corresponding to the five c3-QTLs identified within an interval of 375 kB on chr17 by the eQTL analysis (see above). This cluster also corresponded to one of the "cis-eQTL cluster/gene co-expression domains" we reported previously in this RIS panel [172]. The second group comprised 12 other genes originating from other loci on chr 17. The third group comprised all other module genes. Altogether, a total of 34 genes (out of a total of 51) thus originated from chr17. Given that the ENTREZ database reports that 1059 out of all 20,369 genes originate from chr17, one expects only 1.6 genes to originate from chr17 in a module of 51 genes. As noted for the thistle2 module, the level of connectivity between genes was directly proportional to the value of correlation of each gene with LVM values. However, the overall correlation of plum2 genes with LVM was not as high as that of thistle2 genes. Both connectivity and correlation decreased progressively across groups in the following order: 1) the physical cluster of genes on chr17; 2) other genes from chr17; and 3) all other network genes.

5.4.4 Comparisons of co-expression modules

We further tested whether other modules had properties similar to that of thistle2 and plum2. Out of the 49 modules detected by WGCNA, 27 had a clear genetic component, since they showed linkage to one main "module QTL" (mQTL) (with, for 5 of them, at least one additional mQTL that had a lower LOD score (Table Supp. Table V.I). For

each module, we calculated the proportion of genes originating from the one predominant chromosome that contributed the highest number of genes to the module. Out of the 27 modules showing a genetic component, 21 had their mQTL on the same chromosome than the one identified as "predominant". In these 21 modules, genes from that predominant chromosome clustered within an interval averaging 18.3 $\pm$ 10.3 Mb, a value that was significantly smaller that interval containing the genes from the predominant chromosome in the 6 other modules (51.5 $\pm$ 12.1) (Table Supp. Table V.I). Since this suggested that genes in these 21 modules originated from a restricted domain rather than from the entire chromosome, we defined these 21 modules as being "chromosome domain-driven" (CDD). We further compared the properties of CDD and non-CDD genetic modules to those of the other 22 "non-genetic" modules (Table 5.I, Table Supp. Table V.II and Table V.III). Although the 2 types of genetic modules contained (when compared to non-genetic modules) a higher proportion of genes that could be defined as cis-eQTLs, the abundance of cis-eQTLs was higher in CDD than in non-CDD modules (Tables S2 and S3). Both types of genetic modules could be linked to one main mQTL; however, their corresponding LOD scores were significantly higher in CDD than in non-CDD genetic modules. In all modules, we calculated the relative levels of connectivity of genes from the predominant chromosome by dividing their mean connectivity by that of module genes originating from other chromosomes. In CDD modules, the relative connectivity of genes originating from the predominant chromosome was significantly higher than that of similar genes in the non-genetic modules (Table 5.I). Although non-CDD genetic modules also contained (compared to non-genetic modules) a higher proportion of genes from one predominant chromosome, these genes did not originate (as we had observed for CDD modules) from restricted domains, nor did they show increased levels of relative connectivity (Table 5.I).

Table 5.I: Properties of different types of modules. All values are mean $\pm$ SD. The last column lists the P values for the ANOVA tests, followed by those of the post-hoc tests (when more then 3 groups are tested; groups compared are represented by letters in superscript). CDD: chromosome domain-driven.

Characteristics	**CDD genetic modules (a)**	**Genetic non-CDD modules (b)**	**Non-genetic modules (c)**	**ANOVA / post hoc tests**
Mean distance between genes from pred. chrom. (Mb)	18.3 ± 10.3	51.5 ± 12.2	45.21 ± 17	P =4.2e-08 Pab =1.6e-05 Pac = 2.75e-07 Pbc = 5.97e-01
% of cis-eQTL genes	24.3 ±10	6.09± 3.96	2.7 ± 1.7	P = 3.76e-13 Pab = 1.84e-06 Pac < 2e-16 Pbc = 5.3e-01
Mean LOD of main mQTL	11.5 ± 3.3	4.3 ± 0.7		P = 2.24e-09
% of genes from pred. chromosome	41.3 ± 11.5	12.3± 2.08	11 ± 2	P < 2e-16 Pab = 5.93e-10 Pac <2e-16 Pbc = 9.1e-01
Relative connectivity of genes from pred. chrom. (ratios)	3.2 ± 0.9	1 ± 0.1	0.95± 0.13	P = 5.08e-15 Pab: 4.17e-09 Pac<2e-16 Pbc: 9.91e-01

The "cis-eQTL cluster/gene co-expression domains" that we described previously (using we the same expression data from this RIS panel) contained in average 4.23 ± 1.9 highly co-expressed genes with genomic regions averaging 221.9 ± 130 kb [172]. We

compared the locations of the cis-eQTL clusters to that of the peaks of the CDD modules mQTLs (Suppl. Table S1). In 14/21 CDD modules, the peak of the mQTL coincided very closely with to the locus of the previously reported cis-eQTL clusters. Five of these CDD modules also contained genes from additional cis-eQTL clusters located on the same chromosome, but at some further distance from the mQTL peak (Suppl. Table S1).

5.4.5 Structural variants in chromosome domains

Given that a great number of genes in CDD modules appeared to originate from physical domains on the chromosome that contributed most genes to the module, we tested whether the regions corresponding to these domains had particular physical properties. We thus examined the abundance of either indels or polymorphic SINEs in regions surrounding the mQTL peaks of CDD modules, and compared it to that of regions surrounding either the mQTL peak of non-CDD genetic modules or 500 random polymorphic SNPs from across the genome (Fig. 5.5). In CDD modules, both indels and SINEs were significantly more abundant in regions of 10 Mb on both sides of the mQTL peaks than in the other 2 types of regions. The abundance of these polymorphic elements were distributed in a progressively decreasing gradient fashion around a maximum that coincided with the mQTL of the CDD module.

5.5 Discussion

We have previously identified on chr13 one major QTL (*Lvm1*) linked to LVM in AxB/BxA mouse RIS [57]. We hereby show that the profile of this phenotypic QTL is closely matched by that of: 1) eight cis-eQTLs (that behave as "c3-QTLs"); 2) five trans-eQTLs; 3) the QTL of one entire gene co-expression module, where the above cis-eQTLS and trans-eQTLs are among the most connected genes in the module. Moreover, genes from the above module correlate with LVM in a fashion that is directly proportional to their connectivity. Altogether, these findings have special consequences on the interpretation of how particular loci may connect to a phenotype via alterations of gene expression, as outlined below.

After identification of either SNPs presenting association or QTLs showing linkage to a phenotype, one commonly used strategy is to nominate candidate genes on the basis of their physical proximity to the polymorphisms [166]. However, the success of this approach in the elucidation of the genetic determinants of complex disorders has so far been very limited [76]. Functional genomics, which studies the functional consequences of genetic variations on intermediate molecular traits such as gene expression, has in some cases facilitated the identification of allelic gene variants to quantitative traits. Some examples for cardiovascular traits include (among others) cardiac left ventricular mass (LVM) [58, 176], fatty acid and glucose metabolism [177], hypertension [178], and dystrophic cardiac calcifications [179]. However, the utility of this approach is still limited to cases where the effects of single allelic variants are highly penetrant, which is unlikely to apply to the majority of complex quantitative traits [162, 180].

Gene co-expression network analyses have been proposed as one possible alternative [95, 168, 169]. After constructing gene co-expression modules and identifying which ones correlate with complex traits of interest, one can use different follow-up strategies. One avenue has been to identify within the module cis-eQTL genes (as they may be easier to identify within modules than at the level of the whole genome), and then focus on those that have known functions of particular interest [181–183]. A second way has been to has been to "traverse back" regulatory cascades to identify "master regulator" genes, as successfully demonstrated in some recent studies. [184, 185]. However, the ultimate goal of these approaches still remains to identify one SNP associated with a causal gene having a predominant effect, and therefore does not depart much from the "single gene" perspective.

Assuming that it may be easier to predict the function of a module than that of individual genes [168], another often used strategy has been to rely on annotations (either gene ontology or pathway information) to test whether the module shows enrichment for genes related to annotated functions. However, the drawbacks of this strategy [184, 186] are that: 1) "canonical" pathways are unfortunately still often incomplete, and in fact are "oversimplifications" that do not represent accurate models of the complex interplay of molecules in regulatory networks; 2) enrichment analyses are biased toward what

we already know; and 3) these analyses fail to provide information on the relationships between associated genes. Another difficulty relates to the validity of the modules themselves: since the construction of gene networks is based on mathematical calculations and since high-throughput data are intrinsically noisy, one challenge is to demonstrate that the modules are biologically significant and not just biostatistical flukes [95, 168].

One way to increase our confidence in the validity and biological pertinence of modules has been to test whether they can show linkage to mQTLs. In some cases, such mQTLs have been shown to have profiles that match that of phenotypic QTLs, suggesting that the same genetic determinants link to both the phenotype and the expression levels of genes within the modules [182, 183]. This suggests that the same genetic determinants may link to both a phenotype and the expression levels of genes within the associated module, although the nature of the such genetic determinants has not been determined yet.

By combining c3-eQTL and weighted gene co-expression network analysis, we have detected one module that: 1) correlates highly with LVM in our panel of mouse RIS, and 2) links to a mQTL whose profile matches closely that of *Lvm1*, (i.e. the QTL we have identified previously for LVM in the same strains [57]. Similar observations were made for a second module that also correlated (but to a lesser extent) with LVM. In both cases, the modules contained a disproportionately high number of genes originating from one chromosome that also contained the mQTL. We therefore tested whether this could apply to other modules, which led us to the novel observation that a large proportion of modules identified in the hearts of AxB/BxA RIS mice had such properties. Moreover, genes from a "predominant" chromosome did not originate randomly from the entirety of the chromosome, but rather from more restricted genomic regions averaging less than 20 Mb (whereas the size of mouse chromosomes averages 142 Mb). Several lines of evidence indicated that these restricted chromosome regions represent specialized "domains": 1) genes from these chromosome regions showed much higher connectivity among themselves than other module genes originating from other chromosomes; and 2) these regions showed enrichment for structural variants and polymorphic SINEs. Consequently, we considered corresponding modules as being driven by a "chro-

mosome domain". Of note, using the same heart expression data from this same mouse RIS panel, we have previously reported we have previously we detected 42 loci linked to the expression levels of 3 or more highly co-expressed neighboring genes, and have called these regions "cis-eQTL clusters" [172]. Strikingly, the mQTL peaks of 16/21 CDD modules coincided very closely with the loci of such previously described cis-eQTL clusters. Moreover, the abundance of structural variants and SINEs was maximal at the peaks of these mQTLs , and distributed around it a progressively decreasing gradient fashion. This suggests that the cis-eQTL clusters/gene coexpression domains we have reported previously could in fact be at the center of a larger group of co-expressed genes on the same chromosome.

What could coordinate expression of several genes within chromosome domains? Using data from both mice and humans, Woo et al. have reported previously that co-expressed genes in mammalian genomes could cluster both at short-range (< 1Mb) and long-range (>10 Mb) levels of co-expression [187]. Interestingly, many of the mouse co-expression domains were concordant with syntenic human domains. Moreover, there is growing evidence that chromosome territory organization is involved in regulating gene expression [188]. Using the "Hi-C" method, a spatial proximity map has recently been constructed to reveal the three-dimensional architecture of the whole genome [189]. Woo et al. showed that regions showing close spatial proximity often overlapped with the co-expression domains they detected [187]. This suggests that co-expression within domains may relate to recruitment of chromosome domains within nuclear subdomains, possibly via chromatin modifications. Accordingly, the SINEs (which showed enrichment both in the cis-eQTL clusters and in the larger chromosome domain reported herein) show enrichment for the binding site motifs for chromatin organizing CCCTC-binding factor [151, 172]. Our data further extend these concepts by showing that: 1) structural variants in chromosome domains that show polymorphic abundance between strains match changes in gene regulation of genes within corresponding domains; and 2) that such changes in gene expression at the level of a chromosome domain may correlate closely with a quantitative complex trait, in this case cardiac LVM.

Up until now, the majority of GWAS have been interpreted in a very "gene-centric"

fashion, by focusing on candidate genes that are physically close to the associated SNPs. However, very few GWAS studies to date have identified genes causally associated with complex traits. Retrospectively, this may not be all that surprising, since the majority of SNPs associated to complex traits in humans fall in either intergenic or intronic regions [76]. Unfortunately, "gene-centric" biases are often present in the very way GWAS are designed: SNPs in protein-coding regions are heavily overrepresented on genotypic arrays, and some studies are even performed using gene-centric arrays focusing on particular sets of genes [76]. Others have suggested that, rather than resulting from the additive effects of multiple gene mutations, complex traits may relate more to the way in which normal genes interact with each other [95]. This would be in line with our observation that genes in the chr13 domain (and to a lesser extent, in the chr17 domain) correlate with LVM not only by their expression level, but also by their level of inter-connectivity.

One possible explanation for the observation the expression levels of several genes within chromosome domains can be affected in a coordinate fashion would be the existence of structural variants leading to changes in corresponding chromatin environments. Accordingly, we found that the abundance of structural variants and polymorphic SINEs correlated closely with the chromosome domains that appear to drive co-expression in certain modules. In such cases, SNPs might not affect gene expression via their regulatory effects on individual genes, but rather be proxy markers for more extensive structural changes in corresponding regions. Unfortunately, since structural variants are poorly captured by current genotyping methodologies [76], alternative methodologies might need to be developed. When tissues are available, combined c3-eQTL and weighted gene co-expression network analyses might thus constitute a useful adjunct. Additional studies are needed to determine to which extent polymorphisms in other crosses and/or populations can affect coordinate gene co-expression within chromosome domains. Nonetheless, we have already reported that cis-eQTL clusters/gene co-expression domains can be detected in other tissues from other mouse RIS [172], and others have shown that many mouse co-expression domains were concordant with syntenic human domains [187]. If confirmed, one consequence of our findings is that the interconnectivity of genes within CDD modules greatly extends the number of candidate

genes to consider as possible contributors to a phenotype within QTLs (or around SNPs associating with complex traits). One additional challenge will consist of sorting out to which extent phenotypic variation necessitates coordinate changes in the expression of several genes within co-expression modules, or whether complex traits may result from changes in expression of individual genes within co-expression modules.

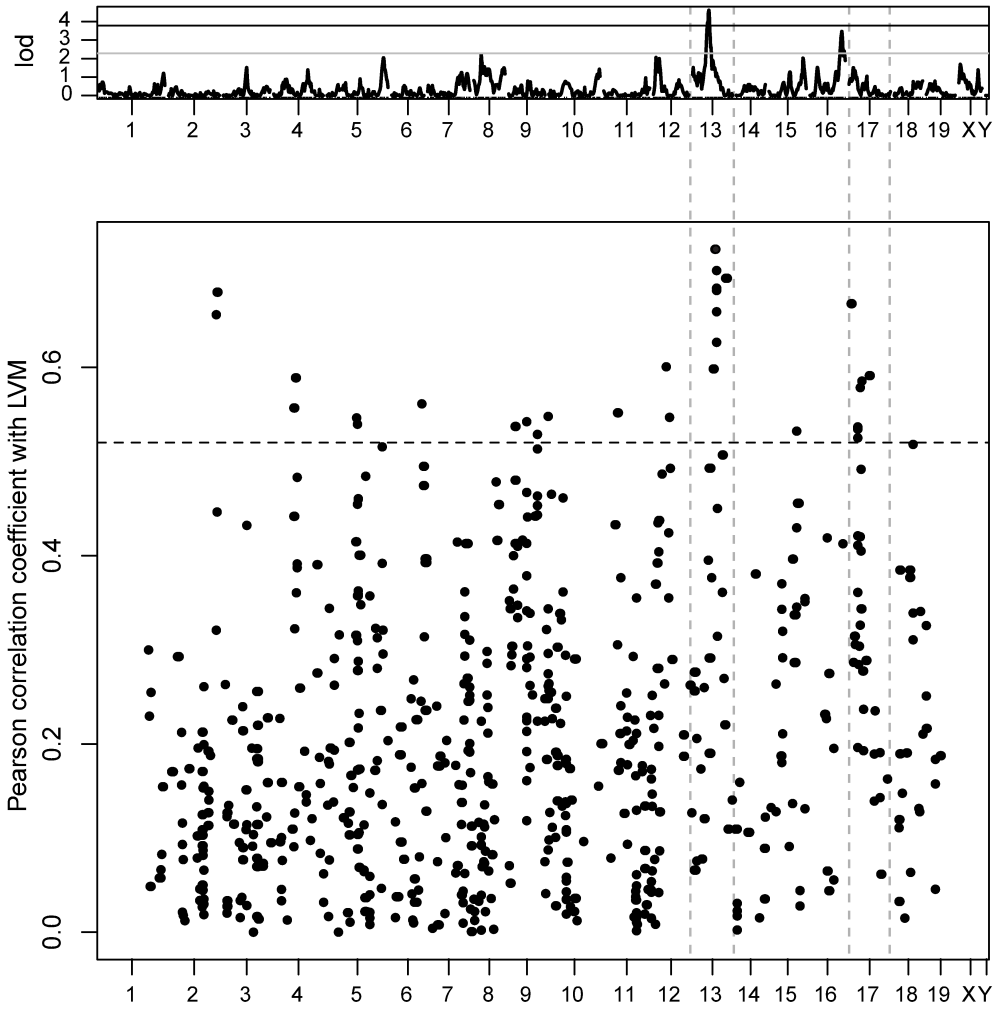


Figure 5.1: Top part: QTL mapping analysis of LVM in mouse AxB/BxA RIS. The number of each chromosome is indicated on the x axis; the LOD scores are indicated on the y axis. The strongest QTL is *Lvm1* on chr13. Bottom part: c3-eQTL analysis of cardiac cis-eQTLs in mouse AxB/BxA RIS. Expression levels of 777 cis-eQTLS were correlated with values of normalized LVM. Similarly as in the top figure, the number of each chromosome is indicated on the x axis. The absolute values of Pearson correlation coefficients are indicated on the y axis. The dashed horizontal line represents significance threshold level, as calculated by permutation tests. The dashed vertical lines represent the boundaries of chr13 and chr17, respectively, where clusters of c3-QTLs are detected. On chr13, there is a clustering of eight c3-QTLs, each having their peak within the confidence interval of *Lvm1* (see Supp. Figure V.1). On chr17, there is a clustering of five c3-QTLs, each having a profile matching closely that of a QTL showing weak linkage with LVM (see Supp. Figure V.2).

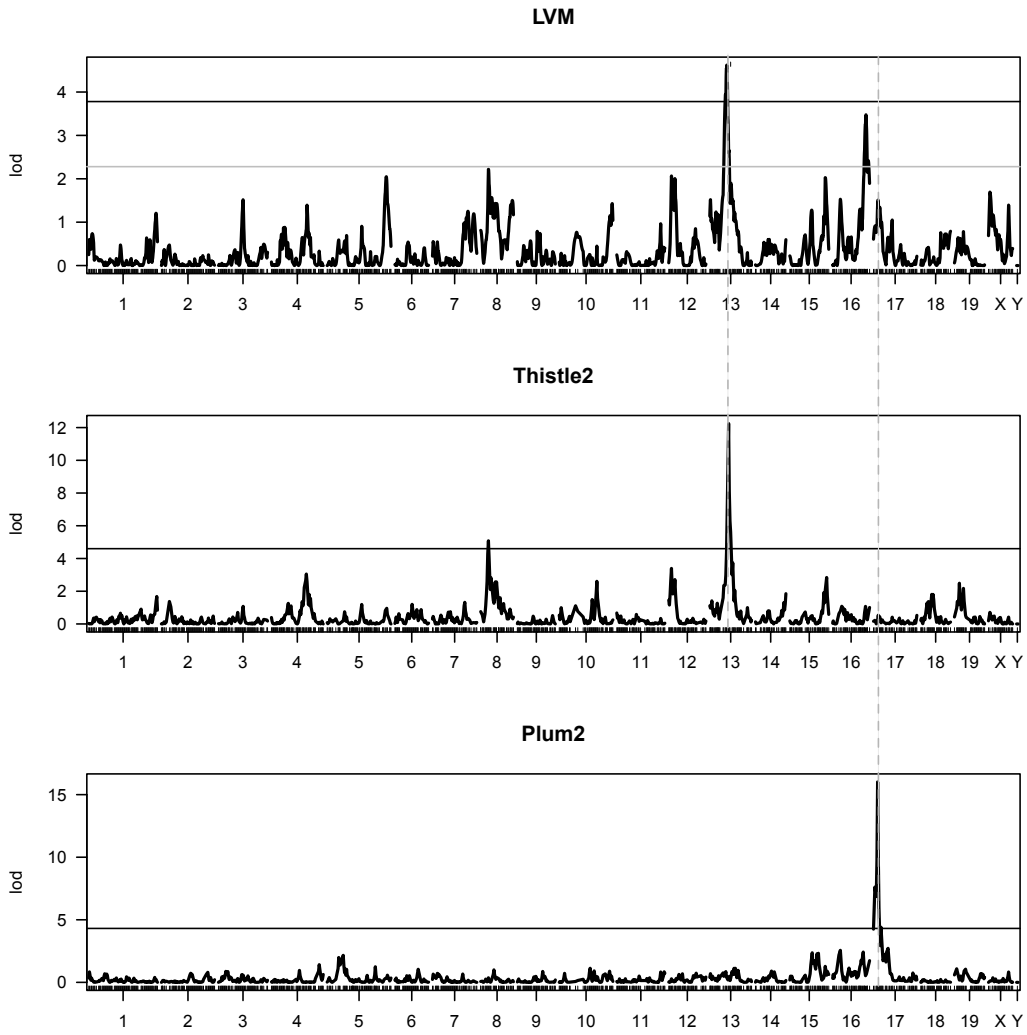


Figure 5.2: QTL mapping of the thistle 2 and plum 2 modules. The graphs represent the QTL mapping profiles for 1) LVM (top graph), the thistle 2 module (middle graph), and the plum 2 module (bottom graph), respectively. The major mQTL for thistle 2 on chr13 (LOD = 12.2) has a profile matching closely that of *Lvm1* on chr13. The major mQTL for plum2 on chr17 (LOD = 16) had a profile that matched closely that of a minor (and non-significant) LVM QTL on chr17.

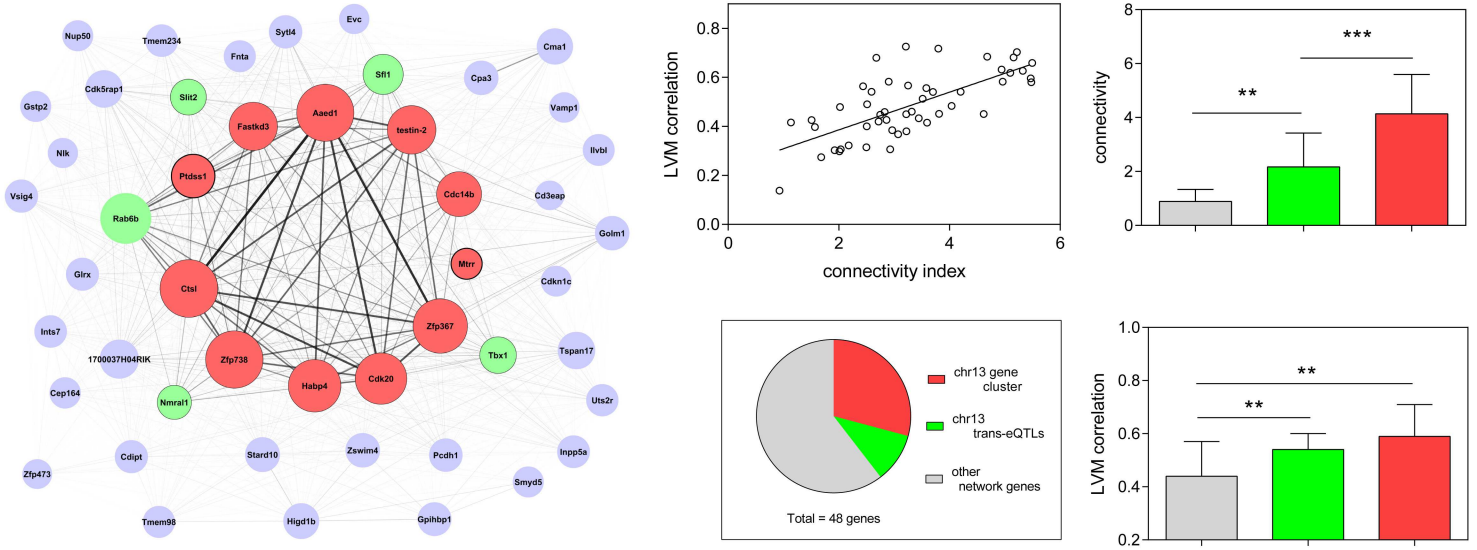


Figure 5.3: Diagram representation and properties of the thistle2 co-expression module. The size of each node is proportional to the connectivity of each corresponding gene; the width of each edge is proportional to the strength of correlation between the two corresponding genes. Each node is color-coded in the following fashion: the red nodes comprised a physical cluster of 11 eQTL genes all contained within a 8 Mb interval on chr 13 (from positions 60.8 to 68.7 Mb); the green nodes represent 5 trans-eQTL genes, each having a profile also matching that of *Lvm1* on chr 13(Supp. Figure V.2); the grey nodes represent all other module genes. The linear regression shows that each module gene correlates with LVM in a fashion that is directly proportional to their connectivity index (defined as the log2 transformation of the connectivity value calculated by WGCNA) (r2 = 0.49, P < 0.0001). The pie chart shows that as much as 14 (out of the total 48) module genes physically originated from chr13. The bar graphs (mean ś SD) show that the connectivity of module genes and their correlation with LVM is proportional to their classification in the three respective groups (**P < 0.01; ***P < 0.001).

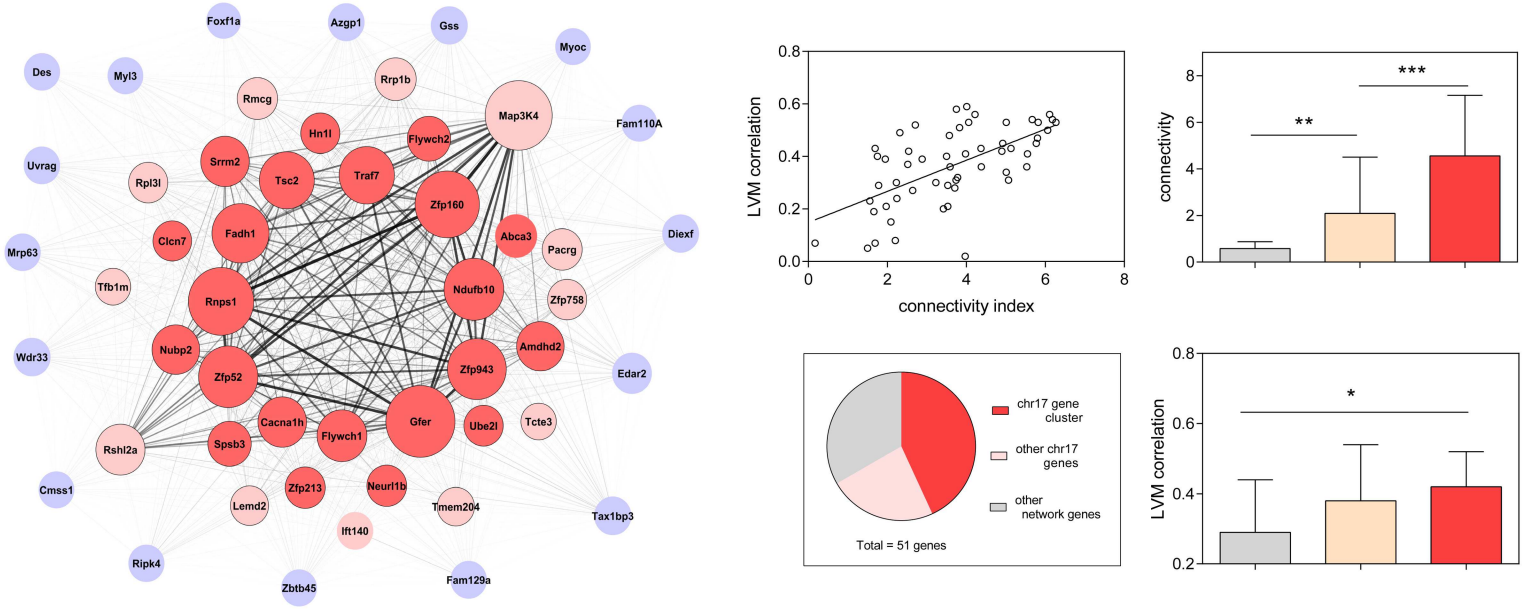


Figure 5.4: Diagram representation and properties of the plum2 the co-expression module. The size of each node is proportional to the connectivity of each corresponding gene; the width of each edge is proportional to the strength of correlation between the two corresponding genes. Each node is color-coded in the following fashion: the red nodes comprised a physical cluster of 22 eQTL genes all contained within a 6 Mb interval on chr 17 (from positions 21 to 26.5 Mb); the pink nodes represent other genes on chr 17; the grey nodes represent all other module genes. The linear regression shows that each module gene correlates with LVM in a fashion that is directly proportional to their connectivity index (defined as the log2 transformation of the connectivity value calculated by WGCNA) (r2 = 0.37, P < 0.0001). The pie chart shows that as much as 34 (out of a total of 51) module genes physically originated from chr 17. The bar graphs (mean ś SD) show that the connectivity of module genes and their correlation with LVM is proportional to their classification in the three respective groups (*P < 0.05; **P < 0.01; ***P < 0.001).

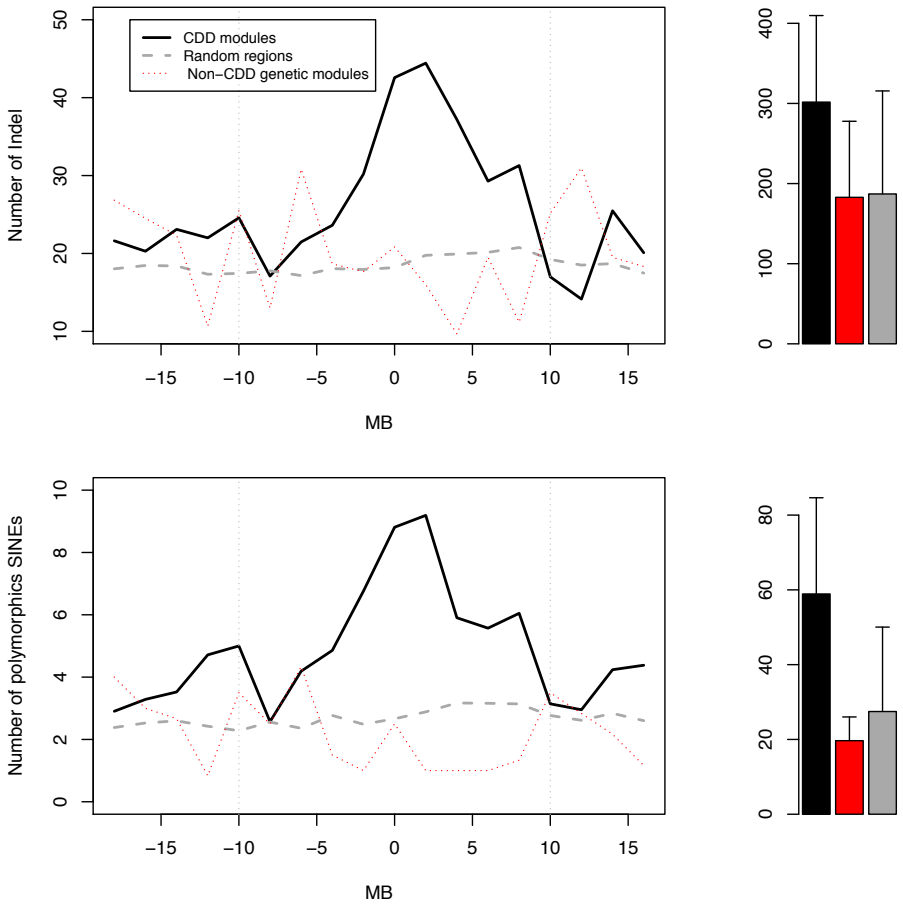


Figure 5.5: Profiles of abundance of structural variants (top graph) or polymorphic SINEs (bottom graph) in three types of regions. The zero Mb position corresponds: 1) for CDD modules (black) and non-CDD genetic modules (red), to the peak of their mQTL; and 2) for random regions (grey), to random SNPs. On right side, the bar graphs represent the mean values of abundance of either structural variants or polymorphic SINEs in regions of 20 Mb centered on the "zero Mb" position. For structural variants, the P-value for the ANOVA test was 0.000327 (post-hoc black vs. grey: $P < 0.0001$). For polymorphic SINEs, the P-value for the ANOVA test was 4.41e-09 (post-hoc black vs. red: P = 5.7e-04; post-hoc black vs. grey: P = 2.9e-09).

CHAPITRE 6

LES «HOTSPOTS » DE TRANS-EQTLS

Dans les chapitres 2 et 3 de cette thèse, nous avons développé un outil, appelé iBMQ, pour l'analyse multivariée des eQTLs. Un des avantages principaux de cette méthode par rapport aux analyses classiques univariées est la détection de ce qu'on appelle des «hotspots» de trans-eQTLs. Au niveau biologique, ces «hotspots» représentent une variation génétique qui affecte l'expression de plusieurs gènes se situant à différents endroits dans le génome. Les «hotspots» de trans-eQTLs ont été très peu étudiés, car les outils actuels n'arrivent pas à bien les détecter. Cette discussion présente une analyse de «hotspots» de trans-eQTLs de tissu cardiaque de souris AXB / BXA avec l'application des iBMQ dans le but de mieux comprendre l'importance fonctionnelle de trans-eQTLs.

6.1 Introduction

Dans le cadre de l'étude des traits complexes, l'expression des gènes est intéressante parce que : 1) l'abondance des transcrits peut être considérée comme un trait intermédiaire entre les variations génétiques et les traits phénotypiques ; 2) l'effet d'une variation d'ADN sur l'expression des gènes est sans doute plus facile à détecter que son lien avec un trait complexe, en raison des nombreuses étapes intermédiaires. Il y a des cas où un seul locus génétique peut être lié avec l'abondance de plusieurs gènes dans le génome, formant ce qu'on appelle un «hotspot» de trans-eQTLs [121, 122]. Nous avons développé et optimisé une nouvelle méthode bio-informatique qui améliore la détection des «hotspots» de trans-eQTLs [120]. Nous avons appliqué cette méthode à des données d'expression de tissu cardiaque de souris AXB / BXA. La prémisse est que généralement dans un groupe de trans-eQTLs, il y a une variation génétique qui modifie l'expression d'un gène particulier et qui à son tour, régule l'expression d'autres gènes. Ces gènes sont d'un intérêt particulier, car ils peuvent être considérés comme des «régulateurs» d'un groupe de gènes et peuvent avoir un impact significatif sur les fonctions biologiques via

la commande coordonnée de plusieurs gènes.

6.2 Matériel et méthode

Nous avons analysé un ensemble de données qui comprennent les profils d'ARNm de 8725 gènes de tissu cardiaque (pour 24 souches de souris AXB-BXA) et d'un ensemble de 977 marqueurs. Avec la fonction eQTL.mcmc de iBMQ nous avons fait 1,000,000 itérations (avec une phase de la vérification* (burn-in) de 50,000 itérations). Ce nombre d'itérations a été estimé suffisant pour calculer les «posteriori » [120].

Pour vérifier si les «hotspots » de trans-eQTLs détectés par iBMQ montrent une cohérence biologique, nous avons testé si les groupes correspondants montrent un enrichissement en gènes appartenant à des catégories de Gene Ontology, en utilisant le site web DAVID.

6.3 Résultats

Nous avons analysé avec iBMQ un ensemble de données qui comprend 8725 profils d'expression de gènes provenant de tissu cardiaque et 977 marqueurs génétiques de 24 souches de souris AXB-BXA. Après l'application de l'approche probabiliste directe postérieure et la détermination d'un seuil correspondant à un FDR de 10% (correspondant a une PPA = 0,693), iBMQ a détecté un total de 1652 eQTLs (278 cis-eQTLs et 1357 trans-eQTLs). Les cis-eQTLs étaient détectés s'il y avait au moins 1 Mo de gènes. Les «hotspots» de trans-eQTLs ont été détectés à plus de 20 gènes. Pour accélérer le calcul, nous avons utilisé 6 processeurs. Sur un Mac 2 3.2 Ghz Quad-Core Intel Xeon, il a fallu 9389,16 minutes.

Nous avons identifié 10 «hotspots » contenant au moins 20 gènes (voir la table 1). Pour vérifier si les «hotspots» détectés par iBMQ montrent une cohérence biologique, nous avons testé si les groupes de gènes correspondants à des trans-eQTLs montrent un enrichissement pour des catégories de Gene Ontology (GO). Nous avons constaté que 9 des 10 «hotspots » présentent un enrichissement significatif pour des termes GO. Nous avons aussi vu que 6 des 10 «hotspots» présentaient au moins un cis-eQTL et que 3

d'entre eux étaient présents dans le terme GO le plus enrichi.

Table 6.I – Liste des 10 «hotspots» de trans-eQTLs de plus de 20 gènes détectés par iBMQ. La première colonne contient l'identifiant de chaque SNP, la deuxième indique l'identité du chromosome et la troisième indique la position associée au «hotspot» de trans-eQTLs. La quatrième colonne contient le nombre de gènes contenus dans chaque «hotspot» . La cinquième, le terme GO le plus enrichi. La sixième, le p-value associé à ce terme GO, avec entre parenthèses, le nombre de gènes du «hotspot» contenus dans ce terme. Et finalement, la dernière colonne contient les cis-eQTLs contenus dans le «hotspot» (ceux en gras sont aussi contenus dans le terme GO).

SNP	Chr	Position génomique	Nombre de gènes	Terme GO le plus enrichis	p-value associé au terme GO	Cis-eQTL parmi les gènes du hots-pot
JAX00006710	1	94893324	173	GO :0016192 vesicle-mediated transport	1.8E-4 (14 gènes)	*Dusp28, LOC100041596, Stk25, Thap4*
JAX00063837	15	76477213	22	GO :0005739 mitochondrion	2.2E-2 (5 gènes)	**C030006K11RIK**, *Lrrc24, Rbm9, Zfp251*
JAX00067891	16	30397320	25	GO :0031406 carboxylic acid binding	5.8E-3 (3 gènes)	*LOC100047583, Ppp1r2*
JAX00079334	17	89889853	20	GO :0000166 nucleotide binding	6.5E-2 (6 gènes)	-
JAX00099053	2	116547769	24	-	-	-
JAX00174813	9	99873152	31	GO :0070469 respiratory chain	3.9E-3 (3 gènes)	-

JAX00187760	12	103524769	53	GO :0044431 Golgi apparatus part	3.6E-3 (5 gènes)	**Golga5**
JAX00378946	14	46218157	26	GO :0004522 pancreatic ribonuclease activity	1.9E-2 (2 gènes)	
JAX00436504	17	33170562	24	GO :0008134 transcription factor binding	2.6E-7 (7 gènes)	*Wdr46,* **Zbtb22**
JAX00445721	17	72449327	192	GO :0006955 immune response	2.7E-13 (26 gènes)	*Clip4, Ypel5*

Nous avons ensuite comparé nos données à celles de l'équipe de E. Petretto. Ces collègues avaient effectué des analyses similaires dans les souches de rats [58] [190]. Chez le rat, le «hotspot » de trans-eQTLs qui a montré le plus haut niveau de connectivité inter-gènes était très similaire dans sa composition à un «hotspot» de trans-eQTLs identifié dans notre panel de souris. Les gènes du «hotspot» de trans-eQTLs identifiés chez le rat et la souris sont des gènes de la réponse inflammatoire et/ou de la réponse aux interférons. De plus, dans le cas du rat, le hotspot appartient à un plus grand réseau de gènes co-exprimés associés à des réponses antivirales et au risque de diabète de type 1 [185]. Les gènes dans le «hotspot » ont été jugés sous le contrôle du gène *Ebi2* (ou *GPR183* qui a un équivalent chez l'homme). Malgré les similitudes entre notre «hotspot» de trans-eQTLs chez la souris et celui détecté chez le rat et chez l'homme, la localisation génomique de notre «trans-eQTL» est différente de la synténique correspondante chez le rat et/ou chez les humains. Cela indique que nous pourrions avoir détecté dans notre panel de souris RIS un régulateur maître qui est différent de la régulation d'un grand nombre des mêmes gènes chez les rats et/ou humains.

Idéalement, les principaux régulateurs au sein d'un «hotspot» de trans-eQTLs peuvent correspondre à un cis-eQTL. Contrairement aux données de nos collègues, l'analyse de

nos données nous a permis de détecter un cis-eQTL au locus du trans-eQTL. Ce cis-eQTL (*Ypel5*) est différent de celui du régulateur putatif de la trans-eQTL chez le rat. On ne sait rien à propos de la fonction du gène *Ypel5*, mais il a été extraordinairement conservé au cours de l'évolution ; en effet, on le trouve chez tous les eucaryotes, et il affiche 100% d'identité en acides aminés chez tous les mammifères (Hosono, 2004), ce qui suggère des fonctions importantes pour ce gène. À ce stade, nous ne savons pas si un phénotype cardiaque pourrait être affecté par des changements dans l'expression de *Ypel5*. Cependant, une étude d'association chez l'homme indique que le gène *Ypel5* a été associé à des maladies auto-immunes [191]. Toutefois, il pourrait être associé à d'autres phénotypes, car les données affichées sur le site «phenotype-genotype Integrator » indiquent que des associations ont été signalées pour la maladie coronarienne et l'hypertension artérielle pour ce gène. Compte tenu de l'abondante littérature documentant les liens entre les voies de l'inflammation et les maladies cardiovasculaires, nous pensons qu'il est justifié de poursuivre les analyses des fonctions de ce gène.

6.4 Conclusion

Nous avons utilisé iBMQ pour détecter des «hotspots » de trans-eQTLs dans les tissus cardiaques de souris. Parmi les 10 «hotspots » que nous avons détectés, nous avons étudié de façon plus détaillée un «hotspot » qui se trouve sur le chromosome 17, et qui est enrichi de façon très significative pour des gènes impliqués dans la réponse inflammatoire et la réponse aux interférons. En particulier, le gène *Ypel5* est un cis-eQTL se trouvant au même locus que le trans-eQTL. De par sa localisation, ce gène est donc un régulateur potentiel des autres gènes contenus dans le «hotspot» ; contrairement aux traits mendéliens, les traits complexes sont généralement considérés comme entraînant des modifications dans l'expression de plusieurs gènes. Cependant, les mécanismes expliquant comment les variantes génétiques peuvent conduire à la dérégulation coordonnée de plusieurs gènes ne sont pas clairs. Ce court chapitre nous donne l'occasion d'explorer les mécanismes alternatifs et complémentaires qui peuvent nous amener à une meilleure compréhension de l'architecture génétique des caractères complexes.

CHAPITRE 7

DISCUSSION ET CONCLUSION

7.1 Discussion

L'intérêt des études de traits complexes peut se situer à divers niveaux, incluant des intérêts d'ordre économique (comme par exemple pour la sélection du bétail et pour l'agriculture), d'ordre fondamental (comme par exemple pour l'étude de l'évolution moléculaire) et d'ordre médical, car les maladies humaines les plus communes peuvent être considérées comme des traits complexes. Une des motivations premières de cette thèse était de mieux comprendre les déterminants de la masse du ventricule gauche. En effet, la masse ventriculaire gauche est un trait héritable qui a une haute valeur prédictive pour la mortalité et morbidité cardiovasculaire. Ainsi, notre laboratoire avait précédemment identifié, dans des croisements génétiques de souris, des QTLs liés à la MVG. Cependant, une des limites des études de QTLs phénotypiques est qu'ils ne constituent qu'une première étape dans l'identification des variants génétiques responsables du phénotype.

Il est donc devenu nécessaire de développer d'autres approches expérimentales pour mieux identifier les mutations causales. Plus récemment, suite au développement des méthodes de profilage d'expression à haut débit, il est devenu possible de détecter des «QTLs d'expression» (eQTL) en considérant les niveaux d'expression tissulaire de milliers de gènes comme des traits quantitatifs. Néanmoins, la majorité des études utilisant les eQTLs visent encore à identifier une mutation causale dans un seul gène. Cette approche ne peut remporter du succès que dans les situations où le gène incriminé a un effet majeur sur le trait complexe, et ne permet donc pas d'élucider les situations où les traits complexes résultent d'interactions entre divers gènes. Cette thèse a pour but de mieux comprendre les traits complexes en mettant l'accent sur la contribution de l'expression d'un groupe de gènes à un trait complexe. Nos diverses contributions (soit méthodologiques, soit expérimentales) peuvent se résumer comme suit :

1. Nous avons montré qu'une analyse multivariée permet de mieux identifier les

«hotspots» de trans-eQTLs. Ceux-ci correspondent à des situations où un locus particulier est lié à l'expression de nombreux autres gènes dont le locus provient de toute autre partie du génome. Tel que discuté, nous avons des évidences préliminaires indiquant que certains de ces «hotspots» peuvent correspondre à des fonctions biologiques particulières. Notre travail a également permis de mettre à la disposition de la communauté scientifique un outil informatique permettant l'analyse multivariée des eQTLs. Ceci montre l'importance d'une bonne modélisation pour l'analyse des données à haute dimensionnalité. Il est à noter que le temps de calcul de notre méthode est plus élevé que certains outils non-bayésiens, mais nous croyons que ce temps de calcul supplémentaire permet un gain intéressant pour la détection des «hotspots» de trans-eQTLs.

2. Nous avons mis en évidence un mécanisme possible par lequel des polymorphismes non codants peuvent affecter l'expression coordonnée de plusieurs gènes voisins.

3. Nous avons montré que des domaines chromosomiques peuvent faire varier l'expression des modules de co-expression de gènes.

4. En utilisant une combinaison d'analyse de eQTLs et d'approche de réseau dans des tissus cardiaques, nous avons identifié (au sein d'un domaine chromosomique) un groupe de gènes candidats liés collectivement à la masse du ventricule gauche.

Dans les prochains paragraphes, je discuterai plus en détails de quelles manières certaines approches bio-informatiques peuvent faciliter la recherche de déterminants génétiques de traits complexes.

7.1.1 Les eQTLs comme phénotypes intermédiaires

Dans cette thèse, nous avons particulièrement exploré le potentiel des eQTLs comme des phénotypes intermédiaires pour aider à la compréhension des traits complexes. Ce type d'expérience est particulièrement faisable dans les modèles génétiques animaux étant donné l'accessibilité des tissus expérimentaux. Les eQTLs peuvent être classifiés

en deux catégories : les cis-eQTLs (aussi appelés eQTLs proximaux ou locaux) et les trans-eQTLs (aussi appelés eQTLs distaux). Il est à noter que la différence entre les cis-eQTLs et les trans-eQTLs est définie par un seuil de distance, mais que cette procédure est en partie arbitraire. Pour notre étude, nous avons défini les cis-eQTLs comme étant les eQTLs dont le pic était situé à moins de 1 Mb du site de début de trnascription du gène correspondant. Fait à noter, les études de séquences d'ARN (RNA-Seq) permettent présentement de définir les cis-eQTLs (et de les différencier de façon plus rigoureuse des trans-eQTLs) en étudiant l'effet de l'expression d'allèles hétérozygotes. Cependant, cette approche ne serait pas possible avec les souris de notre population, car elles sont homozygotes et ont donc deux allèles identiques à chaque locus.

Plusieurs études ont montré que les analyses de eQTLs peuvent être utiles pour mieux cerner les gènes qui sous-tendent les locus génétiques liés à des traits complexes [165] [166]. Par exemple, on peut trouver des gènes associés à des caractères complexes en identifiant des «cis-eQTLs» au même emplacement qu'un QTL phénotypique et dont les niveaux d'expression corrèlent avec le phénotype d'intérêt. Ces cis-eQTLs ont été appelés «c3-eQTLs» (pour Cis-eQTL Colocalizing with trait QTL and Correlating with the trait) et ont été utilisés pour hiérarchiser les gènes candidats dans les études de traits complexes. Contrairement à certains exemples de cis-eQTLs [58, 177, 178], il n'y a jusqu'à présent que très peu d'exemples où des trans-eQTLs sont liés à des traits complexes [192].

Chez l'humain, il est généralement considéré que les trans-eQTLs sont plus difficiles à détecter que les cis-QTLs [66]. Ceci est dû, en partie, au fait que la région génomique intérrogée pour détecter des cis-eQTLs est petite, contrairement aux trans-eQTLs qui peuvent être associés à des gènes situés n'importe où dans le génome ; en conséquence, les tests requis pour détecter des effets distaux sont soumis à une pénalité statistique plus grande en raison de la multiplicité des comparaisons statistiques. De plus, les trans-eQTLs détectés jusqu'à présent semble avoir un plus petit effet que les cis-eQTLs [66]. La plupart des eQTLs rapportés dans la littérature ont été détectés en utilisant des méthodes classiques d'analyse univariée. Cependant, ce type d'outil analyse seulement «un gène à la fois» , et ne tient donc pas compte des interactions possibles entre les gènes

et/ou d'autres effets combinatoires possibles. Donc, en ayant un outil plus approprié, on pourrait augmenter la puissance de détection de eQTLs où l'effet de locus génomiques sur l'expression des gènes est relativement faible

Ces considérations nous ont menés à développer un outil, iBMQ, permettant l'étude multivariée des eQTLs (chapitres 2 et 3). Notre approche repose sur un modèle hiérarchique bayésien qui utilise conjointement des données génétiques et l'expression de gènes pour l'analyse des eQTLs. Nous avons choisi d'utiliser une approche bayésienne, car cette méthode statistique a l'avantage de combiner plusieurs signaux et de permettre une meilleure détection des interactions faibles. Les modèles bayésiens sont également une solution intéressante aux problèmes de haute dimensionnalité des données et ils fournissent un moyen efficace pour surmonter le fardeau de calcul imposé par les tests multiples.

7.1.2 iBMQ : notre nouvel outil bio-informatique

Jusqu'à maintenant, plusieurs stratégies ont été proposées pour effectuer des analyses de eQTLs. Leurs caractéristiques respectives et la manière par laquelle ils se comparent à notre outil iBMQ sont résumées ci-dessous.

Tel que discuté dans le chapitre 2, iBMQ offre plusieurs avantages de performance et/ou d'utilisation par rapport aux méthodes qui étaient disponibles jusqu'ici. Ceci inclut R/qtl [103] (l'analyse univariée de QTLs la plus fréquemment utilisée), QTLBIM [109] et BAYES [110] (deux méthodes bayésiennes), et Sparse Partial Least (SPLS) [70] (une technique de régression regroupant les gènes en fonction de leurs profils d'expression.

Pendant que le manuscrit décrivant notre travail était en processus de révision, une autre équipe a également publié un modèle hiérarchique bayésien, HESS, pour la détection des eQTLs [107]. Ce modèle se distingue par sa façon de décomposer la sélection de probabilités d'inclusion d'un marqueur qui incorpore à la fois une composante hiérarchique (comme notre modèle), mais aussi une composante qui permet de prendre en compte d'autres structures et/ou différents types d'informations (par exemple des informations biologiques sur la nature des gènes, comme les régions codantes vs non codantes, la localisation des gènes, etc.). Cette deuxième composante a l'avantage d'in-

clure des informations biologiques pour aider à trouver des régulateurs clés dans le cas de détection de «hotspots» de trans-eQTLs. De façon semblable à notre travail, les concepteurs de HESS ont comparé les performances de leur modèle aux logiciels BAYES [110], MOM [100] et (SPLS) [70], et ont montré que HESS détectait mieux les «hotspots » de trans-eQTLs que les autres méthodes. Pour des raisons chronologiques évidentes, nous n'avons pas pu comparer iBMQ à HESS. Par ailleurs, HESS est un logiciel qui n'est pas encore disponible publiquement, et est programmé pour une plateforme d'utilisation non publique. Ces mêmes problèmes existent pour d'autres logiciels qui proposent des solutions pour la haute dimensionnalité, incluant MOM [100], SBR [102] et VBQTL [101]) . Pour notre part, nous avons programmé l'outil iBMQ en langage R, et l'avons rendu disponible à la communauté scientifique par le biais de la plate-forme Bioconductor. Pour l'efficacité des calculs, le code source de iBMQ a été écrit en langage C, mais l'enveloppe en code R rend son utilisation plus conviviale. Un des avantages du langage R est que les résultats obtenus peuvent ensuite être intégrés à d'autres outils disponibles en langage R, incluant des outils pour la visualisation graphique, la gestion de bases de données et/ou des requêtes d'annotations.

Plusieurs extensions techniques permettraient d'améliorer notre modèle d'analyse. Tel qu'illustré dans le chapitre 2, une des faiblesses de notre modèle est son incapacité à détecter des eQTLs lorsque la corrélation non génétique entre l'expression des gènes est plus grande que celle due à des effets génétiques. Une solution à ce problème serait de détendre l'hypothèse d'indépendance des erreurs dans notre modèle. Dans la pratique, il s'agit d'un défi irréalisable en terme de temps de calcul. Une solution envisageable serait de modéliser cette co-expression avec une distribution de type «Wishart inverse»(une généralisation de Gamma inverse), tel que décrit par Bottolo *et al.*[107] et Petretto et al. [73]. Cette approche fonctionne bien dans les modèles où un indicateur d'association est commun à tous les gènes. Cependant, comme il y a dans notre modèle beaucoup de variables à calculer, cette approche risque de demander des temps de calculs trop grands. Comme alternative, nos travaux futurs consisteront à intégrer les corrélations entre les gènes en tenant compte des blocs de gènes et d'avoir une variable indicateur d'association pour le bloc de gènes. Ces blocs de gènes pourraient par exemple être des

gènes appartenant tous à des voies particulières de régulation. Ceci pourrait améliorer la détection des associations faibles masquées par de fortes corrélations fonctionnelles, tout en réduisant le nombre de variables à gérer.

Les simulations au chapitre 2 ont également montré que la corrélation entre les SNPs pouvait affecter la détection des eQTLs, mais dans une moins grande mesure que la corrélation entre les gènes. Au niveau biologique, cette corrélation entre les SNPs est due au déséquilibre de liaison c'est-à-dire l'association non aléatoire entre des SNPs. Dans nos travaux futurs, deux directions seront envisagées : 1) seulement calculer une probabilité d'inclusion pour les blocs de SNPs qui sont en déséquilibre de liaison et ainsi réduire le nombre de variables à gérer et donc réduire le temps de calcul. 2) Au contraire, prendre en compte la structure de corrélation entre les SNPs pour pouvoir mieux détecter le vrai SNP causal.

Ultimement, notre but est de développer une approche plus globale et unifiée pour l'analyse des traits complexes. L'idéal serait donc d'avoir un modèle qui incorpore les données génétiques, transcriptomiques (expression de gènes) et phénotypiques (ou cliniques). Ce type d'approche pourrait permettre, entre autres, la détection de «hotspots» de trans-eQTLs liés à des traits complexes. Avec notre modèle actuel, le plus simple serait d'analyser les phénotypes cliniques et les expressions de gènes ensemble. Comme en règle générale le nombre de phénotypes à l'étude est petit par rapport au nombre de gènes, cette approche ne serait pas idéale. Les données d'expressions de gènes étant plus nombreuses, ceux-ci auraient plus de poids que les phénotypes cliniques pour affecter les probabilités d'inclusions. Le mieux serait de faire un modèle avec deux niveaux de hiérarchie, un premier niveau pour détecter les marqueurs liés aux phénotypes et un deuxième pour trouver les marqueurs liés aux expressions de gènes.

7.1.3 Intérêts biologiques des «hotspots » de trans-eQTLs

D'un point de vue biologique, notre outil semble avoir une utilité particulière pour la détection de «hotspots» de trans-eQTLs. Tel qu'illustré dans le chapitre 6, l'analyse des eQTLS par iBMQ nous a permis de détecter 10 «hotspots» de trans-eQTLs contenant plus de 20 gènes dans des données d'expression de tissu cardiaque d'une population de

souris. Pour la majorité d'entre eux, nous avons détecté un cis-eQTL au même locus. Ceci est intéressant, car les cis-eQTLs pourraient être des régulateurs de l'expression de plusieurs gènes «trans-eQTLs» , par exemple si ces cis-eQTLs sont des facteurs de transcription se liant à la région régulatrice des gènes trans-eQTLs. Une validation biologique est présentement en cours dans notre laboratoire. Nous devons toutefois garder à l'esprit que ces cis-eQTLs ne sont pas les seuls régulateurs possibles. D'autres éléments, tels que les microARN et/ou les marques épigénétiques (comme la méthylation de l'ADN et l'acétylation des histones) sont des exemples de modifications autres que des protéines codantes et ayant la capacité de réguler l'expression de plusieurs gènes. Par exemple, Parikh et ses collègues ont identifié le microARN *miR-21* comme un candidat pour la régulation d'un réseau macromoléculaire dans l'hypertension artérielle pulmonaire [193].

7.1.4 Trouver les déterminants génétiques des traits complexes

Ces dernières années, beaucoup d'espoirs ont été mis dans les études de liaison et d'association pour élucider les déterminants génétiques des traits complexes. Bien que plusieurs de ces études ont permis l'identification de locus liés ou associés à des traits quantitatifs complexes, les variations trouvées jusqu'ici (principalement des SNPs) n'augmentent que faiblement le risque de maladie, et collectivement n'expliquent qu'une faible partie de l'héritabilité. Plusieurs hypothèses ont été émises pour expliquer le faible taux de succès et l'héritabilité manquante des études d'association [194, 195].

1. Dans la majorité des analyses, on a étudié seulement les variations génétiques communes (c'est-à-dire celles présentes dans plus de 5% de la population). Dans un avenir plus ou moins proche, ces études de séquençage (séquençage de l'exome ou séquençage du génome entier) pourraient améliorer cette limitation en permettant d'inclure des variations génétiques plus rares.

2. L'une des difficultés fondamentales dans l'étude des traits complexes est que les phénotypes étudiés semblent dans la plupart des cas être le résultat de plusieurs perturbations génétiques. Dans de telles situations, les effets individuels de chaque mutation pourraient être faibles, et donc difficiles à découvrir. De plus, très peu

d'études jusqu'à maintenant ont pris en compte les possibilités d'interactions épistatiques. L'épistasie (définie comme la contribution non additive de deux locus génétiques à un phénotype) pourrait être un élément expliquant pourquoi les études effectuées à ce jour n'expliquent encore qu'une faible proportion de l'héritabilité des traits complexes, tel que suggéré par une étude récente [196]. Cette étude suggère qu'il n'y aurait pas en fait d'héritabilité manquante dans les maladies complexes : ce serait plutôt la façon de comptabiliser l'héritabilité totale, basée sur l'hypothèse de contributions génétiques additives et sans interactions, qui serait fausse. Cependant, ce type d'analyse est complexe, et est difficile à réaliser dans la pratique avec les capacités actuelles de calcul. Dans la littérature, on retrouve malgré tout quelques exemples d'épistasie, par exemple dans une étude décrivant quatre interactions significatives gène-gène pour la maladie de Crohn, l'hypertension, l'arthrite rhumatoïde, et les troubles de bipolarité [197].

3. Il y a d'autres formes de variations génétiques, telles que les répétitions polymorphiques ou les variations dans le nombre de copies, qui ont été peu explorées jusqu'à présent. Bien que ces variations peuvent avoir des effets importants, les techniques actuelles de génotypage ne sont pas présentement bien adaptées pur détecter ce type de variation.

4. Plusieurs types de mécanismes épigénétiques (qui sont potentiellement héréditaires) pourraient également avoir des influences significatives sur les phénotypes complexes, car ils peuvent moduler l'expression des gènes. Cependant, ces mécanismes n'impliquent pas nécessairement de changements dans la séquence d'ADN, et ne sont donc pas non plus détectables par des techniques de génotypage classique.

5. La majorité des analyses effectuées à ce jour ne considèrent pas les interactions entre les gènes et l'environnement, car celles-ci sont encore mal comprises et difficiles à étudier systématiquement dans les populations humaines. Le génome interagit avec l'environnement d'un certain nombre de façons qui peuvent aboutir parfois à des modifications chimiques semi-permanentes (mutations somatiques,

remodelage de la chromatine, etc.). Les facteurs environnementaux sont également un facteur important pour comprendre les maladies complexes, mais ceux-ci sont difficiles à inclure dans la majorité des études [198].

Ainsi l'identification des gènes et des mécanismes génétiques responsables du signal génétique dans les études d'association reste encore un défi considérable. La stratégie la plus courante consiste à se concentrer sur les gènes adjacents au signal génétique. Cependant, dans la plupart des cas, cette approche empirique n'a pas permis d'identifier une mutation causale responsable du trait complexe. Ainsi, les études GWAS ont été jusqu'à présent très centrées sur les gènes («gene-centric»), non seulement dans l'interprétation, mais aussi dans leur conception, car les puces utilisées pour le génotype sont enrichies en SNPs proches de régions codantes [199]. Cependant, la majorité des SNPs associés à des traits complexes chez les humains tombent dans des régions intergéniques ou introniques [199]. Les résultats du chapitre 4 suggèrent un mécanisme potentiel par lequel des polymorphismes non-codants peuvent affecter l'expression coordonnée de plusieurs gènes voisins. De plus, les résultats du chapitre 5 semblent démontrer que les régions intergéniques sont aussi très importantes pour la compréhension des traits complexes.

Les études d'association représentent une avancée considérable ; on a réussi à identifier des locus associés à des traits complexes, mais jusqu'à maintenant, les résultats ont été décevants dans la prédiction de l'ensemble de risque des maladies. Très peu d'études ont permis de clarifier l'architecture génétique des traits complexes. Donc, pour vraiment comprendre les mécanismes, nous avons besoin d'autres méthodes ou des analyses complémentaires. Les phénotypes molécules intermédiaires, incluant les eQTLs, pourraient représenter une approche complémentaire. Cependant, la logique de ce genre d'étude repose encore sur le concept de gènes uniques dont l'importance est suffisamment grande pour que des variations affectant leur fonction ou leur niveau d'expression affecte un trait quantitatif complexe.

7.1.5 Les réseaux de co-expression

Les gènes n'agissent pas de façon isolée, mais forment des réseaux d'interactions complexes avec d'autres gènes. Ainsi, une perturbation dans le niveau d'expression d'un gène peut se répercuter sur le niveau d'expression d'autres gènes dans le réseau. Ainsi certains ont émis l'hypothèse que les traits complexes, plutôt que de représenter la somme des effets de plusieurs mutations géniques, pourraient en fait résulter de la manière dont plusieurs gènes «normaux» (ou qui n'ont pas nécessairement de mutation génique délétère) interagissent entre eux [95]. Les données d'expression de gènes pourraient donc permettre d'autres types d'analyses que celles qui permettent de détecter des eQTLs. Nous avons donc complémenté nos études par des analyses de réseaux de co-expression de gènes. En effet, la multiplicité des comparaisons statistiques dans des analyses au niveau du génome entier est telle que même des cis-eQTLs intéressants pourraient échapper à l'analyse, alors qu'ils pourraient être plus facile à identifier au sein d'un module corrélé à un trait complexe.

Bien que des méthodes statistiques permettent de détecter des modules de gènes co-exprimés, il n'y a eu que très peu d'études identifiant une cause biologique responsable de tels modules de co-expression. Dans la plupart des études précédentes établissant une corrélation entre un module de co-expression avec des caractères complexes d'intérêt, les études subséquentes se sont limitées à détecter au sein de ces modules des gènes cis-eQTLs. Sur la base des fonctions connues de ces gènes, il y a moyen de faire des hypothèses concernant certains mécanismes possiblement responsables des traits complexes [181]. Une deuxième façon de vérifier la validité d'un module est de regarder si, parmi les SNPs liés aux cis-eQTLs dans les modules, il existe un enrichissement pour des SNPs liés a un trait trouvé par d'autres études de QTL[200]. Ce type de validation a déjà été utilisé pour limiter le nombre de cis-eQTLs à examiner, et ainsi permettre l'identification d'un possible régulateur-clé. Toutefois, la logique de ces études repose encore sur l'identification d'un SNP associé à un gène causal ayant un effet prédominant. Cette approche ne s'écarte pas beaucoup de la logique du «gène unique» .

Une autre stratégie souvent utilisée repose sur la notion qu'il peut être plus facile

de prédire la fonction d'un module que celle de gènes individuels. Toutefois, les inconvénients de cette stratégie sont les suivants : 1) les voies de signalisation sont malheureusement encore souvent incomplètes, et en fait, sont des «simplifications» qui ne représentent pas de façon précise les interactions complexes entre les molécules des réseaux de régulation, 2) les analyses d'enrichissement sont biaisées par rapport à nos connaissances actuelles, et 3) ces analyses ne parviennent pas toujours à fournir des informations sur les relations entre les gènes associés. Finalement, la validité même des modules représente une difficulté : en effet, la construction des réseaux de gènes est basée sur des modèles mathématiques et des données à haut débit qui sont intrinsèquement bruitées.

Une façon d'augmenter notre confiance dans la pertinence et la validité biologique des modules a été de vérifier si l'on pouvait identifier des «module-QTLs» (mQTLs), c'est-à-dire des SNPs associés à des modules dans leur entièreté. Dans la littérature, certains mQTLs [182, 183] ont montré des profils similaires a ceux d'un QTL phénotypique. Ceci suggère que les mêmes déterminants génétiques peuvent être à la fois responsables d'un phénotype et des niveaux d'expression des gènes dans un module. Cependant, les mécanismes par lesquels les niveaux d'expression des gènes dans le module sont régulés restent encore à être clarifiés.

Dans le chapitre 5, nous avons utilisé une approche de réseaux de gènes et de mQTLs pour étudier les déterminants génétiques liés à la masse des ventricules gauches dans une population de souris. Dans un premier temps, nous avons fait une analyse classique d'eQTLs dans le but de découvrir des cis-eQTLs au même locus que le QTL de la MVG (chromosome 13) et nous avons trouvé 8 gènes de type «c3-eQTL» dans un intervalle génomique de 5.8 Mb.

Voici les gènes identifiés et une courte description de leurs fonctions. Le gène *Habp4* (Hyaluronic acid binding protein 4) a pour fonction de lier l'acide hyaluronique. Le gène *Zfp367* (Zinc finger protein 367) a pour fonction e lier l'ADN et est impliqué dans la régulation de la transcription. La protéine codée par le gène *Golm* (Golgi membrane protein 1) est une protéine transmembranaire de type II de Golgi. Les gènes *Aaed1* (AhpC/TSA antioxidant enzyme domain containing 1) et *Testin* n'ont pas de fonction

connue. Le gène *Ccd14b* (Cell division cycle 14 homolog B) est une phosphatase à double spécificité (impliquée dans la phosphorylation de la tyrosine). Des souris mutantes (sans le gène *Cdc14b*) sont viables, mais ces souris ont développé des signes de vieillissement prématuré comparativement à des souris de type sauvage [201].

Les deux candidats les plus intéressants pour la fonction cardiaque sont *Ctsl* et *Ccrk*. *Ctsl* (Cathepsine-L) est un membre de la famille de la cystéine protéase lysozomale, qui participe au remodelage de divers tissus. Des souris mutantes pour le gène *Ctsl* ont montré que ce gène est impliqué dans la réparation et le remodelage post-infarctus du myocarde cardiaque [202]. Le gène *Ccrk* (*Cdk20*, cyclin-dependent kinase 20) a pour fonction de lier ATP. Dans le coeur, un variant d'épissage de *Ccrk* existe. Le variant d'épissage de ce gène candidat exprimant dans le coeur a été montré pour favoriser la croissance des cellules cardiaques et de survie [203].

La base de données BioGPS [204] a révélé que ces cis-eQTLs sont exprimées principalement dans les macrophages et dans d'autres cellules d'origine myéloïde plutôt que dans les cellules cardiaques. Cependant, dans le laboratoire, nous avons réalisé le fractionnement cellulaire des différents types cellulaires du tissu cardiaque des souris A/J et C57/B6 (soit les cellules endothéliales, les macrophages, les fibroblastes et les cardiomyocytes). Pour tous les gènes cis-eQTLs, les cardiomyocytes représentent l'une des fractions cellulaires où ils montrent des niveaux d'expression les plus élevés.

Pour bonifier notre analyse, nous avons aussi fait une analyse de réseau et avons identifié 49 modules de gènes fortement co-exprimés. Le module présentant la plus forte corrélation avec la taille du coeur a les propriétés suivantes : 1) il contient une proportion plus élevée que prévue de gènes appartenant au chromosome 13 (où se trouve le QTL de la MVG) ; 2) il contient parmi ses gènes les plus connectés les 8 gènes «c3-eQTL» détectés précédemment et cinq gènes trans-eQTLs au même locus ; et 3) il a été lié dans son ensemble à mQTL sur le chromosome13 qui à un profil correspondant à celui de la MVG. De plus, la majorité des modules déterminés génétiquement («genetically-driven» modules) sont aussi déterminés de façon prédominante par un domaine chromosomique (chromosome domain-driven module). Les gènes des modules déterminés de façon prédominante par un domaine chromosomique provenaient d'un chromosome

«prédominant» et ne sont pas distribués de manière aléatoire sur l'ensemble du chromosome, mais plutôt dans des régions génomiques plus restreintes qui font en moyenne moins de 20 Mo. La figure 7.1 nous montre deux exemples de distribution des gènes dans des modules déterminés de façon prédominante par un domaine chromosomique. Ces régions chromosomiques restreintes représentent des «domaines» particuliers : 1) les gènes de ces régions chromosomiques ont montré une connectivité beaucoup plus élevée que les gènes du module provenant d'autres chromosomes, et 2) ces régions ont montré un enrichissement pour les variations structurelles et les rétrotransposons de type SINE polymorphes entre les souches.

Nous nous sommes demandés ce qui pourrait coordonner l'expression de plusieurs gènes dans des domaines chromosomiques. En utilisant les données de souris et d'humains, Woo et al [187] ont signalé que la co-expression de gènes dans les génomes des mammifères pourrait aussi bien être de courte portée (<1Mb) que de longue portée (>10 Mb). De plus, il existe de plus en plus de preuves que l'organisation chromosomique est impliquée dans la régulation de l'expression des gènes [188]. Des expériences de «Hi-C», méthode qui permet de cartographier la proximité spatiale entre des chromosomes, ont permis récemment de révéler l'architecture tridimensionnelle de l'ensemble du génome [189]. En plus, Woo et al. [187] ont montré que les régions présentant une proximité spatiale chevauchaient souvent des domaines de co-expression qu'ils ont détectés. Ceci suggère que la co-expression observée dans les domaines chromosomiques pourrait être due à des possibles modifications de la chromatine. Les SINEs (qui ont montré un enrichissement à la fois dans les groupes cis-eQTLs et dans le domaine chromosomique plus large rapporté ici) montrent un enrichissement pour les motifs du site de liaison pour le facteur CTCF (CCCTC binding factor), responsable de l'organisation de la chromatine [151]. Nos données étendent ces concepts en montrant que : 1) les variantes structurelles dans des domaines chromosomiques pr'esentant un grand nombre de SINEs polymorphes correspondent à des changements dans la régulation des gènes au sein des domaines et 2) que de tels changements dans l'expression des gènes au niveau d'un domaine chromosomique peuvent être en corrélation étroite avec un caractère quantitatif complexe, dans ce cas la masse du ventricule gauche.

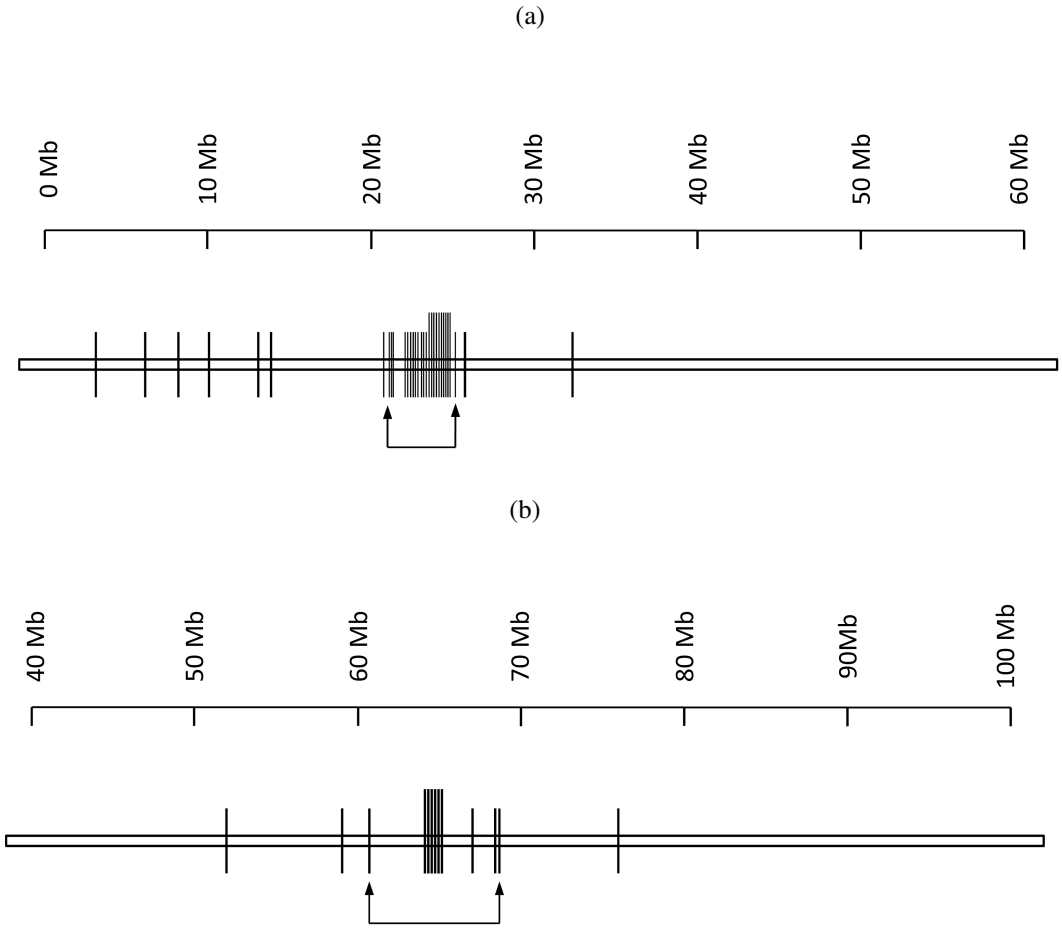


Figure 7.1 – Distribution chromosomique de gènes pour le module plum2 (figure du haut ; chromosome 17) et le module thistle2 (figure du bas ; chromosome 13). Dans les deux cas, il s'agit de modules où un chromosome particulier contribue à un nombre disproportionnellement élevé de gènes du module. On observe que les gènes proviennent d'un intervalle < 30 Mb, et sont particulièrement concentrés dans une région où les gènes appartiennent également à un «cluster» de cis-eQTL (représentés par des lignes verticales plus grandes). Les flèches définissent un intervalle où se retrouvent la majorité des gènes les plus connectés du module.

Les outils de génotypage à haut débit présentement disponibles sont généralement optimisés pour la détection de SNPs. Cependant, nos résultats indiquent que des approches plus globales pourraient être mieux adaptées à l'analyse des traits complexes. Ainsi, il pourrait être important de mieux identifier les variants structuraux, qui ne sont

pas détectés de façon optimale par les techniques actuelles. Les changements de conformation de chromatine pourraient également jouer des rôles importants. Malheureusement, les techniques de criblage à haut débit permettant de détecter des changements de conformation chromosomique ne sont encore qu'à leurs débuts. Néanmoins, nos résultats suggèrent que des variations structurelles peuvent être associées à l'organisation de territoires chromosomiques. Une manière de progresser serait de pouvoir mieux comprendre de quelle manière certaines variations structurelles peuvent mener à des changements au sein de domaines chromosomiques.

7.1.6 Validations biologiques

Bien que nos analyses bio-informatiques aient permis de générer de nouvelles hypothèses, du travail supplémentaire en laboratoire reste nécessaire pour valider ces résultats. Certaines des expériences en cours dans le laboratoire sont détaillées ci-dessous.

1. **Validation des «hotspots de trans-eQTLs »**

   Dans le chapitre 6, nous avons utilisé iBMQ pour détecter des «hotspots » de trans-eQTLs dans les tissus cardiaques de souris. Parmi les 10 «hotspots » que nous avons détectés, nous avons étudié de façon plus détaillée un «hotspot» qui se trouve sur le chromosome 17, et qui est enrichi de façon très significative pour des gènes impliqués dans la réponse inflammatoire et la réponse aux interférons. En particulier, le gène *Ypel5* est un cis-eQTL se trouvant au même locus que le trans-eQTL. De par sa localisation, ce gène est donc un régulateur potentiel des autres gènes contenus dans le «hotspot», ce qui peut être confirmé par d'autres expériences biologiques. Par exemple, on pourrait tester dans une culture de cellules connues pour exprimer ces gènes (comme une culture de macrophages), si la surexpression ou l'inactivation de l'expression du gène *Ypel5* affectent l'expression d'autres gènes contenus dans le «hotspot» de trans-eQTLs. Nous n'avons pas encore identifié de phénotype cardiaque dont le QTL chevauche le locus lié à *Ypel5* et au «hotspot» de trans-eQTLs.

2. **Validation des «clusters» de cis-eQTLs**

Dans le chapitre 4, nous avons montré que des domaines de co-expression de gènes contenant plus de 3 gènes pouvaient être détectés chez les mammifères et que tous les gènes dans ces domaines sont liés à la région de leur lieu propre, formant ainsi des «clusters» de cis-eQTLs. En effectuant des comparaisons avec d'autres régions du génome, nous avons constaté que les régions liées au domaine de co-expression sont caractérisées par un fort enrichissement en rétrotransposons SINEs qui sont polymorphes entre les deux souches parentales du croisement de type RIS. Nous avons donc formulé l'hypothèse que des éléments structuraux polymorphiques pourraient influencer l'expression de plusieurs gènes au sein de certains domaines. La présence de ces éléments structuraux pourrait changer la conformation de la chromatine et ainsi affecter l'expression de plusieurs gènes contigus.

Pour valider cette hypothèse, on pourrait faire des expériences de «chromosome conformation capture » . Cette technique permet de comparer la conformation de la chromatine entre des souches pour ces variations structurales polymorphiques. Comme nous avons également vu qu'il y avait plus de sites CTCF près des domaines de co-expression et que le motif de liaison au CTCF était enrichi au sein des éléments structuraux polymorphiques, on pourrait également faire des expériences de type chip-Seq pour le facteur CTCF et regarder s'il y a une différence entre les souches. Si effectivement les éléments transposables polymorphiques ont un effet sur la régulation des gènes, des technologies pour les découvrir au niveau du génome entier devraient être développées.

3. **Validation des gènes candidats dans le la MVG**

En plus d'avoir montré l'existence de «clusters» de cis-eQTLs, nous avons pu établir qu'un de ces «clusters» pouvait être lié à un QTL dont le profil est très semblable à un QTL lié à un trait quantitatif complexe, soit la MVG (chapitre 5).

Ainsi, le locus lié à la MVG sur le chromosome 13 est lié à un groupe de 6 gènes cis-eQTLs contigus, qui se caractérisent par ailleurs par de forts niveaux de co-expression et une corrélation élevée avec la variation des valeurs de la MVG. Il reste encore beaucoup de travail à faire au niveau de la validation biologique des gènes candidats pour la MVG. Les travaux futurs dans le laboratoire tenteront de répondre aux questions suivantes : 1) Est-ce que la co-expression est due à l'activité d'un gène en particulier ? 2) Est-ce que la co-expression peut être le résultat d'un changement structural au niveau du domaine de co-expression du chromosome 13, par exemple un changement de la chromatine ? 3) Est-ce qu'un gène en particulier peut causer un changement dans la MVG ? 4) Est-ce que les gènes contribuent collectivement ou indépendamment au changement dans la MVG ? 5) Quelle est la fonction des trans-eQTLs ? Nous espérons qu'avec la validation de ces gènes nous pourrons mieux comprendre les facteurs de risques pour les maladies cardiovasculaires.

7.1.7 Travaux futurs

En plus des validations biologiques, il y a d'autres travaux pour exploiter les résultats obtenus de cette thèse.

1. **Concept de réseau et mQTL**

   Un nouvel aspect de nos recherches est qu'une proportion importante des modules identifiés dans les coeurs de souris RIS AXB / BXA contient un nombre disproportionnellement élevé de gènes provenant d'un même chromosome, et que ce dernier correspond au chromosome où est localisé le mQTL. De plus, ces gènes ne sont pas distribués de façon aléatoire dans le chromosome, mais proviennent d'une région génomique plutôt restreinte ; sa taille est en moyenne moins de 20 Mb alors que la taille moyenne des chromosomes de souris est de 142 Mb. Les gènes provenant de ce chromosome «prédominant» ont des valeurs de connectivité beaucoup plus élevées que celles des gènes du module provenant d'autres chromosomes. Par conséquent, nous avons considéré la possibilité que les modules

correspondants étaient causés par un «domaine chromosomique» . Cette notion de «domaine chromosomique» est encore renforcée par l'observation qui montre un enrichissement pour des variations structurelles ainsi que des SINEs polymorphes.

Nous pensons pouvoir utiliser cette observation pour pouvoir mieux caractériser certains modules, comprendre leur fonction biologique, et obtenir des informations qui vont au-delà de ce qui est possible par des analyses d'enrichissement de termes «Gene Ontology». Dans le futur, nous voulons catégoriser les modules en fonction de leurs concepts fondamentaux de réseau (également connus sous le nom «statistique du réseau»). Les concepts de réseau les plus courants sont la densité, la centralisation et l'hétérogénéité. La densité est une valeur représentant la connectivité moyenne du réseau. La centralisation est une mesure qui prend la valeur 0 si le réseau a une topologie en étoile, et la valeur 1 si tous les noeuds ont la même connectivité. L'hétérogénéité est le coefficient de variation de la connectivité. Au sein des modules, où il existe une sur-représentation de gènes d'un même chromosome, la connectivité qui résulte du phénomène d'organisation d'un domaine chromosomique, pourrait se réflètert au niveau des valeurs de concepts de réseaux. Ceci pourrait permettre de différencier ce type de modules d'autre modules où les gènes interagissent pour des raisons différentes.

2. **Les autres molécules intermédiaires :**

Jusqu'à présent, la majorité des études qui ont analysé des phénotypes moléculaires intermédiaires sont des études de eQTLs. Cependant, la variation de l'expression de gènes codants ne représente qu'une petite partie de la variation phénotypique. Il y a d'autres types de molécules qui peuvent influencer un phénotype, tels que les métabolites, les ARN non-codants et les protéines, Ceux-ci peuvent affecter les fonctions génomiques par des mécanismes indirects (par exemple la méthylation de l'ADN, les modifications des histones, l'épissage et/ou les modifications post-translationnelles). Ainsi, il sera important d'inclure ces divers événements moléculaires dans la liste de phénotypes intermédiaires possibles. Dans un

monde parfait, on pourrait avoir accès à des données concernant plusieurs types de molécules à la fois et concevoir des analyses qui intégreront plusieurs types de données. Ceci ne se fera pas sans difficulté, car comme nous l'avons déjà mentionné, la haute dimensionnalité des données biologiques à haut débit apporte son lot de problèmes.

Ceci est un problème intéressant pour la bio-informatique translationnelle qui a pour but d'identifier des relations entre des données issues de différents niveaux fonctionnels et de comprendre les interactions souvent complexes entre ces différentes composantes. Le volume et la spécificité des jeux de données conduisent à des traitements faisant appel à tout l'éventail méthodologique statistique, tels que l'exploration de données, les tests statistiques pour déterminer les molécules différemment exprimées d'une condition biologique à une autre, la modélisation. La plupart des efforts se rapportent à l'analyse d'un seul type de données ; d'autres efforts pour adapter et développer les outils statistiques pour l'approche intégrative de la biologie sont rendus indispensables pour interpréter les résultats des données complexes.

Il y a quelques améliorations qui pourraient être faites au logiciel iBMQ pour pouvoir intégrer d'autres types de données. Par exemple, on pourrait analyser des données haut débit comme la méthylation de l'ADN. De plus, on pourrait aussi intégrer des phénotypes cliniques pour trouver des «hotspots» associés à des traits complexes.

7.1.8 Des souris et des hommes

Nous savons que 99% des gènes de la souris ont un équivalent chez l'homme [8], ce qui rend les souris un outil intéressant pour étudier la fonction des gènes humains dans le contexte de maladies telles que les maladies cardiovasculaires. Cependant, même si le génome est semblable, les deux espèces ne l'utilisent pas de la même façon. Comme nos études effectuées ont été réalisées chez la souris, nous sommes en droit de nous demander si nos découvertes sont directement ou partiellement transposables chez l'hu-

main. Nos expériences ont été réalisées dans des croisements de souris de type RIS. Les RIS sont un croisement génétique particulier : comme il n'y a que 2 lignées parentales à l'origine du croisement, les variations structurales polymorphiques ne proviennent que de ces deux mêmes lignées. Si les «clusters» de cis-eQTLs observés sont effectivement dus à des variations polymorphiques entre les deux souches parentales, il est fort probable que ce type de regroupement soit plus facile à détecter dans les croisements de type RIS. On peut supposer que les variations structurales polymorphiques auront également un rôle important dans la régulation des groupes de gènes chez l'humain, mais comme on retrouve plus de variations génomiques dans une population humaine, les effets possibles des variations structurales seront peut-être plus difficiles à détecter. Il serait donc utile d'effectuer des analyses similaires avec des données provenant de populations humaines ou pour tester la faisabilité avec des jeux de données provenant de populations de souris présentant plus de diversité génétique(comme par exemple en utilisant des souris provenant du «Mouse Heterogeneous Stock»). La validation moléculaire chez l'humain pourrait malgré tout être plus complexe, car nous n'avons pas accès aussi facilement à des tissus cardiaques.

7.2 Conclusion

Les études présentées dans cette thèse ont permis d'explorer deux voies qui pourraient mener à une meilleure compréhension des déterminants génétiques des traits complexes, soit : 1) la détection de «hotspots» de trans-eQTLs, et 2) l'existence de modules de co-expression résultant de domaines chromosomiques. Le logiciel iBMQ nous a permis de mieux détecter les «hotspots» de trans-eQTLs, ce qui pourrait mener à la découverte de mutations-clés qui régulent plusieurs gènes. Un exemple est celui du «hotspot» de trans-eQTL détecté sur le chromosome 17, où il existe un fort enrichissement pour des gènes liés à l'immunité et la réponse interféron.

Nos approches analytiques ont mis en lumière un autre mécanisme potentiel, où des variations génétiques pourraient affecter l'expression de plusieurs gènes au sein de domaines chromosomiques. Dans le chapitre 4, nous avons montré la présence de «clus-

ters» de cis-eQTLs, que nous avons appelé des domaines géniques de co-expression («co-expression domain»). Par ailleurs, nous avons montré (chapitre 5) que ces domaines géniques de co-expression font partie de modules déterminés de façon prédominante par un domaine chromosomique. Ce mécanisme semble être responsable du locus lié à la masse du ventricule gauche dans une population de souris.

La bio-informatique translationnelle est le champ qui répond aux problématiques actuelles de l'intégration d'un grand nombre de données moléculaires et de données cliniques dans le but de mieux comprendre les bases moléculaires des maladies comme les maladies complexes. Dans le futur, je souhaite pouvoir continuer à utiliser des approches semblables pour y intégrer plusieurs types de données, comme par exemple des données provenant d'études de métabolomique et/ou de protéomique. Ces derniers mois, le projet «Encyclopedia of DNA Elements» (ENCODE) [154] a permis d'identifier plusieurs éléments fonctionnels dans le génome humain, y compris les transcriptions non codantes, les marques de chromatine accessibles et les sites de liaison aux protéines. J'aimerais pourvoir incorporer ces données aux analyses de traits complexes, dans l'espoir d'obtenir une meilleure interprétation fonctionnelle des résultats, et une meilleure compréhension des maladies cardiovasculaires et métaboliques chez l'humain.

BIBLIOGRAPHIE

[1] J D WATSON and F H CRICK. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361) :964–967, May 1953.

[2] J D WATSON and F H CRICK. Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature*, 171(4356) :737–738, April 1953.

[3] A Hamosh, A F Scott, and J S Amberger. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids Res*, 2005.

[4] Matthew B Lanktree, Yiran Guo, Muhammed Murtaza, Joseph T Glessner, Swneke D Bailey, N Charlotte Onland-Moret, Guillaume Lettre, Halit Ongen, Ramakrishnan Rajagopalan, Toby Johnson, Haiqing Shen, Christopher P Nelson, Norman Klopp, Jens Baumert, Sandosh Padmanabhan, Nathan Pankratz, ..., and Brendan J Keating. Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height. *American journal of human genetics*, 88(1) :6–18, January 2011.

[5] K W Broman and S Sen. A Guide to QTL Mapping with R/qtl. 2009.

[6] R W Williams, J Gu, S Qi, and L Lu. The genetic structure of recombinant inbred mice high-resolution consensus maps for complex trait analysis. *Genome Biol.*, 2 (11) :RESEARCH0046, 2001.

[7] J P Rapp. Genetic analysis of inherited hypertension in the rat. *Physiol. Rev.*, 80 (1) :135–172, 2000.

[8] Luanne L Peters, Raymond F Robledo, Carol J Bult, Gary A Churchill, Beverly J Paigen, and Karen L Svenson. The mouse as a model for human biology a resource guide for complex trait analysis. *Nature Reviews Genetics*, 8(1) :58–69, January 2007.

[9] Howard J Jacob and Anne E Kwitek. Rat genetics attaching physiology and pharmacology to the genome. *Nat. Rev. Genet.*, 3(1) :33–42, 2002.

[10] William Valdar, Leah C Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O Cookson, Martin S Taylor, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.*, 38(8) :879–887, August 2006.

[11] David W Threadgill and Gary A Churchill. Ten years of the Collaborative Cross. *Genetics*, 190(2) :291–294, February 2012.

[12] Oduola Abiola, Joe M Angel, Philip Avner, Alexander A Bachmanov, John K Belknap, Beth Bennett, Elizabeth P Blankenhorn, David A Blizard, Valerie Bolivar, Gundrun A Brockmann, Kari J Buck, Jean-Francoise Bureau, William L Casley, Elissa J Chesler, James M Cheverud, Gary A Churchill, ..., and Complex Trait Consortium. The nature and identification of quantitative trait loci a community's view. *Nat. Rev. Genet.*, 4(11) :911–916, 2003.

[13] E S Lander and D Botstein. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1) :185–199, 1989.

[14] K W Broman. Review of statistical methods for QTL mapping in experimental crosses. *Lab Anim (NY)*, 30(7) :44–52, 2001.

[15] C S Haley and S A Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity (Edinb)*, 69(4) :315–324, 1992.

[16] S Sen and G A Churchill. A statistical framework for quantitative trait mapping. *Genetics*, 2001.

[17] G A Churchill and R W Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3) :963–971, 1994.

[18] R W Doerge and G A Churchill. Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1) :285–294, 1996.

[19] R C Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 1993.

[20] Z B Zeng. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.*, 90(23) :10972–10976, 1993.

[21] C H Kao, Z B Zeng, and R D Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 1999.

[22] A Korol, Z Frenkel, O Orion, and Y Ronin. Some ways to improve QTL mapping accuracy. *Animal Genetics*, 2012.

[23] Marjorie Maillet, Jop H van Berlo, and Jeffery D Molkentin. Molecular basis of physiological heart growth fundamental concepts and new players. *Nature reviews. Molecular cell biology*, 14(1) :38–48, December 2012.

[24] Christiana M Schannwell, Tobias Zimmermann, Markus Schneppenheim, Gunnar Plehn, Roger Marx, and Bodo E Strauer. Left ventricular hypertrophy and diastolic dysfunction in healthy pregnant women. *Cardiology*, 97(2) :73–78, 2002.

[25] G D Oakley. The athletic heart. *Cardiology clinics*, 5(2) :319–329, May 1987.

[26] S Penpargkul and J Scheuer. The effect of physical training upon the mechanical and metabolic performance of the rat heart. *The Journal of clinical investigation*, 49(10) :1859–1868, October 1970.

[27] Ola Gjesdal, David A Bluemke, and Joao A Lima. Cardiac remodeling at the population level–risk factors, screening, and outcomes. *Nature reviews. Cardiology*, 8(12) :673–685, December 2011.

[28] Wolfgang Lieb, Vanessa Xanthakis, Lisa M Sullivan, Jayashri Aragam, Michael J Pencina, Martin G Larson, Emelia J Benjamin, and Ramachandran S Vasan. Lon-

gitudinal tracking of left ventricular mass over the adult life course clinical correlates of short- and long-term change in the framingham offspring study. *Circulation*, 119(24) :3085–3092, June 2009.

[29] Bradford C Berk, Keigi Fujiwara, and Stephanie Lehoux. ECM remodeling in hypertensive heart disease. *The Journal of clinical investigation*, 117(3) :568–575, March 2007.

[30] Norbert Frey, Mark Luedde, and Hugo A Katus. Mechanisms of disease hypertrophic cardiomyopathy. *Nature reviews. Cardiology*, 9(2) :91–100, February 2012.

[31] Houman Ashrafian and Hugh Watkins. Reviews of translational medicine and genomics in cardiovascular disease new disease taxonomy and therapeutic implications cardiomyopathies therapeutics based on molecular phenotype. *Journal of the American College of Cardiology*, 49(12) :1251–1264, March 2007.

[32] Ali J Marian. Genetic determinants of cardiac hypertrophy. *Current opinion in cardiology*, 23(3) :199–205, May 2008.

[33] Andrew Sharp and Jamil Mayet. Regression of left ventricular hypertrophy hoping for a longer life. *J Renin Angiotensin Aldosterone Syst*, 3(3) :141–144, 2002.

[34] M J Koren, R B Devereux, P N Casale, D D Savage, and J H Laragh. Relation of left ventricular mass and geometry to morbidity and mortality in uncomplicated essential hypertension. *Annals of internal medicine*, 114(5) :345–352, March 1991.

[35] D E Manyari. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. *The New England journal of medicine*, 323(24) :1706–1707, December 1990.

[36] R B Devereux, G de Simone, A Ganau, and M J Roman. Left ventricular hypertrophy and geometric remodeling in hypertension stimuli, functional consequences and prognostic implications. *J Hypertens Suppl*, 12(10) :S117–27, 1994.

[37] D J Nunez, C P Clifford, S al Mahdawi, and D Dutka. Hypertensive cardiac hypertrophy–is genetic variance the missing link ? *Br J Clin Pharmacol*, 42(1) : 107–117, 1996.

[38] T D Adams, F G Yanowitz, A G Fisher, and J D Ridges. Heritability of cardiac size an echocardiographic and electrocardiographic study of monozygotic and dizygotic twins. *Circulation*, 1985.

[39] H A Verhaaren, R M Schieken, M Mosteller, J K Hewitt, L J Eaves, and W E Nance. Bivariate genetic analysis of left ventricular mass and weight in pubertal twins (the Medical College of Virginia twin study). *Am. J. Cardiol.*, 68(6) :661–668, 1991.

[40] E Bielen, R Fagard, and A Amery. The inheritance of left ventricular structure and function assessed by imaging and Doppler echocardiography. *American heart journal*, 121(6 Pt 1) :1743–1749, June 1991.

[41] Christoph A Busjahn, Jeanette Schulz-Menger, Hassan Abdel-Aty, Andre Rudolph, Jens Jordan, Friedrich C Luft, and Andreas Busjahn. Heritability of left ventricular and papillary muscle heart size a twin study with cardiac magnetic resonance imaging. *European heart journal*, 30(13) :1643–1647, July 2009.

[42] L Swan, D H Birnie, S Padmanabhan, G Inglis, J M C Connell, and W S Hillis. The genetic determination of left ventricular mass in healthy adults. *European heart journal*, 24(6) :577–582, March 2003.

[43] Themistocles L Assimes, Balasubramanian Narasimhan, Todd B Seto, Sangho Yoon, J David Curb, Richard A Olshen, and Thomas Quertermous. Heritability of left ventricular mass in Japanese families living in Hawaii the SAPPHIRe Study. *Journal of hypertension*, 25(5) :985–992, May 2007.

[44] C Garner, E Lecomte, S Visvikis, E Abergel, M Lathrop, and F Soubrier. Genetic and environmental influences on left ventricular mass. A family study. *Hypertension*, 36(5) :740–746, November 2000.

[45] P Palatini, L Krause, J Amerena, S Nesbitt, S Majahalme, V Tikhonoff, M Valentini, and S Julius. Genetic contribution to the variance in left ventricular mass the Tecumseh Offspring Study. *Journal of hypertension*, 19(7) :1217–1222, July 2001.

[46] Jonathan N Bella, Jean W MacCluer, Mary J Roman, Laura Almasy, Kari E North, Lyle G Best, Elisa T Lee, Richard R Fabsitz, Barbara V Howard, and Richard B Devereux. Heritability of left ventricular dimensions and mass in American Indians The Strong Heart Study. *Journal of hypertension*, 22(2) :281–286, February 2004.

[47] Kuo-Liong Chien, Hsiu-Ching Hsu, Ta-Chen Su, Ming-Fong Chen, and Yuan-Teh Lee. Heritability and major gene effects on left ventricular mass in the Chinese population a family study. *BMC cardiovascular disorders*, 6 :37, 2006.

[48] Suh-Hang Hank Juo, Marco R Di Tullio, Hsiu-Fen Lin, Tanja Rundek, Bernadette Boden-Albala, Shunichi Homma, and Ralph L Sacco. Heritability of left ventricular mass and other morphologic variables in Caribbean Hispanic subjects the Northern Manhattan Family Study. *Journal of the American College of Cardiology*, 46(4) :735–737, August 2005.

[49] E Bielen, R B Devereux, Marjorie Maillet, Jonathan N Bella, Xia Yang, S Sen, B M Mayosi, E Bielen, G de Simone, Jop H van Berlo, Jean W MacCluer, G A Churchill, B M Mayosi, R Fagard, A Ganau, Jeffery D Molkentin, Mary J Roman, B Keavney, R Fagard, M J Roman, Laura Almasy, B Keavney, A Amery, Kari E North, A Kardos, A Amery, Lyle G Best, A Kardos, Elisa T Lee, C H Davies, Richard R Fabsitz, C H Davies, Barbara V Howard, P J Ratcliffe, Richard B Devereux, P J Ratcliffe, M Farrall, M Farrall, H Watkins, and H Watkins. Electrocardiographic measures of left ventricular hypertrophy show greater heritability than echocardiographic left ventricular mass. *European heart journal*, 23(24) : 1963–1971, December 2002.

[50] W S Post, M G Larson, R H Myers, M Galderisi, and D Levy. Heritability of left

ventricular mass the Framingham Heart Study. *Hypertension*, 30(5) :1025–1028, November 1997.

[51] Jonathan N Bella and Harald Hh Göring. Genetic epidemiology of left ventricular hypertrophy. *American journal of cardiovascular disease*, 2(4) :267–278, 2012.

[52] H Schunkert, H W Hense, and S R Holmer. Association between a deletion polymorphism of the angiotensin-converting-enzyme gene and left ventricular hypertrophy. *New England Journal of Med.*, 1994.

[53] Tatiana Kuznetsova, Jan A Staessen, Lutgarde Thijs, Christiane Kunath, Agnieszka Olszanecka, Andrew Ryabikov, Valérie Tikhonoff, Katarzyna Stolarz, Giuseppe Bianchi, Edoardo Casiglia, Robert Fagard, Stefan-Martin Brand-Herrmann, Kalina Kawecka-Jaszcz, Sofia Malyutina, Yuri Nikitin, Eva Brand, and European Project On Genes in Hypertension (EPOGH) Investigators. Left ventricular mass in relation to genetic variation in angiotensin II receptors, renin system genes, and sodium excretion. *Circulation*, 110(17) :2644–2650, October 2004.

[54] Y Jamshidi, H E Montgomery, H W Hense, and S G Myerson. Peroxisome proliferator–activated receptor Îś gene regulates left ventricular growth in response to exercise and hypertension. *Circulation*, 2002.

[55] E Poch, D González, E Gómez-Angelats, M Enjuto, J C Paré, F Rivera, and A de La Sierra. G-Protein beta(3) subunit gene variant and left ventricular hypertrophy in essential hypertension. *Hypertension*, 35(1 Pt 2) :214–218, January 2000.

[56] Ramachandran S Vasan, Nicole L Glazer, Janine F Felix, Wolfgang Lieb, Philipp S Wild, Stephan B Felix, Norbert Watzinger, Martin G Larson, Nicholas L Smith, Abbas Dehghan, Anika Grosshennig, Arne Schillert, Alexander Teumer, Reinhold Schmidt, Sekar Kathiresan, Thomas Lumley, Yurii S Aulchenko, ..., and

Stefan Blankenberg. Genetic variants associated with cardiac structure and function a meta-analysis and replication of genome-wide association data. *JAMA*, 302 (2) :168–178, 2009.

[57] Bastien Llamas, Sonia Bélanger, Sylvie Picard, and Christian F Deschepper. Cardiac mass and cardiomyocyte size are governed by different genetic loci on either autosomes or chromosome Y in recombinant inbred mice. *Physiol. Genomics*, 31 (2) :176–182, 2007.

[58] Enrico Petretto, Rizwan Sarwar, Ian Grieve, Han Lu, Mande K Kumaran, Phillip J Muckett, Jonathan Mangion, Blanche Schroen, Matthew Benson, Prakash P Punjabi, Sanjay K Prasad, Dudley J Pennell, Chris Kiesewetter, Elena S Tasheva, Lolita M Corpuz, Megan D Webb, Gary W Conrad, Theodore W Kurtz, Vladimir Kren, Judith Fischer, Norbert Hubner, Yigal M Pinto, Michal Pravenec, Timothy J Aitman, and Stuart A Cook. Integrated genomic approaches implicate osteoglycin (Ogn) in the regulation of left ventricular mass. *Nat. Genet.*, 40(5) :546–552, 2008.

[59] P M Visscher, M A Brown, and M I McCarthy. Five years of GWAS discovery. *American journal of human genetics*, 2012.

[60] J S Witte. Genome-wide association studies and beyond. *Annual review of public health*, 2010.

[61] Jennifer F Kugel and James A Goodrich. Non-coding RNAs : key regulators of mammalian transcription. *Trends in biochemical sciences*, 37(4) :144–151, April 2012.

[62] Ryan Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor Pugh, Helen McDonald, Richard Varhol, Steven Jones, and Marco Marra. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1) :81–94, July 2008.

[63] Gisele Passador-Gurgel, Wen-Ping Hsieh, Priscilla Hunt, Nigel Deighton, and Greg Gibson. Quantitative trait transcripts for nicotine resistance in Drosophila melanogaster. *Nat. Genet.*, 39(2) :264–268, 2007.

[64] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568) : 752–755, 2002.

[65] Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, Peter S Linsley, Mao Mao, Roland B Stoughton, and Stephen H Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929) : 297–302, 2003.

[66] Yoav Gilad, Scott A Rifkin, and Jonathan K Pritchard. Revealing the architecture of gene regulation the promise of eQTL studies. *Trends in genetics TIG*, 24(8) : 408–415, August 2008.

[67] Jacek Majewski and Tomi Pastinen. The study of eQTL variations by RNA-seq from SNPs to phenotypes. *Trends in genetics TIG*, 27(2) :72–79, February 2011.

[68] Stephen B Montgomery and Emmanouil T Dermitzakis. From expression QTLs to personalized transcriptomics. *Nature Reviews Genetics*, 12(4) :277–282, March 2011.

[69] Lun Li, Xianghua Zhang, and Hongyu Zhao. eQTL. *Methods in molecular biology (Clifton, N.J.)*, 871 :265–279, 2012.

[70] Hyonho Chun and Sündüz Keles. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, 182(1) :79–90, 2009.

[71] Nengjun Yi, Brian S Yandell, Gary A Churchill, David B Allison, Eugene J Eisen, and Daniel Pomp. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 170(3) :1333–1344, 2005.

[72] Samprit Banerjee, Brian S Yandell, and Nengjun Yi. Bayesian quantitative trait loci mapping for multiple traits. *Genetics*, 179(4) :2275–2289, 2008.

[73] Enrico Petretto, Leonardo Bottolo, Sarah R Langley, Matthias Heinig, Chris McDermott-Roe, Rizwan Sarwar, Michal Pravenec, Norbert Hubner, Timothy J Aitman, Stuart A Cook, and Sylvia Richardson. New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol.*, 6(4) :e1000737, 2010.

[74] Chenwu Xu, Xuefeng Wang, Zhikang Li, and Shizhong Xu. Mapping QTL for multiple traits using Bayesian statistics. *Genet Res (Camb)*, 91(1) :23–37, 2009.

[75] Liang Chen, Tiejun Tong, and Hongyu Zhao. Considering dependence among genes and markers for false discovery control in eQTL mapping. *Bioinformatics*, 24(18) :2015–2022, September 2008.

[76] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461 (7265) :747–753, 2009.

[77] Y A Kim and T M Przytycka. Bridging the gap between genotype and phenotype via network approaches. *Frontiers in Genetics*, 2012.

[78] D Y Cho, Y A Kim, and T M Przytycka. Network Biology Approach to Complex Diseases. *PLOS Computational Biology*, 2012.

[79] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95 (25) :14863–14868, December 1998.

[80] P D'haeseleer, S Liang, and R Somogyi. Genetic network inference from co-expression clustering to reverse engineering. *Bioinformatics*, 2000.

[81] N Friedman, M Linial, I Nachman, and D Pe'er. Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, 7(3-4) :601–620, 2000.

[82] V Emilsson, G Thorleifsson, B Zhang, and A S Leonardson. Genetics of gene expression and its effect on disease. *Nature*, 2008.

[83] S Horvath and J Dong. Geometric interpretation of gene coexpression network analysis. *PLOS Computational Biology*, 2008.

[84] I King Jordan, Leonardo Mariño-Ramírez, Yuri I Wolf, and Eugene V Koonin. Conservation and coevolution in the scale-free human gene coexpression network. *Molecular biology and evolution*, 21(11) :2058–2070, November 2004.

[85] P Tsaparas and L Mariño-Ramírez. Global similarity and local divergence in human and mouse gene co-expression networks. *BMC evolutionary biology*, 2006.

[86] Albert-László Barabási. Scale-free networks a decade and beyond. *Science*, 325 (5939) :412–413, July 2009.

[87] AL Barabasi and R Albert. Emergence of scaling in random networks. *Science*, 286(5439) :509–512, October 1999.

[88] C Castillo and R Baeza-Yates. Effective web crawling. *ACM SIGIR Forum*, 2005.

[89] Réka Albert. Scale-free networks in cell biology. *Journal of cell science*, 2005.

[90] Peter Langfelder and Steve Horvath. WGCNA an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1) :559, 2008.

[91] Trudy F C Mackay, Eric A Stone, and Julien F Ayroles. The genetics of quantitative traits challenges and prospects. *Nat. Rev. Genet.*, 10(8) :565–577, August 2009.

[92] Fabrício M Lopes, Roberto M Cesar, and Luciano Da F Costa. Gene expression complex networks synthesis, identification, and analysis. *J. Comput. Biol.*, 18 (10) :1353–1367, October 2011.

[93] Anatole Ghazalpour, Sudheer Doss, Bin Zhang, Susanna Wang, Christopher Plaisier, Ruth Castellanos, Alec Brozell, Eric E Schadt, Thomas A Drake, Aldons J Lusis, and Steve Horvath. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS genetics*, 2(8) :e130, August 2006.

[94] Xianghua Zhang and Hongyu Zhao. An Evaluation of Gene Module Concepts in the Interpretation of Gene Expression Data. In *Frontiers in Computational and Systems Biology*, pages 331–349. Springer London, London, 2010.

[95] James N Weiss, Alain Karma, W Robb MacLellan, Mario Deng, Christoph D Rau, Colin M Rees, Jessica Wang, Nicholas Wisniewski, Eleazar Eskin, Steve Horvath, Zhilin Qu, Yibin Wang, and Aldons J Lusis. "Good enough solutions" and the genetics of complex diseases. *Circulation research*, 111(4) :493–504, August 2012.

[96] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, 106(23) :9362–9367, June 2009.

[97] Harald H H Göring, Joanne E Curran, Matthew P Johnson, Thomas D Dyer, Jac Charlesworth, Shelley A Cole, Jeremy B M Jowett, Lawrence J Abraham, David L Rainwater, Anthony G Comuzzie, Michael C Mahaney, Laura Almasy, Jean W MacCluer, Ahmed H Kissebah, Gregory R Collier, Eric K Moses, and John Blangero. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.*, 39(10) :1208–1216, 2007.

[98] J Zhu, M C Wiener, C Zhang, and A Fridman. Increasing the power to detect

causal associations by combining genotypic and expression data in segregating populations. *PLoS computational biology*, 2007.

[99] A L Dixon, L Liang, M F Moffatt, W Chen, and S Heath. A genome-wide association study of global gene expression. *Nature*, 2007.

[100] C M Kendziorski, M Chen, M Yuan, H Lan, and A D Attie. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*, 2006.

[101] O Stegle, L Parts, R Durbin, and J Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.*, 2010.

[102] S Richardson, L Bottolo, and J S Rosenthal. Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics 9*, 2010.

[103] K W Broman, H Wu, S Sen, and G A Churchill. R/qtl QTL mapping in experimental crosses. *Bioinformatics*, 2003.

[104] J Peng, J Zhu, A Bergamaschi, and W Han. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 2010.

[105] J Lucas, C Carvalho, Q Wang, and A Bild. Sparse statistical modelling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics*, 2006.

[106] A Zellner and A Siow. Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, 1980.

[107] Leonardo Bottolo, Marc Chadeau-Hyam, David I Hastie, Sarah R Langley, Enrico Petretto, Laurence Tiret, David Tregouet, and Sylvia Richardson. ESS++ a C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics*, 27(4) :587–588, 2011.

[108] F Liang, R Paulo, and G Molina. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 2008.

[109] Brian S Yandell, Tapan Mehta, Samprit Banerjee, Daniel Shriner, Ramprasad Venkataraman, Jee Young Moon, W Whipple Neely, Hao Wu, Randy von Smith, and Nengjun Yi. R/qtlbim QTL with Bayesian Interval Mapping in experimental crosses. *Bioinformatics*, 23(5) :641–643, 2007.

[110] N Yi and D Shriner. Advances in Bayesian multiple quantitative trait loci mapping in experimental crosses. *Heredity (Edinb)*, 2007.

[111] A E Gelfand and AFM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 1990.

[112] W R Gilks, A E Gelfand, W K Hastings, H A Verhaaren, P Wild, AFM Smith, R M Schieken, M Mosteller, J K Hewitt, L J Eaves, and W E Nance. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 1992.

[113] A E Raftery and S Lewis. How many iterations in the Gibbs sampler. *Bayesian statistics*, 1992.

[114] M A Newton, A Noueiry, D Sarkar, and P Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 2004.

[115] K Y Yeung, C Fraley, A Murua, A E Raftery, and W L Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17 (10) :977–987, 2001.

[116] Eldon E Geisert, Lu Lu, Natalie E Freeman-Anderson, Justin P Templeton, Mohamed Nassr, Xusheng Wang, Weikuan Gu, Yan Jiao, and Robert W Williams. Gene expression in the mouse eye an online resource for genetics using 103 strains of mice. *Mol. Vis.*, 15 :1730–1763, 2009.

[117] BTS Da Wei Huang and R A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 2008.

[118] L Dagum and R Menon. OpenMP an industry standard API for shared-memory programming. *Computational Science & Engineering*, 1998.

[119] R C Gentleman, V J Carey, and D M Bates. Bioconductor open software development for computational biology and bioinformatics. *Genome Biol.*, 2004.

[120] Marie Pier Scott-Boyer, Gregory C Imholte, Arafat Tayeb, Aurelie Labbe, Christian F Deschepper, and Raphael Gottardo. An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical applications in genetics and molecular biology*, 11(4), 2012.

[121] Ian C Grieve, Nicholas J Dickens, Michal Pravenec, Vladimir Kren, Norbert Hubner, Stuart A Cook, Timothy J Aitman, Enrico Petretto, and Jonathan Mangion. Genome-wide co-expression analysis in multiple tissues. *PloS one*, 3(12) :e4033, 2008.

[122] Chunlei Wu, David L Delano, Nico Mitro, Stephen V Su, Jeff Janes, Phillip McClurg, Serge Batalov, Genevieve L Welch, Jie Zhang, Anthony P Orth, John R Walker, Richard J Glynne, Michael P Cooke, Joseph S Takahashi, Kazuhiro Shimomura, Akira Kohsaka, Joseph Bass, Enrique Saez, Tim Wiltshire, and Andrew I Su. Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS genetics*, 4(5) :e1000070, May 2008.

[123] Pawel Michalak. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91(3) :243–248, March 2008.

[124] Leah I Elizondo, Paymaan Jafar-Nejad, J Marietta Clewing, and Cornelius F Boerkoel. Gene clusters, molecular evolution and disease a speculation. *Current genomics*, 10(1) :64–75, March 2009.

[125] Paul T Spellman and Gerald M Rubin. Evidence for large domains of similarly expressed genes in the Drosophila genome. *Journal of biology*, 1(1) :5, 2002.

[126] Martin J Lercher, Araxi O Urrutia, and Laurence D Hurst. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.*, 31(2) :180–183, June 2002.

[127] Marie Sémon and Laurent Duret. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Molecular biology and evolution*, 23(9) : 1715–1723, September 2006.

[128] Antje Purmann, Joern Toedling, Markus Schueler, Piero Carninci, Hans Lehrach, Yoshihide Hayashizaki, Wolfgang Huber, and Silke Sperling. Genomic organization of transcriptomes in mammals Coregulation and cofunctionality. *Genomics*, 89(5) :580–587, May 2007.

[129] Hong Zhang, Kuang-Hung Pan, and Stanley N Cohen. Senescence-specific gene expression fingerprints reveal cell-type-dependent physical clustering of up-regulated chromosomal loci. *Proc. Natl. Acad. Sci. U.S.A.*, 100(6) :3251–3256, March 2003.

[130] J D Marshall, J L Mu, Y C Cheah, M N Nesbitt, W N Frankel, and B Paigen. The AXB and BXA set of recombinant inbred mouse strains. *Mamm. Genome*, 3(12) : 669–680, 1992.

[131] Bastien Llamas, Ricardo A Verdugo, Gary A Churchill, and Christian F Deschepper. Chromosome Y variants from different inbred mouse strains are linked to differences in the morphologic and molecular responses of cardiac cells to post-pubertal testosterone. *BMC genomics*, 10 :150, 2009.

[132] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3 :Article3, 2004.

[133] Hyuna Yang, Yueming Ding, Lucie N Hutchins, Jin Szatkiewicz, Timothy A Bell, Beverly J Paigen, Joel H Graber, Fernando Pardo-Manuel de Villena, and Gary A

Churchill. A customized and versatile high-density genotyping array for the mouse. *Nature methods*, 6(9) :663–666, September 2009.

[134] John P Didion, Hyuna Yang, Keith Sheppard, Chen-Ping Fu, Leonard McMillan, Fernando Pardo-Manuel de Villena, and Gary A Churchill. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC genomics*, 13 :34, 2012.

[135] P SIMECEK, J Forejt, R W Williams, L Lu, and Thomas E et al.. Johnson. High Resolution Genomic Architecture of Genetic Reference Populations Chromosome Substitution Panels and Recombinant Inbred Strains. *Abstracts of the Meeting of the Genetics Society of America, June 22-25, 2011,Washington, DC.*, 2011.

[136] E Lander and L Kruglyak. Genetic dissection of complex traits guidelines for interpreting and reporting linkage results. *Nat. Genet.*, 11(3) :241–247, November 1995.

[137] Claudia Schurmann, Katharina Heim, Arne Schillert, Stefan Blankenberg, Maren Carstensen, Marcus Dörr, Karlhans Endlich, Stephan B Felix, Christian Gieger, Harald Grallert, Christian Herder, Wolfgang Hoffmann, Georg Homuth, Thomas Illig, Jochen Kruppa, Thomas Meitinger, Christian Müller, Matthias Nauck, Annette Peters, Rainer Rettig, Michael Roden, Konstantin Strauch, Uwe Völker, Henry Völzke, Simone Wahl, Henri Wallaschofski, Philipp S Wild, Tanja Zeller, Alexander Teumer, Holger Prokisch, and Andreas Ziegler. Analyzing illumina gene expression microarray data from different tissues methodological aspects of data analysis in the metaxpress consortium. *PloS one*, 7(12) :e50938, 2012.

[138] Christoffer Nellåker, Thomas M Keane, Binnaz Yalcin, Kim Wong, Avigail Agam, T Grant Belgard, Jonathan Flint, David J Adams, Wayne N Frankel, and Chris P Ponting. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol.*, 13(6) :R45, 2012.

[139] Keiko Akagi, Jingfeng Li, Robert M Stephens, Natalia Volfovsky, and David E

Symer. Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome research*, 18(6) :869–880, June 2008.

[140] Keiko Akagi, Robert M Stephens, Jingfeng Li, Evgenji Evdokimov, Michael R Kuehn, Natalia Volfovsky, and David E Symer. MouseIndelDB a database integrating genomic indel polymorphisms that distinguish mouse strains. *Nucleic acids research*, 38(Database issue) :D600–6, January 2010.

[141] Hadassa Brunschwig, Liat Levi, Eyal Ben-David, Robert W Williams, Benjamin Yakir, and Sagiv Shifman. Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics*, 191(3) :757–764, July 2012.

[142] Jenny Schlesinger, Markus Schueler, Marcel Grunert, Jenny J Fischer, Qin Zhang, Tammo Krueger, Martin Lange, Martje Tönjes, Ilona Dunkel, and Silke R Sperling. The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS genetics*, 7(2) :e1001313, February 2011.

[143] Matthew J Blow, David J McCulley, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, James Bristow, Bing Ren, Brian L Black, Edward M Rubin, Axel Visel, and Len A Pennacchio. ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, 42(9) :806–810, September 2010.

[144] Markus Schueler, Qin Zhang, Jenny Schlesinger, Martje Tönjes, and Silke R Sperling. Dynamics of Srf, p300 and histone modifications during cardiac maturation in mouse. *Molecular bioSystems*, 8(2) :495–503, February 2012.

[145] Elissa J Chesler, Lu Lu, Siming Shou, Yanhua Qu, Jing Gu, Jintao Wang, Hui-Chen Hsu, John D Mountz, Nicole E Baldwin, Michael A Langston, David W Threadgill, Kenneth F Manly, and Robert W Williams. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.*, 37(3) :233–242, March 2005.

[146] S Durinck, P T Spellman, E Birney, and W Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, 2009.

[147] Binnaz Yalcin, Kim Wong, Avigail Agam, Martin Goodson, Thomas M Keane, Xiangchao Gan, Christoffer Nellåker, Leo Goodstadt, Jérôme Nicod, Amarjit Bhomra, Polinka Hernandez-Pliego, Helen Whitley, James Cleak, Rebekah Dutton, Deborah Janowitz, Richard Mott, David J Adams, and Jonathan Flint. Sequence-based characterization of structural variation in the mouse genome. *Nature*, 477(7364) :326–329, September 2011.

[148] J Jurka, O Kohany, A Pavlicek, V V Kapitonov, and M V Jurka. Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. *Cytogenetic and genome research*, 110(1-4) :117–123, 2005.

[149] Elena M Pugacheva, Teruhiko Suzuki, Svetlana D Pack, Natsuki Kosaka-Suzuki, Jeongheon Yoon, Alexander A Vostrov, Eugene Barsov, Alexander V Strunnikov, Herbert C Morse, Dmitri Loukinov, and Victor Lobanenkov. The structural complexity of the human BORIS gene in gametogenesis and cancer. *PloS one*, 5(11) : e13872, 2010.

[150] Dominic Schmidt, Petra C Schwalie, Michael D Wilson, Benoit Ballester, Angela Gonçalves, Claudia Kutter, Gordon D Brown, Aileen Marshall, Paul Flicek, and Duncan T Odom. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, 148(1-2) :335–348, January 2012.

[151] Guillaume Bourque, Bernard Leong, Vinsensius B Vega, Xi Chen, Yen Ling Lee, Kandhadayar G Srinivasan, Joon-Lin Chew, Yijun Ruan, Chia-Lin Wei, Huck Hui Ng, and Edison T Liu. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome research*, 18(11) :1752–1762, November 2008.

[152] D H Kass, J Kim, A Rao, and P L Deininger. Evolution of B2 repeats the muroid explosion. *Genetica*, 99(1) :1–13, 1997.

[153] Jennifer E Phillips and Victor G Corces. CTCF master weaver of the genome. *Cell*, 137(7) :1194–1211, June 2009.

[154] ENCODE Project Consortium, Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, Carrie A Davis, Francis Doyle, Charles B Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R Lajoie, Stephen G Landt, Burn-Kyu Lee, Florencia Pauli, Kate R Rosenbloom, and ... An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74, September 2012.

[155] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398) :376–380, May 2012.

[156] Elena Gogvadze and Anton Buzdin. Retroelements and their impact on genome evolution and functioning. *Cellular and molecular life sciences CMLS*, 66(23) : 3727–3742, December 2009.

[157] Louie N van de Lagemaat, Josette-Renée Landry, Dixie L Mager, and Patrik Medstrand. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in genetics TIG*, 19(10) : 530–536, October 2003.

[158] Jingfeng Li, Keiko Akagi, Yongjun Hu, Anna L Trivett, Christopher J W Hlynialuk, Deborah A Swing, Natalia Volfovsky, Tamara C Morgan, Yelena Golubeva, Robert M Stephens, David E Smith, and David E Symer. Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. *Genome research*, 22(5) :870–884, May 2012.

[159] A A Palmer and S C Dulawa. Murine warriors or worriers the saga of Comt1, B2 SINE elements, and the future of translational genetics. *Frontiers in neuroscience*, 2010.

[160] Tatyana Chernova, Fiona M Higginson, Reginald Davies, and Andrew G Smith. B2 SINE retrotransposon causes polymorphic expression of mouse 5-aminolevulinic acid synthase 1 gene. *Biochemical and biophysical research communications*, 377(2) :515–520, December 2008.

[161] Li Teng, Hiram A Firpi, and Kai Tan. Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. *Nucleic acids research*, 39(17) :7371–7379, September 2011.

[162] Robert Plomin, Claire M A Haworth, and Oliver S P Davis. Common disorders are quantitative traits. *Nature Reviews Genetics*, 10(12) :872–878, December 2009.

[163] Andreas Rohrwasser, Paul Lott, Robert B Weiss, and Jean-Marc Lalouel. From genetics to mechanism of disease liability. *Advances in genetics*, 60 :701–726, 2008.

[164] Kent W Hunter and Nigel P S Crawford. The future of mouse QTL mapping to diagnose disease in mice in the age of whole-genome association studies. *Annual review of genetics*, 42 :131–141, 2008.

[165] Alexandra C Nica, Stephen B Montgomery, Antigone S Dimas, Barbara E Stranger, Claude Beazley, Inês Barroso, and Emmanouil T Dermitzakis. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS genetics*, 6(4) :e1000895, April 2010.

[166] Xia Yang. Use of functional genomics to identify candidate genes underlying human genetic association studies of vascular diseases. *Arteriosclerosis, thrombosis, and vascular biology*, 32(2) :216–222, February 2012.

[167] Catherine Morrissey, Ian C Grieve, Matthias Heinig, Santosh Atanur, Enrico Petretto, Michal Pravenec, Norbert Hubner, and Timothy J Aitman. Integrated genomic approaches to identification of candidate genes underlying metabolic and cardiovascular phenotypes in the spontaneously hypertensive rat. *Physiol. Genomics*, 43(21) :1207–1218, November 2011.

[168] Dong-Yeon Cho, Yoo-Ah Kim, and Teresa M Przytycka. Chapter 5 Network biology approach to complex diseases. *PLoS Comput. Biol.*, 8(12) :e1002820, December 2012.

[169] Andrea Califano, Atul J Butte, Stephen Friend, Trey Ideker, and Eric Schadt. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.*, 44(8) :841–847, August 2012.

[170] Ricardo A Verdugo, Christian F Deschepper, Gloria Muñoz, Daniel Pomp, and Gary A Churchill. Importance of randomization in microarray experimental designs with Illumina platforms. *Nucleic acids research*, 37(17) :5610–5618, September 2009.

[171] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1) : 118–127, January 2007.

[172] Marie Pier Scott-Boyer and Christian F Deschepper. Genome-wide detection of gene co-expression domains showing linkage to regions enriched with polymorphic retrotransposons in recombinant inbred mouse strains. *G3 (Bethesda)*, 3(4), 2013.

[173] Ani Manichaikul, Josée Dupuis, Saunak Sen, and Karl W Broman. Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. *Genetics*, 174(1) :481–489, September 2006.

[174] Bin Zhang and Steve Horvath. A general framework for weighted gene co-

expression network analysis. *Statistical applications in genetics and molecular biology*, 4 :Article17, 2005.

[175] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11) :2498–2504, November 2003.

[176] C F Deschepper, S Masciotra, A Zahabi, I Boutin-Ganache, S Picard, and T L Reudelhuber. Functional alterations of the Nppa promoter are linked to cardiac ventricular hypertrophy in WKY/WKHA rat crosses. *Circulation research*, 88 (2) :223–228, February 2001.

[177] T J Aitman, A M Glazier, C A Wallace, L D Cooper, P J Norsworthy, F N Wahid, K M Al-Majali, P M Trembling, C J Mann, C C Shoulders, D Graf, E St Lezin, T W Kurtz, V Kren, M Pravenec, A Ibrahimi, N A Abumrad, L W Stanton, and J Scott. Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nat. Genet.*, 21(1) :76–83, January 1999.

[178] Michal Pravenec, Paul C Churchill, Monique C Churchill, Ondrej Viklicky, Ludmila Kazdova, Timothy J Aitman, Enrico Petretto, Norbert Hubner, Caroline A Wallace, Heike Zimdahl, Vaclav Zidek, Vladimir Landa, Joseph Dunbar, Anil Bidani, Karen Griffin, Nathan Qi, Martina Maxova, Vladimir Kren, Petr Mlejnek, Jiaming Wang, and Theodore W Kurtz. Identification of renal Cd36 as a determinant of blood pressure and risk for hypertension. *Nat. Genet.*, 40(8) :952–954, August 2008.

[179] Haijin Meng, Iset Vera, Nam Che, Xuping Wang, Susanna S Wang, Leslie Ingram-Drake, Eric E Schadt, Thomas A Drake, and Aldons J Lusis. Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proc. Natl. Acad. Sci. U.S.A.*, 104(11) :4530–4535, March 2007.

[180] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R Wood, Robert J Weyant, Ayellet V Segrè, Elizabeth K Speliotes, Eleanor Wheeler, and ... Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317) :832–838, October 2010.

[181] Yanqing Chen, Jun Zhu, Pek Yee Lum, Xia Yang, Shirly Pinto, Douglas J MacNeil, Chunsheng Zhang, John Lamb, Stephen Edwards, Solveig K Sieberts, Amy Leonardson, Lawrence W Castellini, Susanna Wang, Marie-France Champy, Bin Zhang, Valur Emilsson, Sudheer Doss, Anatole Ghazalpour, Steve Horvath, Thomas A Drake, Aldons J Lusis, and Eric E Schadt. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452(7186) :429–435, March 2008.

[182] Richard C Davis, Atila van Nas, Lawrence W Castellani, Yi Zhao, Zhiqiang Zhou, Pingzi Wen, Suzanne Yu, Hongxiu Qi, Melenie Rosales, Eric E Schadt, Karl W Broman, Miklós Péterfy, and Aldons J Lusis. Systems genetics of susceptibility to obesity-induced diabetes in mice. *Physiol. Genomics*, 44(1) :1–13, January 2012.

[183] Magalie S Leduc, Rachael Hageman Blair, Ricardo A Verdugo, Shirng-Wern Tsaih, Kenneth Walsh, Gary A Churchill, and Beverly Paigen. Using bioinformatics and systems genetics to dissect HDL-cholesterol genetics in an MRL/MpJ x SM/J intercross. *Journal of lipid research*, 53(6) :1163–1175, June 2012.

[184] Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, Robert J Bollo, Xudong Zhao, Evan Y Snyder, Erik P Sulman, Sandrine L Anne, Fiona Doetsch, Howard Colman, Anna Lasorella, Ken Aldape, Andrea Califano, and Antonio Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279) :318–325, January 2010.

[185] Matthias Heinig, Enrico Petretto, Chris Wallace, Leonardo Bottolo, Maxime Rotival, Han Lu, Yoyo Li, Rizwan Sarwar, Sarah R Langley, Anja Bauerfeind, Oliver Hummel, Young-Ae Lee, Svetlana Paskas, Carola Rintisch, Kathrin Saar, Jason

Cooper, Rachel Buchan, Elizabeth E Gray, Jason G Cyster, Cardiogenics Consortium, Jeanette Erdmann, Christian Hengstenberg, Seraya Maouche, Willem H Ouwehand, Catherine M Rice, Nilesh J Samani, Heribert Schunkert, Alison H Goodall, Herbert Schulz, Helge G Roider, Martin Vingron, Stefan Blankenberg, Thomas Münzel, Tanja Zeller, Silke Szymczak, Andreas Ziegler, Laurence Tiret, Deborah J Smyth, Michal Pravenec, Timothy J Aitman, Francois Cambien, David Clayton, John A Todd, Norbert Hubner, and Stuart A Cook. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, 467 (7314) :460–464, September 2010.

[186] Charles R Farber. Systems-level analysis of genome-wide association data. *G3 (Bethesda, Md.)*, 3(1) :119–129, January 2013.

[187] Yong H Woo, Michael Walker, and Gary A Churchill. Coordinated expression domains in mammalian genomes. *PloS one*, 5(8) :e12158, 2010.

[188] Peter Fraser and Wendy Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143) :413–417, May 2007.

[189] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950) :289–293, October 2009.

[190] Chris McDermott-Roe, Junmei Ye, Rizwan Ahmed, Xi-Ming Sun, Anna Serafín, James Ware, Leonardo Bottolo, Phil Muckett, Xavier Cañas, Jisheng Zhang, Glenn C Rowe, Rachel Buchan, Han Lu, Adam Braithwaite, Massimiliano Mancini, David Hauton, Ramon Martí, Elena García-Arumí, Norbert Hubner, Howard Jacob, Tadao Serikawa, Vaclav Zidek, Frantisek Papousek, Frantisek Kolar, Maria Cardona, Marisol Ruiz-Meana, David García-Dorado, Joan X Comella, Leanne E

Felkin, Paul J R Barton, Zoltan Arany, Michal Pravenec, Enrico Petretto, Daniel Sanchis, and Stuart A Cook. Endonuclease G is a novel determinant of cardiac hypertrophy and mitochondrial function. *Nature*, 478(7367) :114–118, October 2011.

[191] Alexandra Zhernakova, Eli A Stahl, Gosia Trynka, Soumya Raychaudhuri, Eleanora A Festen, Lude Franke, Harm-Jan Westra, Rudolf S N Fehrmann, Fina A S Kurreeman, Brian Thomson, Namrata Gupta, Jihane Romanos, Ross McManus, Anthony W Ryan, Graham Turner, Elisabeth Brouwer, Marcel D Posthumus, Elaine F Remmers, Francesca Tucci, Rene Toes, Elvira Grandone, Maria Cristina Mazzilli, Anna Rybak, Bozena Cukrowska, Marieke J H Coenen, Timothy R D J Radstake, Piet L C M van Riel, Yonghong Li, Paul I W de Bakker, Peter K Gregersen, Jane Worthington, Katherine A Siminovitch, Lars Klareskog, Tom W J Huizinga, Cisca Wijmenga, and Robert M Plenge. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS genetics*, 7(2) :e1002004, February 2011.

[192] RSN Fehrmann, R C Jansen, J H Veldink, and H J Westra. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS genetics*, 2011.

[193] Victoria N Parikh, Richard C Jin, Sabrina Rabello, Natali Gulbahce, Kevin White, Andrew Hale, Katherine A Cottrill, Rahamthulla S Shaik, Aaron B Waxman, Ying-Yi Zhang, Bradley A Maron, Jochen C Hartner, Yuko Fujiwara, Stuart H Orkin, Kathleen J Haley, Albert-László Barabási, Joseph Loscalzo, and Stephen Y Chan. MicroRNA-21 integrates pathogenic signaling to control pulmonary hypertension results of a network bioinformatics approach. *Circulation*, 125(12) : 1520–1532, March 2012.

[194] Edwin K Silverman and Joseph Loscalzo. Network medicine approaches to the

genetics of complex diseases. *Discovery medicine*, 14(75) :143–152, August 2012.

[195] Theodore W Kurtz. Genome-wide association studies will unlock the genetic basis of hypertension. con side of the argument. *Hypertension*, 56(6) :1021–1025, December 2010.

[196] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.*, 109(4) :1193–1198, January 2012.

[197] Mathieu Emily, Thomas Mailund, Jotun Hein, Leif Schauser, and Mikkel Heide Schierup. Using biological networks to search for interacting loci in genome-wide association studies. *European journal of human genetics EJHG*, 17(10) : 1231–1240, October 2009.

[198] Duncan Thomas. Gene–environment-wide association studies emerging approaches. *Nature Reviews Genetics*, 11(4) :259–272, April 2010.

[199] Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine*, 363(2) :166–176, July 2010.

[200] Hua Zhong, John Beaulaurier, Pek Yee Lum, Cliona Molony, Xia Yang, Douglas J MacNeil, Drew T Weingarth, Bin Zhang, Danielle Greenawalt, Radu Dobrin, Ke Hao, Sangsoon Woo, Christine Fabre-Suver, Su Qian, Michael R Tota, Mark P Keller, Christina M Kendziorski, Brian S Yandell, Victor Castro, Alan D Attie, Lee M Kaplan, and Eric E Schadt. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS genetics*, 6(5) : e1000932, May 2010.

[201] Z Wei, S Peddibhotla, H Lin, X Fang, and M Li. Early-onset aging and defective DNA damage response in Cdc14b-deficient mice. *Mol Cell Biol.*, 2011.

[202] M Sun, M Chen, and Y Liu. Cathepsin-L contributes to cardiac repair and remodelling post-infarction. *Cardiovascular Res.*, 2011.

[203] H Qiu, H Dai, K Jain, R Shah, C Hong, and J Pain. Characterization of a novel cardiac isoform of the cell cycle-related kinase that is regulated during heart failure. *Journal of Biological Chemestry*, 2008.

[204] C Wu, I MacLeod, and A I Su. BioGPS and MyGene. info : organizing online, gene-centric information. *Nucleic acids research*, 2013.

[205] M Plummer, N Best, K Cowles, and K Vines. CODA Convergence diagnosis and output analysis for MCMC. *R news*, 2006.

[206] P L'ecuyer. Good parameters and implementations for combined multiple recursive random number generators. *Operations Research*, 1999.

[207] W K Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 1970.

Annexe I

Glossaire

Analyse de liaison Ce terme réfère à des marqueurs génétiques polymorphes qui ont une proximité physique sur un chromosome, de telle sorte qu'un allèle spécifique à un locus a une importante valeur prédictive pour les allèles spécifiques aux autres locus.

Analyse en composantes principales (PCA) Une technique permettant de simplifier des ensembles de données complexes et multidimensionnelles à un nombre réduit de dimensions, soit les principaux composants. Cette procédure conserve les caractéristiques des données qui se rapportent à sa variance.

Cis Une molécule est décrite comme agissant en cis lorsqu'elle affecte d'autres gènes qui sont physiquement adjacents, sur le même chromosome, ou sont génétiquement liés ou à proximité (pour l'expression des ARNm, typiquement un promoteur).

Centimorgan Le centimorgan est une unité utilisée en génétique pour la mesure des distances génétiques. Le centimorgan correspond à un intervalle sur lequel le nombre moyen de recombinaisons se produisant pendant la méiose.

Eigengene Un eigengene est défini comme la première composante principale d'un module. Il peut être considéré comme un représentant des profils d'expression génique dans un module.

Épigénétique L'épigénétique est le domaine qui étudie les changements de l'expression sans altération des séquences nucléotidiques.

Épissage alternatif L'épissage est un processus qui consiste en l'excision des introns et en la ligature des exons. L'épissage alternatif permet à un gène de coder plusieurs ARNm matures ou protéines différentes.

Étude d'association (GWAS) Une analyse de la variation génétique et les relations entre les variants génétiques (allèles) et phénotypes dans une population qui n'ont pas de lien entre eux.

Héritabilité La variation phénotypique entre les individus d'une population est attribuable à la variation génétique entre les individus et à des facteurs environnementaux. L'héritabilité est la variation génétique et elle est généralement exprimée en pourcentage.

Gene Ontology (GO) Une ontologie de termes biologiques. Il existe trois catégories de terme Gene Ontology : le processus biologique, la fonction moléculaire et le compartiment cellulaire. Chaque gène peut être associé à ces termes d'annotation de chacune des trois catégories.

Génétique quantitative Une branche de la recherche en génétique qui emploie des principes statistiques et probabilistes pour identifier les facteurs génétiques et non génétiques d'un trait complexe.

Monte Carlo Markov Chain (MCMC) L'algorithme de Monte Carlo Markov Chain (MCMC) tent de simuler des tirages directs d'une certaine distribution complexe. L'approche MCMC est ainsi nommée parce qu'on utilise les valeurs d'échantillons précédents pour générer aléatoirement le prochain échantillon par une chaîne de Markov.

Phase de la vérification (burn-in) La phase de la vérification ou burn-in est un terme qui décrit la pratique de générer quelques itérations au début d'un long MCMC. Celles-ci ne seront pas considérées dans le calcul final. Il s'agit de la période de rodage. Après le burn-in, on exécute normalement, en utilisant chaque itération des calculs MCMC.

Priori Une distribution a priori est une distribution d'une quantité p, qui exprime une incertitude à propos de p avant que les "données" soient prises en compte.

Promoteur Une séquence d'ADN régulatrice, généralement située en amont d'un gène exprimé, qui de concert avec d'autres éléments de la régulation souvent éloignés, dirige la transcription d'un gène donné.

Posteriori Une distribution a posteriori est une distribution de probabilité d'une quantité inconnue, considérée comme une variable aléatoire, sous réserve des éléments de preuve obtenus à partir d'une expérience.

Post-transcriptionelle Une modification post-transcriptionelle est une modification chimique après la synthèse des protéines. Certaines modifications, comme la phosphorylation, ont un effet sur les propriétés de la protéine et permettent de réguler leurs fonctions.

Puce à ADN (microarray) Les puces à ADN sont utilisées pour le profilage de la transcription génique.

Recombinaison La recombinaison est un phénomène assurant le mélange de matériel génétique entre chromosomes.

Single nucleotide polymorphism (SNP) La variation d'une seule paire de bases dans le génome entre individus d'une même espèce.

Synténique Une région synténique est une région d'une forte similitude dans l'organisation des génomes.

Test de permutation Une méthode statistique qui utilise une distribution empirique de la statistique acquise par test en permutant l'échantillon original afin d'établir quelle est la probabilité d'obtention d'une certaine valeur simplement par le hasard.

Trans Un facteur ou un gène qui agit sur un autre gène, mais qui est séparé physiquement (pour l'expression d'ARNm, typiquement un facteur de transcription).

Transposon Des séquences d'ADN capables de se déplacer à de nouveaux postes au sein du génome d'une cellule unique.

Annexe II

Supplementary Materiel : An integrated Bayesian hierarchical model for multivariate eQTL mapping

Appendix

Appendix A : Full conditional posterior distributions

The set of parameters of the model is $\theta = (\mu_g, \sigma_g^2, \gamma_{jg}, \beta_{jg}, \omega_{jg}, p_j, a_j, b_j)$. The posterior distribution of the parameter set θ is given by the product of the prior distributions $\pi(\theta)$ with the likelihood $L(y|x,\theta)$, that is

$$\pi(\theta|y,x) \propto \pi(\theta) L(y|x,\theta) \tag{II.1}$$

where

$$L(y|x,\theta) \propto \prod_{g=1}^{G}\prod_{i=1}^{n} \frac{1}{\sigma_g} \exp\left(-\frac{1}{2\sigma_g^2}\left(y_{ig} - \mu_g - \sum_{j=1}^{S} x_{ij}\gamma_{jg}\beta_{jg}\right)^2\right).$$

For a specific parameter θ_k, the full conditional $\pi(\theta_k|\dots)$ is obtained by conditioning the posterior $\pi(\theta|y,x)$ in (II.1) on the remaining parameters.

- The full conditional of μ_g is $\mu_g \sim \mathcal{N}(m'_g, \tau'^2_g)$, where m'_g and τ'^2_g are obtained by updating the prior parameters m_g and τ_g as follows :

$$m'_g = \frac{\sum_{i=1}^{n}\left(y_{ig} - \sum_{j=1}^{S} x_{ij}\gamma_{jg}\beta_{jg}\right)/\sigma_g^2 + m_g/s_g^2}{n/\sigma_g^2 + 1/s_g^2} \text{ and } \tau'^2_g = (n/\sigma_g^2 + 1/s_g^2)^{-1}.$$

- The full conditional of σ_g^2 is $\sigma_g^2 \sim \mathcal{I}n\mathcal{G}a(d'_g, e'_g)$, an Inverse Gamma distribution

with parameters d'_g and e'_g, where $d'_g = \frac{1}{2}(n+1+S_g)$ and

$$e'_g = \frac{1}{2}\sum_{i=1}^{n}\left(y_{ig} - \mu_g - \sum_{j=1}^{S} x_{ij}\gamma_{jg}\beta_{jg}\right)^2 + \frac{1}{2c}\sum_{j=1}^{S}\left(\gamma_{jg}\beta_{jg}^2\sum_{i=1}^{n} x_{ij}^2\right).$$

- The parameters γ_{jg} and β_{jg} require special attention. These two parameters are updated simultaneously using their joint full conditional $\pi(\gamma_{jg},\beta_{jg}|\dots)$. We first sample γ_{jg} from the marginal posterior $\pi(\gamma_{jg}|\dots)$ obtained by integrating out β_{jg} in $\pi(\gamma_{jg},\beta_{jg}|\dots)$ and then β_{jg} is simulated from the conditional distribution $\pi(\beta_{jg}|\gamma_{jg},\dots)$. The joint full conditional $\pi(\gamma_{jg},\beta_{jg}|\dots)$ is given by

$$\pi(\gamma_{jg},\beta_{jg}|\dots) \propto L(\gamma_{jg},\beta_{jg}|\dots)\pi(\gamma_{jg}|\omega_j)\pi(\beta_{jg}|\gamma_{jg}), \tag{II.2}$$

where $L(\gamma_{jg},\beta_{jg}|\dots)$ is the part of the likelihood containing γ_{jg} and β_{jg} (*i.e.* the contribution of gene expression g) and is given by

$$L(\gamma_{jg},\beta_{jg}|\dots) \propto \prod_{i=1}^{n}\frac{1}{\sigma_g}\exp\left(-\frac{1}{2\sigma_g^2}\left(y_{ig} - \mu_g - \sum_{j'=1}^{S} x_{ij'}\gamma_{jg'}\beta_{jg'}\right)^2\right).$$

Furthermore, in equation (II.2), $\pi(\gamma_{jg}|\omega_{jg}) = \omega_{jg}^{\gamma_{jg}}(1-\omega_{jg})^{1-\gamma_{jg}}$ is the Bernoulli prior of γ_{jg} and $\pi(\beta_{jg}|\gamma_{jg})$ is the prior distribution of β_{jg} conditional on γ_{jg} such that :

$$\pi(\beta_{jg}|\gamma_{jg}) = \delta_0(\beta_{jg})\mathbb{I}_{(\gamma_{jg}=0)} + \mathcal{N}(0,v_{jg}^2)(\beta_{jg})\mathbb{I}_{(\gamma_{jg}=1)}.$$

In order to sample γ_{jg} from $\pi(\gamma_{jg}|\dots)$, we integrate out β_{jg} and we let

$$\begin{aligned} p_0 &= \int L(\gamma_{jg}=0,\beta_{jg}|\dots)\pi(\beta_{jg}|\gamma_{jg}=0)d\beta_{jg} = L(\gamma_{jg}=0,\beta_{jg}=0|\dots), \\ p_1 &= \int L(\gamma_{jg}=1,\beta_{jg}|\dots)\pi(\beta_{jg}|\gamma_{jg}=1)d\beta_{jg}. \end{aligned}$$

It follows that $\pi(\gamma_{jg}=0|\dots) \propto (1-\omega_{jg})p_0$ and $\pi(\gamma_{jg}=1|\dots) \propto \omega_{jg}p_1$. Further

computation leads to $p_1 = Cp_0$, where the quantity C is equal to

$$C = \frac{1}{(1+c)^{1/2}} \exp\left(\frac{1}{2} \frac{c}{(1+c)\sigma_g^2 \sum_{i=1}^{n} x_{ij}^2} \left[\sum_{i=1}^{n} x_{ij} \left(y_{ig} - \mu_g - \sum_{j' \neq j} x_{ij'} \gamma_{jg'} \beta_{jg'} \right) \right]^2 \right).$$

Finally, the parameter γ_{jg} is sampled from

$$\pi(\gamma_{jg} = 0 | \ldots) = \frac{1 - \omega_{jg}}{C\omega_{jg} + (1 - \omega_{jg})} \text{ and } \pi(\gamma_{jg} = 1 | \ldots) = \frac{C\omega_{jg}}{C\omega_{jg} + (1 - \omega_{jg})}.$$

As we mentioned earlier, the parameter β_{jg} is sampled from the conditional posterior distribution $\pi(\beta_{jg}|\gamma_{jg})$. Precisely, $\beta_{jg} = 0$ if γ_{jg} is sampled as 0 and β_{jg} is generated from a $\mathcal{N}(m'_{jg}, {\nu'}^2_{jg})$ if γ_{jg} is sampled as 1. The quantities m'_{jg} and ${\nu'}^2_{jg}$ are given by

$$\begin{aligned} m'_{jg} &= \frac{c}{(1+c)\sum_{i=1}^{n} x_{ij}^2} \sum_{i=1}^{n} x_{ij} \left(y_{ig} - \mu_g - \sum_{j' \neq j} x_{ij'} \gamma_{jg'} \beta_{jg'} \right) \\ {\nu'}^2_{jg} &= \frac{c\sigma_g^2}{(1+c)\sum_{i=1}^{n} x_{ij}^2}. \end{aligned}$$

- The full conditional of ω_{jg} is $\omega_{jg} \sim r\delta_0(\omega_{jg}) + (1-r)\mathscr{B}eta(a'_j, b'_j)(\omega_{jg})$, which is a mixture of a Dirac mass in 0 and a Beta distribution with parameters $a'_j = a_j + \gamma_{jg}$ and $b'_j = b_j + 1 - \gamma_{jg}$ and with respective weights r and $1-r$, where r is given by

$$r = \frac{p_j^{\mathbb{I}_{(\gamma_{jg}=0)}}}{p_j^{\mathbb{I}_{(\gamma_{jg}=0)}} + (1-p_j)\dfrac{\mathscr{B}(a'_j, b'_j)}{\mathscr{B}(a_j, b_j)}},$$

and $\mathscr{B}(.,.)$ is the Beta function.

- The full conditional of p_j is $p_j \sim \mathscr{B}eta(a',b')$, with $a' = a_0 + \sum_{g=1}^{G} \mathbb{I}_{(\omega_{jg}=0)}$ and $b' = b_0 + \sum_{g=1}^{G} \mathbb{I}_{(\omega_{jg}>0)}$, where $\sum_{g=1}^{G} \mathbb{I}_{(\omega_{jg}=0)}$ represents the number of genes for which SNP j has zero probability to be an eQTL and $\sum_{g=1}^{G} \mathbb{I}_{(\omega_{jg}>0)}$ represents the number of genes with positive probability to have an eQTL at SNP j.

- Full conditionals for a_j and b_j are not available in closed form but are given by

$$\begin{aligned}
\pi(a_j|\dots) &\propto \prod_{g=1}^{G} \Big[p_j \delta_0(\omega_{jg}) + (1-p_j)\mathscr{B}(a_j,b_j)(\omega_{jg}) \Big] \exp(-\lambda_a a_j) \text{ and} \\
\pi(b_j|\dots) &\propto \prod_{g=1}^{G} \Big[p_j \delta_0(\omega_{jg}) + (1-p_j)\mathscr{B}(a_j,b_j)(\omega_{jg}) \Big] \exp(-\lambda_b b_j).
\end{aligned}$$

  Therefore, if $\omega_{jg} = 0$ for all g, the parameters a_j and b_j are simply sampled from their corresponding priors $\mathscr{E}xp(\lambda_a)$ and $\mathscr{E}xp(\lambda_b)$. When $\omega_{jg} \neq 0$ for at least one g, we employ the adaptive rejection sampling algorithm of Gilks et al. [112] to sample from $\pi(a_j|\dots)$ and $\pi(b_j|\dots)$.

Appendix B : Choice of the hyperparameters of the model

Reasonable prior guesses for λ_a, λ_b, a_0 and b_0 can be obtained by computing the *a priori* expected number of eQTLs by gene, namely $\mathbb{E}(n_{g,eQTL})$ and the variance of the number of eQTLs $\mathbb{V}(n_{g,eQTL})$. Given the conditional indepence structure of the model, it can be seen that after integrating out w_{jg}, a_j, b_j, p_j, the distribution of γ_{jg} is again Bernoulli with probability $w^* = b_0/(a_0+b_0)I$ where $I = \iint_{a,b} a/(a+b)\lambda_a \exp\{-\lambda_a a\}\lambda_b \exp\{-\lambda_b b\} da db$.

It follows that $\mathbb{E}(n_{g,eQTL}) = Sw^*$ and $\mathbb{V}(n_{g,eQTL}) = Sw^*(1-w^*)$ where S is the number of SNPs. For example, if $S = 1000$ and $\lambda_a = a_0 = 10$ and $\lambda_b = b_0 = 0.1$, which we used here, we have $\mathbb{E}(n_{g,eQTL}) \simeq 0.37$ and $\mathbb{V}(n_{g,eQTL}) \simeq 0.36$. These values correspond to a scenario where the prior number of eQTLs per gene lies between $\mathbb{E}(n_{g,eQTL}) \pm 2\sqrt{\mathbb{V}(n_{g,eQTL})} = (0, 1.6)$, thus privileging the null model (the model without eQTLs). Note that we used these equations as guidelines only, and the resul-

ting priors are midly informative. More informative prior values could be derived from previous experiments, assuming that such experiments are available. Alternatively, one could use a simple one transcript vs. one SNP regression approach (*e.g.* R-QTL) to provide reasonable estimates on the number of eQTL per transcript.

Appendix C : Comparison of computation times

Table II.I compares computational times for the different methods used in our comparison. The average times are based on the simulation study where $n = 50$.

Table II.I – Computation times for the different tools used in this paper when applied to the simulated data with $n = 50$. Times reported are the means of 50 simulations with standard deviations between parentheses. iBMQ was performed on 4 CPU and the other methods were perform using 1 CPU.

iBMQ	QTLBIM	MSPL	RQTL	remMAP
25.3 mn (1.9)	4.2 mn (.44)	$\leq$ 1 mn	$\leq$ 1 mn	$\leq$ 1 mn

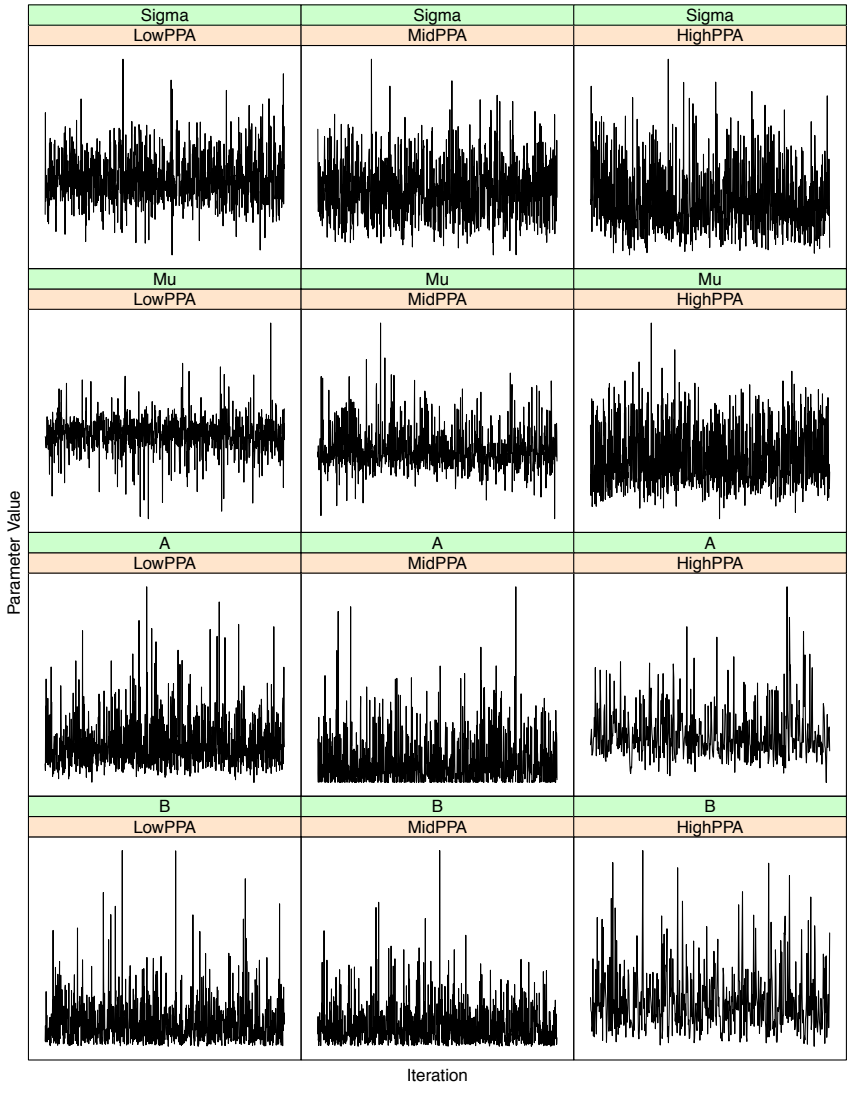
Appendix D : MCMC convergence diagnostics

Given the large number of parameters in our model, it is impossible to report diagnostic convergence results for each one. Here, we have opted to report results for three gene$\times$SNP combinations, with low, medium and large posterior probabilities of association. We feel that by looking at a range of posterior probabilities from low to high, the three gene/SNP combinations reported are well representative of all other combinations. Table II.II shows the results of the Raftery and Lewis [113] convergence test as implemented in the coda package [205] applied to our experimental dataset. The calculations show that the number of iterations used for the results reported here is clearly sufficient. This result is also confirmed by the trace plots included here (Figure II.1). Note that given the large number of simulations performed, we did not perform any diagnostic for the simulated data. However, the ROC curves presented and the diagnostic on the much larger experimental data suggests that convergence was not an issue for these data.

Table II.II – Recommended chain run lengths to estimate .025 quantiles within an error margin of .005, based on the Raftery-Lewis convergence diagnostic for Markov Chain Monte Carlo. Run lengths are calculated on four different parameters at three levels of the highest estimated PPA . The column N gives the run length required to achieve the desired margin of error, Nmin gives the minimum run length when no autocorrelation is present, and the Dependence Factor indicates the inflation factor from Nmin to N, representing the effect of autocorrelation. In all cases the recommended N is less than half of the total number of iterations sampled in our chain.

Parameter	PPA level	N	Nmin	Dependence Factor
	Low	74960	74920	1
A	Mid	228840	74920	3.05
	High	177720	74920	2.37
	Low	76480	74920	1.02
B	Mid	552300	74920	7.37
	High	78060	74920	1.04
	Low	237840	74920	3.17
σ^2	Mid	816000	74920	10.90
	High	969540	74920	12.90
	Low	76980	74920	1.03
μ	Mid	157480	74920	2.10
	High	89360	74920	1.19

Figure II.1 – Trace plot of parameters a_j, b_j, μ_g and σ_g for three gene/SNP combination with low, medium and high PPA. These plots are based on 2,000,000 iterations after 40,000 burn-in.



Annexe III

Supplementary Materiel : iBMQ : a R/Bioconductor package for Integrated Multivariate eQTL Mapping

iBMQ Implementation iBMQ is a hierarchical Bayesian model for the detection of cis and trans-acting expression quantitative trait loci (eQTL). The iBMQ model simultaneously assesses interactions between thousands of genes and SNPs via the linear model

$$y_{ig} = \mu_g + \sum_{j=1}^{S} x_{ij}\gamma_{jg}\beta_{jg} + \varepsilon_{ig}, \tag{III.1}$$

where i indexes subjects, $g = 1,\ldots,G$ indexes genes, and $j = 1,\ldots,S$ indexes SNPs. Thus, gene expression y_{ig} is modelled as a grand mean plus the additive effects of all SNPs x_{ij}. Not all SNPs affect the expression of every gene, so the γ_{jg} parameters are 0/1 indicators for the presence of an interaction between SNP j and gene g. One unique aspect of iBMQ is that each γ_{jg} is controlled by its own inclusion probability parameter ω_{jg}. Hierarchical parameters help induce appropriate shrinkage while sharing information across genes and SNPs ; further details about the model may be found in Scott-Boyer et al. [120].

Our model is more flexible than other eQTL mapping methods, but this flexibility comes at a cost. For several thousand genes and several thousand SNPs, the number of parameters in the model can easily number in the tens of millions. The model's posterior distribution cannot be computed directly, and we employ Markov Chain Monte Carlo (MCMC) to sample the posterior. Our MCMC algorithm is written in C for improved speed, and interfaces with R for convenient data management. Efficient programming allows us to overcome the otherwise tremendous computational burden that comes with updating so many parameters. We outline techniques used to improve the stability, quality, and speed of our MCMC algorithm.

OpenMP® parallelization The MCMC algorithm proceeds via Gibbs sampling of full conditional distributions. The iBMQ model admits conditional independence among

many parameters that index across genes or across SNPs, hence a great number of Gibbs updates can be performed concurrently within each iteration. To take advantage of this model structure, the iBMQ package employs the OpenMP® API to parallelize parameter updates. OpenMP® is a shared memory parallel computing platform, and our algorithm scales well with the number available processors. In a parallel computing environment, care must be taken with pseudorandom number generation. We use Pierre L'Ecuyer's *RngStream* package to ensure that streams of random numbers generated between threads are independent [206].

Parameter updates For most model parameters, conditionally conjugate distributions within our model allow for Gibbs updates using well known probability distributions. Full conditional distributions for most model parameters can be found in Scott-Boyer et al. [120]. The full conditional distributions for parameters a_j and b_j (related to the distribution of parameters ω_{jg}) are not from a well known family of probability distributions. A more general Metropolis-Hastings update [207] can be a great substitute in such cases, but the thousands of a_j and b_j parameters assume a great variety of posterior distributions. Because of this variety, we were unable to find a Metropolis-Hastings proposal mechanism that allowed our Markov chain to efficiently mix across all a_j and b_j. We instead implement the adaptive rejection sampling (ARS) algorithm presented in Gilks et al. [112]. ARS works for log-concave densities, and creates a hull over the target (possibly un-normalized) density from which samples are drawn. Rejected samples are incorporated into the hull, refining the hull and improving the acceptance rate of subsequent samples. Parameters a_j and b_j have log-concave full conditional densities and are updated via a Gibbs sample generated by ARS.

Parameters $\omega_{jg} \in [0,1)$ are converted to the *logit* scale for numerical stability (a separate indicator variable notes whether ω_{jg} equals 0 exactly). Care is needed when sampling the full conditional distribution of ω_{jg}, which is $\omega_{jg}|\ldots =_d \text{Beta}(a_j+\gamma_{jg}, b_j+1-\gamma_{jg})$. Because a_j and b_j can become extremely small, random number generators can occasionally generate values of exactly zero or one, even when these values are theoretically impossible. Values of exactly zero or one create problems in further computations. Sampling ω_{jg} on the *logit* scale eliminates this issue. One can show that the

logit transformation of ω_{jg} has full conditional distribution that is a difference of the logs of independent gamma random variables :

$$\text{logit}(\omega_{jg})|\ldots =_d \log\left[\text{Gamma}(a_j+\gamma_{jg},1)\right]-\log\left[\text{Gamma}(b_j+1-\gamma_{jg},1)\right]. \quad \text{(III.2)}$$

Sparse matrix β representation Our model allows coefficients β_{jg} to be exactly zero. Out of thousands of possible SNPs, relatively few are expected to interact with a given gene g. In practice, the matrix β of model coefficients β_{jg} is sparsely populated at any given MCMC iteration. The matrix β also frequently changes between iterations. Because matrix elements are frequently added and deleted (i.e. set to zero), we use a linked-list sparse representation for β. To prevent stack fragmentation via frequent allocation and deallocation of memory, and to improve speed, matrix elements from the linked list are 'drawn' from thread-safe memory pools.

R code To execute the code below, we have first prepared the "snp" object, which corresponds to a *SnpSet* object containing the 977 informative SNPs for the 24 RIS AXB-BXA population. The genotype is coded with 0 and 1. The "gene" is an *ExpressionSet* object containing the normalized and pre-processed gene expression of 8725 genes for the 24 RIS AXB-BXA population. The "snppos" and "genepos" are dataframes containing the information about the SNP and gene positions respectively. All objects are available at https ://github.com/raphg/iBMQ in the data_application_note folder.

```
library(iBMQ)

load("data_application_note.R")

PPA <- eqtlMcmc(snp, gene, n.iter=1000000, burn.in=50000,
n.sweep=20, mc.cores=6, RIS=TRUE, write.output=FALSE)

cutoff <- calculateThreshold(PPA, 0.1)

eqtl <- eqtlFinder(PPA, cutoff)
```

```
eqtltype <- eqtlClassifier(eqtl, snppos, genepos,1000000)

hotspot <- hotspotFinder(eqtltype, 10)

p <- ggplot(eqtl.type, aes(y=GeneStart, x=MarkerPosition)) +
geom_point(aes(y=GeneStart, x=MarkerPosition, color = PPA), size = 1.5) +
facet_grid(GeneChrm~MarkerChrm)+theme_bw(base_size = 12, base_family = "") +
theme(text=element_text(size=16), panel.margin = unit(0.01, "lines")) +
theme(axis.ticks = element_blank(), axis.text.x = element_blank(),
axis.text.y = element_())+ scale_x_reverse()

p+scale_colour_gradientn(colours=c("grey", "black")
```

Additional information about the GO term enrichment analysis

Gene Ontology (GO) term enrichment was tested using the DAVID Bioinformatics Resources analysis [117]. Table III.I provides additional information for the trans-eQTL hotspots detected both by R/QTL and iBMQ and where corresponding genes showed enrichment for one same GO term. For each GO term, we indicate its full descriptive term, its corresponding head term, and the level of the child term it corresponds to (considering that the head term is at level 1). In each case, the child term was fairly specific, as it belonged to a sub-term 2-6 levels below the head term. Of note, the GO :term "Immune response" for the trans-eQTL on the chr 17 contained many of the same genes as a trans-eQTL hotspot detected by others using heart from a panel of rat RIS [185].)

Table III.I – Additional information concerning GO terms showing enrichment in trans-eQTL hotspots.

GO term #	GO term designation	Head term	Level of child term
GO :0012505	Endomembrane system	Cellular component	4
GO :0007167	Enzyme-linked receptor protein signalling pathway	Biological process	7
GO :0006955	Immune response	Biological process	3

Annexe IV

Supplementary Materiel : Genome-wide detection of gene co-expression domains showing linkage to regions enriched with polymorphic retrotransposons in recombinant

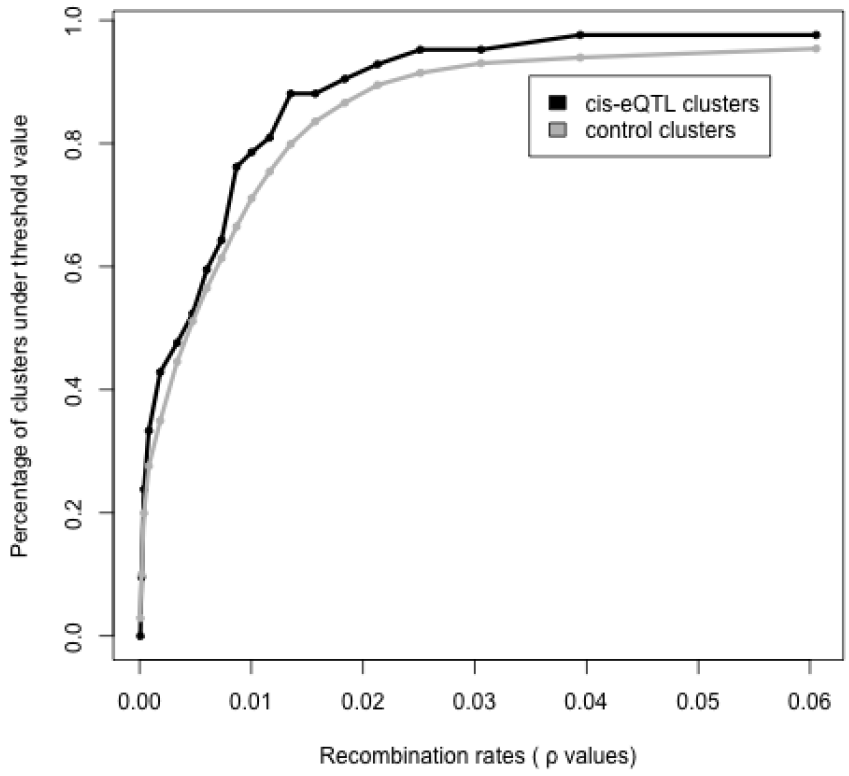


Figure IV.1 – Distribution of recombination rates in regions corresponding to either cis-eQTL or control clusters. The recombination rate values corresponded to the ρ values, as calculated and reported by Brunschwig H et al. (Genetics 191 : 757-764, 2012). According to the Chi-square test, there was no significant difference in the distributions of recombination rates in the 2 types of regions ($P > 0.3$).

Table IV.I – Abundance of polymorphic and total TEs in mouse genomes. Lists of polymorphic TEs were obtained from either the publication of Nellåker et al (2012) or from the MouseIndelDB database. Lists of fixed TEs were obtained from the Transpogene database (n/a : non available.)

Database	Genetic origin	full LINEs	LINE frag.	LTR-TEs	SINEs
Nellåker et al.	Both strains	1808	2969	4734	4303
(polymorphic TEs)	C57(+) / A/J(-)	606	1015	1901	2378
	C57(-) / A/J(+)	1202	1954	2833	1925
MouseIndelDB	Both strains			2413	1512
(polymorphic TEs)	C57(+) / A/J(-)			1436	1512
	C57(-) / A/J(+)			977	0
Transpogene (fixed TEs)		78,002	n/a	84,724	190,057

Table IV.II – Summary of gene expression datasets from mouse RIS.Each line provides information about which tissue and which RIS panel was used for each gene expression dataset, as well as about the number of strains that were profiled and the nature of the microarray platform. The GN access numbers corresponds to the identification number of corresponding datasets in GeneNetwork.

Tissue	RIS panel	# of strains used	Microarray platform	GN access #
Eye	AxB/BxA	26	Illumina MouseRef-6	GN210
Eye	BxD	68	Affymetrix Mouse Genome 430	GN207
Kidney	BxD	54	Affymetrix Mouse Genome430	GN240
Hippocampus	BxD	67	Affymetrix Mouse Genome430	GN112
Hypothalamus	BxD	33	Affymetrix MoGene 1.0 ST	GN281
Cerebellum	BxD	28	Affymetrix Mouse Genome 430	GN72

Table IV.III – Properties of cis-eQTL and control clusters (defined using three different window sizes).In addition to the sizes of the intervals between detected genes, some additional criteria were used in the selection of control clusters in order to have them match cis-eQTL clusters for size and density of total genes. For the 250 kB intervals, only clusters <400 kB and containing more than 3 genes were kept ; for the 500 kB, only clusters <650 kB and containing at least 6 genes were kept ; for the 750 kB intervals, only clusters of <800 kb and containing at least 8 genes were kept.

	250kB			500kB			750kB		
	Cis-eQTL clusters	Control cluters	p-val	Cis-eQTL clusters	Control cluters	p-val	Cis-eQTL clusters	Control cluters	p-val l
Size of boxes(in kb)	221.9 ±130	248±	0.2	467.1±486	456±119	0.87	658.8 ±553	561±123	0.2
Number of detected genes	4.2 ±1.9	4.75±	0.1	4.9±3.5	5.23±2.46	0.63	5.1 ±3.6	6.6±3.89	0.11
Total number of genes	7.2 ±6.6	6.30 ±2.9	0.38	12.3 ±16	0.1±3.45	0.35	15.70±17.6	14.9±11.1	0.81
Number of clusters	42	188		53	59		61	21	

Table IV.IV – Normalized abundance of polymorphic and fixed TEs in several sizes of genomic regions around "250 kB" clusters. Normalized abundance of TEs was calculated by dividing the abundance of each element with that found in random regions. Test regions corresponded to those containing cis-eQTLs (either clustered or single) or control clusters, each region being augmented by flanking regions of either 250, 500, 1000 kb.

		Polymorphic				Fixed	
		full LINEs	LINE frag.	LTR-TEs	SINEs	LTR-TEs	SINEs
Size of flanking regions	250 kB						
Normalized abundance of TEs / MB (mean ± SD)	Cis-eqtl clusters (n=42)	0.51 ± 1.39	0.77 ± 1.42	2.27 ± 1.75	4.78 ± 3.51	1.55 ± 2.17	2.85 ± 3.01
	CTL clusters (n=188)	0.46 ± 1.16	0.55 ± 1.01	1.28 ± 1.58	1.45 ± 2.07	1.53 ± 1.3	2.05 ± 1.62
	Single cis-eQTL regions (n=232)	0.59 ± 1.32	0.94 ± 1.33	1.8± 1.53	1.97±1.84	1.53±1.26	1.71±1.78
	Random regions (n=500)	1 ± 2.04	1 ± 1.72	1 ± 1.39	1 ± 1.72	1 ± 1.27	1 ± 1.42
Anova P values	0.00039	0.0059	1.22E-13		1.11E-33	3.15E-08	4.40E-20
Post-hocTukey P values	Cis-eqtl clusters vs CTL	0.99	0.83	0.00053	<e-5	0.99	0.02306
	Cis-eqtl clusters vs single cis-eQTL regions	0.99	0.90	0.22	<e-5	0.99	0.00023
	Cis-eqtl clusters vs random	0.28	0.77	<e-5	<e-5	0.048	<e-5
	Control vs single cis-eQTL regions	0.8	0.042	0.0019	0.029	0.99	0.15
	CTL vs random	0.001	0.003	0.12	0.03	2.00E-05	<e-5
	Random vs single cis-eQTL regions	0.015	0.96	<e-5	<e-5	<e-5	<e-5
Size of flanking regions	500 kB						
Normalized abundance of TEs / MB (mean ± SD)	Cis-eqtl clusters (n=42)	0.54 ±1.42	0.98 ±1.69	1.88 ± 1.43	4.14 ± 2.84	1.39 ± 1.22	2.46 ± 1.86
	CTL clusters (n=188)	0.61 ± 1.16	0.65 ± 1.05	1.10 ± 1.25	1.27 ± 1.59	1.41 ± 1.10	1.79 ± 1.39

	Single cis-eQTL regions (n=232)	0.75±1.24	0.93±1.07	1.62±1.21	1.7±1.45	1.39±1.1	1.53±1.48
	Random regions (n=500)	1 ± 1.77	1 ± 1.51	1 ± 1.20	1 ± 1.6	1 ± 1.10	1 ± 1.29
Anova P values		0.0068	0.024	1.12E-11	2.22E-31	1.01E-06	6.32E-17
Post-hocTukey P values	Cis-eqtl clusters vs CTL	0.99	0.47	0.001	<e-5	0.99	0.024
	Cis-eqtl clusters vs single cis-eQTL regions	0.83	0.99	0.58	<e-5	1	0.00041
	Cis-eqtl clusters vs random	0.23	0.99	5.00E-05	<e-5	0.12352	<e-5
	Control vs single cis-eQTL regions	0.75	0.13	8.00E-05	0.03	0.99	0.22
	CTL vs random	0.014	0.013	0.79	0.21	1.00E-04	<e-5
	Random vs single cis-eQTL regions	0.18	0.92	<e-5	<e-5	6.00E-05	1.00E-05
Size of flanking regions	1 MB						
Normalized abundance of TEs / MB (mean ± SD)	Cis-eqtl clusters (n=42)	0.59 ± 1.25	0.93 ± 1.27	1.71 ± 1.17	3.47 ± 2.36	1.26 ± 0.93	2.12 ± 1.54
	CTL clusters (n=188)	0.76 ± 1.26	0.72 ± 1.02	1.06 ± 0.92	1.14 ± 1.17	1.33 ± 0.99	1.53 ± 1.1
	Single cis-eQTL regions (n=232)	0.48±1.71	0.81±1.91	1.76±2.28	2.69±3.3	1.7±1.58	2.02±2.02
	Random regions (n=500)	1 ± 1.47	1 ± 1.39	1 ± 1.03	1 ± 1.39	1 ± 0.87	1 ± 1.11
Anova P values		0.00013	0.1	4.95E-11	7.17E-31	1.75E-13	3.84E-20
Post-hocTukey P values	Cis-eqtl clusters vs CTL	0.91	0.82	0.037	<e-5	0.98	0.069
	Cis-eqtl clusters vs single cis-eQTL regions	0.96	0.96	0.99	0.1	0.083	0.97
	Cis-eqtl clusters vs random	0.32	0.99	0.011	<e-5	0.46	1.00E-05
	Control vs single cis-eQTL regions	0.22	0.91	<e-5	<e-5	0.0034	0.0024
	CTL vs random	0.22	0.11	0.96	0.86	0.0034	6.00E-05
	Random vs	7.00E-05	0.37	<e-5	<e-5	<e-5	<e-5

	single cis-eQTL regions			

Table IV.V – Normalized abundance of polymorphic and fixed TEs in several sizes of genomic regions around "500 kB" clusters. Normalized abundance of TEs was calculated by dividing the abundance of each element with that found in random regions. Test regions corresponded to those containing cis-eQTLs (either clustered or single) or control clusters, each region being augmented by flanking regions of either 250, 500, 1000 kb.

		Polymorphic				Fixed	
		full LINEs	LINE frag.	LTR-TEs	SINEs	LTR-TEs	SINEs
Size of flanking regions	250 kB						
Normalized abundance of TEs / MB (mean ± SD)	Cis-eqtl clusters (n=53)	0.66±1.33	0.96±1.59	2.24±1.73	4.00±2.84	1.57±1.99	2.85±2.8
	CTL clusters (n=59)	0.89±1.71	0.72±0.9	1.08±1.1	1.29±1.48	1.36±1.17	1.91±1.43
	Single cis-eQTL regions (n=232)	0.69±1.26	0.97±1.19	1.69±1.29	1.76±1.6	1.43±1.14	1.64±1.6
	Random regions (n=500)	1±1.9	1±1.56	1±1.28	1±1.6	1±1.19	1±1.33
Anova P values		0.099	0.56	8.05E-16	3.05E-32	9.33E-06	1.16E-18
Post-hocTukey P values	Cis-eqtl clusters vs CTL	0.88	0.8	2.00E-05	<e-5	0.81	0.0078
	Cis-eqtl clusters vs single cis-eQTL regions	0.99	1	0.028	<e-5	0.88	<e-5
	Cis-eqtl clusters vs random	0.5	0.99	<e-5	<e-5	0.008	<e-5
	Control vs single cis-eQTL regions	0.84	0.63	0.0071	0.22	0.98	0.62
	CTL vs random	0.96	0.48	0.97	0.59	0.14	0.00012
	Random vs single cis-eQTL regions	0.09	0.99	<e-5	<e-5	8.00E-05	<e-5
Size of flanking regions	500 kB						
Normalized abundance of TEs / MB (mean ± SD)	Cis-eqtl clusters (n=53)	0.69 ± 1.38	1.08 ± 1.73	1.80 ± 1.32	3.41 ± 2.19	1.36 ± 1.16	2.39 ± 1.77

	CTL clusters (n=59)	1.10 ± 1.9	0.88 ± 1.05	1.08 ± 1.07	1.14 ± 1.16	1.32 ± 1.1	1.60 ± 1.28
	Single cis-eQTL regions (n=232)	0.8± 1.2	1± 1.11	1.56± 1.1	1.58± 1.25	1.36± 1.06	1.49± 1.41
	Random regions (n=500)	1 ± 1.67	1 ± 1.49	1 ± 1.15	1 ± 1.54	1 ± 1.03	1 ± 1.22
Anova P values		0.19	0.8	2.65E-11	5.09E-27	3.93E-05	2.80E-14
Post-hocTukey P values	Cis-eqtl clusters vs CTL	0.49	0.86	0.0046	<e-5	0.99	0.0092
	Cis-eqtl clusters vs single cis-eQTL regions	0.96	0.97	0.51	<e-5	0.99	5.00E-051
	Cis-eqtl clusters vs random	0.50	0.97	1.00E-05	<e-5	0.088	<e-5
	Control vs single cis-eQTL regions	0.52	0.93	0.019	0.17	0.99	0.93
	CTL vs random	0.96	0.91	0.96	0.9	0.117	0.0054
	Random vs single cis-eQTL regions	0.35	0.99	<e-5	1.00E-05	9.00E-05	2.00E-05
Size of flanking regions	1 MB						
Normalized abundance of TEs / MB (mean ± SD)	Cis-eqtl clusters (n=53)	0.76 ± 1.21	1.06 ± 1.34	1.58 ± 1.04	2.93 ± 1.87	1.18 ± 0.86	2.02 ± 1.45
	CTL clusters (n=59)	1.01 ± 1.43	0.94 ± 1.08	1.03 ± 0.91	1.06 ± 1.02	1.16 ± 0.85	1.31 ± 0.95
	Single cis-eQTL regions (n=232)						
	Random regions (n=500)	1 ± 1.4	1 ± 1.36	1 ± 0.99	1 ± 1.35	1 ± 0.84	1 ± 1.07
Anova P values		1.97E-05	0.42	8.00E-10	3.84E-26	3.43E-13	4.94E-19
Post-hocTukey P values	Cis-eqtl clusters vs CTL	0.78	0.97	0.16	1.00E-05	0.9	0.04
	Cis-eqtl clusters vs single cis-eQTL regions	0.47	0.7	0.9	0.75	0.012	0.99
	Cis-eqtl clusters vs random	0.65	0.99	0.022	<e-5	0.65	<e-5
	Control vs single cis-eQTL	0.034	0.9	0.004	0	0.0047	0.0036

regions						
CTL vs random	0.99	0.99	0.99	0.99	0.72	0.39
Random vs single cis-eQTL regions	1.00E-05	0.39	<e-5	<e-5	<e-5	<e-5

Table IV.VII – Comparisons for respective normalized abundance of binding sites for regulatory factors in several sizes of genomic regions around "250 kB" clusters. Normalize abundances of binding sites for regulatory factors was calculated in regions containing cis-eQTL cluster, control cluster, single cis-eQTL region or random regions, and augmented by flanking region of either 250, 500, 1000 kb.

		CTSF	h3ac	gata4	MEF2A	NKX2_5	SRF	TBX5	P300
Size of flanking regions	250 kB								
Normalized abundance of TFs / MB (mean ±SD)	Cis-eqtl clusters (n=42)	2.81 ±0.94	4.20 ±2.89	2.10 ±1.65	2.67 ±4.8	2.11 ±1.65	4.02 ±2.92	3.65 ±2.52	2.43 ±2.22
	CTL clusters (n=188)	1.76 ±0.69	1.85 ±1.19	1.72 ±1.29	1.47 ±3.27	1.86 ±1.62	1.87 ±1.51	1.91 ±1.67	1.64 ±1.8
	Single cis-eQTL regions (n=232)	1.63±0.73	1.81±1.6	1.68±1.4	1.34±2.53	1.64±1.41	1.9±1.85	1.85±1.73	1.7±1.89
	Random regions (n=500)	1 ±0.89	1 ±1.57	1 ±1.46	1 ±4.16	1 ±1.63	1 ±1.58	1 ±1.57	1 ±1.61
Anova P values		4.4E-56	7.65E-37	6.20E-14	0.026	8.29E-13	5.89E-31	3.68E-27	1.48E-10
Post-hoc Tukey P values	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.40684	0.22	0.77	<e-5	<e-5	0.039
	Cis-eqtl clusters vs single cis-eQTL regions	<e-5	<e-5	0.29376	0.140	0.27	<e-5	<e-5	0.059
	Cis-eqtl clusters vs random	<e-5	<e-5	1.00E-05	0.025	7.00E-05	<e-5	<e-5	<e-5
	Control and single cis-eQTL regions	0.41	0.99	0.98	0.98	0.49	0.99	0.97	0.98
	CTL vs random	<e-5	<e-5	<e-5	0.44	<e-5	<e-5	<e-5	0.00014
	Random vs single cis-eQTL regions	<e-5	<e-5	<e-5	0.65	<e-5	<e-5	<e-5	<e-5
	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.4	0.22	0.77	<e-5	<e-5	0.039
Size of flanking regions	500 kB								

Normalized abundance of TFs / MB (mean ±SD)	Cis-eqtl clusters (n=42)	2.43 ±0.78	3.60 ±2.36	2.06 ±1.5	2.67 ±4.69	2.06 ±1.69	3.47 ±2.58	3.24 ±2.17	2.03 ±1.64
	CTL clusters (n=188)	1.57 ±0.59	1.63 ±1.14	1.60 ±1.07	1.42 ±3.16	1.79 ±1.56	1.69 ±1.33	1.72 ±1.54	1.33 ±1.27
	Single cis-eQTL regions (n=232)	1.5±0.67	1.61±1.33	1.59±1.18	1.26±2.11	1.58±1.32	1.7±1.65	1.66±1.56	1.4±1.3
	Random regions (n=500)	1 ±0.84	1 ±1.42	1 ±1.26	1 ±3.44	1 ±1.44	1 ±1.47	1 ±1.48	1 ±1.31
AnovaP values		1.88E-41	1.06E-30	9.36E-15	0.0081	5.40E-13	6.99E-25	6.61E-22	1.34E-07
Post-hoc Tukey P values	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.12	0.099	0.69	<e-5	<e-5	0.0094
	Cis-eqtl clusters vs single cis-eQTL regions	<e-5	<e-5	0.1	0.04	0.195	<e-5	<e-5	0.02
	Cis-eqtl clusters vs random	<e-5	<e-5	<e-5	0.0063	4.00E-05	<e-5	<e-5	1.00E-05
	Control and single cis-eQTL regions	0.76	0.99	0.99	0.95	0.45	0.99	0.979	0.94
	CTL vs random	<e-5	<e-5	<e-5	0.41	0<e-5	<e-5	<e-5	0.019
	Random vs single cis-eQTL regions	<e-5	<e-5	<e-5	0.73	<e-5	<e-5	<e-5	8.00E-04
	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.12	0.099	0.690	<e-5	<e-5	0.009
Size of flanking regions	1 MB								
Normalized abundance of TFs / MB (mean ±SD)	Cis-eqtl clusters (n=42)	2.15 ±0.7	3.16 ±1.98	1.87 ±1.18	2.63 ±5.5	1.82 ±1.43	2.99 ±2.2	2.90 ±1.94	1.69 ±1.11
	CTL clusters (n=188)	1.45 ±0.56	1.48 ±0.97	1.45 ±0.87	1.38 ±3.2	1.59 ±1.33	1.51 ±1.19	1.53 ±1.35	1.23 ±0.96
	Single cis-eQTL regions (n=232)	1.4±0.6	1.44±1.12	1.44±0.96	1.16±1.71	1.45±1.13	1.49±1.35	1.47±1.34	1.29±0.95
	Random regions (n=500)	1 ±0.76	1 ±1.27	1 ±1.01	1 ±2.92	1 ±1.17	1 ±1.32	1 ±1.34	1 ±1.08
Anova P values		4.82E-33	4.59E-27	7.69E-14	0.004	6.77E-11	1.98E-20	2.12E-18	5.31E-06
Post-hoc Tukey P values	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.05	0.05	0.65	<e-5	<e-5	0.038
	Cis-eqtl clusters vs single cis-eQTL regions	<e-5	<e-5	0.04	0.013	0.25	<e-5	<e-5	0.087
	Cis-eqtl clusters vs random	<e-5	<e-5	<e-5	0.0028	0.00014	<e-5	<e-5	0.00017

	Control and single cis-eQTL regions	0.82	0.97	0.99	0.86	0.65	0.99	0.97	0.92
	CTL vs random	<e-5	3.00E-05	<e-5	0.42	<e-5	6.00E-05	4.00E-05	0.05
	Random vs single cis-eQTL regions	<e-5	5.00E-05	<e-5	0.9	2.00E-05	4.00E-05	9.00E-05	0.002
	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.059	0.056	0.65	<e-5	<e-5	0.038

Table IV.VIII – Comparisons for respective normalized abundance of binding sites for regulatory factors in several sizes of genomic regions around "500 kB" clusters. Normalize abundances of binding sites for regulatory factors was calculated in regions containing cis-eQTL cluster, control cluster, single cis-eQTL region or random regions, and augmented by flanking region of either 250, 500, 1000 kb.

		CTSF	h3ac	gata4	MEF2A	NKX2_5	SRF	TBX5	P300
Size of flanking regions	250 kB								
Normalized abundance of TFs / MB (mean ±SD)	Cis-eqtl clusters (n=53)	2.50 ±1.01	3.60 ±2.53	1.96 ±1.44	2.15 ±4.14	1.91 ±1.5	3.30 ±2.59	3.18 ±2.31	2.22 ±1.81
	CTL clusters (n=59)	1.49 ±0.56	1.53 ±1.06	1.47 ±1.37	1.30 ±2.43	1.45 ±1.27	1.37 ±1.23	1.20 ±0.94	1.31 ±1.42
	Single cis-eQTL regions (n=232)	1.58±0.71	1.7±1.45	1.63±1.29	1.27±2.21	1.62±1.39	1.81±1.75	1.75±1.63	1.53±1.52
	Random regions (n=500)	1 ±0.86	1 ±1.47	1 ±1.34	1 ±3.71	1 ±1.55	1 ±1.51	1 ±1.52	1 ±1.41
Anova P values		6.89E-41	2.25E-30	3.74E-11	0.1	1.80E-08	3.74E-23	2.03E-22	1.15E-09
Post-hoc Tukey P values	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.21	0.53	0.35	<e-5	<e-5	0.0062
	Cis-eqtl clusters vs single cis-eQTL regions	<e-5	<e-5	0.37	0.30	0.56	<e-5	<e-5	0.011
	Cis-eqtl clusters vs random	<e-5	<e-5	<e-5	0.079	0.00014	<e-5	<e-5	<e-5
	Control and single cis-eQTL regions	0.86	0.86	0.83	0.99	0.86	0.25	0.08	0.73
	CTL vs random	1.00E-04	0.058	0.053	0.91	0.12	0.35	0.79	0.42
	Random vs single cis-eQTL regions	<e-5	<e-5	<e-5	0.74	<e-5	<e-5	<e-5	4.00E-05
	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.21	0.53	0.35	<e-5	<e-5	0.0062

Size of flanking regions	500 kB								
Normalized abundance of TFs / MB (mean ±SD)	Cis-eqtl clusters (n=53)	2.21 ±0.78	3.08 ±2.2	1.79 ±1.34	2.19 ±4.77	1.72 ±1.35	2.82 ±2.36	2.81 ±2.03	1.87 ±1.38
	CTL clusters (n=59)	1.26 ±0.45	1.15 ±0.75	1.30 ±1.1	1.23 ±2.17	1.30 ±1.26	1.11 ±0.95	0.99 ±0.79	1.07 ±1.05
	Single cis-eQTL regions (n=232)	1.46 ±0.66	1.56 ±1.26	1.54 ±1.1	1.25 ±1.99	1.52 ±1.24	1.62 ±1.54	1.59 ±1.49	1.33 ±1.12
	Random regions (n=500)	1 ±0.81	1 ±1.36	1 ±1.16	1 ±3.21	1 ±1.33	1 ±1.42	1 ±1.44	1 ±1.2
Anova P values		2.53E-31	1.46E-24	2.40E-10	0.0467	1.93E-07	1.39E-17	2.76E-18	4.06E-07
Post-hoc Tukey P values	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.10	0.32	0.33	<e-5	<e-5	0.0022
	Cis-eqtl clusters vs single cis-eQTL regions	<e-5	<e-5	0.45	0.165	0.76	<e-5	<e-5	0.016
	Cis-eqtl clusters vs random	<e-5	<e-5	1.00E-05	0.03	0.0008	<e-5	<e-5	<e-5
	Control and single cis-eQTL regions	0.28507	0.18261	0.51	0.99	0.65	0.094	0.025	0.43
	CTL vs random	0.05502	0.84625	0.22	0.94	0.32	0.94	0.99	0.97
	Random vs single cis-eQTL regions	<e-5	<e-5	<e-5	0.71	<e-5	<e-5	<e-5	0.002
	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.1	0.32	0.33	<e-5	<e-5	0.002
Size of flanking regions	1 MB								
Normalized abundance of TFs / MB (mean ±SD)	Cis-eqtl clusters (n=53)	1.97 ±0.65	2.61 ś1.79	1.64 ±1.03	1.97 ±4.7	1.53 ±1.15	2.43 ±1.9	2.48 ±1.79	1.69 ±1.06
	CTL clusters (n=59)	1.14 ±0.47	1.02 ±0.66	1.16 ±0.81	0.96 ±1.59	1.16 ±0.97	0.96 ±0.77	0.88 ±0.67	1.02 ±0.2
	Single cis-eQTL regions (n=232)	1.38±0.6	1.42±1.1	1.42±0.91	1.1±1.58	1.42±1.07	1.46±1.32	1.45±1.32	1.3±0.95
	Random regions (n=500)	1 ±0.74	1 ±1.27	1 ±0.97	1 ±3.22	1 ±1.15	1 ±1.3	1 ±1.32	1 ±1.06
Anova P values		2.61E-25	1.28E-18	1.80E-09	0.13	3.93E-06	4.28E-14	3.91E-15	6.14E-07
Post-hoc Tukey P values	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.042	0.25	0.28	<e-5	<e-5	0.0025

	Cis-eqtl clusters vs single cis-eQTL regions	<e-5	<e-5	0.43	0.19	0.9	1.00E-05	<e-5	0.055
	Cis-eqtl clusters vs random	<e-5	<e-5	2.00E-05	0.092	0.0056	<e-5	<e-5	1.00E-05
	Control and single cis-eQTL regions	0.087	0.12	0.25	0.98	0.38	0.04	0.016	0.21
	CTL vs random	0.41	0.99	0.59	0.99	0.73	0.99	0.91	0.99
	Random vs single cis-eQTL regions	<e-5	0.00014	<e-5	0.97	2.00E-05	9.00E-05	1.00E-04	0.0009
	Cis-eqtl clusters vs CTL	<e-5	<e-5	0.04	0.25	0.28	<e-5	<e-5	0.0025

Table IV.X – Control region identity.

chr	Start	end	length	Number of gene	chr	Start	end	length	Number of gene
1	9892083	10126800	234717	3	1	34364784	34530250	165466	6
1	37933994	38110953	176959	3	1	42962332	43180339	218007	4
1	43946256	44237787	291531	4	1	55118125	55295867	177742	8
1	58463097	58648085	184988	4	1	59543333	59925390	382057	4
1	66766171	67053560	287389	5	1	82717459	82892697	175238	4
1	87995938	88246283	250345	4	1	133609747	133879153	269406	4
1	134229735	134467179	237444	6	1	135033439	135279122	245683	3
1	192994483	193176877	182394	3	1	194939988	195169842	229854	5
2	5715751	6110209	394458	6	2	26127339	26491428	364089	10
2	26761175	27038435	277260	9	2	35037708	35256012	218304	4
2	35892083	36244889	352806	5	2	38860274	39052525	192251	4
2	70925385	71279462	354077	4	2	76218480	76542387	323907	7
2	92221188	92455260	234072	7	2	93675795	93873115	197320	3
2	101527448	101741505	214057	3	2	112102168	112471031	368863	4
2	128627826	128959590	331764	5	2	130895173	131122079	226906	8
2	131763975	132133410	369435	6	2	150409024	150760540	351516	6
2	167302162	167515635	213473	5	2	172840872	173050858	209986	4
3	10204363	10418866	214503	4	3	58918604	59237149	318545	4
3	79407119	79617315	210196	3	3	97364702	97725625	360923	4
3	102861766	103158174	296408	5	3	103539478	103820603	281125	6
3	104585562	104806876	221314	6	3	116251960	116621511	369551	5
3	137949893	138220095	270202	4	4	3476502	3869959	393457	5
4	34516673	34752483	235810	6	4	43956907	44050029	93122	3
4	44944692	45045162	100470	4	4	48445993	48686348	240355	3
4	53635123	53873159	238036	3	4	56763945	56923001	159056	4
4	59704132	59912499	208367	3	4	63075995	63247174	171179	3
4	88826509	89023842	197333	3	4	94541218	94716099	174881	4
4	106720872	107086911	366039	7	4	134769923	134968375	198452	3

4	138614863	138903887	289024	6	4	147808906	147986910	178004	5
5	29864943	30133122	268179	4	5	30398188	30595065	196877	3
5	34240926	34530041	289115	4	5	37138706	37357924	219218	3
5	52541483	52929301	387818	4	5	72548411	72691358	142947	3
5	104284099	104508518	224419	4	5	108090538	108336983	246445	3
5	110533358	110791671	258313	5	5	114063418	114292521	229103	5
6	30316447	30698055	381608	6	6	38452236	38648244	196008	3
6	42200310	42308373	108063	4	6	47746681	47882602	135921	4
6	56651234	56832801	181567	4	6	57456714	57702727	246013	4
6	85026874	85418430	391556	7	6	112280143	112558565	278422	4
6	115527976	115626805	98829	4	6	116359204	116602648	243444	4
6	118333700	118637979	304279	4	6	119170128	119382978	212850	3
6	128271612	128594973	323361	6	6	146498454	146727298	228844	4
7	3571094	3657386	86292	6	7	13478456	13844665	366209	9
7	25086929	25393804	306875	5	7	97357127	97617517	260390	4
7	104498341	104867087	368746	7	7	105200706	105365189	164483	3
7	119495339	119735112	239773	3	7	128050342	128220181	169839	3
7	132605660	132989336	383676	5	7	140863620	140912909	49289	5
8	4207930	4288486	80556	5	8	11448895	11611935	163040	3
8	13779259	14030618	251359	6	8	111864581	112195339	330758	7
8	123104860	123280378	175518	5	9	14553084	14810365	257281	7
9	30957213	31228775	271562	4	9	37345694	37504507	158813	4
9	39935080	40105631	170551	4	9	50404638	50704820	300182	8
9	53301318	53492487	191169	3	9	56746850	57003281	256431	7
9	57261790	57598872	337082	9	9	66119969	66366225	246256	3
9	66623305	66870640	247335	5	9	89970621	90097287	126666	4
9	92161335	92387120	225785	3	9	99288578	99608787	320209	5
9	102989862	103255747	265885	5	9	109744464	109986898	242434	3
9	110613607	110957687	344080	6	9	119835773	120042954	207181	6
9	122745295	123066464	321169	5	10	23589350	23908373	319023	4
10	75814298	76171807	357509	5	11	16871508	17133167	261659	4
11	29409084	29633841	224757	4	11	30780090	31001177	221087	4
11	32089230	32418463	329233	7	11	45766325	46007422	241097	4
11	49985828	50203244	217416	9	11	71967710	72264972	297262	4
11	74483874	74722410	238536	5	11	75859848	76071590	211742	5
11	81860686	82132260	271574	5	11	83218276	83517653	299377	8
11	87864773	88158143	293370	6	11	109259312	109534382	275070	6
12	4720186	4898178	177992	4	12	16600956	16995887	394931	5
12	21275379	21396395	121016	6	12	70845208	71066515	221307	3
12	72002929	72224332	221403	4	12	73092659	73313164	220505	4
12	105459191	105652458	193267	4	13	12361849	12701141	339292	4
13	24844034	25003469	159435	4	13	43233689	43507889	274200	6
13	58277574	58497600	220026	3	13	60864053	61040558	176505	3
13	67003679	67240520	236841	3	13	67591637	67806544	214907	3
13	75988411	76139184	150773	3	13	104813558	104991968	178410	3

14	51440784	51765021	324237	12	14	69984618	70187083	202465	3
14	122353116	122558694	205578	3	15	34070876	34418401	347525	5
15	66681124	66937975	256851	4	15	74499639	74875995	376356	9
15	84383914	84770211	386297	5	15	102983992	103361366	377374	9
16	8409388	8691249	281861	8	16	10282071	10530484	248413	3
16	10959689	11220728	261039	6	16	13702461	13839396	136935	4
16	14192467	14396520	204053	3	16	21934746	22059399	124653	3
16	35825402	36042858	217456	3	16	37510125	37647753	137628	3
16	57154411	57302273	147862	3	16	87367848	87730871	363023	4
16	91425732	91695575	269843	4	16	93593507	93919655	326148	5
16	94351302	94589223	237921	3	17	5924620	6085325	160705	3
17	20952082	21165221	213139	3	17	23694083	23964537	270454	10
17	36996204	37114756	118552	5	17	45528969	45835242	306273	8
17	56096233	56418524	322291	13	17	65787575	66001307	213732	3
18	20761001	21160616	399615	5	18	31740039	32118161	378122	5
18	34726697	35024602	297905	5	18	35741609	35966750	225141	7
18	36664986	36923514	258528	8	18	46664613	46873043	208430	3
18	67366537	67712294	345757	6	19	11661384	12019235	357851	5
19	12664044	12907966	243922	3	19	24752192	24995031	242839	3
19	41937464	42269037	331573	7	19	43574732	43810008	235276	3
19	44420721	44624813	204092	3	19	60833888	60914715	80827	3

Table IV.VI – Most significantly enriched binding sites in polymorphic SINEs.

Motif Name (family)	CHIP-Seq cell line	Consensus	P-value	% of Target Sequences with Motif	% of Background Sequences with Motif
CTCFL /BORIS (Zinc finger)	K562	CNNBRGCGCCCCCTGSTGGC	1e-1283	63.96%	3.33%
ZNF143 (Zinc finger)HPC7	CUTTL	ATTTCCCAGVAKSCY	1e-1273	73.63%	5.91%
c-Myc	mES	VVCCACGTGG	1e-1213	75.98%	7.34%
n-Myc (Helix-Loop-Helix)	mES	HPC7VRCCACGTGG	1e-1116	76.39%	8.77%
Max (Helix-Loop-Helix)	K562	RCCACGTGGYYN	1e-1094	77.54%	9.57%
RUNX (Runt)	HPC7	SAAACCACAG	1e-1000	76.44%	10.61%
Gfi1b (Zinc finger)	HPC7	MAATCACTGC	1e-996	72.28%	8.82%
RUNX2 (Runt)	PCa	NWAACCACADNN	1e-906	76.44%	12.37%
RUNX1 (Runt)	Jurkat	AAACCACARM	1e-820	76.70%	14.39%
Cdx2 (Homeobox)	mES	GYMATAAAAH	1e-623	63.70%	11.95%
Olig2 (Helix-Loop-Helix)	Neuron	RCCATMTGTT	1e-613	79.10%	22.14%
HNF4a (Nuclear Receptors)	HpeG2	CARRGKBCAAAGTYCA	1e-359	33.54%	4.58%
GATA3 (Zinc finger)	iTreg	AGATSTNDNNDSAGATAASN	1e-322	23.19%	1.97%
Erra(NR)	HepG2	CAAAGGTCAG	1.00E-304	62.35%	22.53%
FOXA1 (Forkhead)	MCF7	WAAGTAAACA	1.00E-304	53.46%	16.26%
Smad3 (MAD)	NPC	TWGTCTGV	1.00E-265	68.28%	29.69%
FOXA1 (Forkhead)	LNCaP	WAAGTAAACA	1.00E-265	54.08%	18.54%
AR-halfsite (Nuclear Receptors)	LNCaP	CCAGGAACAG	1.00E-232	78.00%	41.54%
Nkx2.5 (Homeobox)	HL1	RRSCACTYAA	1.00E-182	56.89%	25.78%
Nanog (Homeobox)	mES	RGCCATTAAC	1.00E-106	69.53%	44.63%
NeuroD1 (Helix-Loop-Helix)	Islet	GCCATCTGTT	1.00E-102	25.07%	8.50%
CTCF (Zinc finger)	CD4+	AYAGTGCCMYCTRGTGGCCA	1.00E-67	9.26%	1.82%

Table IV.IX – Descriptive information concerning the 42 250 kB cis-eQTL clusters.

Chr	Start	end	Length	Number of cis-eQTsL	Symbols of cis-eQTL genes
1	172981698	173434853	453155	11	FCGR3,1700009P17RIK,SDHCAPOA2,FCER1G B4GALT3,PPOX,UFC1,KLHDC9,F11R,REFBP2
1	173934520	174206100	271580	5	VANGL2,NCSTN,COPA,PEX19,ATP1A2
1	182837968	183096385	258417	4	PYCR2,LEFTY1,TMEM63A,CNIH4
2	25207840	25354300	146460	4	DPP7,UAP1L1,ENTPD2,C8G
2	103921214	103964818	43604	3	CD59B,CD59A,A930018P22RIK
2	152590596	153048274	457678	6	COX4I2,FKHL18,PDRG1,BC020535,TM9SF4,TSPYL3
3	35794225	35988326	194101	3	LOC100046841,MCCC1,ACAD9
3	87719934	87861011	141077	3	HDGF,NES,APOA1BP
4	41585069	41791875	206806	3	DNAIC1,CCL27,CCL19
4	62160379	62363247	202868	3	HDHD3,ALAD,RGS3
4	129193683	129525601	331918	4	LOC100046039,X2510006D16RIK,CCDC28B,PTP4A2
4	132086525	132456695	370170	4	ATPIF1,EYA3,XKR8,BC008163
4	132754041	133147532	393491	7	WASF2,MAP3K6SLC9A1,4732473B16RIK, 2300002D11RIK,NUDCGPN2
4	133912418	134093953	181535	4	EXTL1,STMN1,2410166I05RIK,SEPN1
4	155207073	155364772	157699	6	AURKAIP1,DVL1,ACAP3,LOC545056,FAM132A,B3GALT6
5	147765528	147890516	124988	3	GTF3A,MTIF3,POLR1D
6	126875000	127079686	204686	3	RAD51AP1,X9630033F20RIK,CCND2
6	145123111	145168681	45570	4	LRMP,CASC1,LYRM5,KRAS
7	19441998	19831693	389695	4	MILL2,IRF2BP1,DMWD,D630048P19RIK
7	30975506	31426230	450724	5	CAPNS1,TBCB,TYROBP,HSPB6,RBM42
7	87461806	87505819	44013	4	RCCD1,LOC675567,UNC45A,MAN2A2
8	32207411	32371718	164307	3	DUSP26,RBM13,FUT10
8	87194527	87621384	426857	11	NACC1,TRMT1,NFIX,FARSA,GCDH,PRDX2,HOOK2, ASNA1,DHPS,1500041N16RIK,MAN2B1
8	124990160	125154233	164073	3	RNF166,TRAPPC2L,CBFA2T3H
9	34933919	35024383	90464	3	DCPS,FOXRED1,SRPR
9	44199978	44212071	12093	3	HYOU1,SLC37A4,TRAPPC4
9	44806893	44962140	155247	3	CD3E,AMICA1,SCN4B
9	106097518	106372827	275309	7	PPM1M,TWF2,DUSP7,RPL29,ACY1,ABHD14A,PARP3
11	58804677	59235079	430402	4	TRIM11,GJC2,MRPL55,1110031B06RIK
11	59589470	59689992	100522	3	MPRIP,COPS3,NT5M
11	82619501	82756537	137036	3	RFFL,LOC100044934,UNC45B
11	94830377	95136060	305683	5	SGCA,SAMD14,PDK2,ITGA3,MYST2
11	96799199	96910230	111031	3	PNPO,SCRN2,MRPL10
11	114844007	115044431	200424	3	4732429D16RIK,SLC9A3R1,LOC100044159
13	64234749	64540754	306005	6	ZFP367,HABP4,CDC14B,1110018J18RIK,CTSL,CCRK
13	113657734	114008245	350511	5	PPAP2A,SKIV2L2,GPX8,2310016C16RIK,ESM1
15	76545807	76682760	136953	3	LRRC24,C030006K11RIK,ZFP251
15	85727285	85967913	240628	3	TRMU,CELSR1,GRAMD4
16	20651876	20742404	90528	3	PSMD2,EIF4G1,CHRD
17	24645834	25029013	383179	5	TRAF7,GFER,NDUFB10,FAHD1,SPSB3
17	34056177	34163052	106875	3	ZBTB22,WDR46,H2.KE6
18	37847239	37909257	62018	3	PCDHGA4,PCDHGB2,PCDHGA10

Annexe V

Supplementary Materiel : Network analyses reveal strong contributions of chromosome domains to gene coexpression modules and a cardiac quantitative trait in mice

Table V.I – Characteristics of module QTLs (mQTLs) of "genetic" modules.The modules showing linkage to one main mQTL were divided into those showing evidence (or not) of being "chromosome-domain driven"(CDD). The names of each module correspond to those given by the WGCNA program. The characteristics of each mQTL correspond to the number of the chromosome harboring the mQTL, followed by its position (in Mb) on the chromosome. For 16/21 CDD modules, the peak of their mQTL was in very close vicinity of the position of a "cis-eQTL cluster", as reported by us previously (doi :10.1534/g3.112.005488).

Type of module	Module Names	Main mQTL position	mQTL LOD	Matching cis-eQTL clusters
CDD	bisque4	15@76.4	12.64	15@76.6
	brown4	9@44.5	11.67	9@44.2 / 9@44.8
	darkolivegreen	11@98.8	10.15	11@96.8
	darkred	6@144.9	6.43	6@145.1
	darkturquoise	9@105.9	17.9	9@106
	lightcyan1	13@113.2	9.9	13@113.7
	midnightblue	4@133.4	11.7	4@113.9
	orangered4	17@33.1	16.76	17@34
	palevioletred3	4@154.5	9.32	4@155.2
	plum1	2@26.1	6.83	2@25.2
	royalblue	11@57	7.35	11@58
	skyblue3	1@172.9	16	1@172.9
	thistle2	13@64.8	12.2	13@64.3
	cyan	2@113.4	14.1	2@103.9
	floralwhite	7@105.8	11.8	
	ivory	12@110.3	11.7	
	lightsteelblue1	13@47	6.8	
	mediumpurple3	7@50.6	9.3	
	plum2	17@9.08	16	
	sienna3	14@46.2	13.53	
	thistle1	5@119.3	10.35	
genetic non-CDD	darkslateblue	1@94.8	4.13	
	orange	1@116.9	3.9	
	pink	17@84.1	3.85	
	saddlebrown	17@89.2	4.1	

	salmon	12 @101.7	5.8	
	skyblue	16@20.4	4.16	

Table V.II – Properties of genes from "genetic" modules.

Type	Module Names	% of cis-eQTLs among module genes	% of module genes from pred. chr.	Predom. Chrom.	Mean Interval between genes from pred. chr. (MB)	Connectivity of genes from pred. chr vs. others
CDD	bisque4	22.41%	50.00%	15	8.13	2.56
	brown4	42.37%	54.24%	9	11.02	2.87
	darkolivegreen	32.38%	56.19%	11	13.34	2.38
	darkred	15.17%	27.59%	6	16.13	2.14
	darkturquoise	24.65%	42.25%	9	16.50	3.53
	lightcyan1	14.08%	22.54%	13	14.25	4.34
	midnightblue	40.12%	30.23%	8	22.58	1.57
	orangered4	16.46%	58.23%	17	12.67	3.94
	palevioletred3	27.91%	34.88%	4	11.89	2.95
	plum1	26.25%	51.25%	2	50.13	3.58
	royalblue	17.01%	26.53%	11	19.21	2.53
	skyblue3	31.87%	41.76%	1	19.26	5.18
	thistle2	32.65%	30.61%	13	13.13	3.23
	cyan	15.61%	39.31%	2	38.05	3.25
	floralwhite	22.39%	41.79%	7	21.66	3.95
	ivory	10.29%	32.35%	12	5.55	2.81
	lightsteelblue1	36.11%	36.11%	13	18.03	2.59
	mediumpurple3	29.11%	56.96%	7	29.18	2.96
	plum2	27.27%	56.36%	17	19.31	2.68
	sienna3	3.13%	30.21%	14	8.61	5.17
	thistle1	22.92%	47.92%	5	15.99	2.11
mean		42.29%	41.30%		18.3	3.2
sd		9.97%	11.48%		10.3	0.9
genetic non-CDD	darkslateblue	3.57%	10.71%	8	58.56	0.88
	orange	2.82%	10.56%	9	41.67	1.09
	pink	2.95%	10.82%	5	55.89	0.88
	saddlebrown	10.26%	14.53%	11	34.92	1.09
	salmon	5.06%	11.80%	4	68.43	1.00
	skyblue	11.86%	15.25%	9	49.27	0.76
mean		6.09%	12.28%		51.5	1.0
SD		3.97%	2.08%		12.1	0.1

Table V.III – Properties of genes from "non-genetic" modules.

Type	Module Names	% of cis-eQTLs among module genes	% of module genes from pred. chr.	Predom. Chrom.	Mean Interval between genes from pred. chr. (MB)	Connectivity of genes from pred. chr vs. others
non-genetic	black	2.29%	11.11%	11	32.14	1.04
	blue	1.60%	8.68%	11	41.88	0.77
	brown	4.09%	8.64%	2	52.58	1.15
	darkgreen	6.29%	9.79%	2	62.61	1.15
	darkgrey	4.93%	14.08%	10	54.85	1.04
	darkmagenta	3.49%	8.33%	11	35.08	0.86
	darkorange	3.67%	9.29%	10	43.45	0.91
	darkorange2	0.00%	8.96%	14	24.44	0.85
	green	3.37%	10.67%	7	62.93	1.01
	greenyellow	2.67%	10.67%	3	44.76	0.99
	grey60	1.86%	9.32%	9	35.42	1.12
	lightcyan	1.54%	12.62%	11	40.86	0.90
	lightyellow	6.49%	14.29%	2	62.59	0.81
	paleturquoise	0.00%	11.32%	5	57.00	0.91
	salmon4	2.17%	13.04%	11	19.50	0.89
	steelblue	3.48%	13.04%	5	49.19	0.80
	tan	1.64%	14.21%	10	30.12	0.96
	turquoise	3.82%	8.79%	9	45.22	0.73
	violet	0.94%	10.38%	5	58.83	0.81
	white	2.48%	8.26%	19	6.27	1.02
	yellow	2.42%	10.48%	3	52.77	1.13
	yellowgreen	1.09%	11.96%	1	82.19	1.04
mean		2.74%	10.81%		45.2	1.0
SD		1.74%	2.01%		17.0	0.1

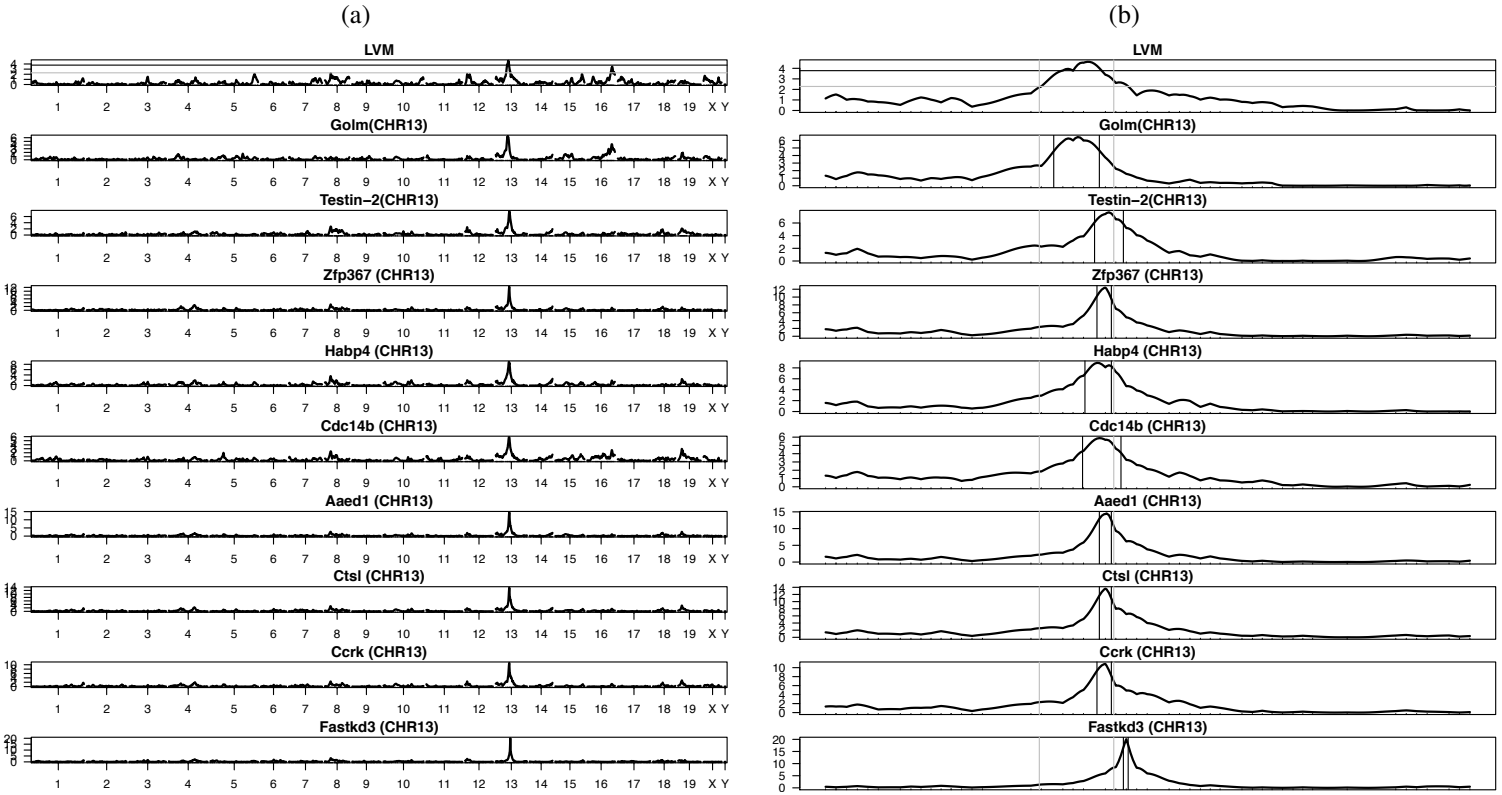


Figure V.1 – QTL mapping profiles for LVM and 9 cis-eQTL genes from chr13 whose expression correlates significantly with LVM. a) genome-wide profiles ; b) profiles at the level of chr13. For LVM, the horizontal black and grey lines correspond to the threshold levels of significant and suggestive QTLs, respectively. The light grey vertical lines represent the confidence interval for the phenotypic QTL Lvm1. The darker vertical lines correspond to the confidence intervals for the QTL of each respective gene. The confidence intervals of the first 8 cis-eQTLs fell within the boundaries of the confidence interval for Lvm1 ; the confidence interval of the eQTL for Fastkd3 fell just outside of these boundaries.

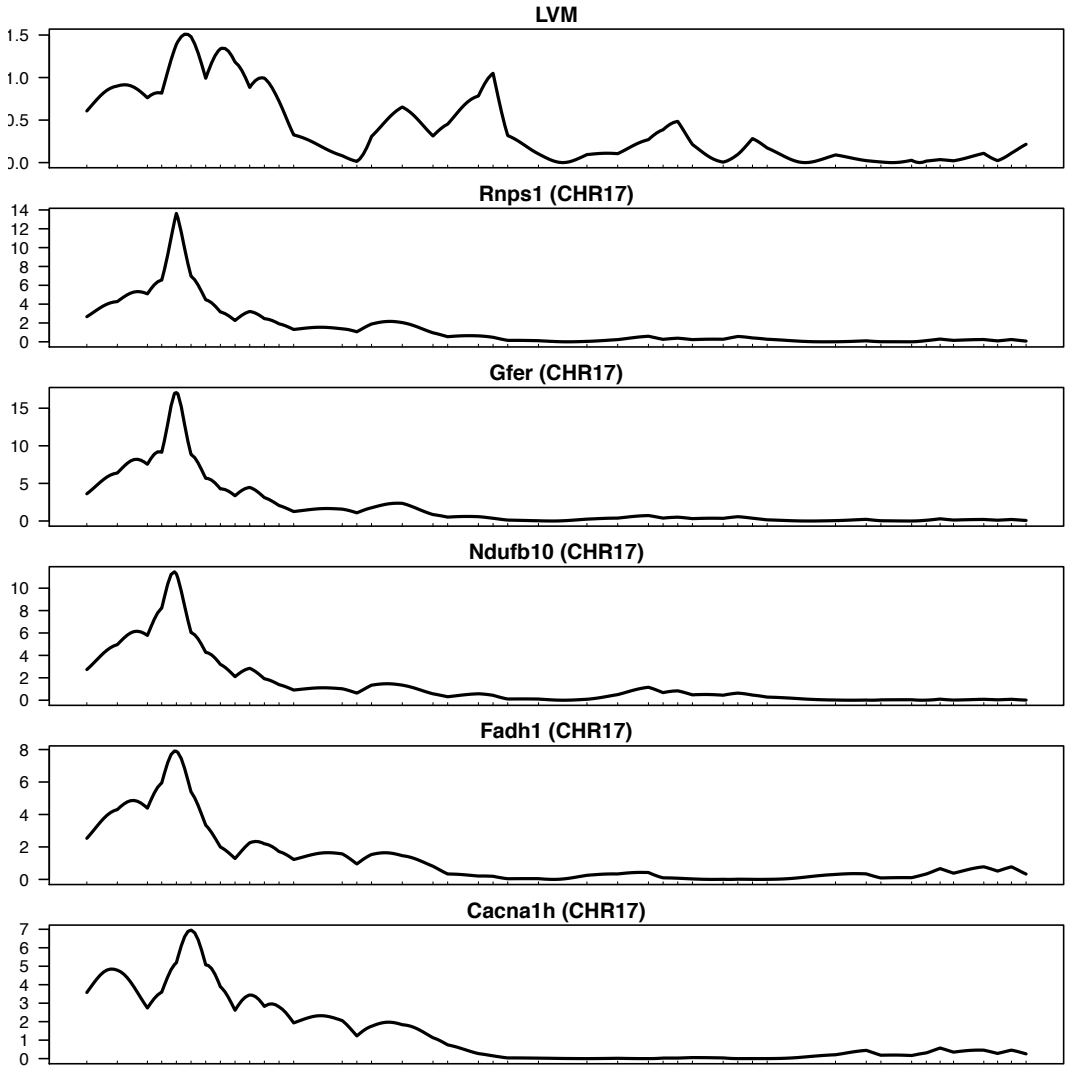


Figure V.2 – QTL mapping profiles (at the level of chr17) for LVM and the expression levels of 5 cis-eQTL genes from chr17 whose expression correlates significantly with LVM. Although the QTL for LVM was not significant, its profile was matched by that of all five cis-eQTL genes

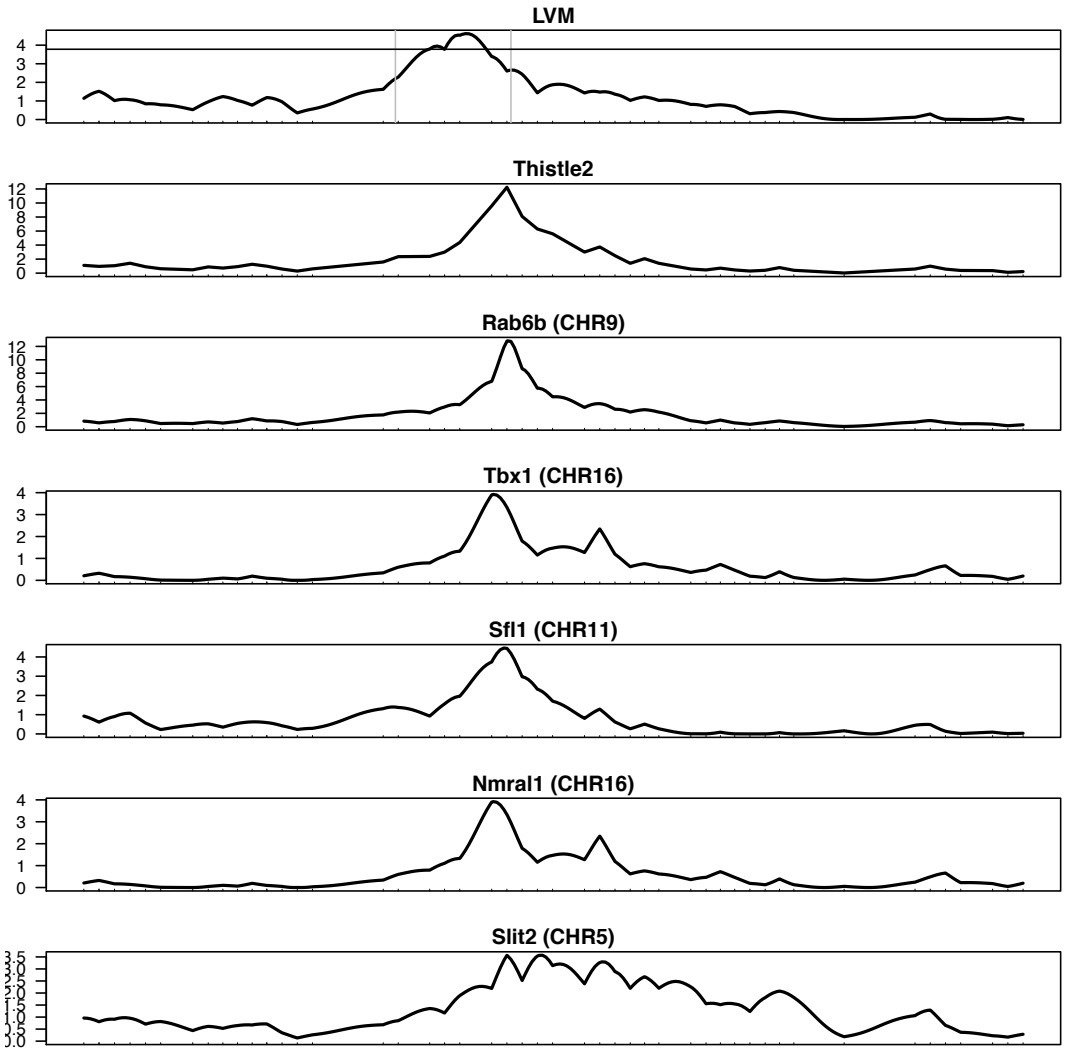


Figure V.3 – QTL mapping profiles (at the level of chr13) for LVM, for the thistle2 module, and 5 trans-eQTLs. The peak of the mQTL and that of all 5 trans-eQTLs fell within the boundaries of the confidence interval for Lvm1.