Approches intégrées du génome et du transcriptome dans les maladies complexes humaines

Maxime Rotival

l'Université Paris 11

N° attribué par la bibliothèque

THÈSE

pour obtenir le grade de

Docteur de l'Université Paris 11

Spécialité : **Statistique génétique**

préparée au laboratoire **INSERM UMRS 937: "Génomique cardiovasculaire"**

dans le cadre de l'École Doctorale **de santé publique de Paris-Sud**

présentée et soutenue publiquement

par

Maxime ROTIVAL

le 27 juin 2011

Titre:

Approches intégrées du génome et du transcriptome dans les maladies complexes humaines

Directrice de thèse: **Laurence TIRET**

Jury

M. Alexandre ALCAÏS,	Président
M. Pascal ROY,	Rapporteur
M. Jean-François ZAGURY,	Rapporteur
M. Stéphane ROBIN,	Examinateur
M. Philippe BROËT,	Examinateur
Mme Laurence TIRET,	Directrice de thèse

Aussi dis-je que celui-là s'expose à un grand danger qui se décide à publier un livre, car il est complètement impossible de le composer tel qu'il satisfasse tous ceux qui le liront.

Don Quichotte, Cervantès

Remerciements

Il parait que les remerciements sont l'élément central d'une thèse, autour duquel se structure tout le reste du travail effectué. Je vais donc tâcher de ne pas faillir et de n'oublier personne ! En tout premier lieu, je voudrais remercier mes rapporteurs, Pascal Roy et Jean-François Zagury qui ont accepté de relire cette thèse et de me donner un avis critique sur mon travail malgré leur emploi du temps chargé. Ensuite, je voudrais remercier "monsieur le président" Alexandre Alcaïs, et mes examinateurs Stéphane Robin et Philippe Broët, d'avoir accepté de prendre sur leur temps pour faire partie de mon Jury de thèse. Mes remerciement vont également à ma directrice de thèse, Laurence Tiret, qui a su me guider dans la jungle de la recherche publique et des collaborations internationales et m'aider à structurer mon travail, parfois chaotique, pour en extraire l'essentiel. Je voudrais également remercier François Cambien, le directeur de l'unité 937 pour m'avoir acceuilli dans son laboratoire et pour son flegme à toute épreuve.

Je tiens ensuite à remercier mes collègues de bureau, passés et présents. Viviane pour avoir été pendant longtemps la mémoire de ce bureau. Marie-Lise pour avoir su animer le labo jusque dans sa dernière ligne droite. Bénédicte pour m'avoir donné envie d'aller découvrir le Québec. Guillemette pour son entrain et sa simplicité. Nico pour ses répliques cultes ("On fait le 23 ! !"). Je le remercie également pour sa participation intensive à la mise en page de cette thèse [1]. Je remercie enfin ma collègue Raph, qui n'est pas vraiment de mon bureau, mais qui depuis le temps a bien fini par y gagner un siège honorifique. Un grand Merci à elle, pour toutes les discussions scientifiques et moins scientifiques [2] que nous avons eu qui ont participé à rendre ma thèse plus agréable. Votre bonne humeur à tous, m'a été d'un grand soutien pour la réalisation de ma thèse. Je tiens également à remercier la ribambelle des stagiaires qui ont fréquentés mon bureau (Caroline, Isabelle, Antoine, Charlotte.) et qui nous ont même parfois apporté des pains au chocolat ! La liste est longue, des gens que j'ai eu le plaisir de cotoyer pendant cette thèse, et si je ne m'étendrai pas davantage sur les cas de Marc, Nathalie, Ricardo, Ulrike, Marie, Françoise, Vinh,

1. Il serait temps d'arrêter de procrastiner et de faire avancer ta thèse maintenant !
2. Qui allaient des conseils capillaires aux raisonnements franchement capillo-tractés.

Nadjim, Tiphaine, Marine, Sophie, Christine, Carole, Claire, Sonia, Jean-Marc, Monique, Rajaï, Henri, Hervé, Maria, Lynda, Ewa ou David Tréchouët [3], ce n'est pas faute d'avoir matière à les remercier.

Je tiens ensuite à remercier mes amis. Fimon, bien fûr, mais aussi Jéjé, Emilie, Rominou, Julie, Mâche, Nouille, Rico, Obi, Bénédicte, Mélanie, Juliette, Robin et la Cecilia, Cédric, Emilien, Méloche, Sana, ...La grande famille des ENSAE (et consorts) avec Nav' et Poop's, Mr Jo, Le Brodeur (Allez Montferrand !), le Chat (Miaou !), et Bruno. Mais aussi mes comparses Fantinou, Célinou et Tim. Il y a aussi toute l'équipe du Havre : PF et Nath, Charlie et Claire, le gros JC, Nath et Céline, KQ, le Beubeu (meeeercciiiiiiiiiiii !), ou encore le Périgourdino (c'est lui le plus beau, c'est lui le plus fort !).

Enfin, la famille, sans qui je ne serais de toute évidence pas là. J'ai d'abord une pensée toute particulière pour mes grand-mères, à qui je ne donne que trop rarement de mes nouvelles, mais à qui je pense souvent. J'ai ensuite une pensée pour ma mère, qui est toujours disponible lorsque j'en ai besoin [4]. Son soutien compte beaucoup pour moi, même si je ne le montre que rarement. Je remercie également mon frère et ma soeur, pour leur présence tout au long de ma thèse, et pour me donner régulièrement de nouvelles occasions de voyager, et parfois des jolis neuveux et nièces. J'ai enfin un pensée pour mon père, qui aurait sûrement été content de me voir faire une thèse dans le domaine de la bio-info, après quelques "égarements" à faire des maths.

Ces remerciements, ne seraient pas tout à fait complets sans un mot pour tout ceux qui ont relu des morceaux de ma thèse et corrigés quelques-unes [5] de mes nombreuses fautes de frappes et de grammaire. Ils se reconnaitront. J'ai aussi une pensée pour tous ceux qui m'ont soutenu au quotidien tout au long de ma thèse, sans le savoir et sans même me connaitre : Tom Selleck, Jorge Cham, Emmanuel Cabut, Pierre Haski, Jimmy Wales, et tant d'autres ...

3. Liste fournie dans le désordre, et probablement non exhaustive.
4. Comme par exemple lorsqu'il s'agit d'organiser mon pot de thèse
5. La majeure partie j'espère

Liste des abbréviations

ACI Analyse en composantes indépendantes

ACP Analyse en composantes principales

WGCNA Weighted Gene Correlation Network Analysis

MDS Multidimensional Scaling

FDR False Discovery Rate

FWER Family Wise Error Rate

VST Variance Stabilizing Transform

IBS Identity by Descent

GWAS Genome Wide Association Study

eQTL expression Quantitative Trait Loci

SNP Single Nucleotide Polymorphism

MAF Minor Allele Frequency

RSAE Relative Strength of Association with Expression

IMC Indice de masse corporelle (BMI - Body Mass Index en anglais)

CRP C réactive Protein

LDL Low Density Lipoprotein

HDL High Density Lipoprotein

Table des matières

Introduction **1**

I Motivations 3

1 Transcriptome et maladies multifactorielles 5

1 Le transcriptome 5
1.1 De l'ADN à la protéine 5
1.2 Variabilité du transcriptome 7
1.3 Variabilité génétique et héritabilité du transcriptome 8
2 Études d'association en génétique 11
2.1 Les maladies multifactorielles 11
2.2 Stratégies de recherche 12
2.3 Utilisation du transcriptome 12

2 Le transcriptome en épidémiologie cardiovasculaire 15

1 Les pathologies cardiovasculaires 15
1.1 Incidence des pathologies cardiovasculaires 16
1.2 Etiologie et facteurs de risque de l'athérosclérose 16
2 GHS et Cardiogenics : étude du transcriptome à grande échelle 21
2.1 Objectif et caractéristiques des études 21
2.2 Le choix des monocytes 22
3 De l'utilité des atlas d'expression 24

II Mesure du génome et du transcriptome par biopuces 25

3 Généralités sur les biopuces 27

1 Principe de fonctionnement 28
2 Puces d'expression 29

2.1 Principe détaillé 29
2.2 Caractéristiques techniques 31
2.3 Différents types de puces 32
3 Puces de génotypage 32
3.1 Principe détaillé 32
3.2 Choix des SNP ciblés 34
3.3 Différents types de puces 34
4 Sources de biais dans les biopuces 35

4 Acquisition des mesures d'expression 37
1 Normalisation intra-puce 37
1.1 Contrôles de la qualité de l'hybridation 37
1.2 Agrégation des mesures 38
1.3 Traitement du bruit de fond 39
1.4 Contrôle de la variabilité des mesures 40
2 Normalisation inter-puces 44
2.1 Normalisation par quantiles 44
2.2 Normalisation par splines/lowess 45
2.3 Repérage d'échantillons atypiques 47
2.4 Classification des échantillons 47
3 Sources d'erreur non prises en compte par le prétraitement 49
3.1 Biais expérimentaux 49
3.2 Biais de construction 50

5 Acquisition des génotypes 55
1 Etapes de prétraitement 55
1.1 Hybridation et Normalisation 55
1.2 Détermination des génotypes 56
2 Contrôle qualité 58
2.1 Contrôle des échantillons 58
2.2 Contrôle des SNP 60

6 Problèmes liés à l'analyse des données de biopuces 63
1 Stratégies d'analyse des biopuces 63
2 Problématique des tests multiples 63
2.1 Rappels sur les tests d'hypothèse 63
2.2 Cas des tests multiples 65
2.3 Méthodes de correction de l'erreur de type I 65
3 Réduction de la dimension des données 68

3.1 Sélection des transcrits 68
3.2 Sélection des SNP 71
4 Biologie en grande dimension 71

III Analyse de la variabilité du transcriptome et intégration en génétique humaine 73

7 Modulation du transcriptome par la variabilité génétique 75
1 Mécanismes de régulation du transcriptome 75
1.1 Les facteurs de transcription 75
1.2 Le rôle de la variabilité génétique 77
2 Etude des eQTL 78
2.1 Stratégies d'étude 78
2.2 Choix de la distance *cis*/*trans* 78
2.3 Analyses univariées pour la recherche d'eQTL 80
3 Enseignements sur la régulation du transcriptome 81

8 Apport des eQTL dans les études d'association pan-génomiques 85
1 Association entre eQTL et loci de prédisposition identifiés par les GWAS 86
1.1 Méthode d'analyse 86
1.2 Co-localisation des eQTL et des loci de GWAS 87
1.3 Effet de la densité de gènes sur la co-localisation 89
2 Quelques exemples de co-localisation 91
3 Utilisation de la densité des gènes pour faciliter la recherche de loci de prédisposition dans les GWAS 92
3.1 Pondération dans les GWAS 93
3.2 Application à une GWAS de l'homocystéine plasmatique 93

9 Identification de modules de gènes *trans*-régulés 95
1 Principe de l'approche factorielle 95
1.1 Les méthodes factorielles 95
1.2 Extraction des composantes 97
1.3 Identification de modules 102
1.4 Association aux génotypes 104
2 Application de l'ACI à la recherche d'effets génétiques à grande échelle sur le transcriptome 106
2.1 Simulations 106
2.2 Application sur les données de GHS 113

3 Comparaison avec l'approche WGCNA 117
3.1 Principe général 117
3.2 Application aux données de GHS 118
4 Une application à l'étude du diabète de type I 119

10 Déconvolution de signaux dans un mélange de types cellulaires 123
1 La contamination dans l'étude GHS 124
1.1 L'origine de la contamination 124
1.2 Impact de la contamination 127
2 Estimation des proportions des différents types cellulaires 128
2.1 Méthode utilisée 128
2.2 Application aux données de GHS 133
3 Impact de la contamination sur les associations du transcriptome avec le génotype 135
3.1 Loci associés aux patterns d'expression et contamination 135
3.2 Association entre SNP et patterns de co-expression après ajustement sur les quantités des différents types cellulaires 137
3.3 Analyse détaillée de l'effet du SNP sur les gènes les plus associés . . 138
4 Bilan et perspectives 142

Conclusion et perspectives 145

Bibliographie 160

Annexes 163

Liste des articles 163
Article 1 180
Article 2 197
Article 3 212
Article 4 219
Article 5 238

Table des figures

1.1 Structure des deux brins complémentaires et antiparallèles. 5
1.2 Formation des protéines à partir de l'ADN. 6
1.3 Variabilité du transcriptome. 7
1.4 Polymorphisme génétique. 9
1.5 Distribution de l'héritabilité des transcrits dans une étude familiale. 10
1.6 Distribution de l'héritabilité des niveaux de transcrits dans du tissu rénal de rat. 11

2.1 Les différentes couches de la paroi artérielle. 17
2.2 Le processus de l'athérogénèse. 18

3.1 Protocole expérimental d'analyse d'une puce ARN. 30

4.1 Visualisation de l'image des intensités et qualité de l'hybridation. 38
4.2 Incertitude des mesures (écart-type) en fonction du niveau d'expression (moyenne des billes). 41
4.3 Variabilité de l'expression des gènes en fonction de leur niveau d'expression, selon la transformation utilisée. 42
4.4 Normalisation par quantiles. 45
4.5 Normalisation par splines/loess. 46
4.6 Contôle qualité : classification des échantillons. 48
4.7 Effet de la dégradation de l'ARN. 52

5.1 Visualisation des signaux associés aux sondes d'un SNP. 57
5.2 Affectation des individus aux trois génotypes d'un SNP. 57
5.3 Repérage d'individus atypiques et d'effets de stratification par MDS. 60

6.1 Expression des sondes du gène ZFY en fonction du sexe. 70

7.1 Régulation de l'expression des gènes par un facteur de transcription spécifique. 76

7.2 Mécanismes de modulation du transcriptome par les SNP. 77
7.3 Répartition chromosomique des eQTL observés dans GHS. 79
7.4 Distance du SNP au gène dans les eQTL. 80
7.5 Evolution du nombre d'associations en fonction de la part de variance de l'expression expliquée par le SNP. 83

8.1 Lien entre eQTL et loci de GWAS selon le type de trait étudié : 89
8.2 Localisation des loci de GWAS présentant un eQTL 90

9.1 Principe de la décomposition obtenue par l'ACI pour $K = 2$. 98
9.2 Screeplot des valeurs propres obtenues par l'ACP sur les données de GHS. 103
9.3 Définition du module à partir de la signature d'une composante. 105
9.4 Modèle de simulation utilisé. 110
9.5 QQ-plot des p-values d'association des 675 350 SNP avec les patterns extraits par l'ACI et WGCNA. 120

10.1 L'hématopoïèse. 125
10.2 Principe simplifié du tri cellulaire par sélection positive et négative. 126
10.3 Proportions estimées des différents types cellulaires dans les échantillons de HaemAtlas. 132
10.4 Distribution des quantités estimées des différents types cellulaires dans les données de GHS et Cardiogenics 134

Liste des tableaux

2.1 Principaux gènes associés à une prédisposition aux maladies cardiovasculaires. 20

3.1 Les différents types de biopuces et leur usage. 27
3.2 Caractéristiques des puces à ARN Illumina et Affymetrix. 33
3.3 Caractéristiques des puces de génotypage Illumina et Affymetrix. 35

6.1 Nombre de positifs et négatifs attendus pour n tests sous l'hypothèse nulle ou alternative. 65

7.1 *Cis* et *trans*-eQTL dans l'étude GHS. 82
7.2 Réplicabilité inter-tissus des eQTL. 84

8.1 Association des eQTL avec les loci de prédisposition aux traits complexes . 88
8.2 Association des eQTL avec les loci de prédisposition aux traits complexes, ajustée sur la densité de gènes 91

9.1 Nombre de composantes identifiées selon la méthode choisie. 107
9.2 Qualité de reconstruction selon la méthode de factorisation et le nombre de composantes extraites. 108
9.3 Qualité de reconstruction, puissance et erreur de type I du teste global $\mathcal{H}_0^g$. 112
9.4 Puissance de détection des gènes associés au sein du module et taux de faux positifs 113

10.1 Proportions estimées des différents types cellulaires dans GHS et Cardiogenics. 133
10.2 Contamination et loci associés à des modules de gènes co-régulés. 136
10.3 Effet sur l'association SNP-pattern de l'ajustement par les proportions du mélange. 138
10.4 Liste des sondes les plus fortement associées au locus CD8A avant et après ajustement sur la quantité de lymphocytes T. 139

10.5 Liste des sondes les plus fortement associées au locus ARHGEF3 avant et après ajustement sur la quantité de plaquettes. 141

Introduction

Depuis plusieurs décennies, un intérêt croissant a été attribué en médecine à la compréhension des facteurs génétiques influençant le développement des maladies communes. Ces dernières années, les analyses en génome entier (GWAS) ont permis l'identification de nombreux loci de prédisposition aux maladies complexes. Malgré ces progrès, l'utilisation des loci identifiés par les GWAS pour la découverte de nouveaux mécanismes causaux reste un des défis majeurs de la génomique. L'intégration de données génomiques issues de sources multiples (génome, transcriptome, protéome, ...) constitue donc une voie de recherche majeure pour l'identification de tels mécanismes.

Le travail présenté ici s'inscrit dans ce cadre et tente d'étudier dans quelle mesure l'expression des gènes peut aider à la compréhension des mécanismes impliqués dans le développement des maladies cardiovasculaires et plus généralement des maladies complexes.

Le présent document est le fruit de mon travail de thèse au cours duquel j'ai eu l'occasion d'explorer de multiples aspects de la génétique humaine et de la biologie moléculaire. Il est organisé en trois parties :

Dans une première partie, nous reviendrons sur le contexte dans lequel s'inscrit ce travail de thèse. Pour cela, nous introduirons tout d'abord quelques notions de base sur le transcriptome et sur les études d'association génétique, avant de revenir sur l'étiologie des maladies cardiovasculaires et de présenter deux grandes études de génomique épidémiologue, GHS (Gutenberg Heart Study) et Cardiogenics, visant à relier la variabilité du génome et du transcriptome et à étudier leur rôle dans les maladies cardiovasculaires.

Dans une deuxième partie, nous décrirons les technologies de puces à ADN et à ARN et présenterons les diverses étapes de traitements liés à ces puces ainsi que les méthodes statistiques classiques utilisées pour l'analyse des données de puces.

Enfin dans une troisième partie, nous étudierons le lien entre génome et transcriptome et son utilisation pour l'étude des traits complexes. Il est largement établi que l'expression de nombreux gènes est génétiquement régulée, le plus souvent par des variants situés à proximité des gènes régulés. Dans ce travail, nous verrons comment il est possible d'iden-

tifier les variants génétiques modulant l'expression des gènes dans le monocyte, et nous discuterons de l'utilisation de cette connaissance pour l'interprétation des loci trouvés dans les GWAS et la découverte de nouveaux loci de prédisposition. Outre la régulation directe de l'expression des gènes par le génotype, l'effet des variants génotypiques sur la maladie peut également passer par l'activation de voies de signalisation spécifiques. Nous verrons comment il est possible d'identifier des modules fonctionnels à partir des structures de co-régulation de l'expression des gènes et d'utiliser cette connaissance pour la recherche de nouveaux loci de prédisposition aux maladies complexes. Lors de ce travail, nous avons découvert que la présence de contamination des échantillons par des types cellulaires non désirés pouvait perturber la reconstruction des réseaux et fausser l'interprétation des résultats. Nous proposons ici une approche permettant d'identifier et de quantifier de tels artefacts, et discutons de la possibilité de contrôler ce genre de biais lors de l'analyse du transcriptome.

Première partie

Motivations

CHAPITRE 1

Apports du transcriptome pour l'étude des maladies multifactorielles

L'objet de cette partie est de définir le transcriptome et son rôle dans le développement des maladies multifactorielles.

1 Le transcriptome

1.1 De l'ADN à la protéine

L'acide désoxyribonucléique (ADN), support de l'information génétique, est composé de près de trois milliards de paires de bases azotées [1], disposées en deux brins antagonistes selon des séquences orientées en sens inverse [2] depuis leur extrémité dite 5' vers leur extrémité dite 3'.

```
5'-AGCTTCCAGGTCAATCGCGA-3'
 3'-TCGAAGGACCTCAATCGGCT-5'
```

FIGURE 1.1 – **Structure des deux brins complémentaires et antiparallèles.**

Cette séquence, identique pour toutes les cellules d'un même individu, est répartie entre les 23 paires de chromosomes humains [3] qui constituent le génome. Sur le génome sont répartis près de 25 000 gènes codant chacun pour une ou plusieurs protéines (macromolécule composée d'acide aminés et remplissant les fonction vitales essentielles de la cellule). La lecture des gènes aboutit à la formation de protéines via des ARN (acides ribonucléiques) aussi appelés transcrits. L'ensemble des ARN issus de la transcription des gènes constitue le transcriptome. Le transcriptome est ainsi un reflet de l'ensemble des protéines produites par la cellule (protéome). La synthèse des protéines à partir du

1. Adénine (A) ,Thymine (T), Guanine (G) et Cytosine (C).
2. Cette orientation est due à la forme asymétrique des bases azotées.
3. Vingt-deux paires d'autosomes plus la paire de chromosomes sexuels XX ou XY (ADN mitochondrial non compris).

gène comprend 3 étapes majeures représentées sur la figure 1.2 auxquelles s'ajoute la dégradation des ARN qui termine le cycle de vie des ARN.

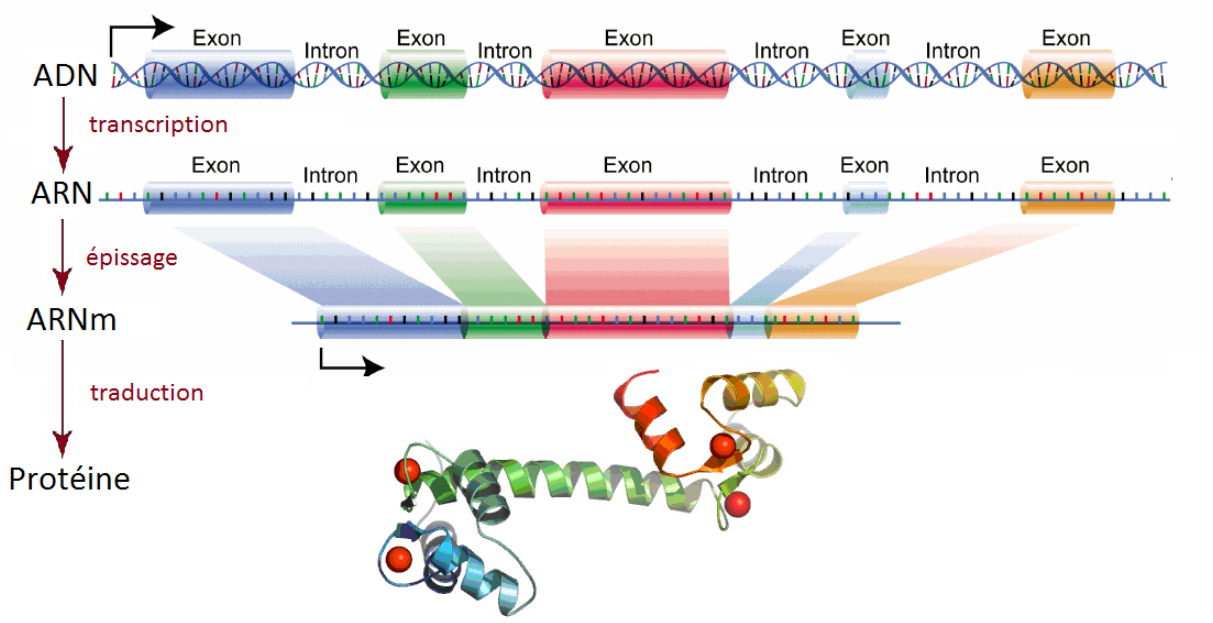


FIGURE 1.2 – **Formation des protéines à partir de l'ADN.**

Transcription : l'ADN double brin est séparé en ADN simple brin autour du gène puis transcrit dans le sens 5'-3' par l'enzyme d'ARN polymérase pour former un pré-ARN ou ARN primaire. Cet ARN est formé des 4 bases ribonucléiques AUGC [4].

Epissage : l'ARN primaire est ensuite coupé pour en retirer les parties non codantes appelées introns. Seules les parties codantes (exons) et les régions de début et de fin de gène (nommées parties 5' non traduite (5'UTR) et 3' non traduite (3'UTR)) sont conservées dans l'ARN messager (ARNm) résultant. A ce stade pour les gènes composés de multiples exons, il arrive que certains exons soient également retirés induisant des *épissages alternatifs* menant à la formation de protéines distinctes (qualifiées d'*isoformes*) à partir d'un même gène.

Traduction : les ARN messagers sont ensuite reconnus par des ribosomes qui vont permettre la synthèse de protéines par lecture successive de triplets de bases ribonucléiques (les codons) et assemblage progressif d'une chaine d'acides aminés. Chaque ribosome fixé à l'ARNm va ainsi synthétiser un exemplaire de la protéine codée par le gène.

Dégradation : après leur synthèse, les ARN sont détruits dans la cellule par un cocktail d'enzymes de dégradation. Cette dégradation a lieu en continu dans la cellule et

4. La Thymine est remplacée par l'Uracile dans l'ARN.

ne prend généralement que quelques heures assurant ainsi une forte réactivité du transcriptome. La plupart des enzymes de dégradation lysent l'ARN en commençant par l'extrémité 5'. La dégradation des transcrits n'est donc pas uniforme [5].

1.2 Variabilité du transcriptome

Si l'ADN est le même pour toutes les cellules d'un même organisme, il n'en va pas de même du transcriptome (cf. figure 1.3).

Au cours du développement embryonnaire, les cellules communiquent par des messages chimiques et se différencient en plusieurs tissus spécialisés. Ceux-ci forment ensuite différents organes assurant chacun une fonction spécifique. Cette différenciation cellulaire est permise par la modification de l'utilisation des gènes par la cellule. Par exemple les cellules du système immunitaire exprimeront des gènes liés à la réponse immunitaire tels que les gènes du Complexe Majeur d'Histocompatibilité (CMH, ou MHC en anglais) tandis que les cellules du pancréas se spécialiseront dans la production d'hormones telles que l'insuline ou les sucs pancréatiques. La variation des niveaux protéiques nécessaires à la spécialisation cellulaire se fait en amont par la modification des quantités d'ARN par les différents gènes. La différenciation cellulaire est donc la première source de variabilité du transcriptome au sein d'un organisme.

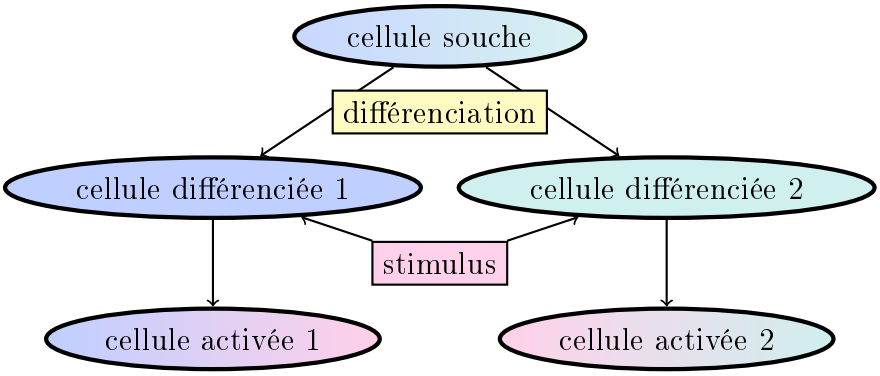


FIGURE 1.3 – **Variabilité du transcriptome.**

Une deuxième source de variabilité provient des réactions de la cellule aux stimuli environnementaux. Ces stimuli environnementaux couvrent un large spectre allant des nutriments reçus par la cellule (glucose, lipides, calcium, métaux,...), aux changements de conditions environnementales (température, pression, ...) en passant par les messages de communication inter-cellulaire qu'ils soient locaux (cytokines, neurotransmetteurs, ...) ou à distance (hormones). Ainsi le transcriptome d'un tissu dépend de plusieurs facteurs :

5. Ce qui a pour conséquence d'entraîner des biais dans les mesures d'expression comme nous le verrons dans la section 4 du chapitre 3.

- les différents types cellulaires qui le composent.
- l'état du tissu et de son environnement immédiat (dysfonctionnements hormonaux, changements de température...).
- dans une moindre mesure, les influences environnementales auxquelles est soumis l'individu (alimentation, exposition aux toxiques) ou de l'état de santé de l'individu. Le plus souvent de telles influences indirectes ne sont décelables qu'à condition d'étudier les tissus appropriés pour le trait considéré.

Ces multiples influences font du transcriptome une mesure extrêmement variable et potentiellement instable et sont susceptibles d'entraîner des biais que nous verrons dans le chapitre 3. Mais elles en font aussi une mesure très complète de l'état d'un tissu et des influences auxquelles il est soumis, justifiant son utilisation dans l'étude des maladies complexes.

1.3 Variabilité génétique et héritabilité du transcriptome

On constate lorsqu'on étudie le même type cellulaire dans plusieurs populations humaines (asiatiques, africaines ou caucasiennes), une grande disparité des niveaux d'expression [1] avec près de 30% des transcrits qui montreraient des différences d'expression entre populations. Cette disparité peut s'expliquer par l'influence du mode de vie mais également par la diversité génétique de ces populations.

1.3.1 Variabilité génétique du génome humain

Bien que l'ADN soit majoritairement conservé au sein d'une espèce donnée, il existe au sein d'une population des variations de la séquence d'ADN nommées polymorphismes. Ces variations sont dûes à des erreurs de copie se produisant lors de la réplication de l'ADN. Ces modifications de la séquence sont ensuite transmises de génération en génération et se fixent dans la population pour former un polymorphisme. On appelle allèle, chacune des variations possibles de l'ADN en un point donné du génome. Pour chaque polymorphisme, le génotype d'un individu est défini comme la combinaison des deux allèles présents sur chacun des brins d'ADN hérités de ses parents (voir figure 1.4).

Parmi les mutations de l'ADN amenant à la formation de polymorphismes on peut distinguer :

- les *substitutions* d'une partie de la séquence par une séquence alternative de la même longueur. L'immense majorité des mutations par substitution se font par le simple remplacement d'une base azotée par une autre. On parle alors de polymorphisme à un seul nucléotide (SNP en anglais pour "Single Nucleotide Polymorphism"). Ces polymorphismes sont les plus courants et les plus étudiés.

– les *insertions/délétions* de séquences pouvant aboutir à des répétitions de certains éléments de séquence et à la formation de polymorphismes multi-alléliques (plus de 2 allèles possibles). On parle alors de variations du nombre de copies (CNV en anglais pour “Copy Number Variation”).

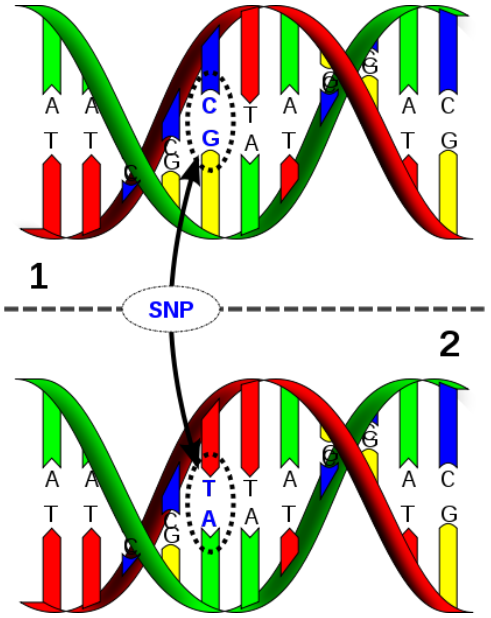


FIGURE 1.4 – **Polymorphisme génétique.**

Près de trois millions de SNP ont aujourd'hui été recensés dans le projet HapMap visant à cartographier la variabilité du génome humain, auxquels s'ajoutent des dizaines de milliers de polymorphismes plus complexes comme les CNV[6]. Ces différences se répercutent ensuite sur les protéines et le fonctionnement de la cellule, donc en particulier sur le transcriptome.

1.3.2 Héritabilité du transcriptome

On définit l'héritabilité d'un trait phénotypique comme la part de la variabilité de ce trait dans la population qui peut être expliquée par la génétique. Cette héritabilité, est le plus souvent estimée à partir d'études familiales. La variabilité phénotypique V_P est décomposée entre la variabilité génotypique V_G et la variabilité environnementale V_E supposées indépendantes

$$V_P = V_G + V_E$$

6. Le nombre plus faible de CNV provient essentiellement de la plus grande difficulté à détecter ces polymorphismes, mais on estime qu'au total les variations du nombre de copies affecteraient près 12% du génome [2].

L'héritabilité est alors donnée par

$$h^2 = \frac{V_G}{V_P}$$

Une étude basée sur des familles mexicaines regroupant 1100 individus [3] a montré une héritabilité très forte des niveaux d'expression des transcrits dans les lymphocytes. La figure 1.5 montre que plus de la moitié des transcrits présentent une héritabilité supérieure à 20%.

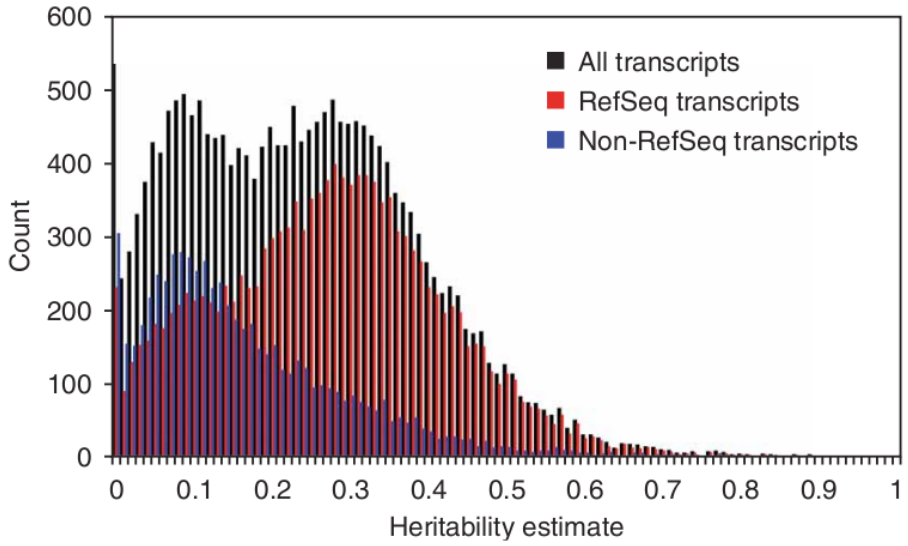


FIGURE 1.5 – **Distribution de l'héritabilité des transcrits selon leur statut RefSeq (présence ou absence dans la base de donnée RefSeq) dans une étude familiale :** La différence constatée s'explique par le fait que la base RefSeq référence uniquement des transcrits pour lesquels une protéine connue existe. A l'inverse, les transcrits absents de cette base ne sont que rarement associés à une protéine et peuvent correspondre à des pseudo-gènes résultant de l'histoire évolutive et n'étant pas forcément exprimés par l'organisme.

Ces résultats doivent cependant être nuancés par le fait que l'héritabilité est généralement sur-estiméee dans les études familiales en raison de l'environnement commun partagé par les individus d'une même famille (habitudes alimentaires, culturelles,...). Des études d'héritabilité menées chez le rat [4] où la variabilité environnementale peut être efficacement contrôlée, confirment une forte héritabilité (bien que diminuée) du transcriptome (figure 1.6 - près de 20% des transcrits ont une héritabilité supérieure à 0.2)).

Ces résultats montrent le lien étroit du transcriptome avec le génome et suggèrent que les différences phénotypiques dues à la variabilité génétique pourraient en partie être expliquées par d'importantes variations dans l'expression des gènes. Ces différences, observables au niveau du transcriptome, se répercutent ensuite sur les concentrations

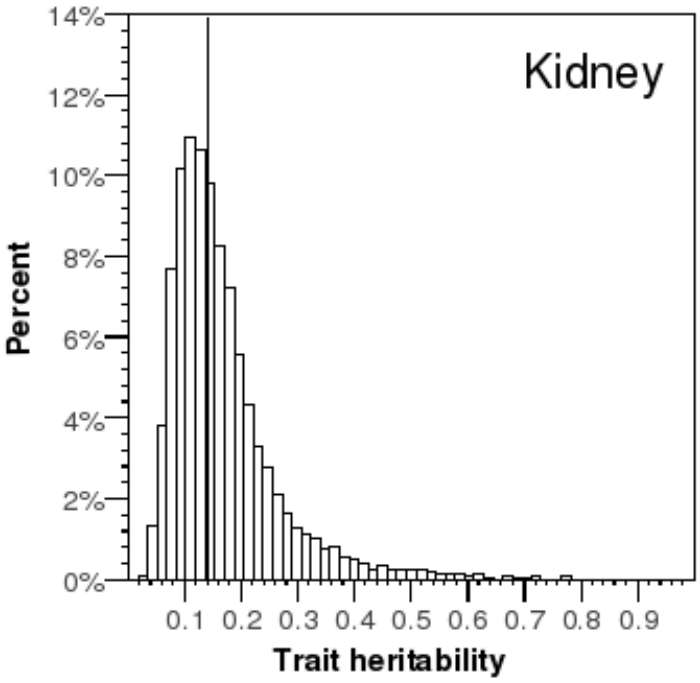


FIGURE 1.6 – **Distribution de l'héritabilité des niveaux de transcrits dans du tissu rénal de rat.**

protéiques et pourraient ainsi expliquer une part importante de la variabilité phénotypique observée entre individus.

2 Études d'association en génétique

2.1 Les maladies multifactorielles

La plupart des maladies humaines, sinon toutes, présentent une composante génétique plus ou moins forte. Classiquement, on considère deux grandes catégories en fonction de l'importance de la composante génétique (bien qu'en réalité il existe un large spectre continu entre ces deux extrêmes) :

- Les maladies *monogéniques* ou *mendéliennes* qui résultent de la présence chez les individus d'une version déficiente du gène responsable de la maladie (généralement suite à une mutation transmise de parent à enfant). Ces maladies sont rares dans la population et le plus souvent très invalidantes voire fatales (ex : mucoviscidose, myopathies,...).
- Les maladies *multifactorielles* ou *complexes* qui présentent à la fois une composante génétique et environnementale et résultent généralement d'interactions complexes entre ces deux facteurs. Ces maladies sont généralement des maladies fréquentes

(obésité, maladies cardiovasculaires ou inflammatoire, diabètes,...). Leur développement est facilité par la présence simultanée de nombreux allèles de prédisposition (fréquents dans la population) ayant un impact individuel modéré.

La génomique épidémiologique a pour objectif d'identifier les gènes de prédisposition aux maladies multifactorielles et de mieux comprendre leur rôle en relation avec l'environnement. La connaissance des mécanismes génétiques à l'origine de la maladie permet une meilleure compréhension de la physiopathologie et contribue à la découverte de nouvelles voies thérapeutiques.

2.2 Stratégies de recherche : de l'approche "gène candidat" aux études "génome entier"

Pendant longtemps, les contraintes physiques et techniques des procédures de génotypage ainsi que la connaissance parcellaire du génome [7] ont amené les chercheurs à adopter une approche dite "gène candidat" en concentrant leurs efforts sur des gènes connus et potentiellement impliqués dans l'étiologie de la maladie étudiée. Par exemple l' étude des gènes du métabolisme des lipides pour l'étude de la dyslipidémie. Malgré quelques succès, comme la découverte de l'importance du gène de l'apolipoprotéine E dans le développement de l'athérosclérose [5], l'approche gène candidat a montré une efficacité limitée pour la recherche des variants génétiques liés aux maladies complexes.

Depuis quelques années, l'émergence des nouvelles technologies de génotypage à haut débit (décrites dans le chapitre 3) a révolutionné les stratégies de recherche de l'épidémiologie génétique. Ainsi la recherche de polymorphismes génétiques prédisposant au développement d'une maladie se fait maintenant en balayant l'ensemble du génome à la recherche de signaux d'association indiquant la présence d'un locus de prédisposition [8]. L'objectif n'est plus de confirmer l'implication d'un gène ou d'une protéine connue dans une maladie mais de trouver de nouveaux gènes de prédisposition à la maladie sans aucune hypothèse a priori. Ces études sont appelées études d'association génome entier ou GWAS [9].

2.3 Utilisation du transcriptome en épidémiologie génétique

Dans l'étude de la composante génétique des maladies, le transcriptome peut jouer un rôle de deux façons :

7. Le premier séquençage complet du génome humain commencé en 1990, a été terminé en 2001. Un séquençage intégral prend aujourd'hui moins de 48h.

8. C'est-à-dire un point du génome présentant un lien avec la maladies.

9. De l'anglais Genome Wide Association Studies.

1. Tout d'abord, le transcriptome constitue un phénotype intermédiaire intimement lié à l'ADN et au fonctionnement de la cellule. Il offre donc souvent les premiers éléments de compréhension des mécanismes reliant les nouveaux loci mis en évidence par les GWAS aux maladies multifactorielles. L'étude au niveau du génome entier de la régulation génétique de l'expression des gènes [10] dont il sera question dans les chapitres 7 et 8 est donc un axe majeur de développement de la génomique moderne.
2. Ensuite, le transcriptome peut être utilisé comme un biomarqueur fournissant à la fois des informations sur l'état de santé du patient [11] et sur certaines expositions environnementales [12]. Ainsi à condition de savoir caractériser l'impact des divers facteurs environnementaux, il est possible d'utiliser le transcriptome soit comme variable d'ajustement, soit comme un trait phénotypique à part entière, comme nous le verrons dans les chapitres 9 et 10.

10. Connue en anglais sous le terme "genetics of expression".

11. Par exemple pour le diagnostic des cancers, où le transcriptome est déjà utilisé cliniquement pour identifier divers types de tumeurs.

12. Comme par exemple le tabagisme [6].

CHAPITRE 2

Le transcriptome en épidémiologie cardiovasculaire

Dans cette thèse, nous évoquerons l'utilisation du transcriptome en génétique humaine en nous focalisant sur l'étude Gutenberg Heart Study (GHS) et dans une moindre mesure sur le projet Cardiogenics. Ces deux études ont pour objectif d'identifier de nouveaux loci de prédisposition à l'athérosclérose et de déterminer le lien entre le transcriptome des monocytes et l'athérosclérose et ses facteurs de risque. Dans ce chapitre, nous reviendrons sur la nature des pathologies cardiovasculaires avant d'aborder le rôle des monocytes dans ces pathologies puis de décrire le contenu et les objectifs des études GHS et Cardiogenics. Nous évoquerons les jeux de données publiques d'expressions utilisés dans cette thèse.

1 Les pathologies cardiovasculaires

Les pathologies cardiovasculaires constituent l'une des principales causes de mortalité en France et dans les pays industrialisés. La compréhension des mécanismes biologiques à l'oeuvre dans ces maladies représente un enjeu majeur de santé publique tant en matière de prévention que de traitement. De nombreuses études ont mis en évidence le caractère multifactoriel des maladies cardiovasculaires par l'identification de nombreux facteurs de risque environnementaux [7–9] autant que génétiques [10, 11].

L'athérosclérose, dont nous décrivons les mécanismes dans la section 1.2, est à l'origine de la majeure partie des pathologies cardiovasculaires acquises [1]. Elle se manifeste par la formation de plaques d'athérome pouvant obstruer les artères et conduisant à diverses pathologies selon la nature des artères touchées et la gravité des atteintes. Ainsi une obstruction des artères coronaires entraînera des complications cardiaques telles que l'angine de poitrine ou l'infarctus du myocarde tandis qu'une obstruction des artères cérébrales provoquera des accidents vasculaires cérébraux. Cette affection peut également atteindre les plus gros tronc artériaux tels que la crosse aortique ou l'artère fémorale entraînant ischémies des membres antérieurs (artérite) et anévrismes.

1. Par opposition aux pathologies congénitales innées.

1.1 Incidence des pathologies cardiovasculaires

Après avoir longtemps constitué la première cause de mortalité en France, les maladies cardiovasculaires sont passés aujourd'hui au deuxième rang après les tumeurs, grâce à l'importance des efforts de prévention réalisés et à l'amélioration des traitements. L'Institut National de Veille Sanitaire estime en effet à environ 150 000 le nombre de décès par an en France dus aux maladies cardiovasculaires, soit près de 30% du nombre total de décès. Parmi ces décès, on compte en 2006 :

- 39 000 décès liés à des cardiopathies ischémiques, soit 27% du nombre de décès total. Ces cardiopathies regroupent l'ensemble des dysfonctionnements résultant d'un arrêt ou d'une réduction de l'irrigation sanguine du muscle cardiaque.
- 33 000 décès liés à des accidents vasculaires cérébraux (AVC) qui peuvent être soit *ischémiques* (arrêt de l'oxygénation d'une partie du cerveau suite à l'obstruction d'un artère) soit *hémorragiques* (rupture d'un vaisseau sanguin.).
- 21 000 décès liés à une insuffisance cardiaque.

Ces 3 types de pathologies qui représentent plus de 60% des pathologies cardiovasculaires sont principlament attribuables à l'athérosclérose et illustrent la réelle nécessité de développer des stratégies de prévention et de traitement de cette pathologie. Des efforts ont déjà été fait dans ce sens entraînant une réduction de 30% des décès cardio-vasculaires en 15 ans.

Les pathologies cardiovasculaires représentent également un enjeu au niveau mondial puisqu'elles totalisent à elles seules 16 millions de décès en 2002 soit près d'un quart des décès dans les pays membres de l'OMS. Une forte disparité existe toutefois entre pays du fait des différences sociales, alimentaires et génétiques. On observe ainsi des incidences atteignant les 915 cas pour 100 000 habitants en Finlande contre 79 cas pour 100 000 habitants en Chine d'après le registres MONICA [12].

1.2 Etiologie et facteurs de risque de l'athérosclérose

Le vieillissement des artères ou athérosclérose s'accompagne le plus souvent d'un épaississement de la paroi artérielle, d'une accumulation de lipides et de tissus fibreux dans les couches superficielles du tissu artériel que sont l'intima et la média (Figure 2.1). Ce phénomène, observé pour la première fois en 1740 par le médecin allemand Krell, a fait l'objet de multiples descriptions ("plaques osseuses", "durcissement de la paroi artérielle", "endartérite déformante",... [14]) avant d'arriver à son acceptation actuelle présentée par l'OMS comme "association variable de remaniement de la couche interne des artères [consistant] en une accumulation focale de graisses (les lipides), de glucides complexes (les sucres), de sang et de produits sanguins, de tissu fibreux et de dépôts calcaires. Le tout est accompagné de modifications de la structure interne de l'artère."

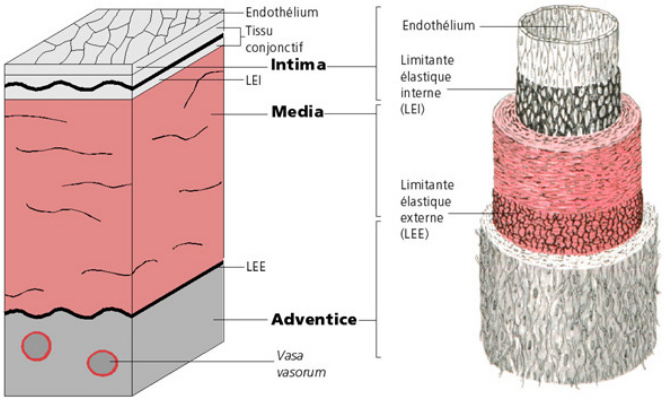


FIGURE 2.1 – **Les différentes couches de la paroi artérielle.**
Source : Stevens & Lowe, 1997 [13]

1.2.1 Athérogénèse

Le développement et l'évolution de la plaque d'athérome sont le résultat d'un processus complexe qu'il est possible de décomposer en 11 étapes majeures proposées par Cohen en 1997 (Figure 2.2, [15]).

Dans un premier temps, des lipoprotéines basse densité (LDL en anglais) circulant dans le sang vont pénétrer dans l'intima (1) où elles seront oxydées (2). Cette oxydation a pour effet d'activer les cellules endothéliales pour permettre l'adhésion des monocytes à l'endothélium (3) et permettre le recrutement des monocytes dans l'intima et leur transformation en macrophages (4). Les macrophages vont ensuite se transformer en cellules spumeuses chargées en lipides qui constitueront le cœur de la plaque d'athérome (5). A ce stade, l'athérosclérose se manifeste sous la forme de légers renflements le long de l'artère appelés "stries lipidiques". Des cellules musculaires lisses (CML) de la média vont ensuite migrer vers l'intima (6) et sécréter collagènes, fibres élastiques et protéoglycanes (7) qui auront pour effet de rendre la plaque d'athérome fibreuse et de faciliter l'accumulation de tissus conjonctif, de LDL, de CML et de cellules spumeuses (8). Les lipides accumulés forment alors un noyau lipidique extra-cellulaire (9), finalisant la constitution de la plaque d'athérome. Dès lors, deux scénarios sont possibles

– La lésion athérosclérotique peut se calcifier et se solidifier devenant alors plus résistante et paradoxalement moins dangereuse, car son impact sera limité à la réduction

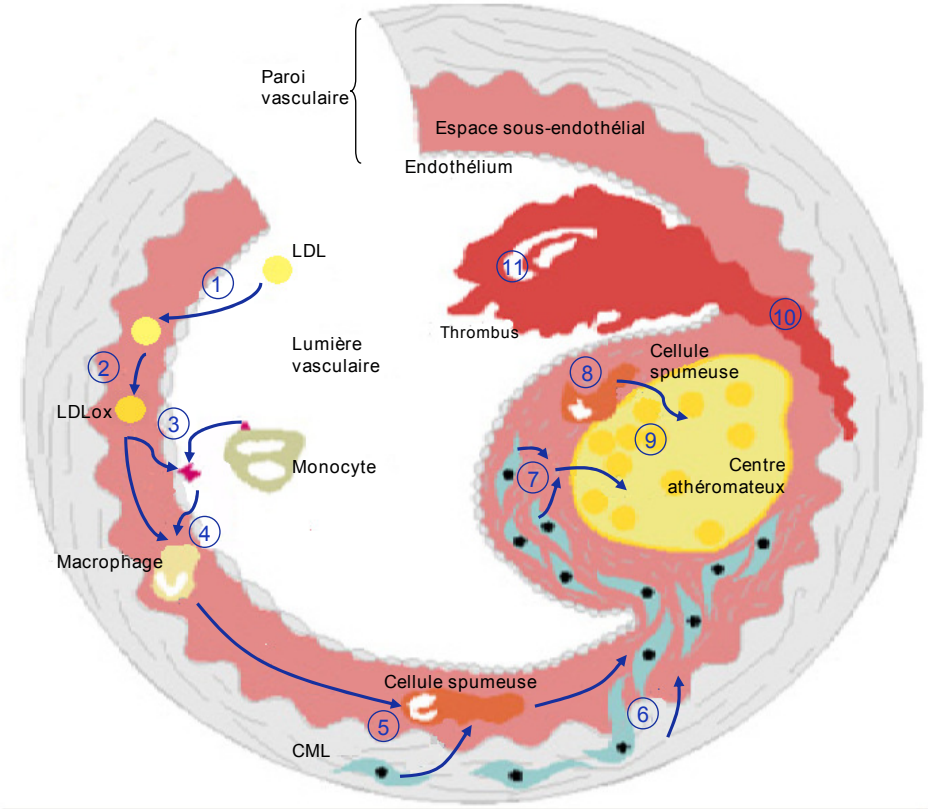


FIGURE 2.2 – **Le processus de l'athérogénèse.** *Source* : Cohen, 1997 [15].

de la lumière vasculaire [2] et donc du flux sanguin.
- Alternativement, une ulcération de la lésion peut se produire (10) libérant au contact du sang le noyau athéromateux contenant des substances pro-coagulantes conduisant à une thrombose (11). Parallèlement la libération de cristaux de cholestérol peut également provoquer des embolies.

Plusieurs théories existent quant aux causes favorisant le développement de la plaque d'athérome. En 1860, Rudolf Virchow pose les bases des théories modernes en suggérant que l'accumulation de lipides dans la paroi artérielle peut être déclenchée par l'inflammation des artères suite à une blessure initiale de l'artère ou à la présence d'agents pathogènes. Après avoir été écartée pendant plus d'un siècle, cette théorie a été confortée par les études anatomopathologiques faites depuis sur des échantillons de plaque d'athérome [16]. En effet, il est désormais admis que l'athérosclérose comporte une dimension inflammatoire dont les LDL oxydés sont les déclencheurs les plus probables. Cette inflammation, présente à tous les stades de l'athérosclérose, est le résultat d'un équilibre entre cytokines pro et anti-inflammatoires. Cet équilibre inflammatoire maintenu dans la plaque d'athérome est déterminant dans l'évolution de la maladie et ses complications [17, 18]

1.2.2 Facteurs de risque environnementaux

En dehors de l'age et du sexe, l'athérosclérose a de nombreux facteurs de risque dont les principaux sont le tabagisme, la consommation d'aliments riches en graisses, le diabète, l'obésité et l'hypertension artérielle. A ces facteurs s'ajoutent d'autres facteurs moins importants tels que la sédentarité, le stress, les niveaux plasmatiques d'homocysteines, de LDL,...

Pour certains de ces facteurs, les mécanismes physiopathologiques ont été identifiés
- Le tabagisme augmente le risque d'infarctus en diminuant le taux de lipoprotéines de hautes densité (HDL) au profit des LDL athérogènes et en favorisant l'apparition de radicaux libres oxydant les LDL.
- Les régimes riches en graisse, la sédentarité et l'obésité augmentent les taux lipidiques sanguins et perturbent l'équilibre HDL/LDL-cholestérol ce qui favorise l'apparition de la plaque d'athérome.
- Une pression artérielle élevée favorise la rigidité artérielle, l'inflammation des parois artérielles et le recrutement des monocytes dans la plaque d'athérome.

Malgré une connaissance avancée de la physiopathologie de l'athéorsclérose, une proportion importante de cas demeure d'étiologie inconnue.

2. Espace disponible pour le flux sanguin.

1.2.3 Facteurs de risque génétiques

Aux facteurs de risque environnementaux s'ajoutent des facteurs de risque génétiques dont un certain nombre ont pu être identifiés par des approches gènes candidats. Le tableau 2.1 emprunté à Annert *et al.* [19] résume les principaux facteurs de risque génétiques initialement identifiés par des approches gènes candidats et récemment confirmés par des GWAS

Gène	Polymorphisme	Risque relatif rapporté
Coronary heart disease		
MTHFR	C677T	1.14-1.21
Cholesterol ester transfer protein (CETP)	TaqIB	0.78
Paraoxonase (PON1)	Q192R	1.14-1.21
Endothelial nitric oxide synthase (eNOS)	T-786C	1.31
Prothrombin	G20210A	1.21
APOB	Ins/Del (DD)	1.30
Glycoprotein IIIa	Pl(A2)	1.10
APOE	e4/e4	1.42
ACE insertion/deletion	DD	1.16-1.21
APOB	SpIns/Del (DD), EcorI (AA)	1.19-1.73
PAI1	4G/5G	1.20
Fibrinogen β-chain	G-455A	0.68
Endothelial nitric oxide	Glu298Asp, Intron-4	1.31-1.34
Arterial ischemic events or overall cardiovascular disease		
MTHFR	C677T	1.20
Prothrombin	G20210A	1.32
ACE insertion/deletion	DD	1.22
Glycoprotein IIIa	Pl(A2)	1.13
Factor V Leiden	Arg506Gln	4.43
APOE	e4/e4	Positive
Cerebrovascular disease		
APOE	e4/e4	1.11
MTHFR	C677T	1.24-1.37
Factor V Leiden	Arg506Gln	1.33
Prothrombin	G20210A	1.44
ACE insertion/deletion	DD	1.21

Tableau 2.1 – **Principaux gènes candidats associés à une prédisposition aux maladies cardiovasculaires et répliqués dans des GWAS :** Seuls les polymorphismes présentant un risque relatif significativement différent de 1.00 dans des méta-analyses d'au moins 1000 sujets, publiées entre 2000 et 2007, sont rapportés. Lorsque plusieurs méta-analyses sont disponibles pour un même gène, le range de risque relatif est rapporté. *Source* : Annert *et al.* [19].

Parmi les gènes rapportés, on trouve :

– des gènes liés au métabolisme des lipides tels que les gènes codant pour les apolipoprotéines E et B (APOE, APOB) et la CETP (CholesterylEster Transfer Protein) ;

- des gènes intéragissant avec les LDL comme le récepteur aux LDL (LDLR) ou la paraoxonase impliquée dans l'oxydation des LDL ;
- des gènes affectant la pression artérielle comme le gène codant pour l'enzyme de conversion de l'angiotensine (ACE), ou la NO synthase endothéliale ;
- des gènes liés à la coagulation (prothrombine, glycoprotéine IIIa , Facteur V , fibrinogène) ;
- des gènes affectant les niveaux plasmatiques d'homocystéine (MTHFR).

A l'exception du facteur V Leiden, les risques relatifs associés aux facteurs génétiques sont modérés (de l'ordre 1.2 en moyenne). La diversité des mécanismes mis en évidence par ces loci illustre le caractère hautement multifactoriel des maladies cardiovasculaires.

2 GHS et Cardiogenics : Deux projets d'étude du transcriptome à grande échelle

La majeure partie des travaux réalisés dans le cadre de cette thèse l'ont été à partir des données issues de la "Gutenberg Heart Study" et ont été répliqués grâce au données du projet européen "Cardiogenics" que nous présentons dans cette section.

2.1 Objectif et caractéristiques des études

2.1.1 Gutenberg Heart Study

L'étude GHS initiée en 2006 par Stefan Blankenberg est un projet collaboratif entre le départment de cardiologie de l'université Johannes-Gutenberg située à Mainz (Allemagne) et l'INSERM. L'objectif principal de cette étude est d'identifier de nouveaux facteurs de risques (génétiques, plasmatiques, environnementaux, psychosociaux) associés à l'athérosclérose et ses phénotypes intermédiaires en population générale. Dans ce but, une cohorte de plus de 3300 sujets a été recrutée dans la population générale des habitants de la région de Mainz. Seuls les sujets âgés de 35 à 75 ans ont été inclus et une stratification a été effectuée lors de l'échantillonnage afin d'avoir un nombre de sujets comparables pour chaque sexe et chaque tranche d'age. Ces sujets ont subi un examen médical très poussé comprenant un large panel de phénotypes cardiovasculaires (échographie cardiaque et vasculaire, fonction endothéliale, ...). Cette étude dispose également d'importantes ressources biologiques (ADN, ARN, Plasma, Serum, ...). Tous les individus de l'étude ont été génotypés à l'aide de la puce génome entier Affymetrix 6.0. De plus, pour la moitié des sujets, l'ARN a été extrait le jour du prélèvement à partir de monocytes fraichement isolés par sélection négative et le transcriptome a été mesuré à l'aide de la puce Illumina

HT-12 v2. Les investigations cliniques et biologiques ont été conduites à Mainz, les analyses statistiques sont réparties entre Lübeck et Paris. Dans le cadre de cette thèse, j'ai été en charge du contrôle qualité et de l'analyse des données d'expression de l'étude GHS. Seuls les individus d'origine européenne ont été inclus dans les analyses. Le détail des étapes de normalisation et de contrôle qualité réalisées sur ces données est décrit dans les chapitres 4 et 5.

2.1.2 Cardiogenics

L'étude Cardiogenics est un projet européen financé par le FP6, regroupant 15 partenaires européens dont le but est la caratérisation de nouveaux gènes et mécanismes de prédisposition à la maladie coronaire. Comme GHS, l'étude Cardiogenics intègre à la fois des données d'expression provenant de monocytes (et de macrophages), et des données génotypiques pan-génomiques. L'étude Cardiogenics comporte deux échantillons de cas et de témoins. Le premier groupe inclue 363 cas de syndrôme coronaire aigu ainsi que défini par la Société Européenne de Cardiologie. Ces cas, âgés de 26 à 87 ans ont été recrutés parmi les cas incidents des hôpitaux de Leicester, Lübeck, Paris et Regensburg. Le deuxième groupe comprend 395 témoins recrutés parmi des donneurs de sang volontaires à Cambridge et appariés par age aux cas. Du fait de la plus forte prévalence des maladies coronaires chez les hommes, d'importantes différences existent entre les proportions d'hommes et de femmes observées chez les cas et les témoins (12% de femmes chez les cas contre 59% chez les témoins). Les individus ont été génotypés à l'aide de puces Illumina Sentrix Human Custom 1.2M et Human 610 quad. L'isolation des monocytes par sélection positive a ensuite été faite séparément dans chaque centre et l'extraction d'ARN a été effectué à partir des monocytes immédiatement après isolation. Les ARN ont ensuite été analysés dans l'U937 à l'aide de puces d'expression Illumina Ref8 v3. Tous les individus inclus dans les analyses sont d'origine européenne. Les étapes standard de normalisation et de contrôle qualité ont été réalisées[3] sur ces échantillons. Dans le carde de ma thèse, j'ai utilisé ces données principalement pour la réplication des résultats obtenus dans GHS.

2.2 Le choix des monocytes

Comme nous venons de le voir, les études GHS et Cardiogenics comportent toutes les deux un volet transcriptomique s'appuyant sur l'étude des ARN monocytaires. Ce choix est lié au rôle prépondérant joué par le monocyte dans l'athérosclérose et au caractère ubiquitaire des fonctions remplies par le monocyte qui en fait un modèle cellulaire pertinent pour l'étude d'un grand nombre de pathologies présentant une composante immuno-inflammatoire. En effet, les monocytes sont des cellules centrales du système immunitaire,

3. Cette partie a été réalisée par Seraya Maouche dans l'U 937.

capables de :

- reconnaitre les antigènes à la surface des cellules avoisinantes (ou des anticorps liés à un agent pathogène), assurant ainsi la reconnaissance du non soi et la spécificité de la réaction immunitaire ;
- phagocyter afin de détruire les corps étrangers en les absorbant au travers de la membrane cellulaire ;
- présenter à sa surface des antigènes du non soi issus des cellules lysées afin de diriger plus efficacement la réponse immunitaire ;
- induire la mort cellulaire programmée (ou apoptose) de cellules avoisinantes par présentation d'anticorps cytotoxiques.

D'autre part, les monocytes jouent un rôle prépondérant dans l'inflammation par la synthèse de nombreuses cytokines pro ou anti-inflammatoires. Ils jouent également un rôle important dans l'athérosclérose via leur capacité à se différencier en cellules dentritiques (spécialisées dans l'adaptation de la réponse immunitaire) et en macrophages (spécialisés dans la réponse immunitaire non spécifique) qui sont des acteurs majeurs de la formation de la plaque d'athérome.

Ainsi, les monocytes constituent un modèle idéal pour l'étude de l'athérosclérose et des maladies inflammatoires. A noter qu'ils sont en outre facilement accessibles par un simple prélèvement sanguin. A l'issue du prélèvement, les monocytes naturellement présents dans le sang peuvent être isolés des autres types cellulaires sanguins par filtrage. Deux types de filtrage existent :

- Le *filtrage positif* consiste à isoler les monocytes en les fixant à des anticorps reconnaissant un marqueur cellulaire spécifique des monocytes (généralement la protéine CD14). Ce type de filtrage présente une très forte spécificité mais a l'inconvénient de modifier légèrement le transcriptome des monocytes par l'activation de la voie de signalisation associée à ce marqueur.
- Le *filtrage négatif* consiste à isoler les monocytes par retrait successif des divers types cellulaires présents dans l'échantillon sanguin (plaquettes, globules rouges, lymphocytes T et B, granulocytes,...). Si elle permet d'isoler des monocytes au repos (sans activation de la voie CD14), cette méthode ne permet pas d'assurer une pureté totale des échantillons de monocytes du fait de la contamination par d'autres types cellulaires résiduels. Cette contamination peut entrainer des biais dans l'étude du transcriptome dont il faut tenir compte (cf. chapitre 10).

Dans un cas comme dans l'autre, les monocytes étudiés doivent être vus comme un modèle fiable permettant de s'approcher au mieux du fonctionnement des monocytes recrutés dans la plaque d'athérome. Il serait en revanche illusoire de prétendre à une mesure fidèle en tout point de la réalité des modifications transcriptomiques présentes dans la plaque.

3 De l'utilité des atlas d'expression

En sus des études d'ARN monocytaires précitées, les travaux présentés dans cette thèse recourent à des atlas d'expression. Ces atlas d'expression, disponibles via des sites tels que ArrayExpress [20] ou Gene Expression Omnibus [21] visent à décrire la variabilité du transcriptome d'un tissu à l'autre. Ils constituent donc une forme de mesure étalon du transcriptome d'un type cellulaire à l'autre. Bien que les différences de protocole d'une expérience à l'autre (choix de la puce, des réactifs,...) soient source de variabilité, ces atlas peuvent être utilisés comme un outil efficace (et disponible facilement) pour l'interprétation des donnés de transcriptomique comme nous le verrons dans le chapitre 10. Dans cette thèse, nous utilisons les données d'HaemAtlas rendues publiques par Watkins *et al.*.

HaemAtlas est un atlas d'expression visant à dresser une carte des profils d'expression des principaux types cellulaires trouvés dans le sang. Il contient 50 échantillons de 8 types cellulaires différents : granulocytes (CD66b+), monocytes (CD14+), lymphocytes B (CD19+), lymphocytes T et cellules tueuses (CD4+, CD8+, CD56+), érythroblastes (proxy pour les globules rouges) et mégakaryocytes (proxy pour les plaquettes) traités avec la puce Illumina WG6 v3.Ces données sont disponibles sur arrayExpress sous l'identifiant `E-TABM-633`. Nous avons utilisés ces profils pour estimer les niveaux de contamination de l'ARN monocytaire par d'autres types cellulaires et corriger pour le biais en résultant.

Deuxième partie

Mesure du génome et du transcriptome par biopuces

CHAPITRE 3

Généralités sur les biopuces

Le terme “biopuces” ou “microarrays” regroupe l'ensemble des technologies de puces qui permettent d'analyser des séquences d'ADN ou d'ARN avec un très haut débit en exploitant les propriétés d'hybridation spontanée de l'ADN dénaturé (faculté de l'ADN simple brin à retrouver sa forme naturelle de double hélice en s'appariant à un brin d'ADN complémentaire). On distingue plusieurs catégories de biopuces selon le type d'information recherchée. Une liste (non exhaustive) des types de biopuces est indiquée dans le tableau 3.1. Nous décrirons dans ce travail le principe des puces à expression et de génotypage. Nous aborderons également les principales analyses réalisées à l'aide de ces biopuces et leurs spécificités.

Type de puce	Cible	Objectif
Expression	transcrits	Mesure de l'expression des gènes
Génotypage	SNP	Génotypage des SNPs
Comparative Genomic Hybridation (CGH)	ADN	Repérage des duplications anormales de l'ADN
Méthylation	CpG islands	Mesure de la méthylation de l'ADN
Chromatin Immuno Précipitation (ChIP)	sites de fixation	Repérage des sites de fixation d'une protéine

Tableau 3.1 – **Les différents types de biopuces et leur usage.**

1 Principe de fonctionnement

Pour comprendre le principe de fonctionnement des biopuces, il est nécessaire de revenir sur les propriétés physico-chimiques des molécules d'ADN et d'ARN. Comme nous l'avons évoqué précédemment, (cf. section 1.1 du chapitre 1) la molécule d'ADN présente une structure spatiale complexe en forme de double hélice. Les deux brins qui la forment présentent des séquences de bases azotées complémentaires, les bases A-T et G-C se faisant face, et sont maintenus ensemble par des liaisons hydrogènes entre bases complémentaires. Ces dernières sont susceptibles d'être rompues par des enzymes (ADN hélicase) ou sous l'effet de la chaleur. Lorsque deux brins d'ADN complémentaires dissociés sont mis en contact, un rappariement se produit : c'est la réaction d'hybridation. La présence de légères différences (SNP par exemple) dans la séquence des deux brins n'empêche pas le rappariement de se faire mais diminue grandement son efficacité et fragilise les liaisons. Plus les séquences sont différentes, plus la force d'hybridation est faible.

Le fonctionnement des biopuces exploite cette propriété d'hybridation des séquences nucléiques en généralisant le principe des méthodes de Southern et Northern blot, appliquées couramment en biologie moléculaire pour la détection et la quantification de séquences nucléiques (ADN ou ARN) dans des échantillons par hybridation d'une séquence complémentaire portant un marquage radioactif. Les biopuces permettent ainsi de détecter et quantifier simultanément plusieurs dizaines (voire centaines) de milliers de séquences. Une puce est un support rigide (lame de verre) sur lequel se trouvent des centaines de milliers d'emplacements d'hybridation (spot) mesurant entre 4 et 20 microns chacun. Sur chaque emplacement sont fixées [1] des séquences d'oligonucléotides spécifiques de cibles choisies par le constructeur. Ces courtes séquences forment les sondes ("probes" en anglais). En hybridant des échantillons biologiques marqués par une molécule fluorescente (fluorochrome Cy3 ou Cy5), on peut alors détecter et quantifier les séquences ciblées par la biopuce : en excitant les fluorochromes à l'aide d'un laser, ceux-ci émettent un rayonnement qui peut être mesuré par un scanner. Les intensités lumineuses ainsi évaluées témoignent du nombre de séquences correspondantes qui se sont hybridées sur la puce, et donnent une approximation de la quantité réelle de telles séquences dans l'échantillon initial.

Depuis leur apparition vers la fin des années 90, les puces ADN n'ont cessé d'être améliorées, pour atteindre aujourd'hui une densité permettant de mesurer plusieurs millions de séquences par puce. Une telle miniaturisation permet d'étudier la quasi-totalité du génome d'un organisme à partir d'une simple lame de verre. Selon les choix de construction des sondes et le type de matériel hybridé sur la puce, de nombreuses applications sont possibles :

1. Par impression jet d'encre ou par synthèse *in situ*.

- Dans le cas des *puces d'expression*, les sondes sont construites de façon à s'hybrider aux ARN synthétisés par un génome. Lors d'une expérience de puce à ARN, on synthétise des ADN complémentaires (ADNc) à partir d'un échantillon d'ARN. Leur hybridation sur la puce permet de connaître, pour les transcrits ciblé par la puce, la quantité des différents ARN dans l'échantillon initial.
- Les *puces de génotypage* intègrent des sondes ciblant la séquence entourant les SNP. Pour chaque SNP, deux sondes correspondant aux deux allèles du SNP sont intégrées. L'ADN de l'échantillon est alors hybridé sur la puce de manière à fournir une mesure pour chaque allèle, marqué par un fluorochrome différent. Le niveau relatif de ces deux mesures permet alors de déterminer le génotype (genotype calling).
- Les *puces de méthylation* ou d'*immuno-précipitation de la chromatine* (Chip) intègrent des sondes couvrant le génome avec régularité. Des anticorps sont ensuite utilisés pour marquer les zones de méthylation ou les zones de fixation d'une protéine et seuls les fragments marqués par des anticorps sont hybridés sur la puce. La lecture des intensités renvoyées par la puce indique l'emplacement sur le génome des zones de méthylation ou de fixation de la protéine (facteurs de transcription).
- les *puces CGH* (comparaison de l'hybridation génomique) couvrent également l'ensemble du génome. L'hybridation de l'ADN sur la puce permet alors de repérer des duplications ou des délétions anormales de l'ADN.

Chaque type de puce nécessite des traitements informatiques et statistiques spécifiques. Nous nous contenterons ici de décrire le fonctionnement des biopuces d'expression et de génotypage telles qu'elles sont construites par les deux constructeurs Illumina et Affymetrix. Nous aborderons ensuite dans les chapitres 4 et 5 les prétraitements relatifs à ces puces, avant de discuter dans le chapitre 6 des problèmes statistiques posés par l'analyse des biopuces.

2 Puces d'expression

2.1 Principe détaillé

Dans le cas des puces d'expression, le protocole expérimental permettant de mesurer l'expression se décompose en 5 étapes (figure 3.1) :

1. Les ARNm issus de la transcription des gènes sont extraits de la cellule ou du tissu. Des ADN complémentaires (ADNc) sont synthétisés par rétro-transcription puis purifiés et amplifiés.
2. Le brin complémentaire des ADNc purifiés est re-synthétisé à partir de nucléotides marqués à la biotine. Les ADN ainsi obtenus fournissent alors une image fidèle de la

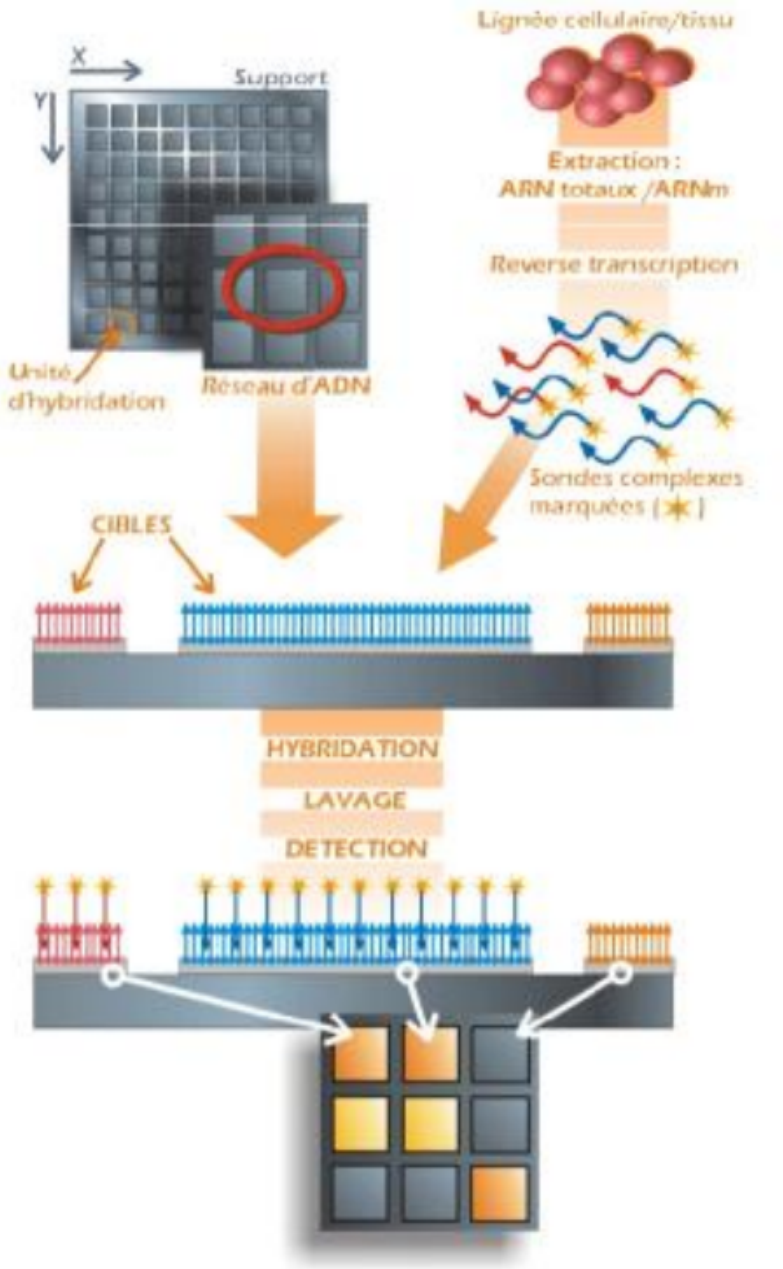


FIGURE 3.1 – **Protocole expérimental d'analyse d'une puce ARN.**

séquence du transcriptome, où chaque morceau de séquence est présent en quantité proportionnelle à la quantité de transcrit correspondant.

3. Les ADN sont alors hybridés sur la puce et la puce est lavée.
4. Un fluorochrome Cy3 est appliqué sur la puce.
5. La puce est scannée. L'intensité lumineuse qui est mesurée sur un point de la puce évalue alors la quantité de l'ARNm correspondant à ce point, dans la cellule ou le tissu, pour l'échantillon étudié.

2.2 Caractéristiques techniques

La forte densité des biopuces (700 000 à 1 million de mesures par puce) relativement au nombre restreint de gènes qui composent le génome permet d'inclure plusieurs mesures par transcrit. Selon les puces, ces mesures peuvent être des réplicats techniques (même séquence) ou simplement des réplicats biologiques (séquence différente ciblant le même transcrit). Les réplicats techniques permettent une quantification fine et précise des erreurs de mesure purement expérimentales, tandis que les réplicats biologiques permettent de pallier en partie les biais de construction liés au choix de la séquence utilisée pour construire les sondes. Notons également que sur les puces les plus récentes (Exon Array Affymetrix), le grand nombre de sondes permet une meilleure couverture de la séquence des transcrits favorisant la recherche de phénomènes d'épissage alternatif.

Une caractéristique importante des puces ARN est la présence de contrôles permettant de vérifier le bon fonctionnement de la puce (contrôles positifs) et de mesurer le niveau d'hybridation non spécifique (contrôles négatifs).

Parmi les contrôles positifs on trouve :

- des *contrôles d'hybridation* ciblant des séquences non humaines ajoutées dans la solution
- des *contrôles de fluorescence* où sont directement fixés des fluorochromes
- des séquences ciblant des gènes de ménage (GAPDH,...)[2]

Les contrôles négatifs plus nombreux que les contrôles positifs (plusieurs milliers de mesures sur chaque puce) sont en général des sondes dont la séquence est formée aléatoirement et ne correspond à aucun transcrit connu. L'hybridation mesurée par ces sondes est alors uniquement non-spécifique.

2. Les gènes de ménage (ou "HouseKeeping Genes" en anglais) sont des gènes codant des protéines essentielles au bon fonctionnement de la cellule et dont le niveau est par conséquent maintenu élevé en permanence.

2.3 Différents types de puces

Bien que fondées sur les principes communs déjà évoqués, les biopuces diffèrent selon les constructeurs, entraînant des différences tant au niveau du traitement que du choix de la terminologie.

Ainsi les puces créées par Affymetrix sont construites en déposant sur une plaque des séquences qui composent les sondes à des emplacements fixes (les *spots*). Ces séquences, différentes pour chaque sonde, sont regroupées par *probesets* d'une dizaine de séquences ciblant le même transcrit. Dans les puces Affymetrix les plus anciennes[3], des versions imparfaites des sondes, appelés "mismatchs" étaient également déposées sur la puce dans le but de capturer l'hybridation non-spécifique liée à chaque séquence[4]. L'efficacité de l'utilisation des mismatchs ayant été largement remise en cause, les puces les plus récentes sont composées uniquement de "perfect matchs" (sondes originales). Les puces Affymetrix contiennent entre 30 000 et 60 000 probesets selon les modèles.

Pour les puces Illumina, les séquences sont assemblées directement sur des billes magnétisées qui sont ensuite réparties aléatoirement sur chaque puce. Chaque séquence est mesurée sur la puce par un nombre aléatoire de billes (entre 15 et 30 en moyenne selon la puce). Ce positionnement aléatoire permet de réduire le risque de biais lié à la répartition spatiale des sondes. Par ailleurs, chaque sonde donnant lieu à un nombre important de réplicats techniques, il est possible de mesurer avec précision le niveau d'expression de chaque ARNm ciblé, et de quantifier l'erreur de mesure spécifique de chaque sonde en considérant la variance entre billes. En revanche, la faible variabilité des séquences rend les puces Illumina plus vulnérables aux biais de construction. En effet, avec un nombre de sondes compris entre 24 000 et 48 000, la majeure partie des transcrits ne sont représentés que par une seule séquence.

Le tableau 3.2 détaille les principales caractéristiques de chaque puce.

3 Puces de génotypage

3.1 Principe détaillé

Le génotypage par biopuce a pour objectif d'identifier simultanément le génotype de plusieurs centaines de milliers de SNP. On peut distinguer parmi les protocoles existants deux principes fondamentaux :

3. Affymetrix HGU95A et HGU133A principalement.

4. La séquence des mismatchs imitait la séquence des sondes originales ("perfect matchs") en faisant varier deux des nucléotides. L'hybridation au transcrit était ainsi empêchée tout en conservant des propriétés chimiques et thermodynamiques proches de la sonde d'origine.

Constructeur	nom de la puce	nombre de sondes	nombre de réplicats par sonde	commentaire
Illumina	Human Ref-8 v3	$> 24\,000$	~ 30	placement aléatoire des billes ; choix restreint de transcrits
Illumina	Human HT-12 v3	$> 48\,000$	~ 15	placement aléatoire des billes
Illumina	Human WG-6 v3	$> 48\,000$	~ 15	placement aléatoire des billes
Affymetrix	hg U133A 2.0	$> 22{,}000$	~ 12	inclut des mismatchs ; choix restreint de transcrits
Affymetrix	hg U133A Plus 2.0	$> 54{,}000$	~ 12	inclut des mismatchs
Affymetrix	Exon Array	1.4 million	~ 4	couvre tous les exons d'un gènes (et donc tous les transcrits)

Tableau 3.2 – **Caractéristiques des puces à ARN Illumina et Affymetrix.**

- Le *génotypage par hybridation directe* est peut-être la méthode la plus spontanée puisqu'elle consiste simplement à construire des puces où chaque SNP est représenté par deux sondes distinctes ciblant les séquences flanquantes de chacun des allèles possibles. Par un protocole classique, on fragmente l'ADN avant de l'hybrider sur la puce. Les fluorescences obtenues correspondent alors à la quantité d'ADN génomique correspondant à chaque allèle. On peut ensuite déduire le génotype à partir du ratio des intensités (cf. section 1.2 du chapitre 5).
- Le *génotypage par marquage différencié* relève d'un principe plus complexe mais permettant de génotyper un nombre plus important de SNP. Après avoir été extrait, l'ADN génomique est fragmenté et mélangé à un cocktail d'oligo-nucléotides dans lequel se trouvent pour chaque SNP deux séquences spécifiques de chaque allèle. Chaque oligo-nucléotide comporte également à l'une de ses extrémités une séquence d'adressage et à l'autre un marqueur fluorescent dont la couleur varie selon l'allèle ciblé (rouge ou vert). Les oligo-nucléotides sont ensuite hybridés à la puce sur laquelle se trouvent des sondes ciblant les séquences d'adressage spécifiques de chaque SNP. Pour chaque SNP, le ratio rouge/vert obtenu par le scanner à la lecture de la sonde correspondante détermine le génotype. Ce protocole, généralisé sur les puces les plus récentes, permet de diviser par deux le nombre de sondes nécessaires au génotypage d'un SNP.

3.2 Choix des SNP ciblés

Depuis leur apparition, la capacité des puces de génotypage n'a cessé de croître, passant de 10 000 SNP en 2004 à près d'un million sur les puces les plus répandues aujourd'hui (Affymetrix 6.0) et plus de 2.5 millions sur la dernière génération de puces Illumina (Illumina Omni). Cette amélioration, rendue possible par l'amélioration des techniques et des protocoles, est également liée à l'accroissement de la connaissance de la variabilité du génome humain résultant projets HapMap (`www.hapmap.org`) et 1000 Génomes (`www.1000genomes.org`).

Le choix des SNP est en effet l'une des questions cruciales qui se posent lors du design de puces de génotypage. La raison principale est la fréquence relativement faible de nombreux variants, qui les rend peu compatibles avec des analyses génome entier[5]. Cela induit un biais vers la sélection de variants fréquents.Une seconde raison vient du déséquilibre de liaison entre les SNP[6].

Les stratégies de choix des SNP cherchent à établir une couverture fine du génome, soit par une forte densité et un espacement régulier permettant d'utiliser les SNP mesurés comme proxys pour les SNP non génotypés, soit par la sélection de "tagSNP" caractérisant les haplotypes et permettant d'inférer au mieux les génotypes des SNP manquants par imputation (Illumina Human Hap550).

3.3 Différents types de puces

Comme pour l'expression, les puces de génotypage sont majoritairement produites par les compagnies Illumina et Affymetrix. Outre les différences de construction des puces déjà décrites pour les puces d'expression (technologie d'hybridation sur billes ou sur plaque, présence de mismatchs, positionnement aléatoire ou fixe,...), les puces de génotypage des deux compagnies diffèrent par le choix de la méthode de génotypage (hybridation directe pour Affymetrix, marquage différencié pour Illumina) et les stratégies de sélection des SNP (sélection de SNP en fonction de leur capacité à décrire le haplotypes sur les puces Illumina, haute densité et bonne couverture par des proxys pour Affymetrix).

Le tabeau 3.3 détaille les caractéristiques des puces de génotypage les plus courantes.

5. La puissance pour détecter l'association d'un variant avec un phénotype diminue avec la fréquence de l'allèle mineur.

6. C'est à dire la forte dépendance qui existe entre des SNP proches sur le génome, du fait de la transmission par blocs (haplotypes) de l'information génétique.

Constructeur	nom de la puce	nombre de SNP génotypés	espace moyen entre SNP (kb)	fréquence moyenne de l'allèle mineur	commentaire
Illumina	Human 660W	$>$ 658,000	4.2	24%	tagSNP : capture de 92% des variations chez les caucasiens
Illumina	Human 1M	$>$ 1 million	2.4	20%	tagSNP : capture de 96% des variations chez les caucasiens
Illumina	Human Omni 2.5	$>$ 2.5 million	1.2	NC[7]	tagSNP
Affymetrix	Affy 5.0	$>$ 500,000	6	22%	inclut des mismatchs
Affymetrix	Affy 6.0	$>$ 940,000	3.1	19.6%	–

Tableau 3.3 – **Caractéristiques des puces de génotypage Illumina et Affymetrix.**

4 Sources de biais dans les biopuces

Au-delà des différences existant entre les différents types de biopuces, il est possible d'isoler un certain nombre de sources de biais communes à l'ensemble des biopuces que nous tenterons de décrire brièvement dans cette partie. Ces sources d'erreur se manifestant surtout lors de la mesure de traits quantitatifs tels que les niveaux d'expression des transcrits, nous détaillerons les méthodes de correction de ces différents biais dans la section 1 du chapitre 2.

- Le *bruit de fond* constitue l'une des premières sources d'erreur dans l'analyse des données de fluorescence. Lors de l'acquisition de l'image, les imperfections de la puce et les phénomènes d'hybridation non spécifiques entraînent l'apparition de bruit dans les mesures d'intensité. Ce bruit peut être purement aléatoire et simplement augmenter la variabilité des mesures ou avoir une cohérence spatiale (ex : image plus claire dans un coin). Il entraîne des difficultés d'estimation des signaux faibles et une distorsion de la distribution des intensités.
- La deuxième source de variabilité est liée à ce qu'on appelle l'*effet "batch"*. Des variations des conditions expérimentales même mineures (temps d'attente entre l'extraction et l'hybridation, température d'hybridation, ...) peuvent entraîner d'importants changements sur les mesures effectuées (dégradation des ADN/ARN, activation de voies de signalisation des cellules, ...). Des mesures répétées du même échantillon dans des conditions différentes peuvent aboutir à des variations supé-

rieures à celles observées entre des échantillons différents traités ensemble. Cette ressemblance entre échantillons traités dans le même lot ("batch" en anglais) peut être la source de nombreuses confusions dans les expériences de biopuces si elle n'est pas prise en compte de manière adéquate.

– Enfin la dernière source de variabilité est liée aux *biais de construction*, c'est-à-dire au choix des séquences utilisées lors de la construction de la puce. En effet, du fait des propriétés thermodynamiques de l'ADN, certaines séquences s'hybrident mieux que d'autres, résistent mieux à la dégradation, aux changements de température ou sont susceptibles de s'hybrider à des cibles multiples.

CHAPITRE 4

Acquisition des mesures d'expression

Dans ce chapitre, nous présentons les différentes étapes de normalisation et de contrôles qualité standard permettant de limiter l'impact des biais expérimentaux sur les résultats obtenus par biopuces. Il est à noter que les étapes décrites ici et leur ordre peuvent différer selon les données disponibles. Bien que certaines étapes décrites ici puissent être transposées à différents types de puces, nous nous contenterons ici de décrire les étapes nécessaires à la normalisation des puces de type Illumina utilisées dans l'étude GHS.

1 Normalisation intra-puce

1.1 Contrôles de la qualité de l'hybridation

Lors du traitement des puces, un premier contrôle qualité est nécessaire afin de repérer les échantillons présentant des problèmes (et qui nécessitent donc d'être ré-hybridés) et orienter le choix de la normalisation. Les étapes standard de contrôle qualité incluent :

- La *vérification des contrôles positifs et négatifs de la puce* afin de repérer les échantillons sur lesquels l'hybridation ne s'est pas bien déroulée ou qui contiennent un bruit de fond anormalement élevé.
- Le *repérage des biais spatiaux* par la visualisation directe de l'image du microarray (cf. figure 4.1) permet de répérer les éhantillons présentant des défauts d'hybridation et nécessitant d'être réhybridés.
- d'autres *défauts d'hybridation* peuvent être repérés en observant les distributions des intensités sur chaque puce (graphe de la densité ou boite à moustache). En particulier une accumulation de points à la limite supérieure de la distribution peut indiquer une saturation du scanner nécessitant de refaire l'hybridation sur des ARN dilués.

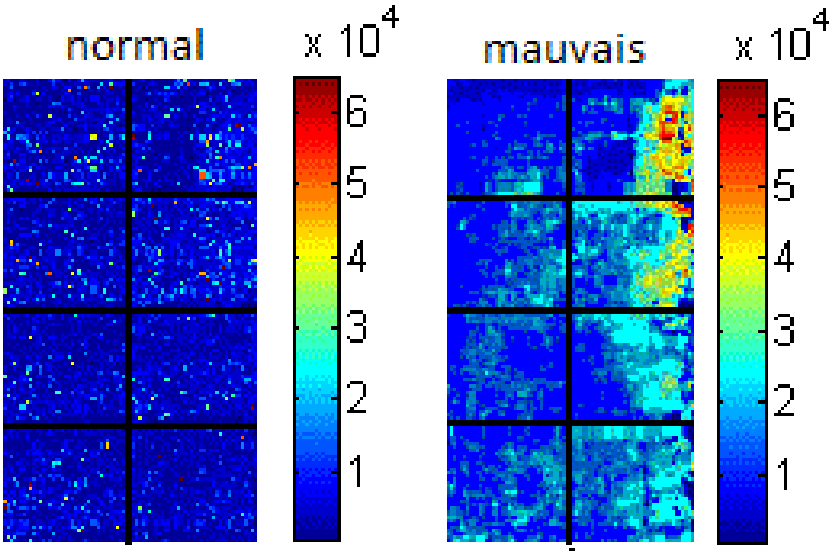


FIGURE 4.1 – **Visualisation de l'image des intensités et qualité de l'hybridation :** on voit à gauche l'image des intensités pour une puce où l'hydridation s'est bien déroulée et à droite une puce présentant un problème d'hybridation.

1.2 Agrégation des mesures

Comme nous l'avons évoqué plus haut, les données de biopuces Illumina présentent un nombre important de réplicats techniques des mesures d'intensité associées à chaque séquence. Dès lors, il est nécessaire d'agréger ces mesures en une mesure unique (avec également une quantification de l'erreur de mesure) qu'il sera ensuite possible d'associer à des phénotypes ou des génotypes [1].

Il serait tentant de s'affranchir de ces questions par le recours à des statistiques d'agrégation standard telles que la moyenne ou l'écart type. Cependant l'observation montre que la distribution des erreurs de mesure liées aux biopuces s'apparente plus à une loi de Laplace [22] (distributions à queues épaisses) qu'à une loi normale. Un choix alternatif généralement adopté dans les expériences de biopuces est le recours aux moyennes robustes. Ces méthodes ont l'avantage de ne pas être sensibles à la présence d'observations extrêmes tout en intégrant la majeure partie de l'information disponible. Les moyennes robustes sont des moyennes pondérées des observations où la pondération $w(\cdot)$ choisie est liée à la place de l'observation dans la distribution. Ainsi pour n billes d'intensité $(X_i)_{i\in\{1\dots n\}}$, la moyenne robuste s'écrit :

$$M_w(X) = \sum_i w(X_i) * X_i \quad \text{avec} \quad \sum_i w(X_i) = 1$$

1. Ce point est particulièrement critique sur les puces Illumina où le nombre de mesures associées à une sonde varie d'une puce à l'autre.

Le logiciel BeadStudio fourni par Illumina agrège les mesures d'expression par des moyennes tronquées où les observations au-delà d'un certain quantile α sont retirées [23], ce qui correspond à la pondération suivante :

$$w(X_i) = \frac{1-\alpha}{n} \mathbb{1}_{q_{\alpha/2} < X_i < q_{1-\alpha/2})}$$

On peut également utiliser des fonctions de pondération plus complexes telles que la fonction "biweight" de Tukey utilisée par défaut sur les puces Affymetrix [24]. Cette fonction introduit une pondération décroissante des observations jusqu'à 5 écart-types du centre de la distribution et ne tient pas compte des observations au-delà de cette limite.

$$w(X_i) = \left(1 - \left(\frac{X_i - m}{5s}\right)^2\right)^2 \mathbb{1}_{|\frac{X_i - m}{5s}|<1}$$

où m et s sont respectivement la médiane et l'écart-type des observations.

Quelle que soit la fonction de pondération choisie, le même principe de pondération peut être généralisé pour estimer l'erreur standard de la moyenne robuste ainsi réalisée :

$$SE_w(X) = \sum_i w(X_i) * (X_i - M_w(X))^2$$

1.3 Traitement du bruit de fond

Le terme bruit de fond désigne la présence d'une hybridation non spécifique aléatoire s'ajoutant aux vrais signaux liés aux quantités d'ARN hybridées sur les puces. Cette hybridation non spécifique se traduit par l'ajout au vrai signal S d'une variable aléatoire B (qu'on suposera généralement positive et distribuée selon une loi gaussienne) ayant pour effet d'augmenter la variabilité des mesures et de mener à une surestimation du signal.Le signal mesuré X est donc modélisé comme la somme de ces deux composantes $X = B+S$.

Comme nous l'avons évoqué, le bruit de fond peut être de deux sortes (cf. Figure 4.1) :
- soit un bruit blanc sans structure, augmentant simplement le niveau moyen des mesures et leur variabilité.
- soit un bruit structuré entraînant des artefacts spatiaux, c'est-à-dire tel que le bruit de fond $\epsilon_{x,y}$ mesuré sur une bille située à la position (x, y) soit corrélé au bruit de fond mesuré à la position $(x + \Delta x, y + \Delta y)$

Selon la modélisation choisie, les méthodes de correction se limiteront à un ajustement global pour tenir compte de l'augmentation de niveau et de variabilité induite par le bruit de fond, ou estimeront localement le niveau de bruit de fond afin de corriger d'éventuels biais spatiaux. Sur les puces Illumina, la nature aléatoire du positionnement des billes

rend les mesures robustes aux biais spatiaux. Par conséquent, nous n'aborderons ici que les méthodes de correction globales qui sont les seules implémentées dans les logiciels standard.

1.3.1 Utilisation des contrôles négatifs

La méthode la plus simple pour mesurer le bruit de fond [25], consiste à mesurer le niveau moyen et la variance des contrôles négatifs présents sur la puce. Par définition ces contrôles sont associés à un signal nul et donnent une image fidèle du bruit de fond. Néanmoins, retrancher le signal moyen des contrôles négatifs pour limiter l'impact du bruit de fond sur les valeurs faibles peut amener à l'apparition de valeurs négatives. Ces valeurs négatives sont ensuite susceptibles d'entraîner l'apparition de valeurs manquantes lors du passage à l'échelle logarithmique. Une solution proposée consiste donc à réintroduire une constante dans le modèle afin d'éviter l'apparition de valeurs manquantes.

1.3.2 Modélisation Norm-Exp

Une alternative proposée par Irizarry [26] consiste à modéliser les intensités observées X comme la résultante d'un signal S distribué selon une loi exponentielle de paramètre λ et d'un bruit de fond B distribué selon une loi normale de moyenne μ_b et de variance σ_b. Estimer le vrai signal pour une sonde donnée revient alors à estimer l'espérance de S sachant l'intensité observée Xdonnée par la formule suivante :

$$\mathbb{E}\left[S|X\right] = a + b\frac{\phi(\frac{a}{b}) - \phi(\frac{x-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{x-a}{b}) - 1}$$

où ϕ et Φ désignent respectivement la densité et la fonction de répartition d'une loi normale centrée réduite et avec $a = x - \widehat{\mu_b} - \widehat{\sigma_b^2}\widehat{\lambda},\ b = \widehat{\sigma_b}$.

σ_b et μ_b pouvant être estimés directement à partir des contrôles négatifs et λ par maximum de vraisemblance, on déduit facilement l'espérance du vrai signal (positif par construction) à partir des observations. On peut noter que cette correction du bruit de fond a pour effet principal de compresser les signaux les plus faibles dans lesquels la variabilité est majoritairement attribuable au bruit de fond. Cette correction est par ailleurs reprise et intégrée avec quelques modifications dans les modèles de stabilisation de la variance présentés dans la section suivante.

1.4 Contrôle de la variabilité des mesures

1.4.1 Le problème

Une fois les mesures agrégées et les erreurs de mesure estimées, on peut voir apparaître un autre artefact présent sur les biopuces. Il s'agit de l'existence d'un "effet taille" dans

la variabilité du signal, c'est-à-dire d'un lien entre niveau du signal et erreur de mesure. Le graphique 4.2 montre que l'incertitude liée au signal augmente avec ce signal. Les conséquences de ce genre d'artefact sont multiples et peuvent aller d'une perte de puissance dans les tests d'association [2] à une sur-pondération des gènes les plus exprimés dans certaines méthodes d'analyse des données [3].

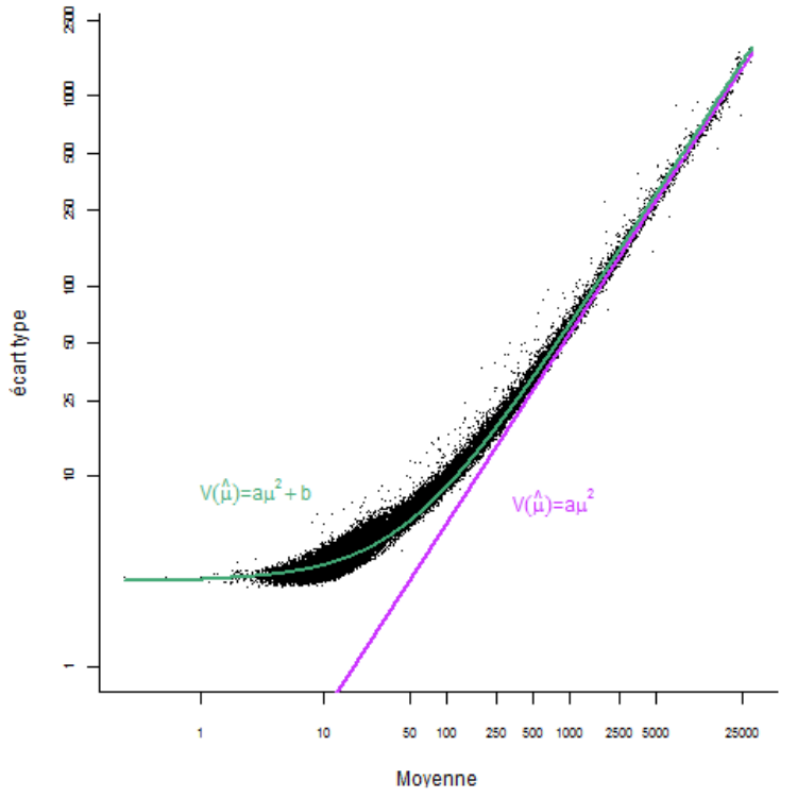


FIGURE 4.2 – **Incertitude des mesures (écart-type) en fonction du niveau d'expression (moyenne des billes).** La droite mauve représente l'incertitude attendue sous un modèle à erreurs multiplicatives, la courbe verte représente l'incertitude attendue sous le modèle avec combinaison d'erreurs multiplicatives et additives supposé par la transformation VST. *source :* Données GHS.

2. Cette perte de puissance est en fait une conséquence de l'hétéroscédasticité des résidus (lien entre la variance des résidus et les covariables) qui pourrait être incluse dans le modèle. Cependant la prise en compte de cette hétéroscédasticité complique fortement la définition et l'estimation des modèles et il est donc d'usage de la traiter à priori.

3. Telles que l'analyse en composantes principales ou le Multi-Dimensional Scaling.

1.4.2 Transformation de stabilisation de la variance

Une solution classique pour pallier cet effet taille, consiste à supposer des erreurs multiplicatives et à considérer le logarithme des données plutôt que les intensités brutes. On évite alors l'effet taille en réduisant la variabilité des gènes les plus fortement exprimés (figures 4.3a. et 4.3b.). Cependant, selon la méthode utilisée pour la correction du bruit de fond cette transformation peut avoir des conséquences néfastes : non seulement elle génère des valeurs manquantes lorsque les mesures d'expression initiales sont négatives mais en plus elle augmente artificiellement la variabilité des gènes les moins exprimés (figure 4.3b.).

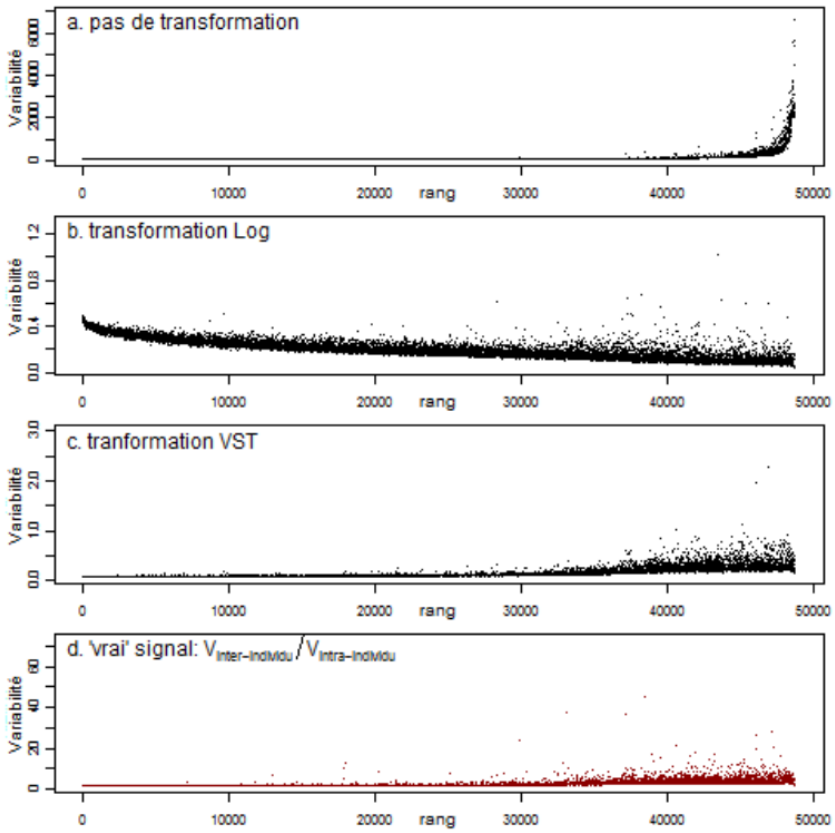


FIGURE 4.3 – **Variabilité de l'expression des gènes en fonction de leur niveau d'expression, selon la transformation utilisée :** Les graphiques a. à c. représentent les écart types des mesures de chaque sondes, rangés par niveau d'expression croissant pour les données brutes (a.), les données transformées par logarithme (b.) et les données transformées par la transformation VST (c.). Le graphique (d.) donne une estimation du "vrai" signal, obtenu en calculant pour chaque sonde le ratio entre la variance calculée sur les différents individus de GHS (données brutes) et l'erreur standard moyenne estimée à partir des réplicats techniques. *source :* Données GHS.

Cette augmentation de la variabilité des signaux les plus faibles provient du fait que l'hypothèse de multiplicativité des erreurs n'est pas vérifiée pour les niveaux d'expression les plus faibles (figure 4.2). Faisant ce constat, Lin et al. [27] ont proposé de modéliser le lien entre le niveau d'expression μ d'un transcrit et le signal $\hat{\mu}$ de la sonde correspondante par la relation

$$\hat{\mu} = \alpha + \mu e^{\eta} + \epsilon$$

avec α une constante représentant le niveau moyen de bruit de fond dû aux phénomènes d'hybridation non spécifiques et η et ϵ deux bruits blancs gaussiens, de variances respectives σ_ϵ^2 et σ_η^2, représentant les erreurs de mesure additives et multiplicatives. Ce modèle peut-être vu comme une extension du modèle Norm-Exp présenté dans la section 1.3.2 dans laquelle on aurait un bruit de fond gaussien $B = \alpha + \epsilon$ et un signal de distribution non précisée, sujet à une erreur multiplicative lognormale $S = \mu e^{\eta}$.

Dans ce nouveau modèle, on a

$$\mathbb{E}\left[\hat{\mu}\right] = \alpha + m_\eta \mu$$

et

$$\mathbb{V}\left[\hat{\mu}\right] = \sigma_\epsilon^2 + \mu^2 s_\eta^2$$

où m_η et s_η^2 désignent l'espérance et la variance de e^η, la variance peut alors s'écrire en fonction de l'espérance comme

$$\mathbb{V}\left[\hat{\mu}\right] = \left(\mathbb{E}\left[\hat{\mu}\right] - \alpha\right)^2 \frac{s_\eta^2}{m_\eta^2} + \sigma_\epsilon^2$$

On voit donc que ce modèle permet de retrouver le lien quadratique entre la variance et l'espérance observé figure 4.2. De plus pour une variance nulle du bruit de fond σ_e^2 on retrouve le modèle logarithmique classique.

Lin et al. montrent que sous ce modèle, il est possible de trouver une transformation $h : x \longrightarrow h(x)$ telle que la variance du signal transformé $h(\hat{\mu})$ soit indépendante de son espérance. Cette transformation appelée transformation de stabilisation de la variance peut s'écrire sous la forme

$$h(x) = \begin{cases} \frac{1}{c_1}\mathsf{arcsinh}\left(\frac{c_1}{\sigma_\epsilon}(x-\alpha)\right) & \text{, si } \sigma_\epsilon > 0 \\ \frac{1}{c_1}\log(c_1(x-\alpha)) & \text{, si } \sigma_\epsilon = 0 \end{cases}$$

où c_1 est le coefficient de variation de l'erreur multiplicative e^η donné par $c_1 = \frac{s_\eta}{m_\eta}$ Il suffit alors d'estimer les paramètres du modèle liant la variance aux expressions à partir de l'observation des niveaux d'expression et des incertitudes pour calculer la transformation adéquate. L'application de cette transformation permet de réduire la variabilité des mesures les plus fortes sans augmenter la variabilité des mesures les plus faibles.

De plus si on décompose la variance d'une sonde dans la population entre une variance intra-individu (obtenue en prenant le carré de l'erreur standard moyenne estimée à partir des réplicats techniques) et une variance inter-individus

$$\mathbb{V}\left[\hat{\mu}\right] = \mathbb{V}_{intra} + \mathbb{V}_{inter}$$

on peut estimer le "vrai" ratio signal/bruit en prenant

$$\frac{\mathbb{V}_{inter}}{\mathbb{V}_{intra}} = \frac{\mathbb{V}\left[\hat{\mu}\right] - \mathbb{V}_{intra}}{\mathbb{V}_{intra}}$$

Ce ratio doit alors être nul pour les sondes contenant uniquement du bruit de fond et être strictement supérieur à 0 pour les sondes ciblant un transcrit dont le niveau est variable entre les individus. On s'attend alors si la correction est satifsfaisante à ce que la variance du signal après correction se rapproche du ratio signal/bruit, ce qu'on observe effectivement avec la transformation VST (figures 4.3c. et 4.3d.).

2 Normalisation inter-puces

Comme nous l'avons vu plus haut, les mesures d'expression peuvent se révéler fortement instables d'une expérience à l'autre à cause de l'effet "batch". En effet, bien que les protocoles expérimentaux rigoureux permettent aujourd'hui de minimiser ces différences, il est courant que lors de la préparation des échantillons la quantité d'ARN déposée sur les puces, les réglages du scanner, ou d'autres conditions expérimentales génèrent des différences dans la distribution des intensités mesurées. Ces différences ont pour conséquence d'augmenter la quantité de bruit présente dans les données et peuvent conduire à des associations fallacieuses entre l'expression et les traits considérés. De telles associations peuvent se produire lorsque l'erreur n'est pas aléatoire. Par exemple, lorsque des cas et des témoins d'une expérience sont traités séparément, des différences dans le traitement des lots (température d'hybridation, temps d'attente après extraction des ARN,...) peuvent induire des différences artificielles entre cas et témoins. Ces biais sont présents à plus forte raison lorsque sont réunis dans un même jeu de données des échantillons traités par des personnes différentes, à des dates différentes, dans des centres différents ou provenant d'études différentes. Il est donc préférable de normaliser les données afin de réduire la variabilité inter-puces. Nous présentons ici les outils standard utilisés dans ce but.

2.1 Normalisation par quantiles

Afin de limiter les différences entre les distributions des expressions d'un échantillon à l'autre, Bolstad a proposé d'appliquer une méthode de normalisation dite par quantile

[28] : cette normalisation vise à rendre identiques les distributions des expressions des sondes d'un échantillon à l'autre, sans modifier l'ordre relatif des transcrits sur chaque puce. Pour cela une procédure en 3 étapes est utilisée :

1. Attribuer un rang aux sondes pour chaque échantillon en fonction de leur niveau d'expression.
2. Calculer pour chaque rang i la moyenne m_i des niveaux d'expression des sondes de rang i entre les différents échantillons.
3. Attribuer à toutes les sondes de rang i le niveau d'expression m_i.

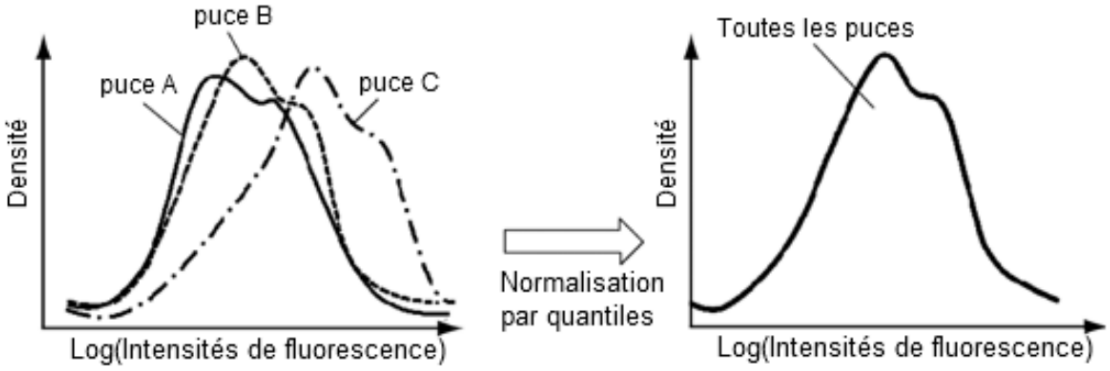


FIGURE 4.4 – **Effet de la normalisation par quantiles sur les distributions des niveaux d'expression.**

La procédure décrite permet de garantir l'égalité des distributions entre les différentes puces (figure 4.4) tout en conservant l'ordre relatif des expressions sur chaque puce. Boldstad *et al.* ont montré que l'application de cette procédure permettait de réduire efficacement la variabilité inter-puces sans biaiser les résultats des tests d'association [29]. Par ailleurs, Qiu *et al.* ont montré que l'application de méthodes de normalisation telles que la normalisation par quantile permettait de réduire fortement la dépendance entre les tests lors des procédures de tests multiples (sans toutefois permettre de s'en affranchir totalement) [30]. Il convient à ce propos de noter que l'atténuation des corrélations entre les gènes liée à la normalisation par quantile, s'applique également aux corrélations d'origine biologique et peut donc perturber la reconstruction des réseaux biologiques [31].

2.2 Normalisation par splines/lowess

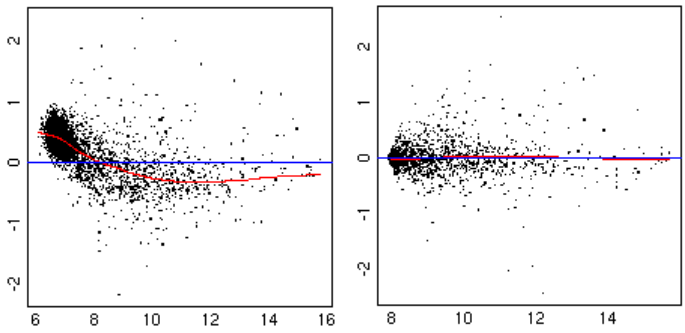
Il a été reproché à plusieurs reprises [32, 33] à la normalisation par quantile d'être trop drastique et susceptible d'effacer certaines différences biologiquement pertinentes entre puces (en particulier pour les sondes présentant les intensités les plus élevées). Afin

de limiter ce risque, d'autres types de normalisation ont été développés. Nous allons ici présenter le principe général des méthodes de normalisation par lowess [34] ou par splines [35, 36] qui figurent parmi les alternatives à la normalisation par quantile les plus fréquemment utilisées.

Ces méthodes se décomposent en quatre étapes :

1. Estimation d'un profil d'expression consensus A en moyennant chaque sonde sur l'ensemble des échantillons
2. Calcul pour chaque échantillon i de l'écart M_i au profil consensus : $M_i = X_i - A$
3. Ajustement pour chaque échantillon i d'un modèle $M_i = f_i(A) + \epsilon$ à l'aide de méthodes de régression non paramétriques (figure 4.5a)
4. Pour chaque échantillon i, la tendance $f_i(A)$ est retranchée (figure 4.5b)

Il est également possible d'appliquer cette correction sans définir un profil consensus. La correction se fait alors la correction sur l'ensemble des paires de biopuces possibles par un processus itératif. Ces méthodes de normalisation permettent de supprimer les écarts systématiques entre distribution sans imposer une égalité stricte des distributions. Elles sont donc moins susceptibles de supprimer des signaux pertinents en queue de distribution.



(a) avant normalisation (b) après normalisation

FIGURE 4.5 – **Normalisation par splines/loess :** Visualisation écarts-consensus (MA plots) avant et après la normalisation inter-puces. En abscisse est représentée la moyenne des signaux des sondes des différents puces (consensus). En ordonnée est représenté l'écart de chaque puce au consensus. La ligne rouge montre la tendance estimée des écarts qui est retranchée pour aboutir aux données normalisées montrées en (b.). La normalisation est ici effectuée sur deux échantillons issus des données de HaemAtlas.

Notons toutefois que ces méthodes font, comme la normalisation par quantiles, l'hypothèse d'une distribution commune des niveaux d'expression entre les échantillons supposés venir d'une source homogène. Or, cette hypothèse n'est pas toujours vérifiée, et peut poser

problème lorsque l'on étudie des échantillons hétérogènes en masquant partiellement certaines différences réelles entre individus (composition d'un mélange cellulaire, différences biologiques importantes).

2.3 Repérage d'échantillons atypiques

Lorsque les échantillons considérés proviennent d'une population homogènes (issus du même centre, un seul type cellulaire,...), il peut être souhaitable de repérer les individus atypiques, pour éliminer les échantillons présentant d'éventuels défauts et éviter de donner trop de poids à des "outliers". Pour cela, on utilise un critère fondé sur les corrélations médianes entre échantillons. En calculant les corrélations 2 à 2 entre échantillons, et en prenant pour chaque échantillon la médiane des corrélations avec les autres échantillons, on obtient un indicateur s de la similarité d'un échantillon au reste des échantillons.

On peut alors définir un seuil de similarité en dessous duquel un échantillon sera considéré comme atypique en appliquant un critère classique de détermination des observations atypiques. Nous considérons donc comme atypiques les observations telles que $s < \overline{s} - a\sqrt{\mathbb{V}[s]}$ avec $a = 3$ ou 4. Sur GHS, l'utilisation de ce critère conduit à retirer les échantillons dont la corrélation mediane est inférieure à 0.98.

Lorsque d'importantes différences existent entre groupes, il est généralement souhaitable de séparer les échantillons en groupes cohérents avant d'appliquer ces procédures.

2.4 Classification des échantillons

Lorsque le design expérimental permet de définir plusieurs sous-groupes (ex : cas-témoins, plusieurs types cellulaires,...), il arrive que certains échantillons soient mal étiquetés ce qui peut mener à une réduction de la puissance dans les études d'association. Pour parer à ce genre d'éventualité, il est parfois utile d'effectuer une classification des échantillons en se basant sur les corrélations entre les profils d'expression et de confronter la classification obtenue aux informations a priori dont on dispose sur les échantillons. La présence d'incohérences à ce niveau peut alors révéler des erreurs d'étiquetage ou mettre en évidence des échantillons atypiques comme le montre la figure 4.6. Une vérification a posteriori permet dans la plupart des cas d'identifier les causes des différences observées.

De même une classification effectuée exclusivement à partir des niveaux d'expression des gènes du chromosome Y peut permettre de vérifier la cohérence des informations concernant le sexe puisque ce chromosome n'est présent que chez les hommes et que les gènes s'y trouvant ne sont donc pas exprimés chez les femmes. Dans GHS ce critère a permis de retirer une dizaine d'échantillons mal étiquetés pour lesquels des incohérences entre le sexe et l'expression du chromosome Y étaient observées.

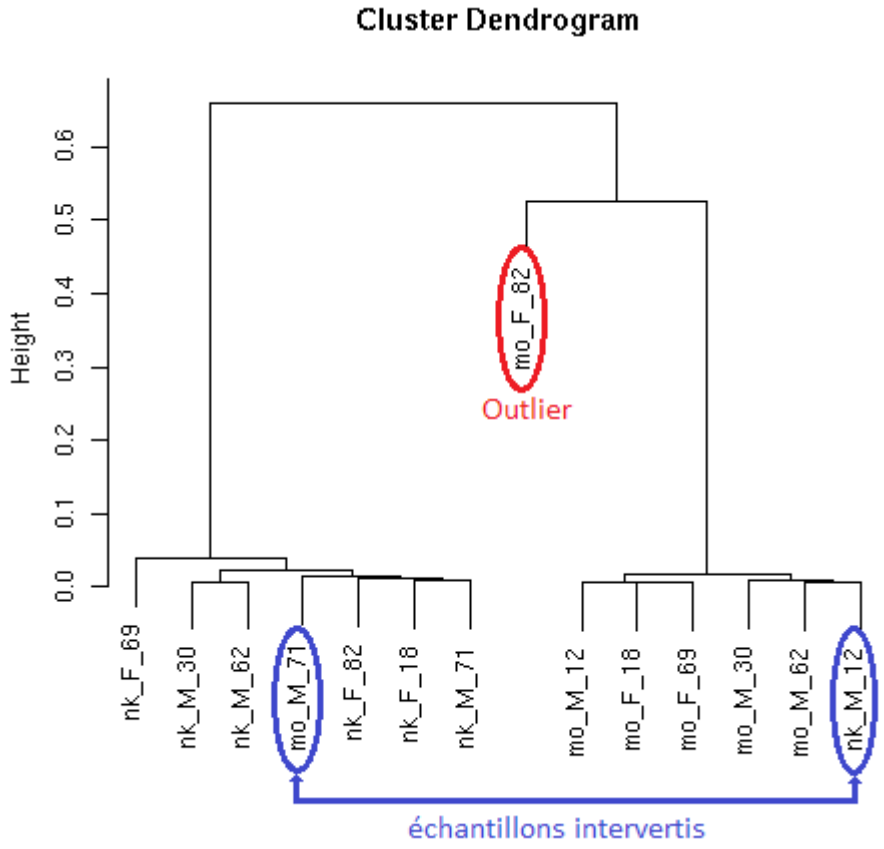


FIGURE 4.6 – **Exemple illustrant l'utilisation de la classification pour le contrôle qualité :** partant d'une étude comparant deux types cellulaires (mo = monocyte, nk= natural killer), la classification met en évidence un échantillon de monocytes se comportant de manière atypique en rouge et une inversion de label entre les deux échantillons marqués en bleu. La classification utilisée est une classification ascendante hiérarchique basée sur la méthode de Ward et la distance 1-cor.

3 Sources d'erreur non prises en compte par le prétraitement

Il existe, en plus des sources d'erreur déjà prises en compte par la normalisation, une large gamme de biais susceptibles de perturber les analyses et de fausser les interprétations. Bien que des méthodes de normalisation ou de correction existent pour certains de ces biais, ces méthodes ne font pas l'objet d'un large consensus. Les différents biais abordés dans cette partie doivent donc généralement être traités au cas par cas en fonction des analyses réalisées. Nous détaillerons ici ces différents types de biais et les erreurs d'interprétation auxquelles ils peuvent mener. Nous distinguerons deux types de biais :

les biais expérimentaux liés au design des expériences et souvent à la base de l'effet "batch" décrit précédemment (et donc parfois partiellement corrigés par la normalisation inter-puces).

les biais de construction liés au design des sondes. Ils peuvent généralement être évités par le recours à une annotation minutieuse des sondes.

3.1 Biais expérimentaux

Les biais expérimentaux regroupent l'ensemble des conditions expérimentales qui par leurs variations conduisent à des variations des mesures d'expression. Le contrôle de ces conditions expérimentales est nécessaire pour limiter les biais dus aux effets "batch". Parmi les facteurs susceptibles d'entraîner des biais expérimentaux, on peut citer :

la qualité des ARN : Entre l'extraction des ARN et leur hybridation sur la puce, les ARN cellulaires sont dégradés par diverses enzymes. Les conditions de stockage [4] des ARN peuvent avoir un impact sur la qualité et la quantité des ARN hybridés. En fonction de la stabilité et de la durée de vie des ARN, ceux-ci seront affectés différemment par la dégradation. Ces différences auront donc un impact sur la composition du transcriptome mesuré.

l'activation différentielle de voies de signalisation : Lors des protocoles expérimentaux visant à extraire des tissus, à isoler des types cellulaires ou même à extraire les ARN, il n'est pas rare que les tissus cellulaires soient soumis à des traitements susceptibles d'en altérer le comportement (changement de température, fixation à des anticorps,...). Certaines voies de signalisation sont alors activées et peuvent déclencher la transcription de gènes "spécifiques" d'une voie. Les niveaux d'expression relatifs peuvent donc être modifiés par le choix d'un protocole expérimental. Par

4. En particulier la température.

exemple dans le cas des monocytes, l'isolation des monocytes par sélection positive grâce à des anticorps fixant le marqueur cellulaire CD14[5] a pour conséquence d'activer la voie de signalisation CD14 et de conduire à la synthèse de protéines inflammatoires telles les interleukines IL6 et IL1β et le facteur de nécrose tumorale TNFα [37].

les différences de composition cellulaire entre échantillons : Lorsqu'on travaille sur un tissu ou une type cellulaire donné, la composition du mélange cellulaire est susceptible de changer d'un individu à l'autre ou d'une condition expérimentale à l'autre. Dès lors, ces changements se répercutent sur la mesure du transcriptome. Cette variation peut conduire à de lourdes erreurs d'interprétation si elle n'est pas connue des expérimentateurs. En effet, les changements de composition peuvent faire apparaître des différences dans l'expression de groupes de gènes liés du point de vue fonctionnel. Cette différence peut alors être interprétée comme l'activation d'une voie biologique interne à un type cellulaire, alors qu'elle résulte d'un simple changement de la composition du mélange cellulaire observé. Le comptage des différents types cellulaires[6] et l'ajustement sur les proportions de cellules de chaque type peut permettre d'identifier de tels artefacts et de s'en prémunir. Néanmoins ces méthodes sont particulièrement lourdes et coûteuses et par conséquent peu applicables à large échelle.

la contamination par des types cellulaires non désirés est une variante du biais précédent dans le cas de l'isolation d'un type cellulaire spécifique, la présence, même à l'état de traces, de certains types cellulaires non désirés dans les échantillons peut entraîner des variations de la composition du transcriptome et générer des associations fallacieuses comme nous le verrons dans le chapitre 10.

3.2 Biais de construction

Les biais de construction sont les biais liés au design des sondes : c'est-à-dire au choix par le constructeur des séquences ciblées par les sondes. La connaissance des différentes caractéristiques des sondes et leurs conséquences sur l'analyse peut s'avérer primordiale dans l'interprétation des résultats fournis par les puces d'expression. Dans GHS un important travail dont j'ai eu la charge a consisté à mettre à jour les annotations des sondes en croisant les fichiers d'annotation fournis par Illumina avec les réannotations proposées par Barbosa-Morais *et al.* [38] et les base de données du National Center for Biotechnology Information (NCBI).

5. Marqueur spécifique des monocytes qui permet de séparer les monocytes des autres cellules circulant dans le sang.

6. Par exemple, par cytométrie de flux.

position sur le génome Le premier biais possible est lié à une mauvaise annotation de la sonde. La séquence du génome et du transcriptome étant mise à jour au fur et à mesure des alignement successifs dans le cadre du Human Genome Project, il arrive que des sondes construites à partir de versions anciennes du génome se révèlent défectueuses, soit parce qu'elles ne correspondent à aucun transcrit, soit parce que le transcrit ciblé n'est pas celui indiqué par le constructeur. Afin d'éviter ce genre de phénomène il est prudent de réaligner les séquences des sondes fournies par le constructeur contre les versions les plus récentes du génome ainsi que l'ont proposé Wilson *et al.* [38]. L'annotation des sondes peut alors être grandement améliorée.

hybridation croisée On parle d'hybridation croisée lorsque plusieurs transcrits sont susceptibles de s'hybrider à la même sonde. L'hybridation croisée est susceptible de se produire dès qu'une sonde contient une séquence présente dans plusieurs transcrits différents. Irizarry *et al.* [26] note que l'hybridation des transcrits aux sondes se fait même lorsque la séquence de la sonde correspond de manière imparfaite à celle du transcrit. Le risque d'hybridation croisée est donc assez élevé dès lors que plusieurs transcrits présentent une forte homologie entre eux et avec la séquence de la sonde (comme c'est le cas pour les pseudogènes). Cette fois encore l'alignement des séquences des sondes sur le génome et le retrait (préventif ou a posteriori) des sondes incriminées sont le seul recours.

position de la sonde par rapport au transcrit La position de la sonde par rapport au transcrit peut générer un autre type de biais. Lors de la dégradation, les brins d'ARNm sont dégradés progressivement en partant de leur site de début de transcription (extrémité 5'). Par conséquent les sondes proches du site de début de la transcription sont plus affectées par la dégradation de l'ARN, et présentent systématiquement un niveau d'expression plus faible que les sondes en 3' comme le montre la figure 4.7.

épissage alternatif Lorsqu'un gène est sujet à épissage alternatif, les sondes visant un exon qui n'existe pas dans certains isoformes peuvent apparaître différentiellement exprimés entre des individus porteurs de ces isoformes. Il s'agit alors d'une différence qualitative peut être interprétée à tort comme une différence quantitative. Il faut donc toujours faire attention à ne pas confondre gène et transcrit.

SNP dans la séquence des sondes Lorsqu'un SNP se trouve dans la séquence de la sonde, les ARN s'hybridant à la sonde existent sous deux forme distinctes correspondant aux deux allèles du SNP. La qualité de l'hybridation est alors susceptible de varier selon que la séquence d'ARN est parfaitement complémentaire de la sonde ou présente une base de différence. Dès lors, le signal mesuré apparaîtra artificiellement différent entre individus porteurs de l'un ou l'autre des allèles. Ce genre d'artefact

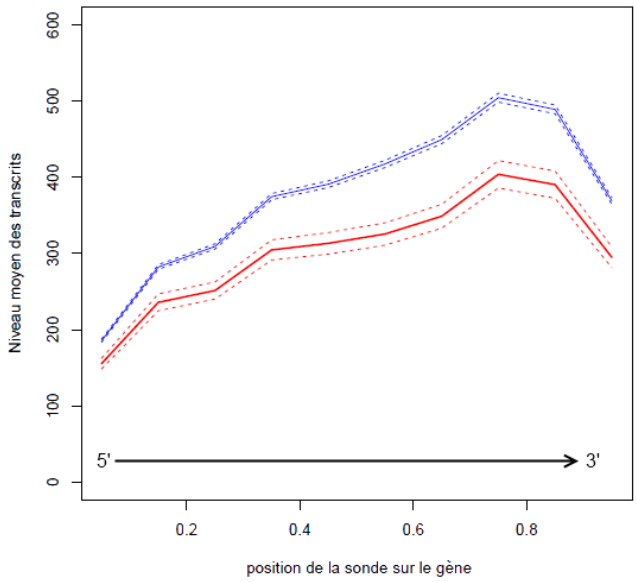


FIGURE 4.7 – **Effet de la dégradation de l'ARN :** niveau moyen d'expression des sondes en fonction de la position en 5' ou en 3' du gène. En bleu est représentée la médiane sur l'ensemble des individus de GHS. En rouge, le niveau mesuré sur des réplicats faits une à deux semaines plus tard. Les intervalles de confiance à 95% sont représentés en pointillés.

est particulièrement gênant lorsque que l'on recherche des associations entre les expressions et les variants génétiques localisés en *Cis* (C'est-à-dire à proximité du gène correspondant) car il génère alors des associations fallacieuses.

teneur en GC la teneur en nucléotides Guanine et Cytosine des ARN détermine en partie leur propriétés thermodynamiques. En particulier les ARN présentant une forte teneur en GC sont plus résistants aux augmentations de la température [39], et résistent donc mieux à la fragmentation des ARN qui précède l'hybridation. Pour cette raison, les sondes à forte teneur en GC ont généralement des niveaux légèrement plus élevés et peuvent apparaître différentiellement exprimés si les protocoles expérimentaux ne sont pas contrôlés.

CHAPITRE 5

Acquisition des génotypes

De même que les puces d'expression, les puces de génotypage sont des outils sensibles dont les résultats peuvent être perturbés par de nombreux artefact techniques. Cependant, l'ADN étant par nature beaucoup plus stable que l'ARN, les biais liés à l'hybridation sont généralement moins complexes. Dans le cadre des études GHS et Cardiogenics, le travail de prétraitement des données génotypiques a été fait en amont du travail de thèse présenté ici. Nous ne développerons donc pas ces aspects et nous nous contenterons d'en donner les principes généraux.

1 Etapes de prétraitement

1.1 Hybridation et Normalisation

Comme nous l'avons déjà évoqué, l'hybridation de l'ADN sur les biopuces et la mesure des intensités lumineuses correspondant aux séquences de chaque SNP est soumise aux mêmes biais que les puces à ARN à savoir :

- La présence d'un bruit de fond non spécifique, présentant éventuellement une structure spatiale.
- Des différences de distribution entre les différentes puces reflétant des paramètres physiques (ex : réglages du scanner) ou biologiques (ex : concentration de l'ADN).
- Des différences de qualité de l'hybridation aux sondes, liées à la séquence (en particulier à la position du SNP dans la séquence).

Ces différences ont cependant un impact moindre sur les résultats du fait de la plus forte stabilité de l'ADN et de la nature qualitative des génotypes.

Il existe également d'autres biais parmi lesquels on peut citer :

- Des biais de coloration sur les puces Illumina (les fluorochromes verts présentent des intensités généralement supérieures et doivent faire l'objet d'un réajustement)
- Des biais d'hybridation liés à la longueur des fragments d'ADN (les fragments les plus longs s'hybrident généralement moins bien)

– Des biais d'orientation (sur les puces Affymetrix) liés au brin sur lequel s'hybride la sonde (pour certaines sondes, les deux brins ne distinguent pas les deux allèles avec la même efficacité)

La procédure de normalisation (décrite plus en détail par Carvalho *et al.* [40]) se déroule donc généralement en 5 étapes :

1. Estimation des différences systématiques dues à la séquence [1] ou à la longueur des fragments ciblés.
2. Correction des intensités (en logarithme) pour ces biais par soustraction des effets estimés.
3. Normalisation par quantile pour limiter les différences entre puces.
4. Agrégation des intensités par allèle et par brin [1] (*sens* ou *anti-sens*).
5. Estimation des log-ratios M des intensités ajustées sur les différents biais possibles (séquence [1] , longueur, position du SNP [1] , biais de coloration [2],...).

Une fois les log-ratios estimés, les génotypes peuvent alors être déterminés.

1.2 Détermination des génotypes

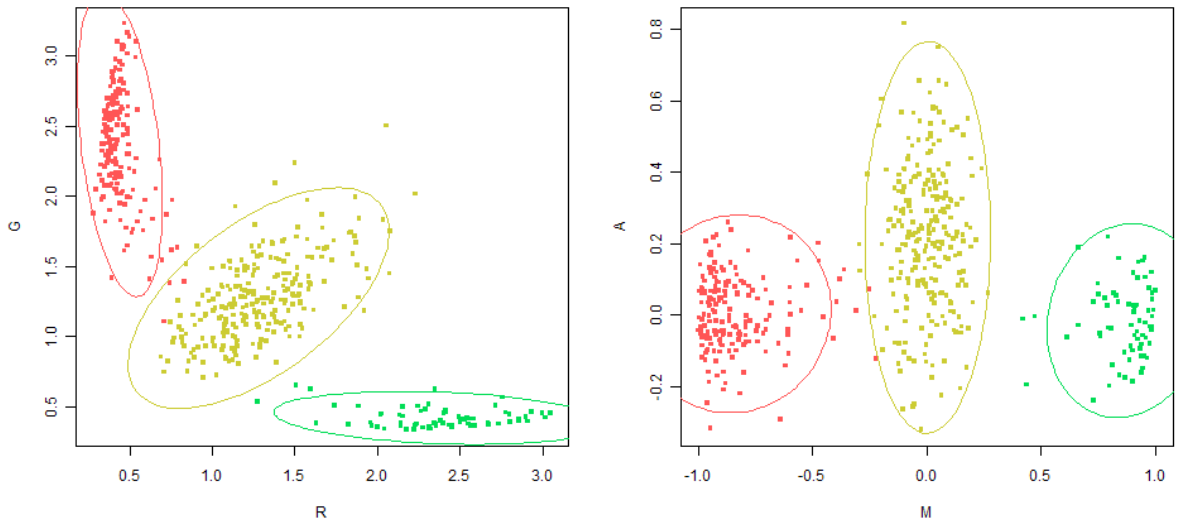
La détermination des génotypes se fait en procédant à un clustering des signaux. Ce clustering est généralement basé sur les log-ratios plutôt que sur les intensités directes des deux sondes. En effet, la majeure partie des différences entre groupes de génotype est portée par les log-ratios (figures 5.1a et 5.1b). Bien que de nombreux algorithmes puissent être utilisés pour effectuer le clustering des intensités, nous présentons ici la méthode retenue par l'algorithme CRLMM, à la base de la plupart des algorithmes utilisés pour déterminer les génotypes.

Cet algorithme repose sur la modélisation de la distribution des log-ratios des différents individus présents dans la population par un mélange de lois gaussiennes centrées sur chaque génotype (cf. figure 5.2).Pour un SNP, la densité f des log-ratios M peut s'écrire

$$f(m) = p_{AA} * \phi(m, \mu_{AA}, \sigma^2_{AA}) + p_{AB} * \phi(m, \mu_{AB}, \sigma^2_{AB}) + p_{BB} * \phi(m, \mu_{BB}, \sigma_{BB} * *2)$$

avec $\mu_{AA} < \mu_{AB} < \mu_{BB}$ les intensités moyennes pour les 3 génotypes, p_{AA}, p_{AB}, p_{BB} les fréquences des 3 génotypes, $\sigma^2_{AA}, \sigma^2_{AB}, \sigma^2_{BB}$ les variances des intensités pour les 3 génotypes, et $\phi(\cdot, m, s)$ la densité d'une loi normale d'espérance m et d'écart type s.

1. Affymetrix uniquement.
2. Illumina uniquement.



(a) Intensités brutes (b) Log-ratios et intensités totales

FIGURE 5.1 – **Visualisation des signaux associés aux sondes d'un SNP :** en intensités brutes R et G (a) et en log-ratio $M = \log(R/G)$ et $A = \log(R.G)$ (b). Les individus homozygotes (AA et BB) sont représentés en rouge et vert, et les individus hétérozygotes (AB) sont en représentés en jaune.

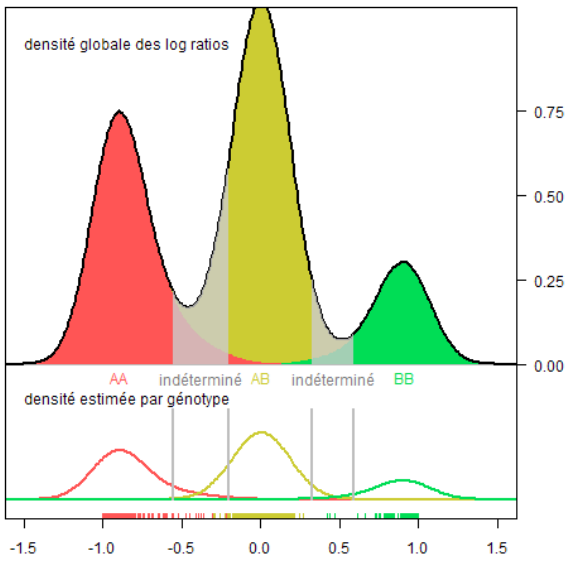


FIGURE 5.2 – **Affectation des individus aux trois génotypes d'un SNP en fonction de la valeur observée du log-ratio :** un modèle est ajusté à la densité globale des observations (en haut) et trois sous-distributions sont identifiées (en bas en rouge, jaune, et vert). Chaque individu est alors affecté au génotype le plus probable. Le génotype des individus situés dans des zones d'incertitude (en gris) reste indéterminé.

Les paramètres de la distribution sont estimés par un algorithme EM et le modèle est comparé à des modèles à un ou deux groupes de génotype afin de pouvoir identifier correctement des SNP rares (pas d'homozygotes pour l'allèle rare) ou n'existant pas dans la population étudiée. Une fois le meilleur modèle choisi, chaque individu est affecté à un génotype si celui-ci a une probabilité supérieure à 95% d'être issu de la distribution associée. Les individus pour lesquels aucun génotype ne peut être déterminé avec une certitude suffisante restent alors indéterminés.

2 Contrôle qualité

Une fois les génotypes déterminés, un certain nombre de contrôles qualité sont nécessaires pour prévenir les biais dans les études d'association subséquentes et éliminer les SNP pour lesquels le processus de détermination n'aurait pas bien fonctionné.

Ces contrôles qualité consistent en un contrôle des échantillons et un contrôle des SNP effectués successivement et de manière itérative. La plupart de ces contrôles qualité peuvent être effectués de manière standard par les logiciels PLINK [41] ou GenABEL [42].

2.1 Contrôle des échantillons

Les études d'association en population reposent sur deux hypothèses majeures :
- les sujets sont indépendants les uns des autres
- les sujets proviennent d'une même population génétiquement homogène.

Ces deux hypothèses permettent en effet d'assurer le caractère "indépendant et identiquement distribué" des variables étudiées, nécessaire à l'application des théorêmes statistiques classiques.

En génétique, ces conditions se traduisent par l'absence de lien de parenté entre les sujets (indépendance) et l'absence de stratification non contrôlée de la population.

2.1.1 Détection de liens de parenté

Afin de détecter les liens de parenté entre individus, on définit l'identité par état ("Identity By State" ou IBS en anglais) d'un SNP entre deux individus comme la moitié du nombre d'allèles partagés par ces deux individus pour ce SNP. L'IBS à un locus reflète l'information génétique partagée par les deux individus à ce locus. Cet IBS peut valoir 1, 1/2 ou 0 selon que les individus ont 2,1 ou aucun allèle en commun.

On peut ensuite définir l'IBS moyen entre deux individus comme la moyenne des IBS sur tous les SNP génotypés chez ces deux individus. On obtient ainsi un score global compris entre 0 et 1 reflétant l'information génétique partagée par deux individus. On peut montrer que pour deux individus indépendants issus d'une même population à l'équilibre

d'Hardy-Weinberg (voir section 2.2.2 pour une définition précise), l'IBS d'un SNP entre ces deux individus aura pour espérance $1 - 2f(1-f) + 2f^2(1-f)^2$ où f est la fréquence du SNP. Si les fréquences alléliques sont réparties uniformément entre 0 et 1, l'espérance de l'IBS se situera autour de 0.75. On peut également montrer que le niveau attendu de l'IBS est plus faible entre individus issus de populations différentes (c'est à dire ayant des fréquences alléliques différentes) et augmente avec le niveau de parenté des individus. Une façon simple de filtrer les individus apparentés est donc de repérer les paires d'individus ayant un IBS trop élevé et de retirer à chaque fois l'individu ayant le plus de génotypes indéterminés. On utilise en général un seuil de 0.95 pour ce filtrage.

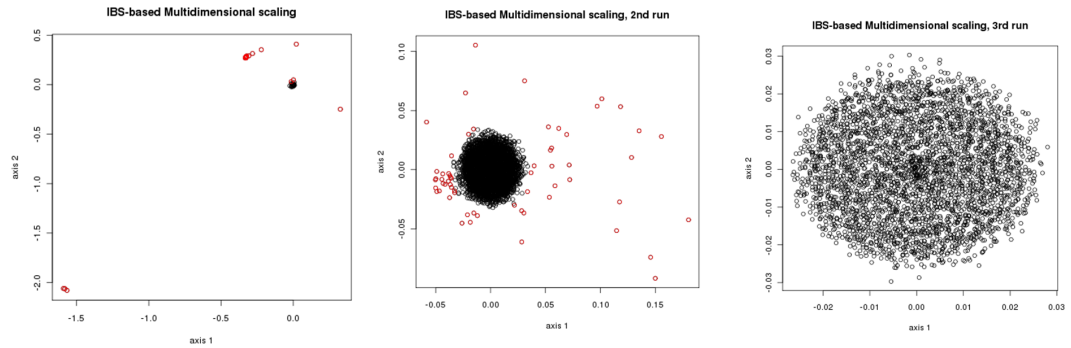
2.1.2 Détection de stratifications de population

L'IBS peut également être utilisé pour étudier la structure de la population et détecter des stratifications sous-jacentes. En effet, les différences de fréquences alléliques entre populations augmentent les différences génétiques entre individus issus de populations différentes. Si on effectue une classification en se servant de l'IBS comme mesure de similarité, il est alors possible de détecter des stratifications de population lorsqu'elles existent.

Le plus souvent on préfère cependant recourir à une méthode de projection non linéaire telle que le *MultiDimensionnal Scaling* (MDS). Cette méthode consiste à chercher la projection des données sur des sous-espaces de dimension fixe, de façon à minimiser le *stress*, défini comme l'écart entre les distances mesurées dans ce sous-espace et les vraies distances entre les points [43]. Ce type de méthode permet une visualisation claire des effets de stratification (cf. figure 5.3a). Il est ensuite possible de contrôler l'effet de la stratification en ajustant sur les premiers axes du MDS.

Dans le cas d'une population homogène a priori, le MultiDimensionnal Scaling (MDS) permet de repérer des individus présentant un profil génétique atypique, généralement issu d'une population différente [3]. Dans ce but, on effectue une projection des données sur un plan (2 dimensions). Des simulations montrent qu'en l'absence totale de structure, le MDS distribue les points sur les axes selon deux gaussiennes indépendantes. Pour identifier les individus atypiques, on peut modéliser les points par une gaussienne bivariée et construire des ellipsoïdes de confiance. En procédant ainsi, on considère comme atypiques les individus situés à plus de 3 écart-types du centre du nuage avec un taux d'erreur de $1\,{}^{0}/_{00}$ environ (figure 5.3a). En itérant ce processus, on arrive à une stabilisation autour de points reflétant une population homogène (figure 5.3b et 5.3c).

3. La comparaison aux données d'Hapmap peut ensuite permettre d'identifier l'origine des sujets.



(a) avant filtrage (b) après un premier filtrage (c) après un second filtrage

FIGURE 5.3 – **Repérage d'individus atypiques et d'effets de stratification par MDS :** Chaque point correspond à un individu. Les distances dans le plan reflètent les distances (1-IBS) entre individus. Les points rouges sont les individus suffisamment éloignés du nuage principal pour être considérés comme atypiques. *Source :* GHS

2.1.3 Taux d'hétérozygotie

Un troisième critère est le taux d'hétérozygotie. Ce taux est défini comme le pourcentage de génotypes hétérozygotes dans un échantillon donné. Les individus présentant un taux d'hétérozygotie trop élevé sont retirés des analyses. En effet, connaissant la répartition des fréquences alléliques, on en déduit la distribution attendue du taux d'hétérozygotie des individus (par exemple par des méthodes de ré-échantillonnage) et on exclut les individus différents significativement de cette distribution.

Les individus présentant des taux d'hétérozygotie anormalement élevés sont le plus souvent des individus sur lequel le génotypage n'a pas bien fonctionné et pour lesquels le génotype d'une grande majorité de SNP a été attribué par défaut à la classe des hétérozygotes.

2.2 Contrôle des SNP

2.2.1 Taux de détermination

Le premier contrôle effectué sur les SNP est la vérification du taux de détermination, c'est-à-dire du pourcentage d'individus pour lesquels un génotype a pu être attribué. Un taux de détermination trop bas indique le plus souvent une défaillance du modèle utilisé pour déterminer les génotypes. De telles défaillances sont relativement fréquentes et peuvent avoir plusieurs causes :

– Des biais expérimentaux mal contrôlés peuvent gêner la reconnaissance des classes et donc l'inférence des génotypes (ex : problème d'hybridation).

– Le SNP peut se situer dans une région où des variations du nombre de copies existent. Dans ce cas le nombre de génotypes possibles est supérieur à trois, puisqu'on peut avoir des génotypes fonction du nombre de copies (A, B, AA, AB, BB, AAA, AAB, ABB, BBB,...)

Un seuil élevé de détermination de l'ordre de 95% est donc généralement requis.

2.2.2 Equilibre d'Hardy-Weinberg

Dans une population, sous certaines conditions (population suffisamment grande, panmixie, pas de mutations, de migrations, ni de pressions de sélection), les fréquences génotypiques des SNP (p_{AA} , p_{AB}, p_{BB}) sont fonction uniquement des fréquences alléliques. C'est l'équilibre de Hardy-Weinberg. Les proportions des trois génotypes sont alors données par

$$\begin{aligned} p_{AA} &= p^2 \\ p_{AB} &= 2pq \\ p_{BB} &= q^2 \end{aligned}$$

où p et $q = 1 - p$ sont les fréquences des allèles A et B.

Des écarts aux fréquences d'équilibre peuvent donc être le signe :

1. De la présence de sous-populations dont les fréquences alléliques diffèrent pour le SNP.
2. De pressions de sélection au locus considéré.
3. D'écart à la panmixie (Consanguinité, choix du conjoint par appariement sur certains critères phénotypiques,...).
4. De défaillances du modèle de détermination des génotypes.

Il est donc souvent préférable de retirer lors des analyses d'association les SNP qui échappent à l'équilibre d'Hardy-Weinberg chez des témoins (excès d'homozygotes ou d'hétérozygotes) [4].

2.2.3 Fréquence allélique

Du fait des différences génétiques entre populations, des allèles détectés dans une population peuvent ne pas exister dans une autre ou y être trop peu fréquents pour avoir un effet détectable sur le risque de développement d'une pathologie. Garder les SNP rares lors des études d'association peut amener à une perte de puissance globale à cause de l'augmentation du nombre d'hypothèses testées (ce point sera évoqué plus en détail dans le chapitre suivant) et entraîner une augmentation du nombre de faux positifs dûs

4. Dans une population de malade, les SNP associé au risque peuvent s'écarter de l'équilibre d'Hardy-Weinberg si le modèle génétique sous-jacent s'écarte du modèle additif (ou modèle de codominance sticte).

à l'impact d'un écart à la normalité des résidus. On retire donc généralement les variants rares (MAF [5]>0.01 ou MAF>0.05) des analyses génome entier.

5. Minor Allele Frequency.

CHAPITRE 6

Problèmes liés à l'analyse des données de biopuces

Si les biopuces se sont révélées être un outil formidable pour la génétique, leur apparition a également soulevé de nombreux problèmes d'ordre statistique. Nous abordons dans cette partie les diverses difficultés liées à la dimension des données de biopuces et tentons d'apporter un éclairage sur les stratégies qui peuvent être utilisées pour y remédier.

1 Stratégies d'analyse des biopuces

Une caractéristique fondamentale des données de biopuces réside dans l'importance du nombre de variables observées (SNP ou transcrits). On observe en général plusieurs dizaines à plusieurs centaines de milliers de variables pour seulement quelques centaines à quelques milliers d'individus. En présence d'un tel nombre de variables, il n'est pas possible d'inclure l'intégralité des variables dans un modèle unique. Pour cette raison, l'approche la plus couramment utilisée consiste à tester, pour chaque SNP ou chaque transcrit, l'association univariée avec le phénotype d'intérêt. On sélectionne ensuite le sous-ensemble des SNP ou des transcrits les plus fortement associés pour des analyses plus approfondies.

Des approches alternatives ont également été développées afin de construire des modèles multivariés du phénotype basés sur la sélection d'un sous-ensemble de SNP ou d'expression. Ces approches reposent généralement sur :

- des contraintes de pénalisation imposées sur les coefficients du modèle (Lasso, Elastic Net, ...) ;
- des modèles bayésiens de sélection de variables (Bayesian Sparse Regression).

2 Problématique des tests multiples

2.1 Rappels sur les tests d'hypothèse

En statistique inférentielle classique on définit par test toute procédure qui vise à rejeter ou accepter une hypothèse statistique $\mathcal{H}_0$ (appelée "hypothèse nulle") à partir

de l'observation d'un échantillon de données. Dans les études d'association génétique, l'hypothèse nulle correspond le plus souvent à l'absence de lien entre un génotype et un phénotype. Le rejet de l'hypothèse nulle $\mathcal{H}_0$ conduit implicitement à l'acceptation de l'hypothèse opposée, nommée "hypothèse alternative" et notée $\mathcal{H}_a$.

En réalité, la décision prise n'est pas nécessairement symétrique. En effet, on distingue deux types d'erreurs :

- **Erreur de Type I :** L'hypothèse nulle est rejetée alors qu'elle est vraie. On a donc conclu à tort à un lien statistique entre les variables étudiées. C'est un "faux positif".
- **Erreur de Type II :** L'hypothèse nulle est acceptée alors qu'elle est fausse. On a donc rejeté à tort l'existence d'un lien entre les variables considérées. C'est un "faux négatif".

On note α et β les probabilités respectives de ces erreurs. La quantité $1-\beta$ représente alors la probabilité de rejeter l'hypothèse nulle à raison, c'est-à-dire de détecter un lien lorsqu'il existe. C'est la puissance du test. L'objectif des tests d'hypothèse est le plus souvent de maximiser la puissance (minimiser l'erreur de type II) tout en contrôlant l'erreur de type I à un niveau suffisamment faible (5% dans la plupart des cas).

Le principe général des procédures de test est donc le suivant :

1. On sélectionne une statistique T dont la loi sous $\mathcal{H}_0$, $F_{\mathcal{H}_0}$ est connue
2. On définit une région de rejet W_α telle que $F_{\mathcal{H}_0}(W_\alpha) = \alpha$, avec α l'erreur de type I souhaitée
3. On calcule la statistique T sur les observations X
4. On rejette $\mathcal{H}_0$ avec un risque α si la statistique appartient à la région de rejet W_α.

On note tout de suite qu'il est possible de construire une large gamme de régions de rejet pour une même valeur α. Cependant, la théorie de Neymann et Pearson [44] propose de choisir une famille de régions de rejet permettant de maximiser la puissance des tests. Pour un test d'hypothèse sur un paramètre θ d'hypothèse nulle $\mathcal{H}_0$: "$\theta = \theta_0''$ ces régions sont généralement [1] de la forme :

$$W_\alpha = \left\{ t \text{ tel que } |t - \theta_0| < q_{1-\alpha/2}(F_{\mathcal{H}_0}) \right\}$$

où $q_{1-\alpha/2}(F_{\mathcal{H}_0})$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi sous $\mathcal{H}_0$ de $\hat{\theta}$.

Une fois l'unicité de la famille de régions de rejet W_α assurée, on peut alors définir la valeur critique ou p-value comme :

$$p = \min \{\alpha | T(X) \in W_\alpha\}$$

1. En fait cette forme est valable dès lors que la loi sous $\mathcal{H}_0$ de $\hat{\theta}$, $F_{\mathcal{H}_0}$ est symétrique autour de la valeur θ_0.

C'est cette p-value qui permettra de rejeter (si $p \leq \alpha$) ou d'accepter (si $p > \alpha$) l'hypothèse $\mathcal{H}_0$.

2.2 Cas des tests multiples

Lorsqu'on teste n hypothèses indépendantes simultanément [2] le nombre de faux positifs augmente proportionnellement à n comme le montre le tableau 6.1. Si la proportion d'hypothèses non nulles parmi les hypothèses testées est faible, le nombre de faux positifs risque de dépasser largement le nombre de vrais positifs, réduisant considérablement l'efficacité de la procédure.

	Négatif	Positif	total
$\mathcal{H}_0$	VN= $n\pi_0(1-\alpha)$	FP= $n\pi_0\alpha$	$n\pi_0$
$\mathcal{H}_a$	FN= $n(1-\pi_0)\beta$	VP = $n(1-\pi_0)(1-\beta)$	$n(1-\pi_0)$
total	N =VN+FN	P = VP+FP	n

Tableau 6.1 – **Nombre de positifs et négatifs attendus pour n tests selon l'hypothèse vérifiée (nulle ou alternative) :** π_0 représente la proportion d'hypothèses nulles, α et β sont les erreurs de type I et II. VP, FP sont les nombres de vrais et faux positifs. VN et FN, les nombre de vrais et faux négatifs.

Afin de limiter cet effet, on considère des mesures globales de l'erreur de type I intégrant le nombre de tests réalisés, telles que le taux d'erreur global ("Family Wise Error Rate"-FWER en anglais) et le taux de faux positifs ("False Discovery Rate", FDR en anglais).

2.3 Méthodes de correction de l'erreur de type I

2.3.1 Taux d'erreur global (FWER)

Le taux d'erreur global associé à n tests statistiques est la probabilité sous l'hypothèse nulle globale (tous les tests sont sous $\mathcal{H}_0$) d'obtenir au moins un faux positif. Si on note p_i, $i = 1, \ldots, n$ les p-values associées aux n tests effectués, le FWER associé au seuil α s'écrit alors :

$$FWER(\alpha) = \mathbb{P}\left(\exists i | p_i < \alpha\right) = \mathbb{P}\left(\bigcup_i A_i\right)$$

où A_i est l'évènement $\{p_i < \alpha\}$.

2. De l'ordre de 10 000 ou 100 000 pour les données de transcriptomique et de génomique.

Bonferroni a montré qu'il était possible dans le cas général de trouver une borne supérieure pour le FWER en appliquant l'inégalité de Boole qui stipule que la probabilité d'une union d'évènements est toujours inférieure ou égale à la somme des probabilités des différents évènements pris séparément [45]. On obtient ainsi

$$FWER(\alpha) \leq \sum_i \mathbb{P}(A_i) = \sum_i \alpha = n\alpha$$

En exploitant ce résultat, on voit que pour contrôler le FWER à un niveau α il suffit d'imposer pour chaque test une erreur de type I α' égale à $\frac{\alpha}{n}$.

Dans le cas de tests indépendants, le FWER peut être calculé avec précision et est donné par

$$FWER(\alpha') = 1 - \mathbb{P}\left(\bigcap_i \overline{A_i}\right) = 1 - \prod_i \mathbb{P}\left(\overline{A_i}\right) = 1 - (1-\alpha')^n$$

Sidák a montré que le FWER pouvait être contrôlé efficacement dans le cas de n tests indépendants en utilisant comme seuil

$$\alpha' = 1 - (1-\alpha)^{\frac{1}{n}}$$

L'usage de cette correction plus fine apporte un léger gain de puissance [45]. Le principal défaut des méthodes de correction fondées sur le FWER tient à leur caractère très conservateur. En effet, la correction apportée par ces méthodes ne se fait qu'au prix d'une important perte de puissance. Cette perte est encore plus flagrante lorsque les tests ne sont pas indépendants. En effet, en présence de tests corrélés, ces méthodes de correction surestiment fortement le FWER, entraînant des pertes de puissance. Or, les données de biopuces sont justement caractérisées par de fortes corrélations du fait du déséquilibre de liaison entre SNP et des phénomènes de co-régulation de l'expression de gènes.

2.3.2 Taux de faux positifs (FDR)

Une approche alternative au contrôle de FWER a été proposée par Benjamini et Hochberg [46] avec l'introduction de la notion de taux de faux positifs. Cette notion correspond à un changement de paradigme, puisque l'objectif n'est plus de contrôler au sens strict la présence de faux positifs mais seulement la part des faux positifs parmi les résultats positifs obtenus. Le taux de faux positifs associé à n tests statistiques est l'espérance du ratio $\frac{FP}{FP+VP}$, c'est-à-dire

$$FDR(\alpha) = \mathbb{E}\left[\frac{FP}{FP+VP}\right]$$

Ainsi, contrôler le FDR revient à accepter un certain nombre de faux positifs tant que celui-ci reste faible devant le nombre de vraies découvertes. Alternativement le FDR peut être vu comme la probabilité qu'une hypothèse nulle rejetée le soit à tort.

$$FDR(\alpha) = \mathbb{P}\left(\mathcal{H}_0 \text{ vraie} | p_i \leq \alpha\right)$$

Dans ce cas, en appliquant la formule de Bayes, on obtient :

$$\begin{aligned} FDR(\alpha) &= \frac{\mathbb{P}(p_i<\alpha \cap \mathcal{H}_0)}{\mathbb{P}(p_i \leq \alpha)} \\ &= \frac{\mathbb{P}(p_i<\alpha|\mathcal{H}_0 \text{ vraie})\mathbb{P}(\mathcal{H}_0 \text{ vraie})}{\mathbb{P}(\mathcal{H}_0 \text{ vraie})\mathbb{P}(p_i \leq \alpha|\mathcal{H}_0 \text{ vraie})+\mathbb{P}(\mathcal{H}_0 \text{ fausse})\mathbb{P}(p_i \leq \alpha|\mathcal{H}_0 \text{ fausse})} \\ &= \frac{\alpha\pi_0}{\alpha\pi_0+(1-\beta)(1-\pi_0)} \end{aligned}$$

On voit donc que le taux de faux positifs est une fonction de l'erreur de type I mais également de la puissance correspondant au seuil choisi. Ainsi le FDR peut amener à sélectionner un seuil de significativité α moins strict si la puissance obtenue en retour permet de maintenir un faible taux de faux positifs. L'estimation du FDR se fait par le calcul des q-values [47]. On a

$$FDR(\alpha) = \frac{\alpha\pi_0}{\mathbb{P}\left(p_i \leq \alpha\right)}$$

Si on réordonne les p-values par ordre croissant (en notant $p_{(i)}$ la i^{e} plus petite p-value), et qu'on garde les k premiers résultats, on obtient l'estimateur suivant du FDR :

$$FDR(p_{(k)}) = \frac{p_{(k)}\pi_0}{\mathbb{P}\left(p_i \leq p_{(k)}\right)} = \frac{p_{(k)}\pi_0}{k/n}$$

Si le nombre d'associations vraies est petit par rapport au nombre de tests effectués, π_0 est proche de 1. En prenant $\pi_0 = 1$ et en imposant au FDR d'être inférieur à 1, on est capable d'estimer le FDR quel que soit le nombre k de résultats retenus. Pour contrôler le FDR à un niveau α, il suffit alors de garder le plus grand nombre d'associations possible en s'assurant que le FDR correspondant reste inférieur au niveau α choisi. En pratique, on construit pour cela des q-values par :

$$q_{(k)} = \min\left(q_{(k+1)}, \min\left(\frac{np_{(k)}\pi_0}{k}, 1\right)\right)$$

Ces q-values peuvent alors s'utiliser de façon similaire aux p-values. En rejetant toutes les hypothèses telles que $q_i \leq \alpha$ on contrôle efficacement le FDR au niveau α. De plus, Benjamini et Hochberg ont montré [46] que cette procédure garantissait également le contrôle du FWER au niveau α. Les propriétés de contrôle du FDR de cette prodédure sont conservées en présence de corrélations [3].

3. Les corrélations entraînent toutefois une forte augmentation de la variabilité des estimations du FDR.

Le FDR a fait l'objet d'une littérature très riche et de nombreuses extensions ont été développées. On peut notamment citer :

- Les extensions de FDR local [48, 49] qui cherchent à estimer la probabilité locale d'être sous $\mathcal{H}_0$ lorsque p_i est autour d'une valeur λ fixée plutôt que de décrire la probabilité dans l'intervalle $[0, \lambda]$.
- Les approches d'estimation de la proportion π_0 d'hypothèses nulles [50]. La procédure standard contrôlant le FDR à un niveau $\alpha\pi_0$, l'estimation préalable de π_0 et son intégration dans les procédures de contrôle du FDR permet un gain de puissance dans les cas où π_0 est fortement inférieur de 1.
- Les extensions du FDR fondées sur l'estimation de la distribution empirique sous $\mathcal{H}_0$ permettant de réduire les effets des corrélations sur l'estimation du FDR et d'augmenter la puissance [51, 52].

En 2008, un estimateur du FDR unifiant ces différents points de vue a été proposé par Strimmer *et al.* [53]

3 Réduction de la dimension des données

Du fait de la taille des données générées par les biopuces, il est parfois préférable de retirer a priori les transcrits ou les SNP les moins pertinents dans les analyses. Cette sélection joue un double rôle :

- **Réduction des temps de calcul :** Les temps de calcul sont souvent un problème majeur du traitement des données de biopuces. A plus forte raison lorsque le nombre d'individus est très important comme c'est le cas dans des études telles que GHS ou Cardiogenics (voir section 2 du chapitre 2).
- **Réduction du nombre de tests effectués :** Et donc de la correction effectuée pour les tests multiples, permettant d'augmenter la puissance des tests d'association.

Pour qu'une telle sélection fonctionne correctement, il est cependant primordial que le critère utilisé pour sélectionner les données soit informatif sur la propension des transcrits ou des SNP à être sous l'hypothèse nulle.

3.1 Sélection des transcrits

Lors de l'utilisation de puces à ARN, il est d'usage de retirer les transcrits dont le niveau d'expression est considéré comme n'étant pas significativement différent du bruit de fond. Dans ce but, on utilise le plus souvent les p-values de détection. Ces p-values sont calculées automatiquement lors des prétraitements de la puce à partir des mesures d'expression obtenues sur les contrôles négatifs. Pour chaque transcrit dont l'expression est mesurée par une sonde, et pour chaque individu, on calcule le pourcentage de contrôles

négatifs dont le niveau est supérieur au niveau de la sonde. Ce pourcentage compris entre 0 et 1 est assimilable à une p-value associée à l'hypothèse

$$\mathcal{H}_0 : \mu = \mu_b$$

où μ représente le niveau du transcrit et μ_b le niveau du bruit de fond.

Pour chaque transcrit j et chaque individu i, on obtient une p-value de détection p_{ij}. Un critère du type $p_{ij} < \alpha$ dans une proportion γ de la population permet ensuite de retirer les transcrits considérés comme non exprimés. Bien qu'il n'y ait pas de règle stricte pour le choix des paramètres α et γ, il est d'usage de prendre $\alpha = 0.01$ ou 0.05. Le choix de γ est plus problématique. La plupart des études de micro-arrays anciennes sont basées sur des valeurs élevées de γ afin de ne travailler que sur des transcrits détectés dans une large majorité d'échantillons. Toutefois, un transcrit dont le niveau d'expression varie en fonction d'un facteur (par exemple un SNP) peut n'être détectable que dans une fraction limitée de la population et l'écarter a priori peut s'avérer contre-productif.

Ce filtrage sur les p-values de détection repose sur l'idée communément admise que dans une cellule, seule la moitié environ des gènes sont exprimés. Des études récentes de grand séquençage suggèrent que cette idée est partiellement erronée [54]. La conception binaire de l'expression des transcrits dans un type cellulaire peut donc constituer une source de confusion puisque la non-détection d'un grand nombre de transcrits peut vraisemblablement être attribuée au manque de sensibilité des mesures d'expression par micro-array plutôt qu'à une absence d'expression.

De plus, même lorsque le niveau d'un transcrit est en dessous du niveau moyen du bruit de fond, il arrive que la sonde capture un signal biologique pertinent comme le montre l'exemple du gène ZFY. Pour ce gène, situé sur le chromosome Y et donc exprimé uniquement chez les hommes, deux sondes sont présentes sur la puce Illumina utilisée dans GHS. Seule la sonde ILMN_2090059 donne un signal s'écartant significativement du bruit de fond. Le signal renvoyé par la sonde ILMN_1804958 est considéré comme ne différant pas significativement du bruit de fond chez plus de 95% des individus, sans doute à cause d'une hybridation moins efficace. Un observateur ne considérant que cette deuxième sonde conclurait donc à l'absence d'expression du gène ZFY dans le monocyte. Néanmoins, pour cette sonde, comme pour la sonde ILMN_2090059, on observe une différence très forte entre hommes et femmes (figure 6.1). Ceci suggére que la mesure obtenue avec la sonde ILMN_1804958, même atténuée reflète bien un signal biologique pertinent. Il s'agit ici d'un problèm classique de ration signal/bruit.

Ce point a été abordé dans un article [4] traitant de la comparaison des niveaux d'expression moyens entre le chromosome X (où seule une copie du gène s'exprime) et les autosomes (pour lesquels les deux copies du gènes s'expriment). Dans cette analyse, nous

4. Ecrit en collaboration avec Raphaële Castagné et disponible en annexe.

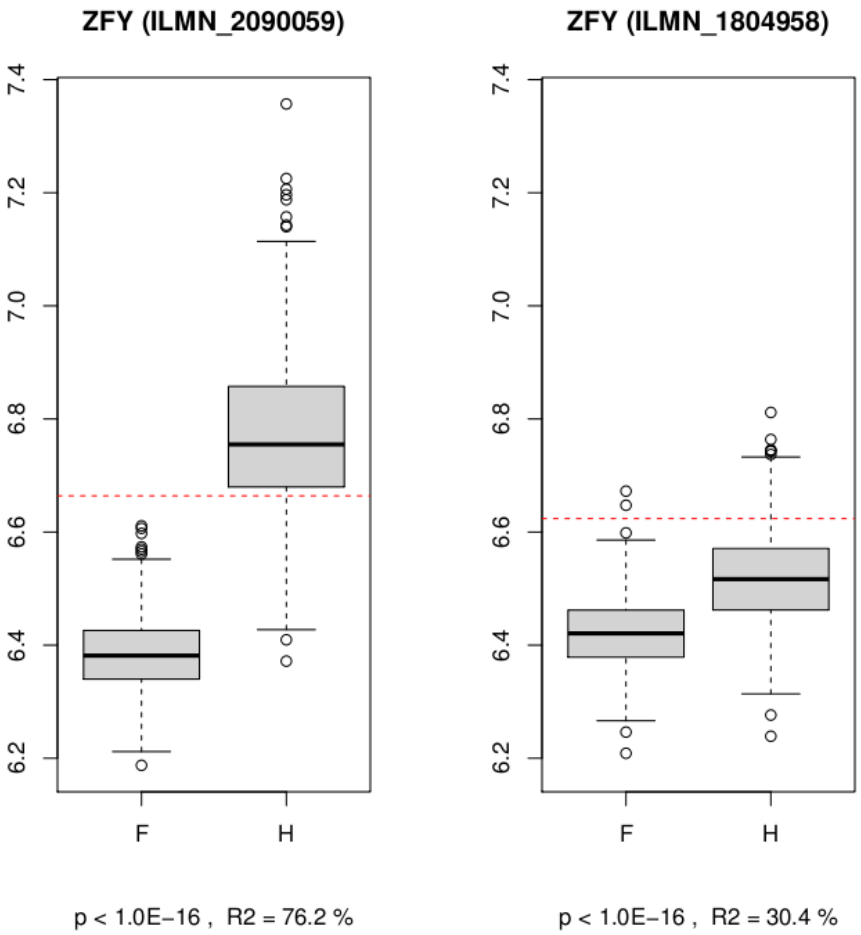


FIGURE 6.1 – **Expression des sondes du gène ZFY en fonction du sexe :** Les boxplots gris donnent la répartition des niveaux d'expression observés pour la sonde ILMN_2090059 (à gauche) et la sonde ILMN_1804958 (à droite) séparément chez les femmes (F) et chez les hommes (H). La droite horizontale en pointillés donne le niveau du 99^{e}percentile de la distribution du bruit fond qui correspond au seuil généralement utilisé pour déclarer un gène exprimé. Pour chaque sonde, on reporte en dessous du graphique la p-value d'association de la sonde avec le sexe et la part de variance de l'expression mesurée par la sonde qui est attribuable au sexe.

avons montré que le filtrage a priori des gènes non détectés et considérés à tort comme non exprimés, pouvaient largement biaiser les résultats et amenait à conclure à tort à la prséence de phénomènes de compensation des niveaux d'expression entre le chromosome X et les autosomes.

Une alternative au filtrage sur la détection consiste à sélectionner les transcrits en fonction de leur variabilité afin de limiter les analyses aux transcrits les plus variables et donc les plus susceptibles de présenter une expression différentielle entre les conditions étudiées. Une telle sélection peut cependant s'avérer délétère si les sources de variabilité non souhaitées sont importantes par rapport aux sources de variabilité étudiées.

3.2 Sélection des SNP

Si dans les analyses génome entier un filtrage des données est moins courant, on peut tout de même noter que le filtrage fait sur la fréquence allélique lors du contrôle qualité a déjà pour fonction d'éviter les tests inutiles pour lesquels la puissance serait de toute façon trop faible.

Bien que rarement réalisé à cause des difficultés de mise en oeuvre, un filtrage basé sur le déséquilibre de liaison [5] entre les SNP a pour effet de limiter la redondance inutile et de diminuer le nombre de tests.

Enfin des stratégies d'étude centrées sur les SNP ayant potentiellement un rôle fonctionnel (liés à l'expression ou codants) ont été proposées [55, 56].

4 Biologie en grande dimension

Outre les problèmes statistiques déjà évoqués, la dimension des donnés de biopuces pose des problèmes d'interprétation. En effet, dans un contexte où des variations, même modérées, des facteurs environnementaux peuvent engendrer des variations des niveaux d'expression de plusieurs centaines voire milliers de transcrits, il n'est généralement pas aisé de donner une interprétation aux résultats. Afin d'y parvenir il est généralement nécessaire de recourir à des analyses d'enrichissement. Ces analyses se fondent sur des bases de données comme KEGG (Kyoto Encyclopédia of Genes and Genomes [57]) ou GO (Gene Ontology [58]) détaillant les caractéristiques des gènes telles que

- La fonction remplie par la protéine (marqueur membranaire, facteur de transcription, protéine de structure, ...)
- La localisation de la protéine dans la cellule (membrane, noyau, cytoplasme,...)

5. Le déséquilibre de liaison désigne la corrélation existant au sein d'une population entre des marqueurs génétiques qui co-ségrègent.

- Les domaines actifs de la protéine (ex : domaine de liaison à l'ADN, d'interaction avec d'autres protéines,...)
- Les processus biologiques dans lesquels la protéine est impliquée (réponse immunitaire, production d'énergie , maintien de la structure cellulaire, coagulation,...)
- Les types cellulaires où la protéine est présente
- Les gènes cibles dans le cas des facteurs de transcription

Ces analyses d'enrichissement reposent sur l'identification de caractéristiques surreprésentées dans une liste de gènes obtenue par une analyse du transcriptome ou une analyse génome entier. Pour ce faire, on considère l'ensemble des gènes inclus dans l'analyse (par exemple : l'ensemble des gènes dont l'expression est mesurée par la puce, ou l'ensemble des gènes pour lesquels une expression est détectée). On teste ensuite pour chaque caractéristique l'hypothèse nulle d'indépendance entre la présence de la caractéristique et l'appartenance à la liste des gènes étudiés.

Plus formellement, parmi l'ensemble des gènes étudiés, si on note G l'ensemble des gènes significatifs à l'issue de l'analyse et F_j la liste des gènes présentant une caractéristique j, on souhaite tester :

$$\mathcal{H}_0 : \mathbb{P}\left(g \in F_j | g \in G\right) = \mathbb{P}\left(g \in F_j | g \notin G\right)$$

Pour ce faire il est possible d'utiliser directement un test d'indépendance du χ^2 entre G et F_j, ou dans le cas où le nombre d'éléments dans G et/ou F_j est faible, un test exact de Fischer [6].

En appliquant une correction pour les tests multiples sur le nombre de fonctions biologiques testées (généralement de l'ordre de plusieurs milliers) on peut alors identifier les fonctions sur- ou sous-représentées parmi la liste des gènes d'intérêt.

6. Qui modélise le nombre de gènes appartenant simultanément aux deux groupes par une loi hypergéométrique.

Troisième partie

Analyse de la variabilité du transcriptome et intégration en génétique humaine

CHAPITRE 7

Modulation du transcriptome par la variabilité génétique

Une partie de mon travail de thèse a été le prétraitement des données d'expression de GHS ainsi que l'étude des liens existants entre transcriptome et variants génétiques. Dans cette partie, nous ferons tout d'abord quelques rappels sur les mécanismes biologiques de régulation de l'expression des gènes, puis nous verrons les différentes stratégies d'études possibles de la variabilité génétique du transcriptome avant d'évoquer les enseignements apportés par l'étude GHS sur le transcriptome.

1 Mécanismes de régulation du transcriptome

Afin d'étudier les facteurs influençant la variabilité du transcriptome, il est important de s'attarder sur le fonctionnement de la régulation de l'expression génique dans les cellules.

1.1 Les facteurs de transcription

Comme nous l'avons vu précédemment, le transcriptome des cellules n'est pas figé mais évolue au cours de la différenciation ou en réponse à des stimuli biologiques et environnementaux. Cette capacité d'adaptation est permise par un système de régulation reposant sur la présence de protéines appelées facteurs de transcription. Les facteurs de transcription se caractérisent par leur capacité à se lier à l'ADN pour permettre la transcription des gènes. On compte sur le génome humain plus de 2600 protéines présentant un site de fixation à l'ADN. On pense aujourd'hui que la majeure partie de ces protéines jouent un rôle dans la transcription. Parmi les facteurs de transcription, on distingue :

- *les facteurs généraux de la transcription* nécessaires au processus de transcription dans son ensemble et exprimés dans toutes les cellules eucaryotes. On trouve notamment parmi ces facteurs les sous-unités qui composent l'ARN polymérase qui synthétise l'ARN selon la séquence originale d'ADN.

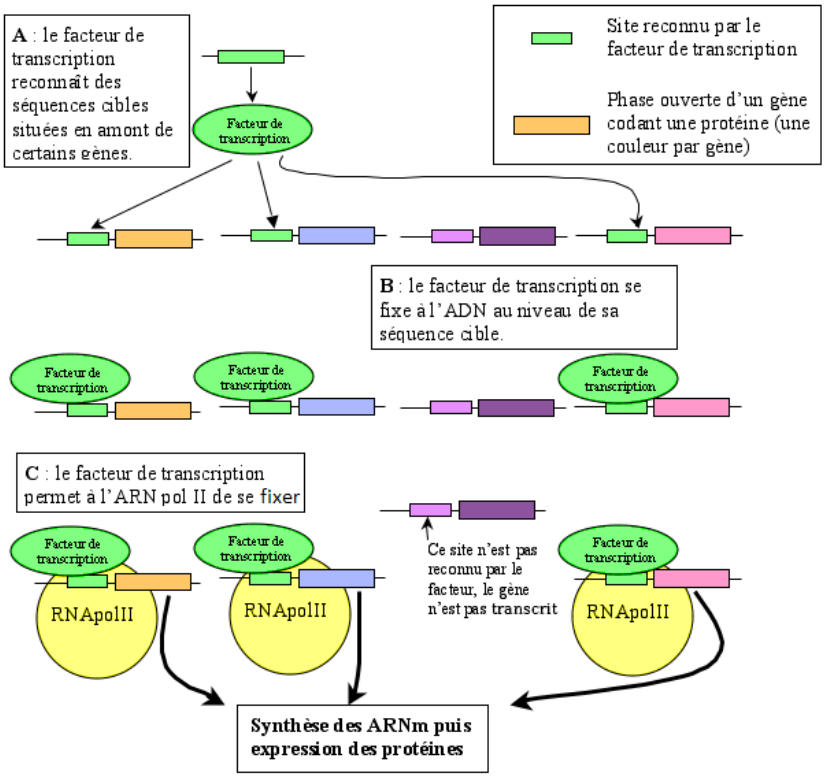


FIGURE 7.1 – **Régulation de l'expression des gènes par un facteur de transcription spécifique.**

– *les facteurs de transcription spécifiques* qui régulent l'expression des gènes en se fixant à des régions dites promotrices situées le plus souvent en amont des gènes (figure 7.1). Ces facteurs spécifiques ne ciblent qu'un nombre restreint de gènes agissant le plus souvent de manière coordonnée pour répondre aux stimuli extérieurs. Leur rôle se cantonne à un rôle d'activation ou de répression de la transcription. La transcription d'un même gène peut être activée par plusieurs facteurs spécifiques.

L'activation des facteurs de transcription par des voies de signalisation permet ainsi d'adapter [1] la transcription en fonction des stimuli perçus par la cellule.

Aux facteurs de transcription s'ajoutent le plus souvent des co-activateurs ou co-répresseurs de la transcription. Ces protéines interagissent avec des facteurs de transcription pour en augmenter ou diminuer l'effet. L'expression d'un gène est donc le fruit des interactions entre les différents facteurs de transcription pouvant se fixer à son promoteur et des co-activateurs de ces facteurs de transcription. De plus, l'expression des facteurs de transcription et de leurs co-activateurs varie également d'un type cellulaire à l'autre

1. La transcription et la maturation des protéines étant des phénomènes relativement lents à l'échelle de la cellule, l'adaptation permise par le transcriptome est une adaptation à long terme, plutôt qu'une réaction immédiate aux stimuli extérieurs.

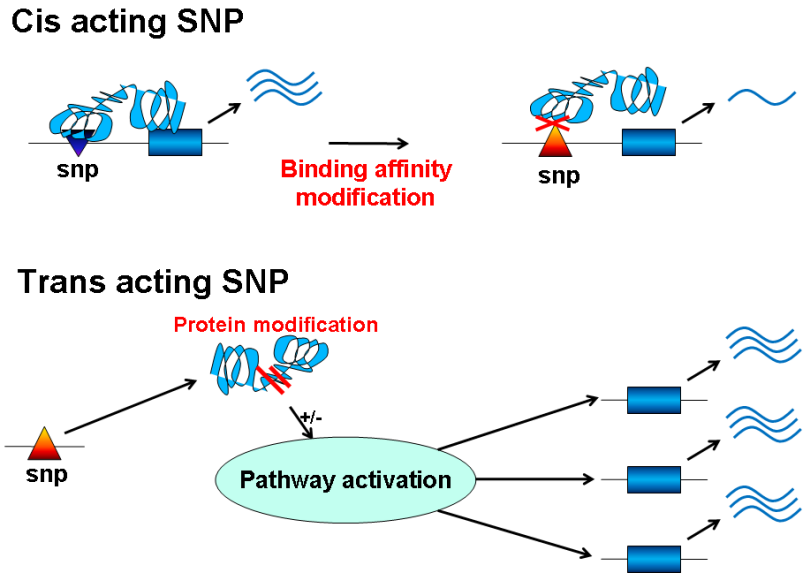


FIGURE 7.2 – **Mécanismes de modulation du transcriptome par les SNP.**

formant ainsi un réseau de régulation extrêmement complexe.

1.2 Le rôle de la variabilité génétique

L'impact des polymorphismes sur l'expression des gènes peut se produire par deux types de mécanismes (figure 7.2) :

- Des mécanismes en *cis* où un polymorphisme se situe à proximité du gène (le plus souvent au niveau du site promoteur) et affecte directement la liaison des facteurs de transcription au gène. Dans ce cas le SNP se situe sur le même chromosome que le gène dont il régule l'expression à une distance relativement modérée du gène.
- Des mécanismes en *trans* où le polymorphisme affecte indirectement l'expression d'un gène par l'intermédiaire d'un élément régulateur (facteur de transcription, co-activateur de la transcription, microARN, ...). Dans ce cas, le SNP responsable de l'association peut se situer n'importe où sur le génome.

Chaque facteur de transcription ou activateur/répresseur de la transcription pouvant avoir plusieurs cibles, il n'est pas rare que les SNP ayant des effets *trans* affectent plusieurs gènes [59, 60]. On parle dans ce cas de modules de gènes *trans*-régulés.

Les loci modulant l'expression d'un gène sont appelés eQTL (expression Quantitative Trait Locus). Par abus de langage, ce terme est parfois utilisé également pour les gènes dont l'expression est modulée par un SNP. On parle alors d'eSNP pour le SNP affectant l'expression.

2 Etude des eQTL

2.1 Stratégies d'étude

On voit se dessiner derrière cette classification plusieurs stratégies d'étude :

1. La recherche de polymorphismes agissant sur l'expression des gènes situés en *cis* (*cis*-eQTL). Ces *cis*-eQTL représenteront le plus souvent des effets sur l'expression via des modifications des sites de liaison des facteurs de transcription ou en favorisant l'épissage alternatif des transcrits.
2. La recherche de polymorphismes agissant en *trans* sur l'expression. Cette recherche est plus complexe pour plusieurs raisons :
   - Les effets *trans* sont généralement des effets indirects sur l'expression, susceptibles d'être beaucoup plus faibles et plus variables que les effets *cis*.
   - Le nombre de *trans*-eQTL possibles est beaucoup plus grand, ce qui a pour effet d'imposer des seuils de significativité très stricts pour limiter les faux positifs et ainsi de diminuer la puissance des tests d'association.
   - La grande variété des mécanismes possibles en *trans* rend très difficile l'interprétation des résultats et leur réplication dans des jeux de données indépendants.
3. La recherche de modules de gènes *trans*-régulés par un même locus (cette dernière stratégie sera abordée séparément dans le chapitre suivant).

Le graphique 7.3 représente les eQTL de l'étude GHS significatifs à un seuil arbitraire de 10^{-6} (bien en dessous du seuil de significativité théorique de Bonferroni qui serait de 10^{-12}). On voit sur ce graphique une majorité de points situés sur la diagonale qui correspondent aux associations en *cis* (stratégie 1). Les points en dehors de la diagonale correspondent aux associations en *trans* (stratégie 2). Enfin certaines bandes verticales peuvent apparaitre et marquent des modules de gènes *trans*-régulés (stratégie 3).

2.2 Choix de la distance *cis*/*trans*

Si l'on définit comme *cis* les eQTL pour lesquels le SNP se situe dans le voisinage du gène, le choix de la définition de la notion de voisinage peut avoir une large influence sur les résultats. Veyreiras *et al.* ont montré dans une étude basée sur les données du projet HapMap [61] que les pics d'associations des eQTL avaient une probabilité inférieure à 5% d'être situé à plus de 20 kb des sites de début ou de fin de la transcription (figure 7.4a). Cependant il arrive que le SNP causal ne soit pas sur la puce. Dans ce cas, du fait du déséquilibre de liaison entre les marqueurs génétiques proches, l'eQTL peut tout de même être détecté, mais le pic d'association peut se trouver assez loin du SNP causal. Il est donc nécessaire de définir le voisinage du gène de façon plus large pour pouvoir espérer capturer

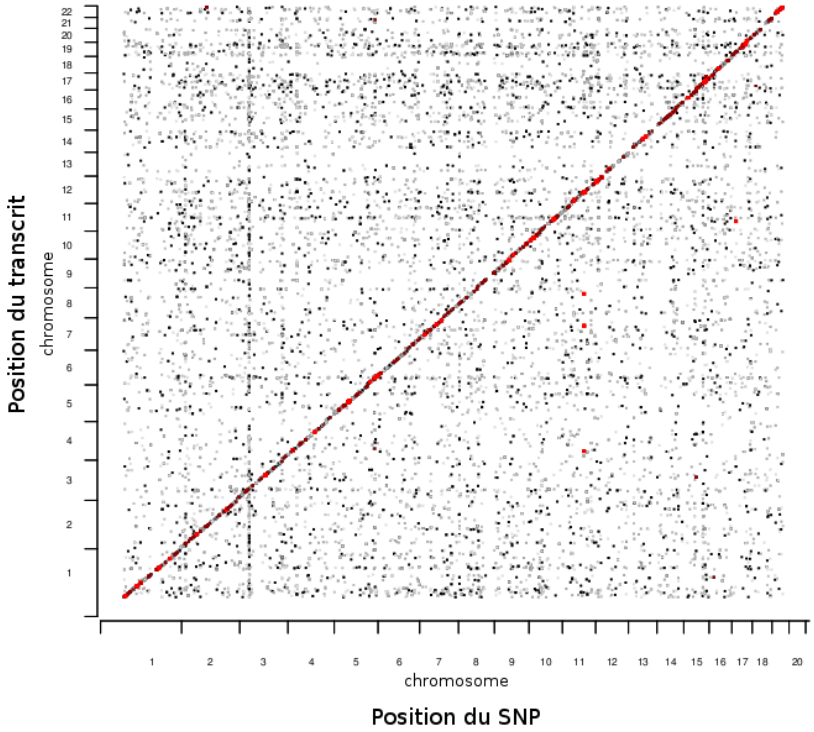
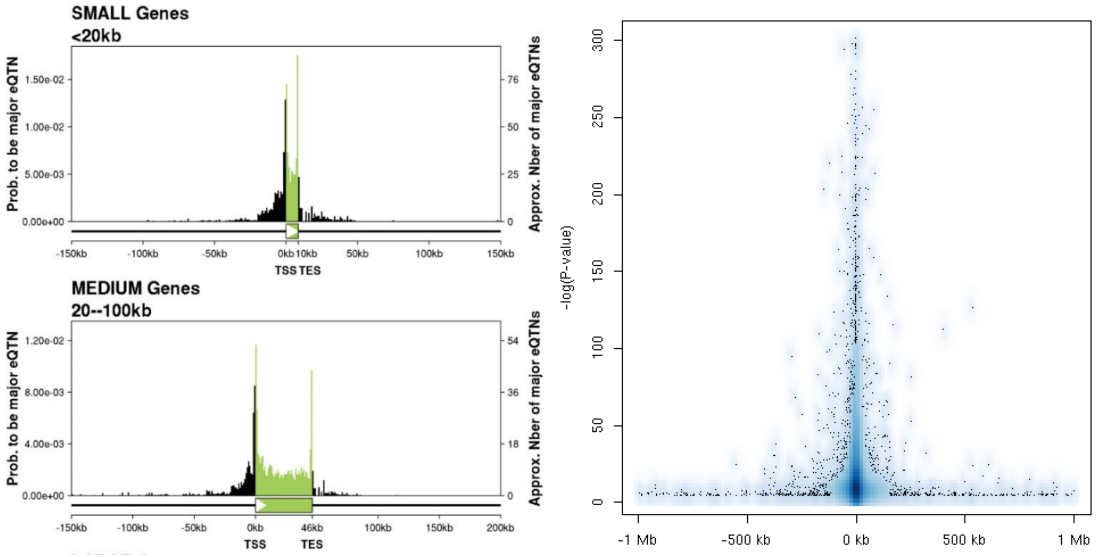


FIGURE 7.3 – **Répartition chromosomique des eQTL observés dans GHS :** Chaque point représente un couple SNP-transcrit. La couleur reflète la force de l'association (gris pâle à gris noir : $p = 10^{-8}$ à $p < 10^{-12}$, rouge sombre : $p < 10^{-12}$ et $R^2 > 20\%$, rouge clair $R^2 > 50\%$).

la majorité des associations en *cis*. Dans le même temps, étudier un voisinage trop grand augmente le nombre d'hypothèses testées et diminue donc la puissance pour détecter des associations en *cis*. Afin de déterminer la distance la plus appropriée pour définir les *cis*-eQTL, nous avons étudié la localisation des pics d'association à chaque transcrit dans GHS (figure 7.4b). Il apparaît clairement que la majorité des pics d'association sont situés dans le voisinage immédiat du gène. Plus de 99% des pics sont situés à moins de 250 kb du début ou de la fin du gène. Une distance de 250kb permet donc d'assurer une puissance importante pour la recherche de *cis*-eQTL tout en capturant la majorité des phénomènes de *cis*-régulation. Dans la suite, on utilisera également un seuil de 1Mb pour la définition des *cis*-eQTL par souci de cohérence avec les études précédentes sur le sujet. De plus, dans certaines régions de fort déséquilibre de liaison telles que la région HLA du chromosome 6, on observe parfois des pics d'association à des distances allant jusqu'à 5Mb du gène (non représentés sur la figure 7.4b).



(a) SNP causal. *Source :* HapMap [61]

(b) Pic d'association(GHS)

FIGURE 7.4 – **Distance du SNP au gène dans les eQTL :** Position par rapport au gène du SNP responsable de l'association dans HapMap (à gauche) et du pic d'association dans GHS (SNP présentant la meilleure p-value.)

2.3 Analyses univariées pour la recherche d'eQTL

Sur l'étude GHS, après les différents prétraitements et filtrages des SNP et des données d'expression [2], on obtient plus de 14 000 sondes correspondant à plus de 11 000 gènes pour

2. C'est-à-dire qu'on ne garde que les SNP à l'équilibre de Hardy Weinberg, avec un taux de détermination de plus de 98%, et une fréquence allélique supérieure à 1%. De plus on sélectionne uniquement les

près de 675 000 SNP. Soit plus de 9 milliards de tests, dont 6 millions environ correspondent à des SNP situés en *cis* du gène associé. Pour chaque expression, on teste l'association avec chaque SNP par une analyse de variance simple.

Sur les données d'expression, la normalité des résidus, nécessaire pour l'analyse de variance, n'est pas toujours vérifiée. On est donc confronté, à distance finie, à une augmentation de l'erreur de type I. Cette augmentation est surtout apparente sur les SNP dont la fréquence de l'allèle mineur est faible [3]. Pour pallier cette augmentation de l'erreur de type I, une solution serait de recourir de façon systématique à des tests non paramétriques tels que le test de Kruskal-Wallis. Dans notre cas, cette solution n'était pas envisageable pour des raisons computationnelles. On se contente donc de vérifier a posteriori, lorsqu'un test est déclaré significatif par ANOVA, que cette association est également significative par un test de Kruskal-Wallis, afin de retirer l'excédent de faux positifs induit par les écarts à l'hypothèse de normalité des résidus.

Du fait de cette approche en deux étapes et à cause de l'impossibilité physique de stocker l'intégralité des p-values résultant des analyses de variance, il n'est pas possible dans les études d'eQTL de contrôler avec exactitude le taux de faux positifs. On utilise donc ici une correction de Bonferroni. On estime également pour le seuil α de significativité utilisé, le taux de faux positifs attendu par le FDR :

$$\widehat{FDR} = \frac{\mathbb{E}\left[FP\right]}{FP + VP} = \min\left(\frac{\widehat{n\alpha}}{\#\{p_i < \alpha\}}, 1\right)$$

où VP et FP désignent les nombres de vrais et de faux positifs identifiés, n le nombre total d'hypothèses testées et les p_i désignent les p-values des analyses de variance.

3 Enseignements sur la régulation du transcriptome

Le tableau 7.1 résume les résultats des analyses d'eQTL en *cis* et en *trans* effectuées sur les données de GHS. Ces résultats ont été déposés dans la base de données GHS_Express qui a été mise à disposition de la communauté scientifique [4]. Ces résultats ont également fait l'objet d'un article publié dans Plos One [6].

sondes bien annotées, ne contenant pas de SNP et dont le niveau est considéré comme significativement supérieur au bruit de fond.

3. C'est seulement lorsque qu'au moins un des génotypes est rare que l'on s'éloigne des conditions nécessaires à l'application des théorèmes de normalité asymptotique garantissant le contrôle de l'erreur de type I.

4. Accessible sur le site `www.genecanvas.org` .

Seuil de significativité	$5.0\,10^{-5}$	10^{-5}	10^{-7}	10^{-9}	10^{-12} (Bonferroni)
nombre de gènes *cis*-régulés (< 250 kb)	4175	3827	3170	2707	2225
(FDR estimé)	0.02	0.004	$4.9\,10^{-5}$	$5.7\,10^{-7}$	$6.9\,10^{-10}$
nombre de gènes *cis*-régulés (< 1Mb)	4303	3890	3195	2730	2238
(FDR estimé)	0.07	0.02	0.0002	$2.2\,10^{-6}$	$2.7\,10^{-9}$
nombre de gènes *trans*-régulés (> 1Mb)	11166	11079	1345	476	340
(FDR estimé)	1	1	0.69	0.02	$2.0\,10^{-5}$
nombre total de gènes régulés génétiquement	11166	11108	4100	3068	2501
(FDR estimé)	1	1	0.23	0.003	$3.9\,10^{-6}$

Tableau 7.1 – ***Cis* et *trans*-eQTL dans l'étude GHS selon le seuil de significativité choisi :** les analyses ont été faites sur les 14329 sondes codant pour 11109 gènes distincts. L'expression d'un gène est considérée comme génétiquement régulée dès lors que l'expression d'au moins une de ses sondes est trouvée associée à un SNP.

On peut faire plusieurs constats à partir de ce tableau.

- Tout d'abord, on observe que plus d'un tiers des gènes exprimés sont *cis*-régulés dans les monocytes. Cette proportion est supérieure aux résultats des études précédentes du transcriptome [1, 62–64]. Avec plus de 4000 eQTL identifiés, GHS est l'étude qui recense le plus grand nombre d'eQTL. Ce résultat s'explique par la taille d'échantillon utilisée pour l'étude GHS. En effet avec plus de 1400 sujets, la puissance pour la recherche d'eQTL atteint 80% pour détecter un eQTL expliquant 3% de l'expression du gène au seuil de Bonferroni de 10^{-12}.
- La très forte augmentation du nombre de *trans* observée pour les seuils de significativité les plus lâches correspond à une augmentation très forte du nombre de faux positifs liées au nombre de tests effectués.
- Ensuite, on voit que pour un même seuil de détection ($p < 10^{-9}$), on a 5 à 6 fois plus de gènes régulés en *cis* qu'en *trans*. Ce chiffre, en accord avec les résultats des études précédentes, montre la rareté des mécanismes *trans* ayant un impact fort sur les expressions. Cette rareté peut s'expliquer par le fait que les associations en *trans* impliquent des mécanismes indirects et sont généralement d'une amplitude plus faible.
- Enfin on observe que la stratégie de recherche de *cis*-eQTL consistant à limiter la zone de recherche à un intervalle de 250kb autour du gène permet bien un léger gain de puissance. Pour un FDR égal à 2%, on obtient 4175 gènes *cis*-régulés en recherchant dans une fenêtre resserrée de 250kb contre seulement 3890 avec la fenêtre de 1Mb.

La figure 7.5 montre que les eQTL expliquent le plus souvent une part modérée de la variabilité des transcrits. Toutefois on trouve une centaine de loci pour lesquels le SNP

explique plus de 50% de la variabilité de l'expression du gène.

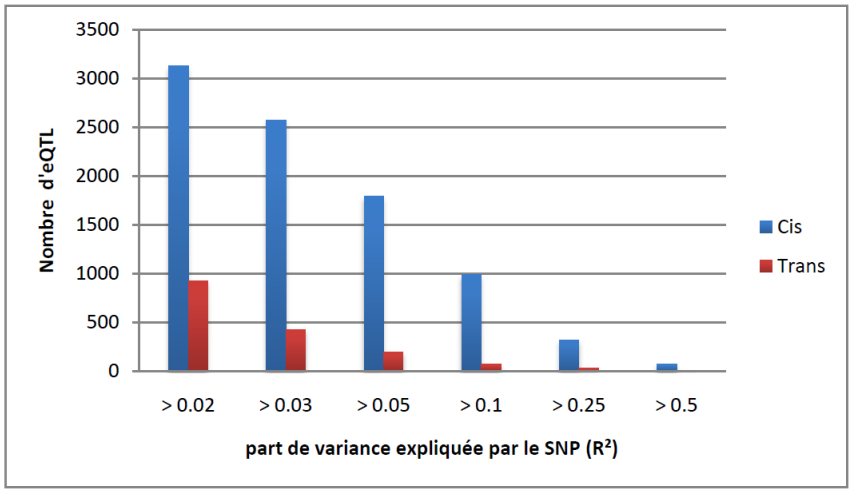


FIGURE 7.5 – **Evolution du nombre d'eQTL en fonction de la part de variance de l'expression expliquée par le SNP**

L'étude dans GHS de la fonction des gènes régulés en *cis* ou en *trans* montre que les facteurs de transcription et les gènes impliqués dans la régulation de l'expression sont nettement sous-représentés parmi les eQTL($p < 10^{-15}$). Ce résultat peut s'expliquer par le fait que les facteurs de transcription sont des protéines jouant un rôle clé dans la survie cellulaire. Des modifications même légères de l'équilibre maintenu entre les facteurs de transcription sont donc à même de mettre en danger la cellule. Aussi les pressions de sélection s'exercent et tendent à maintenir constante l'expression de ces facteurs de transcription [5].

Il a été proposé que la régulation de l'expression se fasse de façon tissu-spécifique [4,66]. Nous avons testé cette hypothèse en estimant la réplicabilité dans GHS des résultats provenant d'autres études basées sur des tissus différents (lignées de cellules hépatiques et de lymphoblastoïdes). Les résultats de cette analyse sont résumés dans le tableau 7.2.

On voit que plus de la moitié des eQTL sont réplicables entre tissus différents à condition de disposer d'une puissance suffisante. Le pourcentage d'eQTL répliqués d'un tissu à l'autre atteint près de 70% pour les associations les plus fortes. Ce résultat montre que les mécanismes de régulation génétique sont fortement partagés entre différents types cellulaires. Toutefois, l'absence de réplication de près de 30% des eQTL parmi les associations les plus fortes indique l'existence de phénomènes de régulations spécifiques à certains types cellulaires.

Le cas du locus CELSR2/SORT1/PSRC1 sur le chromosome 1p13 constitue un exemple

5. Une telle observation est tout a fait cohérente avec l'observation d'une très forte conservation de l'expression des gènes dans les tissus au cours de l'évolution [65].

	Stranger *et al.*	(LCL [a])	Dixon *et al.*	(LCL[a])	Schadt *et al.*	(Foie)
Significativité	nombre d'eQTL	Pourcentage répliqués dans GHS	nombre d'eQTL	Pourcentage répliqués dans GHS	nombre d'eQTL	Pourcentage répliqués dans GHS
$> 10^{-10}$	149	61,7	272	50,4	1096	49,4
$10^{-10} - 10^{-20}$	204	65,2	339	56,6	331	60,7
$< 10^{-20}$	86	79,1	162	66,6	176	71,0
Tous	439	66,7	773	56,5	1603	54,1

a. Lymphoblastoïd Cell Lines

Tableau 7.2 – **Réplicabilité inter-tissus des eQTL.**

de régulation tissu-spécifique. Ce locus été a trouvé associé aux LDL-cholestérol dans plusieurs GWAS [67, 68]. A ce locus se trouve un cluster de 3 gènes parmi lesquels aucun n'était précédemment connu comme influençant les lipides. D'après les données de GHS, PSRC1 était *cis*-régulé dans le monocyte, alors que SORT1 et CELSR2 ne l'étaient pas, faisant de PSRC1 un candidat attractif pour expliquer l'association du locus avec les lipides. Très récemment, Musunuru *et al.* ont montré à partir de modèles cellulaires et animaux que le gène responsable de l'association était SORT1 et que le mécanisme d'action sur les lipides passait par une régulation en *cis* de SORT1 dans le foie [69]. Cet eQTL est lié à la présence d'un SNP dans le site de liaison du facteur de transcription C/EBP exprimé spécifiquement dans le foie. C'est pourquoi il n'était pas observé dans les monocytes de GHS. Cet exemple illustre donc l'importance que peut revêtir le choix du tissu étudié pour l'intégration des eQTL en génétique humaine.

CHAPITRE 8

Apport des eQTL dans les études d'association pan-génomiques

Un des intérêts majeurs des eQTL tient dans leur apport pour l'étude des maladies complexes. Il a en effet été proposé [70] que l'héritabilité de l'expression pourrait expliquer une part importante de l'héritabilité des maladies complexes en agissant comme un médiateur de l'association entre la variabilité génétique et la maladie. Cette hypothèse a été confortée par des études récentes [71, 72] montrant une co-localisation fréquente au sein du génome entre les eQTL et les loci de prédisposition aux maladies complexes. Il a ainsi été suggéré que l'étude des eQTL pouvait permettre de fournir une interprétation biologique aux associations trouvées dans les GWAS [69, 73].

Les études précédentes [71,72] de la co-localisation entre eQTL et loci de GWAS ayant été conduites à partir de lignées cellulaires lymphoblastoïdes [1] et des eQTL d'HapMap, nous avons jugé pertinent de répliquer ces résultats dans les monocytes. En effet, il a été montré que l'utilisation du Virus Epstein-Barr pour la constitution de lignées cellulaires lymphoblastoïdes était susceptible d'induire l'expression des gènes, modifiant l'activité de certains pathways. Par contraste avec ces données *in vitro*, les données d'expression *in vivo* de GHS reflètent une situation plus proche de la réalité. De plus, le monocyte joue un rôle clé dans les processus inflammatoires et immunitaires impliqués dans de nombreuses maladies dont les pathologies cardiovasculaires. Enfin la forte puissance de l'étude GHS pour la recherche d'eQTL comparativement aux études basées sur HapMap laissait espérer la découverte de nouveaux eQTL pertinents pour l'interprétation des résultats de GWAS. Les résultats de cette analyse ont fait l'objet d'un article actuellement soumis à l'European Journal of Human Genetics.

1. Lymphocytes B "immortalisés" par l'ajout du virus Eptsein Barr (EBV).

1 Association entre eQTL et loci de prédisposition identifiés par les GWAS

Pour tester l'hypothèse d'une sur-représentation des eQTL parmi les loci de prédisposition identifiés par les GWAS , nous avons croisé les résultats de l'étude des eQTL présentée dans la partie précédente avec la base de données du National Human Genome Research Institute (NHGRI). Cette base de données catalogue les résultats de GWAS provenant de plus de 800 publications différentes et recense plus de 3900 associations issues de GWAS portant sur une grande variété de traits complexes.

1.1 Méthode d'analyse

Pour analyser ces données, nous avons dans un premier temps extrait les données du catalogue du NHGRI, et identifié les SNP rapportés comme associés à un trait complexe et se trouvant sur la puce Affymetrix 6.0 utilisée dans GHS, ou pour lesquels il était possible d'identifier un proxy sur la puce. Un proxy est défini comme un SNP présentant un r^2 supérieur à 0.8 avec le SNP ciblé. Ces SNP ont ensuite été regroupés au sein de 1 712 loci distincts. Pour cela, on a défini un locus comme un ensemble de SNP associés à un même trait, situés sur le même chromosome et séparés de moins de 500kb les uns des autres. A chaque locus, le SNP présentant la plus forte p-value d'association (ou son proxy sur la puce) a été retenu. On aboutit ainsi à une sélection de 1 597 SNP de la puce Affymetrix 6.0 considérés comme des marqueurs trouvés associés par GWAS à au moins un trait complexe. Afin de tester si les eQTL et les loci de GWAS co-localisent plus souvent que ne le voudrait le hasard, on modélise le lien entre les deux par un modèle logistique :

$$\log \frac{P}{1-P} = \alpha + X\beta$$

où P est la probabilité qu'un SNP soit un marqueur trouvé associé à un moins un trait complexe dans une GWAS, et X est la variable caractérisant l'implication du SNP dans un eQTL. On modélise ainsi la probabilité qu'un SNP tiré au hasard parmi les 675 000 SNP de la puce Affymetrix ayant passé les différentes étapes du contrôle qualité, soit un marqueur de prédisposition génétique pour au moins un trait complexe en fonction de l'association en *cis* de ce SNP avec au moins un transcrit. Les modèle peut ensuite être testé pour plusieurs définitions de la notion de *cis*-eQTL et en ajustant sur d'éventuels facteurs confondants tels que la fréquence du SNP.

La qualité du modèle obtenu est déterminée par le pseudo R^2 de McKelvey et Zavoina défini comme

$$R^2 = \frac{\mathbb{V}[y*]}{\mathbb{V}[y*] + \mathbb{V}[e]}$$

où $y*$ correspond à la variable latente sous-jacente du modèle logistique donnée par

$$\widehat{y*} = X\hat{\beta}$$

et $\mathbb{V}[e]$ est l'erreur résiduelle du modèle qui correspond à la variance d'une distribution logistique standard donnée par

$$\mathbb{V}[e] = \frac{\pi^2}{3}$$

.

1.2 Co-localisation des eQTL et des loci de GWAS

Les résultats de l'analyse indiquent un net enrichissement des SNP de GWAS en SNP significativement liés à l'expression en *cis* (OR = 2,327, IC = [2,030 - 2,669] , $p = 1,19\,10^{-33}$). Cet enrichissement augmente à mesure qu'on se restreint aux SNP pour lesquels l'association avec l'expression est la plus forte (tableau 8.1).

Afin d'écarter les co-localisation dues au déséquilibre de liaison dans la région, nous avons défini un score décrivant la force relative d'association avec l'expression (score RSAE en anglais). Ce score, calculé pour chaque couple SNP-transcrit, reflète le niveau de déséquilibre de liaison entre le SNP d'intérêt et le SNP responsable du pic d'association avec l'expression. Dans ce but, on calcule le ratio entre la part de variance du transcrit expliquée par le SNP d'intérêt et la part de variance maximale du transcrit expliquée par le meilleur SNP au locus [2]. Le tableau 8.1 montre une nette augmentation du lien entre eQTL et loci de prédisposition lorsque le score RSAE augmente. Ce résultat étaye l'hypothèse selon laquelle une part importante des mécanismes causaux sous-jacents aux loci de prédisposition trouvés dans les GWAS impliquent des phénomènes de régulation de l'expression en *cis*. L'ajustement sur la fréquence allélique, ou le retrait de la région HLA [3] des analyses ne change pas les résultats présentés ici.

Comme nous l'avons déjà évoqué, les variations d'expression d'un tissu à l'autre engendrent une spécificité par tissu des eQTL. Bien que la majorité des eQTL puissent être observés dans de multiples tissus, l'intensité de la modulation de l'expression peut fortement varier d'un tissu à l'autre. Ainsi il a été proposé que des eQTL ayant un impact sur les maladies complexes avaient une plus forte chance d'être observés dans des tissus impliqués dans le développement de la maladie (cf. exemple de l'eQTL de SORT1 observé uniquement dans le foie). On s'attend ainsi à ce que les monocytes soient plus appropriés

2. On a donc par construction un score strictement supérieur à 0 et inférieur ou égal à 1, atteignant 1 lorsque le pic d'association trouvé dans la GWAS coïncide parfaitement avec le pic d'association de l'eQTL.

3. Qui contient un très fort déséquilibre de liaison et un fort enrichissement en gènes liés aux maladies auto-immunes.

	significativité	OR [b.inf - b.sup]	p-value	R^2
Estimation brute				
	10^{-5}	2,327 [2,030 - 2,669]	$1,19\,10^{33}$	1,4%
	10^{-7}	2,495 [2,152 - 2,892]	$7,00\,10^{34}$	1,3%
	10^{-9}	2,708 [2,317 - 3,165]	$5,79\,10^{36}$	1,3%
	10^{-12}	2,984 [2,525 - 3,527]	$1,33\,10^{37}$	1,2%
Score RSAE $\geq$ 0.8				
	10^{-5}	2,868 [2,327 - 3,534]	$5,32\,10^{23}$	0,7%
	10^{-7}	2,963 [2,370 - 3,703]	$1,35\,10^{21}$	0,6%
	10^{-9}	3,105 [2,457 - 3,924]	$2,46\,10^{21}$	0,6%
	10^{-12}	3,559 [2,798 - 4,526]	$4,51\,10^{25}$	0,6%
Score RSAE = 1				
	10^{-5}	4,082 [2,902 - 5,743]	$6,67\,10^{16}$	0,3%
	10^{-7}	4,162 [2,878 - 6,021]	$3,64\,10^{14}$	0,3%
	10^{-9}	4,547 [3,102 - 6,665]	$8,37\,10^{15}$	0,3%
	10^{-12}	5,519 [3,764 - 8,094]	$2,22\,10^{18}$	0,3%

Tableau 8.1 – **Association des eQTL avec les loci de prédisposition aux traits complexes en fonction du score RSAE :** Odds ratio (OR) avec son intervalle de confiance, p-value associée, et pseudo R^2 du modèle logistique pour différents choix du seuil de significativité utilisé pour définir les eQTL et différentes valeurs du score RSAE.

pour l'étude des maladies auto-immunes ou infectieuses que pour des maladies neurologiques et psychiatriques. Cette hypothèse se trouve confirmée par les résultats obtenus lorsqu'on regroupe les phénotypes en grandes catégories de pathologies (figure 8.1).

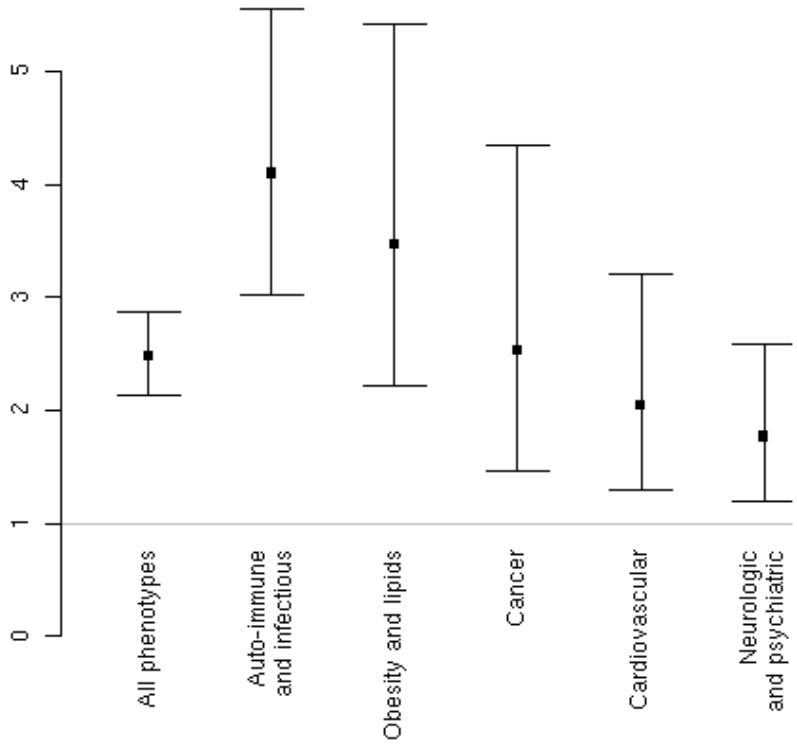


FIGURE 8.1 – **Lien entre eQTL et loci de GWAS selon le type de trait étudié :** Pour chaque type de trait on donne l'odds ratio brut obtenu dans modèle non ajusté avec son intervalle de confiance à 95%

1.3 Effet de la densité de gènes sur la co-localisation entre eQTL et loci de GWAS

Au cours de cette analyse nous nous sommes intéressés à la répartition des co-localisations entre eQTL et loci de GWAS sur le génome. Nous avons observé que ces phénomènes de co-localisation étaient répartis inégalement sur le génome. Outre une nette sous-représentation des loci de GWAS sur les chromosomes sexuels, attribuable à un biais d'analyse [4], de fortes différences sont observées entre les régions autosomiques. Comme

4. Les loci de prédisposition sont sous-représentés sur les chromosomes sexuels du fait de leur exclusion a priori de la plupart des analyses génome entier.

le montre la figure 8.2 les co-localisations surviennent préférentiellement dans les régions particulièrement denses en gènes. Ceci peut s'expliquer par une plus forte propension de ces régions à contenir à la fois des loci de GWAS et des eQTL.

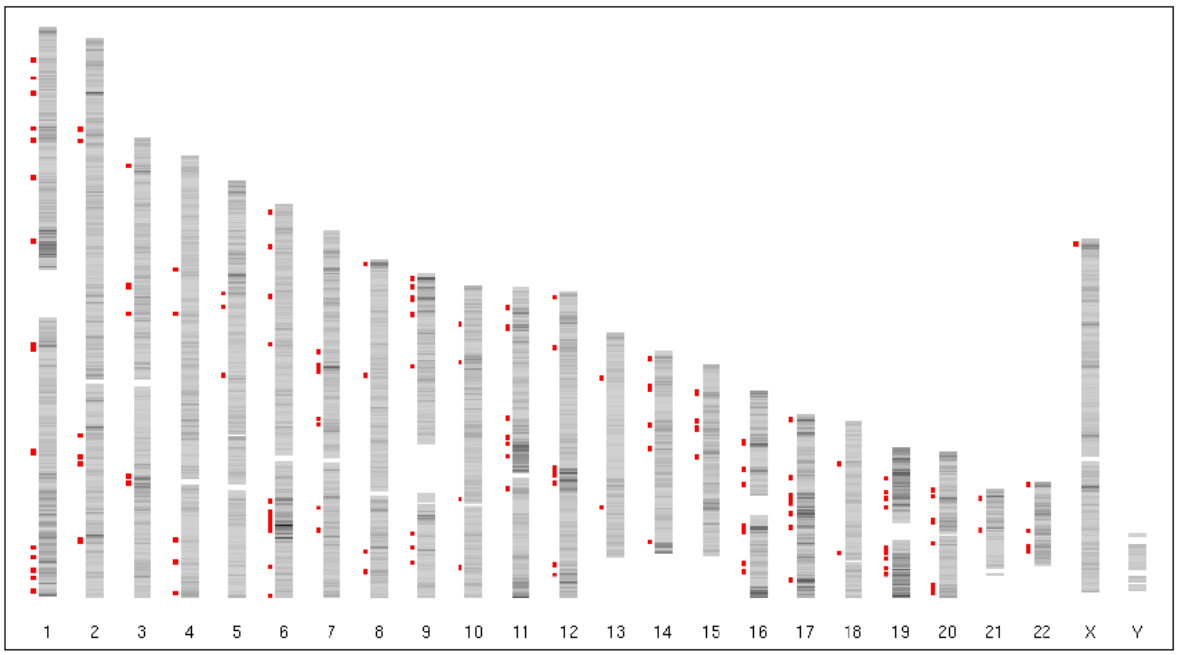


FIGURE 8.2 – **Localisation des loci de GWAS présentant un eQTL :** Chaque point rouge représente un locus de prédisposition trouvé dans une GWAS pour lequel on détecte une association significative avec au moins une sonde en *cis*. Les bandes de couleur figurant sur les chromosomes représentent la densité en gène autour de chaque SNP. Une bande noire indique une région dense en gènes ($\approx$ 40 gènes dans une fenêtre de 250 kb autour du SNP). Une bande gris clair indique un désert de gènes (pas de gène dans une fenêtre de 250 kb autour du SNP).

Pour tenir compte de ce facteur potentiellement confondant, nous avons calculé à partir des bases de données RefSeq, la densité de gènes présents dans une région de 250 kb autour de chaque SNP. Nous avons ensuite intégré la densité de gènes comme une covariable dans le modèle de la probabilité P qu'un SNP soit un marqueur de prédisposition d'une maladie. La densité de gènes est effectivement un facteur prédictif de la probabilité pour un SNP d'être identifé dans une GWAS (OR=1.083, IC = [1,071 - 1,095], $p = 4,4\,10^{-43}$ par gène présent dans la région). Les résultats des analyses ajustées sur la densité de gènes apparaissent dans le tableau 8.2. Bien que l'enrichissement des locus de GWAS en eQTL reste significatif après ajustement sur la densité de gènes, il se trouve presque divisé par deux. Par ailleurs, l'étude des pseudo-R^2 montre que la densité de gènes apporte à elle seule plus d'information que l'inclusion des eQTL dans le modèle ($R^2 = 2.6\%$ contre

	significativité	OR [b.inf - b.sup]	p-value	R^2
Estimation brute				
	10^{-5}	1,424 [1,222 - 1,658]	$5,55\,10^{-6}$	0,1%
	10^{-7}	1,506 [1,28 - 1,771]	$7,90\,10^{-7}$	0,1%
	10^{-9}	1,626 [1,372 - 1,927]	$1,99\,10^{-8}$	0,1%
	10^{-12}	1,782 [1,488 - 2,133]	$3,15\,10^{-10}$	0,1%
Score RSAE $\geq$ 0.8				
	10^{-5}	1,621 [1,300 - 2,022]	$1,83\,10^{-5}$	$<$ 0,1 %
	10^{-7}	1,675 [1,326 - 2,116]	$1,50\,10^{-5}$	$<$ 0,1 %
	10^{-9}	1,743 [1,365 - 2,225]	$8,39\,10^{-6}$	$<$ 0,1 %
	10^{-12}	1,999 [1,556 - 2,568]	$6,03\,10^{-8}$	$<$ 0,1 %
Score RSAE = 1				
	10^{-5}	2,102 [1,478 - 2,989]	$3,52\,10^{-5}$	$<$ 0,1 %
	10^{-7}	2,146 [1,469 - 3,133]	$7,73\,10^{-5}$	$<$ 0,1 %
	10^{-9}	2,346 [1,586 - 3,47]	$1,96\,10^{-5}$	$<$ 0,1 %
	10^{-12}	2,840 [1,919 - 4,203]	$1,81\,10^{-7}$	$<$ 0,1 %

Tableau 8.2 – **Association des eQTL avec les loci de prédisposition aux traits complexes, ajustée sur la densité de gènes et en fonction du score RSAE :** Odds ratio (OR) du modèle ajusté sur la densité de gène, avec son intervalle de confiance et la p-value associée pour différents choix du seuil de significativité utilisé pour définir les eQTL et différentes valeurs du score RSAE. Le pseudo R^2 affiché correspond à la différence entre le pseudo R^2 du modèle complet et le pseudo R^2 du modèle incluant seulement la densité de gènes.

1.3% pour les eQTL) et que la majeure partie de l'information apportée par les eQTL est capturée par la densité de gènes.

2 Quelques exemples de co-localisation

Dans cette section, nous illustrons les résultats précédents par quelques exemples de co-localisation entre des eQTL et loci de GWAS :

Parmi les 3956 SNP associés à l'expression d'au moins un transcrit en *cis*, 101 présentaient un score RSAE supérieur à 0.8 (et 35 présentant un score égal à 1). Parmi ces loci la densité en gènes était de 10 en moyenne, contre 2 en moyenne pour l'ensemble des SNP illustrant bien l'effet de cette variable. Un premier exemple concerne le locus 7q22, trouvé associé au nombre de globules rouges dans une GWAS [74]. A ce locus, on observe une

très forte *cis*-régulation du gène GIGYF1 co-localisant parfaitement avec le signal d'association avec le phénotype. Malgré cela, le meilleur candidat dans cette région demeure le gène EPO, qui synthétise l'hormone de l'érythropoïétine bien connue pour son effet sur la production des globules rouges. Cet exemple, bien que trivial, illustre le risque potentiel d'une sur-interprétation de la présence d'eQTL à un locus où le mécanisme biologique de l'association ne serait encore connu. Il montre également la nécessité de recourir à des modèles statistiques permettant d'écarter de telles co-localisations fortuites en exploitant la connaissance de la structure du déséquilibre de liaison dans la région. Dans ce but, un test formel de co-localisation a été développé par Plagnol *et al.* [75] et un article sur l'application de ce test à l'étude dans GHS, des eQTL liés au diabète de type I est en préparation. Dans le cas du diabète de type I, l'application de ce test sur les eQTL de GHS permet d'écarter la moitié des co-localisations apparentes (38 sur 70).

Inversement, on trouve parmi les co-localisations des cas où l'eQTL a une pertinence biologique pour expliquer l'implication du locus dans le trait complexe. Par exemple, sur le chromosome 11, le SNP rs4752856 associé à une diminution de l'indice de masse corporelle (IMC) [76] est également associé à l'expression du gène C1QTNF4. Ce gène situé à près d'1 Mb du SNP initial n'avait pas été rapporté dans la GWAS initiale et le signal d'association avait été attribué au gène MTCH2 dans lequel est situé le SNP. Le gène C1QTNF4 code pour un paralogue [5] de l'adiponectine, une hormone impliquée dans la régulation du glucose et le métabolisme des acides gras. Cette hormone est associée à une diminution du pourcentage de masse grasse chez l'adulte et est sous-exprimée chez les sujets obèses [77]. Le gène C1QTNF4 apparaît donc ici comme un candidat plausible pour expliquer l'association constatée du locus avec l'IMC.

De même, la présence d'eQTL régulant les niveaux d'expression des gènes LCAT (lécithine-cholestérol-acyl-transférase) et LPL (lipoprotéine lipase), deux gènes impliqués dans le métabolisme des lipides, à deux loci associés aux niveaux de cholestérol sanguins [68] suggère que la régulation génétique de ces gènes joue un rôle déterminant dans le métabolisme du cholestérol et la susceptibilité aux maladies cardiovasculaires.

3 Utilisation de la densité des gènes pour faciliter la recherche de loci de prédisposition dans les GWAS

Nous avons vu dans les sections précédentes que si l'information apportée par les eQTL pouvait être utile dans la recherche par GWAS de loci de prédisposition aux traits

5. Un paralogue est un gène dont la séquence présente de très fortes similarités avec la séquence d'une autre gène connu. On s'attend généralement à trouver des similarités fonctionnelles entre gènes paralogues.

complexes, la densité de gène était un critère plus prédictif de la localisation de ces loci. Une conséquence logique de ce résultat était donc de tenter d'utiliser cette information a priori dans l'analyse des GWAS.

Dans ce but nous avons utilisé une méthode de FDR pondéré proposée par Genovese [78] permettant de pondérér les p-values en fonction d'une information a priori (ici, la densité de gènes). Nous détaillons le principe de cette méthode dans la section suivante, avant de présenter les résultats obtenus par cette méthode sur une GWAS des taux plasmatiques d'homocysteine réalisée sur les individus de GHS.

3.1 Pondération dans les GWAS

Une façon simple d'intégrer des informations extérieures dans la recherche de signaux d'association dans les GWAS consiste à recourir à des méthodes de pondération des procédures de tests multiples. Nous utilisons pour cela la méthode de FDR pondéré proposée par Genovese [78] dont le principe est le suivant :

Etant donné une variable de pondération $w_i > 0$ observée pour chacune des hypothèses testées :

1. Standardiser w_i de façon à avoir $\mathbb{E}[w_i] = 1$
2. Construire les p-values pondérées $p'_i = \min\left(\frac{p_i}{w_i}, 1\right)$
3. Construire les q-values à partir de ces p-values pondérées
$$q_{(k)} = \min\left(q_{(k+1)}, \min\left(\frac{np'_{(k)}\pi_0}{k}, 1\right)\right)$$
4. On contrôle le FDR au seuil α en déclarant significatives les hypothèses dont la q-value est inférieure à α

Genovese a montré que cette procédure permettait de contrôler efficacement le FDR au seuil de 5% et pouvait engendrer un gain de puissance lorsque les valeurs de w_i sont corrélées positivement à la probabilité d'être sous l'hypothèse alternative, tout en assurant une perte limitée de puissance lorsque ce n'est pas le cas.

Dans notre cas, on utilise comme variable de pondération w_i la densité de gènes au locus considéré, en ajoutant 1 au nombre de gènes pour obtenir des poids strictements positifs. Cette variable est par construction liée à la probabilité qu'un SNP appartienne à un locus de prédisposition et est donc appropriée pour la pondération des résultats de GWAS.

3.2 Application à une GWAS de l'homocystéine plasmatique

L'homocystéine qui est un acide aminé ayant un effet pro-inflammatoire et jouant un rôle dans le développement de l'athérosclérose. Les niveaux sanguins de l'homocystéine

sont en partie déterminés génétiquement et le facteur génétique le plus connu pour influencer ces niveaux est le polymorphisme non synonyme C677T (rs1801133) situé dans l'exon 5 du gène MTHFR (5,10-méthylènetétrahydrofolate réductase) qui code pour une enzyme participant au métabolisme de l'homocystéine.

Dans GHS, les niveaux d'homocystéine plasmatique ont été mesurés chez 3306 individus et une analyse d'association génome entier a été conduite pour identifier les loci associés à l'homocystéine. Tandis qu'une analyse classique avec un FDR à 5% ne détecte pas d'association significative, on trouve en intégrant la densité de gènes dans le FDR pondéré deux loci dépassant le seuil de significativité. Parmi ces deux loci, le premier correspond au locus du gène MTHFR. Le SNP causal rs1801133 n'était pas présent sur la puce Affymetrix utilisée et le meilleur proxy présentait un r^2 de 0.36 avec ce SNP, ce qui explique que ce locus n'ait pas été détecté dans la première analyse. Bien que certains des SNP à ce locus soient associés à l'expression du gène MTHFR les scores RSAE associés à ces SNP ne dépassaient pas 0.4 suggérant une dissociation entre les variants influençant l'expression et ceux affectant les taux plasmatiques d'homocystéine. Le deuxième locus associé à l'homocystéine est situé sur le chromosome 6q21. Cette région contient 34 gènes, parmi lesquels se trouve un cluster de 4 gènes appartenant à la famille des co-transporteurs SLC17 (solute carrier 17). L'un de ces gènes, SLC17A2, a été trouvé associé à l'homocystéine dans une étude gène candidat précédente [79] bien que le mécanisme sous-jacent ne soit pas encore identifié. Cet exemple illustre le gain réalisable par l'introduction de connaissance à priori dans les GWAS. L'intérêt d'utiliser la densité de gènes est qu'il s'agit d'un critère facile à obtenir. En revanche, la pondération par ce critère défavorise les régions pauvres en gènes qui peuvent malgré tout jouer un rôle dans la prédisposition aux maladies complexes.

CHAPITRE 9

Identification de modules de gènes *trans*-régulés

Ainsi que nous l'avons évoqué précédemment, l'étude des eQTL montre de façon consistante [1,6,63,64] une large sur-représentation des *cis*-eQTL comparativement aux *trans*-eQTL. Ce déséquilibre peut s'expliquer par le fait que les *trans*-eQTL correspondent le plus souvent à des effets faibles du fait de leur caractère indirect. L'identification de loci affectant en *trans* l'expression de modules de gènes co-régulés (par exemple de gènes cibles d'un même facteur de transcription) est cependant fondamentale pour permettre de mieux comprendre les mécanismes impliqués dans les processus pathophysiologiques des maladies complexes. Plusieurs études ont mis en évidence de tels modules de gènes co-régulés chez la levure [80,81], la drosophile [82], la souris [83,84] et l'homme [85]. Le but du travail présenté ici est d'identifier des loci contrôlant des modules de gène co-régulés.

Dans ce but, j'ai étudié l'application des méthodes de classification et des méthodes à facteurs à la recherche de modules de gènes co-régulés et j'ai développé une approche visant à identifier des SNP contrôlant l'expression de tels modules. Nous présentons tout d'abord cette approche et les résultats obtenus dans GHS. Nous comparerons ensuite les résultats de cette approche, aux résultats fournis par une approche fondée sur la reconstruction de réseaux de régulation et l'identification de modules de gènes fortement inter-connectés. Nous discuterons enfin d'une application de ces approches qui a permis d'identifier un nouveau locus associé au diabète de type I, dans le cadre d'une collaboration européenne.

1 Principe de l'approche factorielle

1.1 Les méthodes factorielles

L'objectif des méthodes factorielles que nous utilisons dans cette partie est d'identifier des variables cachées qu'on appellera composantes et qui représentent des processus latents influençant l'expression des gènes. Nous nous focalisons ici sur la méthode de l'analyse en composantes indépendantes (ACI) que j'ai appliquée à l'extraction de modules de gènes co-régulés.

Dans cette méthode, l'expression de chaque gène est écrite comme une combinaison

linéaire de composantes, qui influencent l'expression des gènes de façon indépendante [1]. Chaque composante peut être caractérisée par un un sous-ensemble de transcrits co-régulés appelé module. Chacune de ces composantes est supposée refléter une source de variabilité biologique (ou expérimentale) telle que

- l'activation de voies de signalisation
- l'effet d'un facteur de transcription
- des phénomènes de régulation post-transcriptionnels

Si on note X la matrice d'expression de taille $p \times n$ où chaque ligne représente un transcrit dont l'expression a été centrée et réduite [2], et chaque colonne représente un échantillon le modèle de décomposition de l'ACI pour K composantes revient à une factorisation matricielle dans laquelle on approxime X par une matrice de rang K de la forme

$$X \approx S.A$$

ce qui peut encore s'écrire pour un transcrit i et un individu j :

$$x_{ij} = \sum_{k=1}^{K} s_{ik} a_{kj}$$

avec

- S, la matrice $p \times K$ dont chaque élément s_{ik} donne la contribution (éventuellement nulle) de la composante k à l'expression du transcrit i. Les colonnes de cette matrice définissent les **signatures** caractéristiques des différentes composantes influençant l'expression.
- A, la matrice $K \times n$ dont chaque élement a_{kj} correspond au "degré d' activation" du processus reflété par la composante k dans l'échantillon j. Les lignes de cette matrice définissent des motifs d'expression ou "**patterns**" dont le niveau caractérisent la composante au sein de la population.

Pour des raisons d'identifiabilité du modèle on impose, sans perdre de généralité, que les signatures extraites vérifient

$$\forall k, \frac{1}{p} \sum_{i} s_{ik}^2 = 1 \,.$$

La part de variabilité expliquée par la composante k peut alors être estimée par la variance de la k^{e} ligne de la matrice A.

Cette décomposition laisse donc apparaître une dualité des composantes extraites qui peuvent être représentées par

1. C'est-à-dire que le fait qu'un processus affecte l'expression d'un gène est indépendant de l'influence éventuelle des autres processus sur le gène.

2. Une telle transformation sert à éviter l'apparition d'une composante dont la signature reflèterait uniquement les différences de niveaux d'expression entre transcrits et permet d'extraire des facteurs latents reflétant la structure de corrélation plutôt que la covariance entre les transcrits.

- Une "**signature**" traduisant la contribution de la composante aux niveaux d'expression des différents transcrits. Dans l'analyse en composantes indépendantes, ces signatures sont estimées en cherchant des combinaisons linéaires des échantillons maximisant un critère d'indépendance basé sur l'information mutuelle (cf. section 1.2.2).
- Un "**pattern**" reflétant le niveau d'activation du processus biologique sous-jacent chez les individus. Ces patterns sont généralement estimés par des combinaisons linéaires des gènes en inversant [3] la matrice des signatures S dans la formule $X \approx S.A$. En ACI, les patterns extraits peuvent être corrélés entre eux. Ceci permettant à l'ACI de mieux capturer des processus biologiques sous-jacents que ne le feraient les méthodes classiques de réduction de dimension telles que l'ACP.

Le principe de la méthode est illustré pour $K = 2$ sur la figure 9.1.

Dans la suite, on utilisera pour désigner les composantes, les termes de "pattern" ou de "signature", selon qu'on se réfère à une des lignes de la matrice A ou à une des colonne de la matrice S.

Une fois les expressions caractérisées par un nombre restreint de composantes, des "**modules**" de gènes sont définis à partir des signatures (cf. section 1.3). Les patterns extraits et les modules sont ensuite étudiés en relation avec le génotype pour extraire des modules enrichis en gènes associés au génotype (cf. section 1.4).

1.2 Extraction des composantes

Nous détaillons dans cette partie le fonctionnement de l'ACI. Cette méthode étant généralement précédée d'une étape de réduction de la dimension par ACP, nous rappelons brièvement le principe de l'ACP avant d'exposer le principe de l'ACI.

1.2.1 L'analyse en composantes principales

L'analyse en composantes principales (ACP) est sans doute la méthode d'analyse factorielle la plus répandue. L'ACP fonctionne en cherchant successivement des composantes de variance maximale non corrélées entre elles, permettant de décrire au mieux les observations. Ainsi, si on visualise les transcrits comme des points dans un espace de dimension n, on peut voir l'ACP comme la recherche d'un petit nombre d'axes orthogonaux entre eux tels que la projection sur ces axes capture une part la plus grande possible de la variance du nuage des transcrits. La recherche de la k^{e} composante principale revient à chercher un vecteur s_k unitaire et orthogonal aux $k-1$ composantes précédentes, tel que la projection des données sur ce vecteur $s_k' X$ soit de variance maximale. Ce qui revient au programme de maximisation suivant :

3. Ou plus exactement en utilisant la pseudo-inverse de la matrice S, et en estimant $\widehat{A} = (S'S)^{-1}S'X$.

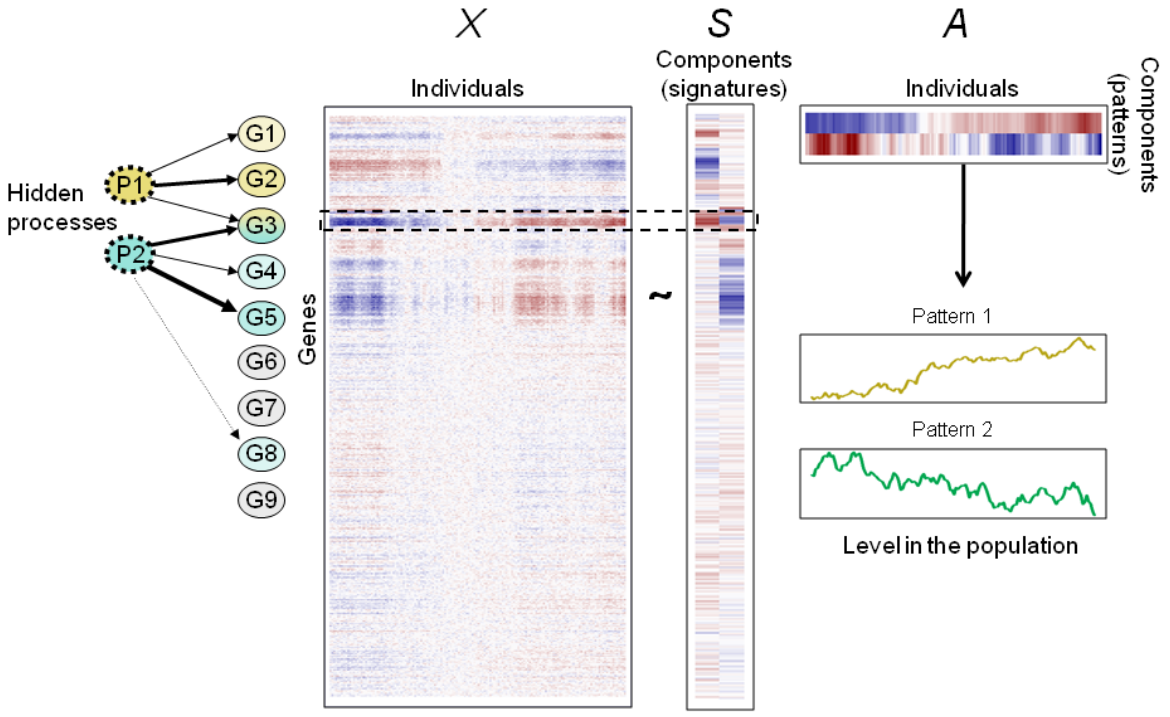


FIGURE 9.1 – **Principe schématique de la décomposition obtenue par l'ACI pour $K = 2$:** On représente les données par une image où les niveaux d'expression sont visibles par un gradient de couleur allant du bleu (minimum) au rouge (maximum). L'ACI décompose les données selon un produit matriciel $X = SA$. Les composantes extraites sont caractérisées par leurs signatures S indépendantes entre elles (indépendance des colonnes de S) et leurs patterns (matrice A). Les composantes extraites peuvent être vues comme le reflet de processus biologiques (P1 et P2). Sur le schéma P1 influence 3 gènes tandis que P2 influence 4 gènes. Certains gènes comme le gène G3 peuvent être influencés par plusieurs processus à la fois. Cela apparaît alors par une forte participation des gènes dans les deux signatures. Ainsi qu'on le voit sur les profils des patterns extraits dans la population qui sont ici corrélés négativement, aucune condition d'indépendance n'est imposée sur la matrice A.

$$s_k = \underset{\substack{s_k \\ s.c. \quad ||s_k|| = 1 \\ et \quad s'_k s_i = 0 \ \forall i \in [1, k-1]}}{\operatorname{argmax}} (s'_k X X' s_k)$$

On peut montrer que la solution de ce programme de maximisation revient à prendre pour signature de la k^{e} composante principale le k^{e} vecteur propre de la matrice de corrélation des expressions. La variance du nuage des transcrits expliquée par cette composante est alors donnée par la k^{e} valeur propre associée à la matrice de corrélation des expressions.

On peut également montrer que faire une ACP revient à effectuer la décomposition en valeurs singulières de la matrice de données. Selon cette décomposition, toute matrice X de rang K peut être écrite comme un produit de trois matrices U, D et V

$$X = U.D.V$$

où U et V sont des matrices de taille $n \times K$ et $K \times p$ vérifiant respectivement $U'U = I_n$, $V'V = UU' = I_K$ et $VV' = I_p$. Et où D est une matrice diagonale de taille K. On obtient alors directement la décomposition en facteurs principaux en prenant $S = U$ et $A = D.V$. Dans ce cas, les parts de variance expliquées par chaque composante sont obtenues directement en prenant les carrés des éléments diagonaux de la matrice D.

1.2.2 L'analyse en composantes indépendantes

L'analyse en composantes indépendantes (ACI) peut être vue comme une extension de l'ACP au cas non gaussien. En effet, dans l'ACP, on projette les variables initiales sur des axes orthogonaux, pour obtenir un faible nombre de motifs d'expression décorrélés et de variance maximale. L'ACI en revanche relâche les contraintes de non-corrélation entre les motifs recherchés, en imposant l'indépendance et la non-gaussianité des signatures. Ainsi l'ACI permet d'identifier des causes de variabilité qui peuvent être corrélées entre elles. Cette méthode a été utilisée avec succès à plusieurs reprises pour l'analyse des donnés de biopuces [86–88].

Nous utilisons ici l'algorithme `fastICA` pour effectuer l'ACI. Cet algorithme se base sur le lien qui existe entre indépendance et gaussianité. Intuitivement, ce lien peut s'expliquer à l'aide du théorème central limite. Puisqu'une somme de variables aléatoires indépendantes converge en loi vers une gaussienne, on s'attend à ce que la distribution de toute combinaison linéaire des profils d'origine tende à se rapprocher d'un gaussienne. Une stratégie pour retrouver les signatures indépendantes consiste donc à rechercher des combinaisons des profils observés qui soient les moins gaussiennes possibles.

De façon plus formelle, le critère d'indépendance des signatures se distingue du critère de non-corrélation imposé par l'ACP par le fait qu'il permet de tenir également compte

des moments d'ordre supérieur, en minimisant l'information mutuelle entre les signatures. L'information mutuelle est définie par :

$$I(s_1, \ldots, s_K) = H(s_1, \ldots, s_K) - \sum_{k=1}^{K} H(s_k)$$

où H désigne l'entropie de Shannon, définie pour une variable aléatoire [4] y de densité f par

$$H(y) = -\int f(y) \log(f(y)) dy$$

Une caractéristique fondamentale de cette entropie est qu'elle est maximale lorsque la variable aléatoire y est gaussienne [89]. Hyvärinen et Oja ont montré [90] qu'on pouvait mettre en évidence le lien entre l'information mutuelle d'un ensemble de variables aléatoires $y_1, \ldots, y_K$ et leur écart à la gaussianité en introduisant le concept de néguentropie. La néguentropie $J(y)$ est définie par

$$J(y) = H(y_{gauss}) - H(y)$$

où y_{gauss} est une variable aléatoire gaussienne de même matrice de variance-covariance que y. La néguentropie est donc une valeur positive ou nulle, valant 0 lorsque la variable aléatoire y suit une distribution gaussienne.

On peut montrer que l'information mutuelle d'un ensemble de variables aléatoires $y_1, \ldots, y_K$ peut se décomposer à une constante près en fonction des néguentropies individuelles de chacune des v.a. y_k. On a alors en notant C cette constante :

$$I(s_1, \ldots, s_K) = C - \sum_{k=1}^{K} J(s_k)$$

On voit donc comment en maximisant la non-gaussianité des signatures extraites, on minimise l'information mutuelle entre ces mêmes signatures. De plus dans notre cas, comme nous l'évoquons plus loin, la non-gaussianité des signatures extraites est favorable à la recherche de facteurs affectant uniquement un sous-ensemble de transcrits. En revanche l'adoption de ce critère de non-gaussianité pour la définition des composantes rend instable l'estimation des composantes dont la signature ne diffère pas significativement d'une gaussienne. Nous retirons donc dans la suite ces composantes de l'analyse (voir section 1.3).

Il convient de noter que l'ACI n'est pas à proprement parler une méthode de réduction de la dimension des données au même titre que l'ACP. Elle est donc généralement combinée à l'ACP dans une procédure en 4 étapes :

4. Eventuellement multidimensionnelle.

1. Décomposition des données par l'ACP et choix du nombre de composantes à analyser K en fonction du screeplot.
2. Reconstruction des données à partir des K premières composantes principales.
3. Extraction des K signatures $s_{\cdot k}$ par ACI.
4. Calcul des K patterns $a_{k\cdot}$ correspondants.

1.2.3 Choix du nombre de composantes

Bien souvent en analyse factorielle, le choix du nombre de composantes est une question cruciale. Si ce choix dépend évidemment de l'objectif recherché, il repose le plus souvent sur l'analyse des valeurs propres renvoyées par l'ACP. Ces valeurs indiquent en effet la part de variance expliquée par chaque composante. L'étude de ces valeurs permet donc d'exclure des analyses les composantes expliquant une part trop faible de la variabilité des données. Afin de déterminer un seuil à partir duquel les facteurs latents pouvaient être considérés comme non informatifs, Horn a proposé de déterminer les valeurs propres attendues en l'absence de structure de corrélation par des méthodes de permutations [91]. La méthode proposée par Horn se décompose en 3 étapes :

1. On calcule les valeurs propres d_k de l'ACP sur les vraies données.
2. On permute chaque variable indépendamment pour éliminer la structure de corrélation des données et on calcule les valeurs propres sur les données permutées.
3. On répète B fois le processus de permutation et on note r_k l'espérance de la k^{e} valeur propre, que l'on estime à partir des données permutées.
4. On compare les valeurs propres observées aux valeurs obtenues sur les données permutées et on garde un nombre de composantes K le plus grand tel que

$$d_k > r_k,\ \forall k \leq K$$

Bien que cette méthode donne d'assez bons résultats sur des données simulées [92], elle peut se montrer trop conservatrice lorsque les premières composantes expliquent une part importante de la variabilité des données. En effet, lorsqu'on permute les données, la variabilité totale des données reste constante. L'estimation de la k^{e} valeur propre calculée sur les données permutées ne tient donc pas compte de l'excédent de variabilité déjà expliqué par les $(k-1)$ premières valeurs propres. Ce phénomène est visible sur la figure 9.2 où on voit une surestimation du niveau attendu des valeurs propres en l'absence de structure.

Ici nous optons donc pour une variante de la méthode de Horn.

1. On calcule les valeurs propres d_k de l'ACP sur les vraies données.
2. On permute chaque variable indépendamment pour éliminer la structure de corrélation et on calcule les valeurs propres sur les données permutées.

3. On répète B fois le processus de permutation et note r_k l'espérance de la k^{e} valeur propre calculée sur les données permutées.
4. On estime l'espérance r'_k de la k^{e} plus grande valeur propre corrigée pour la part de variance déjà expliquée par les $k-1$ premières valeurs propres :

$$r'_k = r_k - \frac{1}{n-k+1}\sum_{i=1}^{k-1} d_i - r_i$$

5. On compare les valeurs propres observées aux valeur obtenues et on garde un nombre de composantes K le plus grand tel que

$$d_k > r'_k,\ \forall k \leq K$$

On voit sur la figure 9.2 que ce critère permet de mieux modéliser la distribution des valeurs propres les plus faibles et tend à augmenter le nombre de composantes considérées. Ce choix permet donc de limiter le risque de sous-estimation du nombre de composantes dont nous verrons les conséquences dans la section 2.1.

1.3 Identification de modules

Afin d'identifier l'ensemble des gènes significativement affectés par une composante, qu'on appellera module, on étudie les signatures des composantes. L'analyse des signatures obtenues sur les données de GHS montre que, sur des données globalement homogènes, les facteurs latents affectent généralement un nombre restreint de gènes (moins de 10% en général) conduisant à des signatures leptokurtiques [5].

Afin d'étudier uniquement les composantes affectant spécifiquement un sous-ensemble de gènes et d'écarter d'éventuelles composantes instables et pour lesquelles l'algorithme d'ACI n'aurait pas convergé, on ne considère dans la suite que les composantes dont le kurtosis est supérieur à 3. Ce seuil, choisi de manière empirique, correspond au kurtosis d'une loi de Laplace.

Des simulations montrent qu'en l'absence de structure de corrélation dans les données, la distribution des signatures $s_{.k}$ estimées suit une loi normale. On va donc considérer les contributions s_{ik} comme des statistiques de test et tester pour chaque transcrit et chaque composante :

– $\mathcal{H}_0$: Le transcrit i n'est pas affecté par la composante k

5. Lorsqu'un certain nombre de sous-groupes d'échantillons très fortement différenciés sont présents (par exemple sur un jeu de données regroupant deux tissus distincts), il arrive que des signatures platikurtiques reflétant la différence entre ces sous-groupes apparaissent. De telles signatures ne sont pas adaptées à la définition de modules telle que nous l'entendons ici et doivent plutôt être analysées en opposant les gènes situés de chaque côté de la distribution.

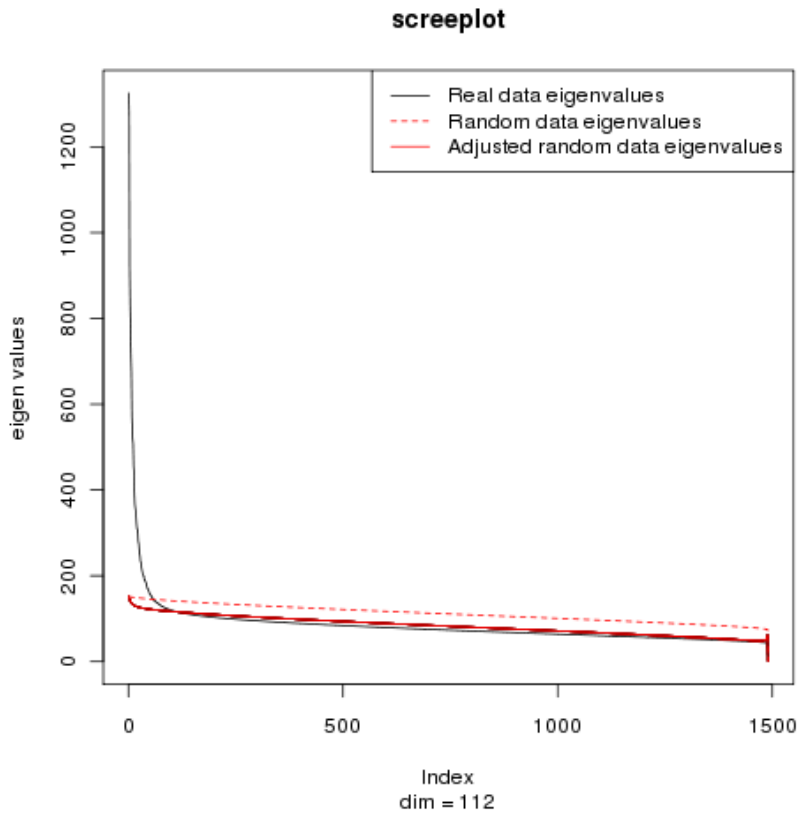


FIGURE 9.2 – **Screeplot des valeurs propres obtenues par l'ACP sur les données de GHS :** On voit en noir la courbe des valeurs propres d_k obtenues par une ACP de la matrice d'expression. En pointillés rouges, on voit la distribution des valeures propres r_k obtenues sur la matrice d'expression permutée. Le trait plein rouge montre les valeurs propres r'_k obtenues en appliquant la correction proposée pour tenir compte de la variance déjà expliquée par les premières composantes. On voit que la correction permet d'augmenter le nombre de composantes à garder par une meilleure estimation du niveau attendu des valeurs propres sous l'hypothèse nulle d'une absence de structure.

– $\mathcal{H}_a$: Le transcrit i est affecté par la composante k

Pour une composante affectant une fraction π_1 des gènes, on s'attend à ce que pour la fraction $\pi_0 = 1 - \pi_1$ des gènes n'étant pas affectés par la composante, les contributions s_ik soit distribuées selon une loi normale [6]. On utilise donc un modèle de mélange adapté du travail de Strimmer [53] pour l'estimation du FDR. Dans ce modèle, les statistiques de tests sous $\mathcal{H}_0$ suivent une loi normale de paramètres (μ, σ^2) et de densité f_0. Les statistiques sous $\mathcal{H}_a$ suivent une distribution alternative f_a non précisée. La distribution des statistiques s est donc donnée par

$$f(s) = \pi_0 f_0(s) + \pi_1 f_a(s)$$

et la probabilité qu'un transcrit i ne soit pas affecté par la composante k, sachant la contribution estimée s_{ik} de cette composante au transcrit i, est donnée par

$$\mathbb{P}\left(\mathcal{H}_0 | s_{ik}\right) = \frac{\pi_0 f_0(s_{ik})}{\pi_0 f_0(s_{ik}) + \pi_1 f_a(s_{ik})}$$

Ici, on estime tout d'abord μ à partir de la médiane des signatures [7]. Puis ainsi que proposé par Strimmer [53], on estime σ^2 et π_0 par maximum de vraisemblance tronquée et f_A par un estimateur non paramétrique de la densité. On est alors capable de fournir pour chaque transcrit i et chaque composante k, la probabilité que le transcrit ne soit pas affecté par le facteur conditionnellement à la contribution s_{ik} estimée. On définit dans la suite le module correspondant à la composante k comme l'ensemble des transcrits tels que $\mathbb{P}\left(\mathcal{H}_0 | s_{ik}\right) < 0.001$. La figure 9.3 illustre sur un exemple la façon dont sont définis les modules.

1.4 Association aux génotypes

Une fois les patterns d'expression extraits et les modules définis, on teste l'association des patterns avec les génotypes. Pour conclure à la présence d'un module de gènes co-régulés par un SNP, un double critère d'association au génotype est requis :

– Une association du pattern avec le SNP, testée par une analyse de variance à 2 degrés de libertés, avec confirmation des résultats positifs par un test de rang robuste.

– La présence d'un enrichissement du module en gènes associés au SNP, testée par un test exact de Fisher comparant le nombre de gènes associé au seuil de 10^{-5} à l'intérieur et à l'extérieur du module. L'utilisation du critère d'enrichissement permet d'écarter les cas où l'association du pattern avec le génotype est uniquement due à la présence d'un ou deux transcrits *cis*-régulés contribuant très fortement à l'estimation du pattern.

6. Dont la moyenne et la variance varient en fonction de la distribution des contributions sous $\mathcal{H}_a$.

7. Ce faisant, on fait implicitement l'hypothèse que π_0 est suffisamment grand pour que la médiane de la distribution des signatures ne diffère pas trop de celle de f_0.

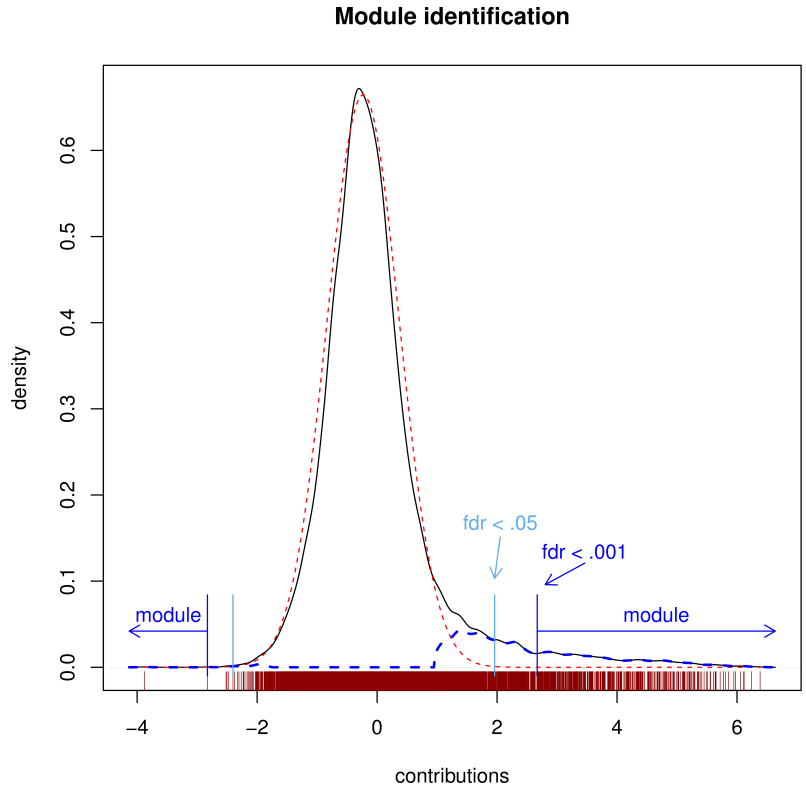


FIGURE 9.3 – **Définition du module à partir de la signature d'une composante :** le graphique montre la constitution du module de gènes à partir des contributions s_{ik}. La signature représentée correspond à la composante expliquant la plus grande part de variance du transcriptome dans les données de GHS. Sur le graphique, les contributions s_{ik} sont représentées en abscisse : chaque trait rouge sombre sous l'axe des abcisses représente un transcrit différent. On voit représentée par la courbe noire la densité des contributions. Cette densité est décomposée selon un mélange de deux lois : une loi gaussienne f_0 (trait rouge pointillé) et une loi f_A (trait bleu pointillé) estimée selon un modèle non paramétrique. Le taux de faux positifs attendu pour chaque valeur de s_{ik} est estimé à partir de ces deux distributions et des seuils sont placés sur les contributions s_{ik} (barres verticales bleues) au delà desquels le gène est considéré comme significativement affecté par la composante. On représente sur le graphique les seuils obtenus en imposant un FDR local de 5% et un FDR local de $1\,{}^0/_{00}$. Les flèches bleues montrent les valeurs des contributions pour lesquelles les gènes sont affectés au module associé à la composante, avec le seuil retenu de $1\,{}^0/_{00}$.

Pour chaque test, un seuil de Bonferroni corrigeant pour le nombre de composantes et de SNP testés a été utilisé dans les simulations. Lors de l'application de la méthode aux données réelles, afin d'augmenter la puissance, un seuil suggestif de 10^{-7} a été retenu pour l'association des patterns au génotype (critère 1) et le seuil de Bonferroni a été appliqué uniquement au test d'enrichissement (critère 2).

L'utilisation de ce double critère permet de garantir simultanément l'association des gènes du module et la spécificité du signal d'association au module considéré.

2 Application de l'ACI à la recherche d'effets génétiques à grande échelle sur le transcriptome

2.1 Simulations

Dans un premier temps, des études de simulations on été réalisées, afin de tester les propriétés de l'ACI. Nous rapportons ici quelques-uns les principaux résultats apportés par ces études.

Nous nous plaçons dans toute cette section dans le cadre décrit précédemment où l'expression peut être décrite (au moins en partie) par une série de facteurs latents. Pour des raisons computationnelles, on simule ici des jeux de données de taille inférieure à celle des données de GHS avec à chaque fois 1000 gènes pour 100 individus. Pour chaque cadre de simulation décrit dans cette section, les résultats rapportés sont calculés sur 300 jeux de données simulés.

2.1.1 Identification du nombre de composantes

Dans un premier temps, nous avons cherché à évaluer la capacité de notre méthode à retrouver le bon nombre de composantes. Dans ce but on fixe une structure de corrélation avec K composantes expliquant une part Λ de la variabilité totale. Pus précisément on impose que la k^{e} composante explique une part λ_k de la variance totale et on fixe les lambda K de façon à ce que

$$\sum_{k=1}^{K} \lambda_k = \Lambda$$

et

$$\forall k \leq K-1, \;\; \lambda_k - \lambda_{k+1} = cste > 0$$

Le reste de la variabilité est ajoutée sous la forme d'un bruit blanc.

Chaque pattern simulé affecte une fraction π_1 des gènes fixée à 0,05. Les contributions s_{ik} des patterns à l'expression des gènes affectés sont tirées dans une loi normale centrée de variance $\frac{1}{\pi_1}$ afin que la variance des signatures soit égale à 1. Les gènes affectés sont

	K	10	20	30	10	20	30	10	20	30
	Λ	0,3	0,3	0,3	0,5	0,5	0,5	0,7	0,7	0,7
critère :		**nombre moyen de composantes (écart-types)**								
Kaiser		31,3 (1,0)	31,5 (0,9)	33,3 (0,8)	22,7 (1,2)	21,8 (0,8)	26,2 (0,6)	14,6 (1,4)	19,3 (0,5)	25,5 (0,6)
Horn		9,5 (0,5)	15,6 (0,8)	19,4 (1,0)	10,0 (0,2)	16,7 (0,6)	20,9 (0,8)	10,0 (0,0)	17,6 (0,6)	22,0 (0,8)
Horn modifié		12,6 (1,8)	22,4 (2,3)	29,7 (2,6)	10,7 (0,9)	19,9 (1,0)	27,8 (1,1)	15,6 (0,2)	19,8 (0,4)	28,0 (0,7)

Tableau 9.1 – **Nombre de composantes identifiées selon la méthode choisie :** K désigne le nombre de composantes simulées et Λ indique la part de la variance totale qui est expliquée par les K composantes simulées.

sélectionnés aléatoirement et indépendamment pour chaque composante, ce qui permet de vérifier la condition d'indépendance des signatures imposée par l'ACI. Les niveaux des K patterns sont tirés dans des lois normales centrées et de variance ${\lambda_k}^{**}$; Formellement ce modèle peut donc s'écrire :

$$x_{ij} = \sum_{k=1}^{K} s_{ik} a_{kj} + \epsilon_{ij} \tag{9.1}$$

avec

$$\begin{aligned} \forall i,j \quad & \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\ \forall i,k \quad & s_{ik} \overset{iid}{\sim} \mathcal{N}\left(0, \tfrac{1}{\pi_1}\right) * \mathcal{B}(\pi_1) \\ \forall k,j \quad & a_{kj} \overset{iid}{\sim} \mathcal{N}(0, \lambda_k) \end{aligned}$$

On teste ensuite pour différentes valeurs des paramètres K et Λ le nombre de composantes retrouvées selon le critère choisi. On compare ici 3 critères :

- Le critère de Kaiser qui préconise de ne garder que les composantes dont la valeur propre est supérieure à 1.
- Le critère de Horn qui est une extension du critère de Kaiser tenant compte des corrélations aléatoires (donc plus conservatrice).
- Le critère de Horn modifié tenant compte de la variance expliquée par les premières composantes (moins conservatrice que le critère de Horn).

Le tableau 9.1 donne les résultats de cette analyse pour $K = 10, 20, 30$ et $\Lambda = 0.3, 0.5, 0.7$.

On voit que la méthode Kaiser donne en général une estimation très grossière du nombre de composantes, et s'avère presque tout le temps la moins bonne. A contrario, la méthode de Horn, se comporte bien lorsque le nombre de composantes est faible. En revanche, lorsque le nombre de composantes augmente (et que par conséquent la part de variance expliquée par chaque composante diminue), elle tend à sous-estimer le nombre de composantes. La méthode de Horn modifiée permet de limiter cet effet et de détecter des composantes expliquant moins de variance, au prix d'une légère surestimation du nombre de composantes lorsque celui-ci est faible (et d'une légère augmentation de la

variabilité). Globalement, la méthode de Horn modifiée est celle qui semble donner les meilleurs résultats.

2.1.2 Impact du nombre de composantes sur la reconstitution des patterns

Dans un deuxième temps, nous avons voulu tester la capacité de l'ACI à reconstituer les patterns simulés et l'impact du nombre de patterns déterminé a priori (surestimation ou sous-estimation par rapport au nombre de patterns simulés) sur cette reconstitution. Dans ce but, nous avons repris le même cadre de simulation que précédemment en ajoutant la possibilité d'une corrélation entre les patterns extraits :

En pratique, les niveaux des K patterns sont tirés dans une loi normale multivariée de matrice de corrélation C définie par

$$C = (1 - \rho)I + \rho R$$

où R est une matrice de corrélation aléatoire construite en prenant la matrice de Gram d'un ensemble de K vecteurs aléatoires tirés uniformément sur la sphère unité selon la méthode proposée par Holmes [93]. On peut ainsi régler le niveau de corrélation souhaité par le choix du paramètre rho ($\rho = 0$ correspond à la situation d'indépendance des patterns supposée par l'ACP). L'indépendance des signatures requise par l'ACI reste vérifiée ici puisque les s_{ik} sont tirés de façon indépendante.

On teste la capacité de l'ACI à retrouver les facteurs latents simulés en mesurant la corrélation absolue moyenne entre chaque pattern simulé et le pattern estimé dont il est le plus proche (qualité de reconstruction). Ce critère est testé pour plusieurs choix du nombre de composantes, allant d'une sous-estimation (nombre de composantes divisé par deux), à une surestimation (nombre de composantes multiplié par 2) par rapport au nombre de composantes simulé. A titre de comparaison, ce critère est également testé sur les motifs extraits par l'ACP lorsque le vrai nombre de composantes est utilisé.

Les résultats de cette analyse sont rapportés dans le tableau 9.2.

	K	10	20	30	10	20	30	10	20	30
	ρ	0	0	0	0,3	0,3	0,3	0,5	0,5	0,5
Méthode de factorisation		**corrélation moyenne avec les facteurs simulés**								
ACP										
	(vrai nombre de composantes)	0,68	0,55	0,47	0,66	0,54	0,47	0,65	0,53	0,48
ACI										
	(vrai nombre de composantes)	0,98	0,93	0,89	0,98	0,93	0,88	0,98	0,92	0,87
	(surestimation du nombre de composantes ($2K$))	0,98	0,94	0,88	0,98	0,93	0,88	0,98	0,92	0,86
	(sous-estimation du nombre de composantes ($K/2$))	0,65	0,64	0,63	0,68	0,65	0,64	0,72	0,68	0,66
	(Estimation par Horn modifié)	0,98	0,93	0,88	0,98	0,93	0,87	0,98	0,91	0,85

Tableau 9.2 – **Qualité de reconstruction selon la méthode de factorisation et le nombre de composantes extraites** (choisi a priori).

La qualité de reconstruction des facteurs latents extraits par l'ACP est ici relativement mauvaise car l'ACP ne peut pas différencier des facteurs expliquant des parts de variance trop proches. En présence de facteurs présentant des parts de variance très différentes, on verrait également la qualité de reconstruction de l'ACP chuter lorsque la corrélation entre les facteurs augmente puisque l'ACP repose sur des hypothèses de non-corrélation. Ces mauvais résultats de l'ACP nous on conduit dans la suite à nous concentrer sur l'ACI, plus adaptée aux situations rencontrées dans l'étude du transcriptome. L'ACI permet ici une reconstruction de bien meilleure qualité. On peut faire plusieurs constats à partir de ce tableau :

- D'abord, on voit que la présence de corrélations entre les facteurs latents n'affecte pas la capacité de reconstruction des patterns. Ce résultat est attendu puisque l'ACI ne fait pas d'hypothèse sur l'indépendance entre les patterns mais repose uniquement sur la non-gaussianité des signatures.
- De façon générale, même lorsque le nombre de composantes utilisé dans la décomposition correspond au nombre de composantes simulé, on voit une légère diminution de la qualité de reconstruction lorsque le nombre de composantes augmente. Cette diminution peut être attribuée à une plus lente convergence de l'algorithme d'ACI lorsque le nombre de composantes augmente. Elle peut être facilement évitée en augmentant le nombre d'itérations de l'algorithme lorsque le nombre de composantes augmente.
- Lorsqu'on surestime le nombre de composantes, l'impact sur la qualité de reconstruction est faible et est largement attribuable aux problèmes de rapidité de convergence évoqués ci-dessus.
- Lorsqu'on sous-estime le nombre de composantes, on voit en revanche une nette diminution de la qualité de reconstruction. Ces résultats montrent donc qu'une surestimation du nombre de composantes est largement préférable à une sous-estimation, en termes de qualité de reconstruction.
- On voit que le critère de Horn modifé permet une qualité de reconstruction acceptable dans la majeure partie des cas.

2.1.3 Détection d'effets génétiques à grande échelle sur le transcriptome

Nous avons ensuite voulu déterminer la capacité de notre méthode à retrouver l'effet d'un SNP affectant un module de gènes co-régulés. Afin de se rapprocher au maximum de la réalité biologique, on simule à présent une structure de corrélation inspirée de celle observée sur les données de GHS. Pour cela, on tire aléatoirement 100 individus et 1000 gènes directement dans la matrice d'expression[8]. On simule ensuite un facteur de trans-

8. Qui comprend près de 14 000 sondes et 1500 individus.

cription dont le niveau d'activité est régulé par un SNP, et qui affecte en cascade une proportion π_1 des gènes (figure 9.4.)

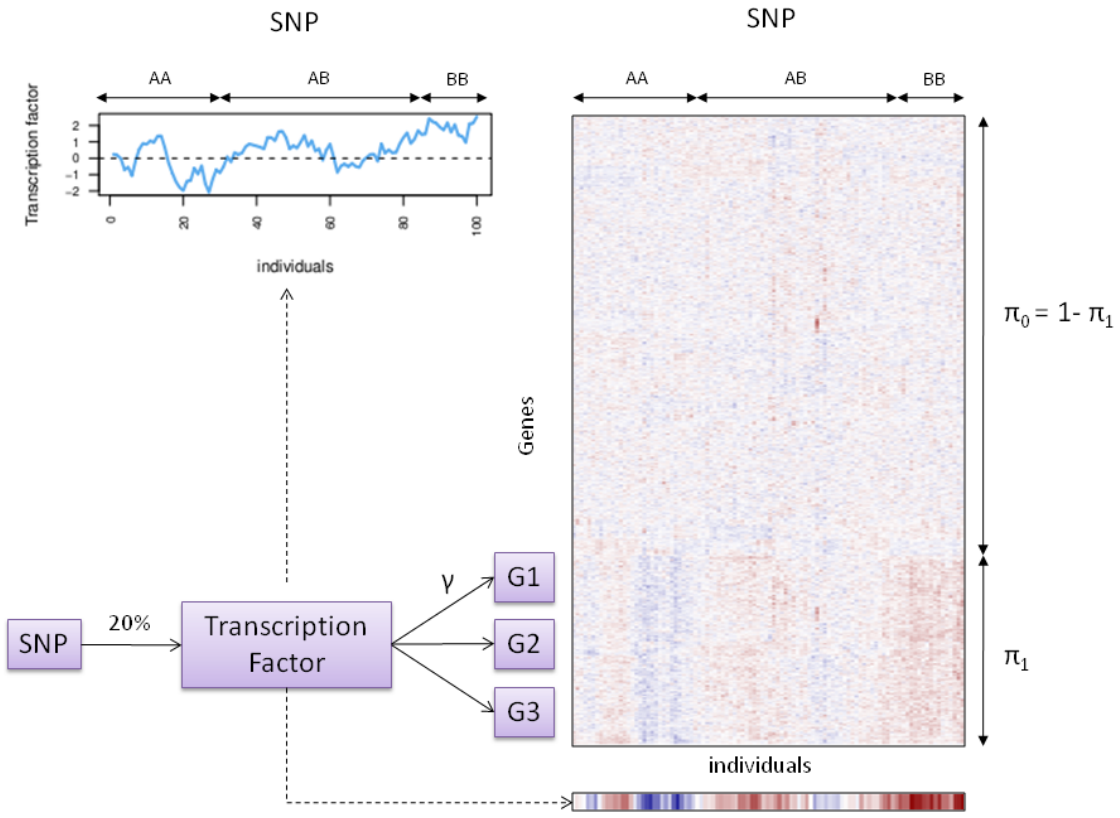


FIGURE 9.4 – **Modèle de simulation utilisé :** On représente les données simulées par une image où les niveaux d'expression sont visibles par un gradient de couleur allant du bleu (minimum) au rouge (maximum). Après avoir extrait un sous-échantillon dans les données réelles, on simule un SNP et un facteur de transcription dont 20% de la variabilité est expliquée par le SNP. On sélectionne ensuite aléatoirement une proportion π_1 de transcrits sur lesquels on ajoute un effet du facteur de transcription expliquant une proportion γ de la variance.

Le polymorphisme, de fréquence fixe égale à 0.4, est simulé à l'équilibre d'Hardy-Weinberg en tirant dans une loi binomiale $\mathcal{B}(2, 0.4)$. On simule ensuite un facteur de transcription (TF) tel que 20% de la variabilité du facteur de transcription soit attribuable au SNP. Pour un individu j, on a :

$$TF_j = \sqrt{0.2} * SNP_j + \sqrt{0.8} * \epsilon_j$$

avec $\epsilon_j \overset{iid}{\sim} \mathcal{N}(0, 1)$.

On sélectionne ensuite aléatoirement une proportion π_1 de transcrits sur lesquels on ajoute un effet du facteur de transcription expliquant une part γ de l'expression des transcrits. L'expression du transcrit i est caluclulée pour l'individu j en prenant

$$\tilde{x}_{ij} = \sqrt{\gamma} * TF_j + \sqrt{1-\gamma} * x_{ij}$$

où x_{ij} représente l'expression du transcrit i chez l'individu j dans l'échantillon initial des données de GHS et $\tilde{x}_{ij}$ représente l'expression du transcrit i chez l'individu j après ajout de l'effet du facteur de transcription.

On compare ensuite pour différentes valeurs de γ et π_1 l'efficacité de l'ACI et de l'ACP pour

1. Trouver un pattern reproduisant l'effet du facteur de transcription sur le transcriptome

2. Tester l'hypothèse nulle globale :

   $\mathcal{H}_0{}^g$: Le SNP n'a d'effet sur aucun pattern d'expression.
   $\mathcal{H}_a{}^g$: Le SNP affecte (au moins) un pattern d'expression.

Pour cela on calcule pour chaque valeur des paramètres et pour chaque méthode :

1. La corrélation absolue moyenne entre le facteur de transcription simulé et le pattern estimé le plus proche (qualité de reconstruction).

2. Le nombre de fois où un effet significatif du SNP sur au moins un des patterns extraits est détecté (Puissance $\mathcal{H}_0{}^g$). Les p-values d'association sont ajustées sur le nombre de composantes testées par un critère de Bonferroni.

3. Le nombre de fois où un SNP simulé indépendamment des données d'expression ressort associé à tort à un des facteurs afin de vérifier que le contrôle de l'erreur de type I est garanti par la procédure utilisée (Erreur Type I $\mathcal{H}_0{}^g$).

Le tableau 9.3 montre les résultats obtenus pour un facteur de transcription affectant entre 10% et 2.5% des gènes et expliquant entre 40% et 10% de la variabilité de l'expression de ses gènes cibles.

	γ	0,4	0,4	0,4	0,3	0,3	0,3	0,2	0,2	0,2	0,1	0,1	0,1
	π_1	0,1	0,05	0,025	0,1	0,05	0,025	0,1	0,05	0,025	0,1	0,05	0,025
méthode de factorisation		**test de l'effet du SNP sur les composantes**											
ACP													
	Qualité de reconstruction	0,80	0,68	0,57	0,74	0,63	0,50	0,66	0,56	0,42	0,56	0,41	0,33
	Puissance $\mathcal{H}_0^g$	0,78	0,62	0,39	0,70	0,49	0,28	0,57	0,37	0,20	0,34	0,17	0,12
	Erreur de type I $\mathcal{H}_0^g$	0,03	0,05	0,06	0,06	0,05	0,04	0,04	0,04	0,05	0,04	0,07	0,05
ACI													
	Qualité de reconstruction	0,98	0,98	0,94	0,98	0,96	0,90	0,96	0,92	0,78	0,63	0,53	0,35
	Puissance $\mathcal{H}_0^g$	0,96	0,96	0,91	0,94	0,92	0,87	0,94	0,88	0,66	0,51	0,33	0,10
	Erreur de type 1 $\mathcal{H}_0^g$	0,03	0,06	0,04	0,04	0,03	0,05	0,05	0,03	0,06	0,06	0,05	0,06

Tableau 9.3 – **Qualité de reconstruction, puissance et erreur de type I du teste global $\mathcal{H}_0^g$** selon la force γ de l'effet du facteur de transcription et le pourcentage π_1 de gènes affectés par ce facteur de transcription.

On voit à partir de ce tableau que la qualité de reconstruction ainsi que la puissance augmentent globalement pour les deux méthodes avec la part de variabilité $\pi_1\gamma$ expliquée par le facteur simulé. En effet si le facteur explique une part trop faible de la variabilité, cette variabilité ne sera pas capturée par l'ACP lors de la réduction de la dimension des données et ni l'ACP ni l'ACI ne pourront retrouver le facteur simulé. Pour l'ACI, un autre facteur important est la part π_1 des gènes affectés par le facteur de transcription simulé. A part de variance expliquée $\gamma\pi_1$ égale [9], on voit que l'ACI a de meilleures performances lorque π_1 diminue [10]. Ce phénomène s'explique par le fait que plus la fraction π_1 de gènes affectés est faible, plus la signature de la composante simulée est leptokurtique, et donc plus l'ACI est efficace pour retrouver le motif simulé. L'ACI apparaît donc plus apte à reconstituer le facteur simulé dans l'ensemble des cas testés. La puissance d'association globale est par conséquent plus forte avec l'ACI dans l'intégralité des cas simulés. Le différentiel de puissance avec l'ACP atteint 50% dans les cas favorables. L'erreur de type I est bien contrôlée à 5% dans l'intégralité des simulations.

2.1.4 Détection des gènes affectés par le génotype

Enfin, on contrôle la capacité de la méthode à identifier correctement les gènes appartenant au module de gènes affecté par le SNP. On veut ici tester pour chaque gène i, l'hypothèse :

$\mathcal{H}_0^i$: Le gène i n'appartient pas à un module de gènes co-régulés par le SNP.

$\mathcal{H}_a^i$: Le gène i appartient à un module de gènes co-régulés par le SNP.

Dans ce but, on reprend les données simulées dans la section précédente, et on définit pour chaque pattern associé au SNP le module caractérisant ce pattern ainsi que décrit

9. Par exemple, pour les simulations avec les paramètres $(\gamma = 0.4, \pi_1 = 0.025)$;$(\gamma = 0.2, \pi_1 = 0.05)$ et $(\gamma = 0.1, \pi_1 = 0.1)$.

10. Tandis que l'ACP affiche des performances comparables dans toutes les configurations.

	γ	0,4	0,4	0,4	0,3	0,3	0,3	0,2	0,2	0,2	0,1	0,1	0,1
	π_1	0,1	0,05	0,025	0,1	0,05	0,025	0,1	0,05	0,025	0,1	0,05	0,025
Méthode de factorisation		**Détection des gènes du module**											
ACP													
	Puissance $\mathcal{H}_0^i$	0,22	0,14	0,03	0,07	0,04	0,01	0,00	0,00	0,00	0,00	0,00	0,00
	FDR $\mathcal{H}_0^i$	0,02	0,09	0,16	0,07	0,05	0,11	0,08	0,29	1,00	1,00	1,00	1,00
ACI													
	Puissance $\mathcal{H}_0^i$	0,96	0,97	0,95	0,94	0,92	0,87	0,66	0,59	0,41	0,04	0,07	0,02
	FDR $\mathcal{H}_0^i$	0,00	0,00	0,01	0,00	0,01	0,01	0,00	0,01	0,08	0,26	0,33	0,88

Tableau 9.4 – **Puissance de détection des gènes associés au sein du module et taux de faux positifs**

dans la section 1.3. On calcule alors sur les simulations :

1. Le pourcentage moyen de gènes qui sont retrouvés dans le module associé au SNP parmi les gènes du module simulé (Puissance $\mathcal{H}_0^i$).
2. Le pourcentage moyen de gènes n'appartenant pas au module simulé parmi les gènes retrouvés dans le module associé au SNP (FDR $\mathcal{H}_0^i$).

Le tableau 9.4 montre la puissance de détection des gènes affectés par le SNP et le taux de faux positifs dans le module, lorsqu'un module associé au SNP a été trouvé. Les résultats sont fournis pour l'ACI et l'ACP pour différentes valeurs des paramètres γ et π_1.

On voit que pour l'ACP les modules trouvés comme associés au SNP ne sont généralement que peu spécifiques du SNP et ne contiennent qu'une partie des gènes réellement associés au SNP. Pour l'ACI en revanche on a une puissance élevée et un faible taux de faux positifs lorsqu'on est dans les conditions assurant une qualité de reconstruction au dessus de 0.8 (cf. tableau 9.3). Lorsque la qualité de reconstruction est faible (pour $\gamma = 0.1$ principalement) on observe une chute de la puissance et une dégradation de la spécificité des modules trouvés. Ce résultat s'explique en partie par le fait que dans ces conditions la majeure partie des associations trouvées entre un facteur et le SNP sont des faux positifs.

2.2 Application sur les données de GHS

L'application de cette approche aux données de GHS a fait l'objet d'un article présent en annexe 4. Nous détaillons ici quelques uns des principaux résultats de cet article, ainsi que certains éléments de réflexion apportés par l'application de l'ACI aux données réelles.

2.2.1 Identification des modules et interprétation

Dans GHS, l'application de l'ACI pour la recherche de modules de gènes co-régulés a permis de mettre en évidence 112 composantes. Parmi celles-ci, 21 étaient créées par

la présence d'individus atypiques et ont été écartées de l'analyse par la suite. Parmi les 91 facteurs restants, 64 ont permis l'identification de modules d'expression et ont donc fait l'objet de recherches plus approfondies. L'ensemble des résultats de ces analyses sont décrits dans l'article associé (voir annexes) et ont été mis à disposition de la communauté scientifique dans une base de données HTML accessible via le site `www.genecanvas.fr`.

Le nombre de gènes dans les modules extraits varie de 14 à 670 (médiane 178). Si certains modules peuvent être interprétés directement, comme par exemple un module impliquant uniquement des gènes des chromosomes X et Y et qui sépare nettement les hommes des femmes, la plupart nécessitent de recourir à des analyses d'enrichissement pour pouvoir être interprétés.

Parmi les modules obtenus, 42 étaient enrichis [11] en gènes appartenant à un pathway des bases de données KEGG (Kyoto Encyclopédia of Genes and Genomes [57]) ou GO (Gene Ontology [58]). Ainsi, on trouve par exemple un module enrichi en gènes liés aux changements de conformation des protéines, avec un nombre important de gènes codant pour des "Heat Shock Proteins". L'expression de ces protéines augmente fortement lorsque la cellule est soumise à des températures élevées ou à un stress important [94]. Il a été montré que ces protéines jouaient le plus souvent un rôle dans la réparation des protéines en maintenant leur structure [95]. Elles jouent également un rôle dans le système cardio-vasculaire où elles sont associées à la vasodilatation [12] et pourraient concourir à l'activation coordonnée des cellules musculaires lisses contenues dans les vaisseaux sanguins [96]. De façon intéressante, ce pattern était significativement associé au rythme cardiaque ($p < 6,4\,10^{-47}$, $R^2 \approx 13\%$) dans l'étude GHS, ce qui semble cohérent avec l'hypothèse d'un rôle des Heat shock protéines dans le contrôle du ryhtme cardiaque.

On note qu'on trouve parmi les fonctions biologiques caractérisant les modules une large proportion de fonctions liées à la réponse immunitaire (réponse aux virus, défense contre les bactéries (2 modules), réponse inflammatoire, présentation d'antigènes du complexe majeur d'histocompatibilité, réponse humorale, ...) ce qui est à première vue cohérent avec le rôle joué par les monocytes dans la réponse immunitaire.

Dans certains cas, l'association avec les facteurs de risque apporte également un éclairage nouveau sur les patterns extraits. Par exemple, on trouve un pattern très fortement associé au tabagisme ($p < 5.8\,10^{-82}$, $R^2 \approx 21\%$). Le module caractérisant ce pattern contient des gènes dont l'expression est très fortement associée au tabagisme. Parmi ces gènes, plusieurs sont significativement associés à la présence de la plaque d'athérome dans la carotide et le restent après ajustement sur le tabagisme. Ces gènes font actuellement l'objet d'une étude plus approfondie afin de déterminer leur éventuel rôle dans la formation de la plaque d'athérome.

11. On se reportera à la table Supplémentaire 1 de l'article associé pour le détail des enrichissements.
12. C'est-à-dire la dilatation des vaisseaux sanguins.

On retrouve également un pattern fortement associé à trois marqueurs de l'inflammation liés au risque cardiovasculaire [97–99] : CRP [13])($p < 6.2\,10^{-55}$, $R^2 \approx 15\%$), IL1Ra [14] ($p < 6.7\,10^{-56}$, $R^2 \approx 15\%$), et MR-proADM [15] ($p < 2.9\,10^{-43}$, $R^2 \approx 11\%$). Ces associations suggèrent que le pattern correspondant reflète un processus lié à l'inflammation.

2.2.2 Association avec le génotype

Nous avons ensuite testé l'association des 64 patterns avec l'ensemble des SNP de la puce Affymetrix 6.0 (675 354 SNP). Parmi ces patterns, 11 étaient associés à un ou plusieurs SNP avec le double critère que nous nous étions fixé (association suggestive à $p < 10^{-7}$ avec le pattern et enrichissement en gènes associés au SNP à l'intérieur du module).

Parmi les loci détectés par cette analyse, le locus du gène ARHGEF3 sur le chromosome 3p21 présente un intérêt tout particulier. Ce locus est trouvé associé à 2 modules différents, tous deux enrichis en fonctions biologiques connues. Avec plus de 140 gènes associés en *trans* au seuil de Bonferroni (10^{-12}), ce locus représentait le cas le plus flagrant d'un SNP affectant l'expression des gènes à grande échelle [16]. Une analyse des fonctions des gènes associés à ce locus révèle un enrichissement en fonctions liées pour le premier module à la coagulation ($p = 7.3\,10^{-12}$), l'hémostase ($p = 2.5\,10^{-11}$) et l'adhésion cellulaire ($p = 2.5\,10^{-9}$), et pour le second module à la régulation de la coagulation ($p = 2.5\,10^{-11}$). Le gène ARHGEF3 (Rho Guanine nucleotide exchange factor 3) code pour une protéine qui se lie aux protéines G pour stimuler les voies de signalisation Rho-dépendantes (Rho A et Rho B). L'hypothèse suggérée par ces résultats était qu'un variant dans le gène ARHGEF3 pouvait modifier la capacité de liaison du facteur GEF3 aux protéines G, perturbant ainsi l'activation de la voie de signalisation Rho kinase et l'expression des protéines effectrices. A la suite de ces résultats, l'hypothèse d'une modification de l'activité de la voie RhoA en fonction du génotype d'ARHGEF3 a été testée dans des modèles cellulaires par les biologistes de Mainz avec qui nous collaborons pour l'étude GHS. Les résultats de ces tests fonctionnels ont été négatifs. En parallèle, nous avons tenté de répliquer le lien entre le génotype et l'expression dans les monocytes de l'étude Cardiogenics (758 sujets). Aucune des associations mises en évidence avec le locus ARHGEF3 n'a pu être répliquée dans Cardiogenics. Ces résultats nous ont poussé à rechercher les causes possibles de cette association. Les études GHS et Cardiogenics diffèrent par la méthode d'extraction des monocytes. Ce constat nous a donc conduit à émettre l'hypothèse que cette asso-

13. Protéine C-Réactive.
14. Récepteur antagoniste à l'interleukine 1.
15. Pro-adrénomédulline.
16. Cet effet peut être observé sur le graphique 7.3, où l'on observe une ligne verticale à l'emplacement du locus

ciation pouvait être due à une contamination des échantillons de GHS par d'autre types cellulaires et en particulier par des plaquettes [17]. Cette hypothèse était largement appuyée par l'existence de résultats issus de GWAS indiquant une association du locus ARHGEF3 avec le nombre de plaquettes dans le sang et le volume plaquettaire moyen [100]. Nous avons donc développé une approche visant à identifier des associations qui pouvaient s'expliquer par la contamination par des types cellulaires autres que le monocyte. Ceci nous a permis de conclure que 4 des 11 associations trouvées dans GHS pouvait être imputables à la contamination par des types cellulaires autres que le monocyte présents à l'état de traces dans nos échantillons. Dans le chapitre 10, nous détaillerons les problèmes liés à la contamination, ainsi que des méthodes permettant de contrôler ce genre de biais et d'identifier la source des signaux d'association à partir d'un mélange composé de plusieurs types cellulaires.

Pour les 7 associations restant significatives après ajustement sur les variables de contamination, nous avons testé la réplication dans l'étude Cardiogenics. Seules les associations de 3 modules ont pu être répliquées. Ces modules contiennent à chaque fois un nombre restreint de gènes (128, 45 et 14 respectivement). Ceci suggère que les mécanismes de *trans*-régulation à grande échelle sont plus rares qu'initialement attendu, et/ou ont des effets faibles difficilement reproductibles du fait de la forte variabilité du transcriptome.

Parmi les associations SNP-pattern répliquées, le locus du gène RPS26 sur le chromosome 12q24 est particulièrement intéressant à plusieurs égards : A ce locus, le SNP rs11171739 situé entre les gènes ERBB3 et RPS26, a été trouvé associé à un risque accru de diabète de type I dans une analyse génome entier [101]. Ce même SNP est associé en *cis* avec l'expression de RPS26 ($p < 10^{-300}$, $R^2 = 80\%$) et de SUOX ($p = 7.7\,10^{-15}$, $R^2 \approx 4\%$). Il est de plus associé en *trans* à une dizaine de transcrits composant avec RPS26 et SUOX le module associé à ce locus. Si l'association avec RPS26 est bien connue et a fait l'objet de débats dans la littérature quant à son lien avec le diabète [63, 75], l'association de ce locus à un module en *trans* constitue ici un nouveau résultat. En réalité, l'étude des sondes associées en *trans* révèle que 6 d'entre-elles au moins ciblent des transcrits qui sont annotés comme étant des pseudo-gènes du gène RPS26 et pour lesquels on peut suspecter des phénomènes d'hybridation croisée. Parmi les autres transcrits du module, on trouve les gènes BEND4, DCFA16 et MADCAM1. Ce dernier gène présente un intérêt tout particulier puisqu'il a déjà été proposé comme gène candidat pour le diabète dans les modèles animaux et pourrait expliquer l'association de ce locus avec le diabète de type

17. Les plaquettes sont de petites cellules sanguines responsables entre autre de la coagulation. Les plaquettes, comme les globules rouges, ne contiennent pas de noyau. L'ARN n'y est donc présent qu'à l'état de traces provenant de la fragmentation des mégakaryocytes, précurseurs des plaquettes. Cependant l'importance du nombre de plaquettes dans le sang (400 000 plaquettes par mg de sang environ, contre seulement quelques milliers de monocytes), rend possible la contamination des échantillons par de l'ARN de plaquettes.

I. L'association de MADCAM1 avec le SNP rs11171739 est très homogène dans GHS et Cardiogenics. Toutefois, le niveau d'expression de MADCAM1 dans les monocytes est très faible et est même en dessous du seuil conventionel de détection dans Cardiogenics. Ce point rejoint les questions déjà discutées dans le chapitre 6 de la sensibilité des méthodes actuelles pour la détection des transcrits exprimés à des niveaux très faibles et illustre la pertinence de méthodes de filtrage des transcrits fondées sur le ratio signal/bruit plutôt que sur le niveau absolu du bruit de fond. Grâce aux nouvelles technologies de séquençage de l'ARN (RNA-seq), beaucoup plus sensibles et capables de détecter les transcrits à partir de quelques copies seulement, il sera possible de confirmer ou d'infirmer la présence du gène MADCAM1 dans les monocytes et l'effet du SNP rs11171739 sur son expression (ou son épissage). Les deux autres associations sont détaillées dans le manuscrit présent en annexe 4.

3 Comparaison avec l'approche WGCNA

Une autre approche pour mettre en évidence des groupes de gènes co-régulés a été proposée par Horvath *et al.* [102]. Nous rappelons ici le principe de cette approche et comparons les résultats obtenus sur GHS par cette approche avec les résultats obtenus par la méthode que nous proposons.

3.1 Principe général

L'approche WGCNA (Weighted Gene Correlation Network Analysis) proposée par Horvath est fondée sur une classification des gènes à partir d'une mesure de similarité reflétant la présence de voisins communs entre les gènes. Les gènes sont ainsi répartis entre un nombre réduit de groupes appelés "modules", par analogie avec l'approche précédente. Pour chaque module, un motif caractéristique des gènes du module est extrait en considérant le premier axe de l'ACP de la matrice d'expression des gènes du module. On nomme ce motif "pattern" par analogie avec la dénomination utilisée pour l'ACI.

3.1.1 Similarité entre les gènes

Dans un premier temps, on construit une mesure de similarité entre gènes s_{ij} basée sur la corrélation. Pour deux gènes $x_{i\cdot}$ et $x_{j\cdot}$, on prend donc

$$s_{ij} = |cor(x_{i\cdot}, x_{j\cdot})|^{\beta}$$

avec $\beta \geq 1$ un paramètre de "soft-thresholding" permettant de réduire le bruit et de favoriser les corrélations les plus fortes.

Dans un deuxième temps, une mesure d'adjacence a_{ij} est calculée en se fondant sur la topologie du réseau ("topological overlap") afin de renforcer la similarité entre transcrits ayant des voisins communs.

$$a_{ij} = \frac{\sum_{k \notin \{i,j\}} s_{ik} s_{kj} + s_{ij}}{\left(\sum_{k \neq i} s_{ik}\right) + \left(\sum_{k \neq j} s_{kj}\right) + 1 - s_{ij}}$$

3.1.2 Construction des modules

Afin de former des modules de gènes co-régulés, on effectue une classification ascendante hiérarchique sur la matrice d'adjacence. Dans ce but, on constitue autant de classes qu'il y a de gènes en notant A_i la classe numéro i. Puis, à chaque étape, on regroupe les deux classes les plus proches et on calcule la distance entre cette nouvelle classe et les autre classes par

$$d(A_i \cup A_j, A_k) = \frac{n_i d(A_i, A_k) + n_j d(A_j, A_k)}{n_i + n_j}$$

où n_i (resp. n_j) est le nombre de gènes présents dans la classe i (resp. j).

Le nombre de modules et la constitution des modules sont alors déterminés à partir du dendrogramme par une méthode d'élagage dynamique [103].

Pour chaque module, un pattern associé est ensuite dégagé en considérant la première composante principale de la matrice d'expression des gènes inclus dans le module.

Tout comme l'ACI, cette méthode peut être vue comme une décomposition matricielle $X \approx SA$, qui diffère par le choix des contraintes imposées sur la matrice S :

$$\begin{cases} \forall i, k & s_{ik} \in \{0, 1\} \\ \forall i & \sum_k s_{ik} = 1 \end{cases}$$

Cette méthode est implémentée dans le package R `WGCNA` que nous avons utilisé pour l'application aux données de GHS. Les différents paramètres de l'algorithme ont été fixés aux valeurs par défaut pour la comparaison avec l'ACI.

3.2 Application aux données de GHS

Sur GHS, la méthode WGCNA conduit à l'identification de 25 modules de gènes co-régulés dont la taille varie de 20 à 756 transcrits (ainsi qu'un module regroupant les 5635 gènes restants, pour lesquels la méthode ne parvient pas à isoler un motif spécifique). Parmi les patterns caractérisant ces modules, 23 (88%) présentaient une corrélation supérieure à 0.8 avec au moins un des facteurs extraits par l'ACI. A l'inverse, seulement 20 facteurs issus de l'ACI (31%) montraient une telle corrélation avec les motifs extraits par WGCNA, suggérant que l'ACI est capable d'identifier davantage de patterns que

WGCNA. Parmi les modules extraits, 11 (42%) étaient significativement enrichis en fonctions biologiques connues, contre 66% des 64 modules extraits avec l'ACI. La part de gènes partagés avec le module le plus proche extrait par l'ACI variait de 4% à 95% (médiane 57%). Une telle variabilité peut s'expliquer par le fait que les modules extraits par WGCNA sont généralement plus grands et présentent le plus souvent une forte hétérogénéité.

Nous avons ensuite testé l'association des 25 modules avec le génotype en utilisant le double critère utilisé pour l'ACI. Un seul des modules extraits par WGCNA était associé significativement au génotype. Le locus responsable de cette association est en fait le locus du gène ARHGEF3 déjà trouvé. Le module associé à ce locus était à 93% contenu dans le module correspondant trouvé par l'ACI. Afin de comparer la puissance globale des deux méthodes, on a calculé pour chaque SNP et chaque méthode d'extraction, une p-value d'association globale à partir des p-values d'association avec chacun des facteurs p_1,...,p_K donnée par

$$p_{globale} = 1 - (1 - \min_i(p_i))^K \tag{9.2}$$

Pour un SNP n'affectant aucun des facteurs latents, sous l'hypothèse d'indépendance entre les p_i, la p-value d'association globale suit une loi uniforme sur $[0, 1]$. On peut donc tracer et superposer les QQ-plots obtenus par les deux approches. On voit sur la figure 9.5 que les significativités obtenues avec l'ACI sont beaucoup plus importantes que celles obtenues avec la méthode WGCNA, ce qui suggère une puissance plus élevée de l'ACI pour détecter des associations entre SNP et groupes de gènes co-régulés.

4 Une application à l'étude du diabète de type I

Nous discutons dans cette partie d'une application de ces approches à la recherche de nouveaux loci de susceptibilité aux maladies complexes. Cette application a fait l'objet d'un article publié dans Nature [104] dans le cadre d'une collaboration avec le Royaume-Uni et l'Allemagne. Dans ce travail, j'ai effectué les analyses dans GHS et Cardiogenics. Ce travail montre la transposition chez l'homme d'un module de gènes *trans*-régulés chez le rat et associé au diabète de type I chez l'homme.

Ce module, identifié chez le rat, est centré sur le facteur de transcription IRF7 (Interferon Regulatory Factor 7) et contient les gènes cibles connus de ce facteur de transcription. Un module similaire est retrouvé chez l'homme dans GHS et Cardiogenics à la fois par l'ACI et par WGCNA. Parmi les gènes du module, on trouve un enrichissement en gènes liés à la réponse immunitaire et en particulier à la réponse aux virus. D'une façon générale, l'analyse des gènes du module mettait en évidence la présence d'un grand nombre de gènes impliqués dans la réponse à la stimulation par l'interféron. Chez le rat, un locus

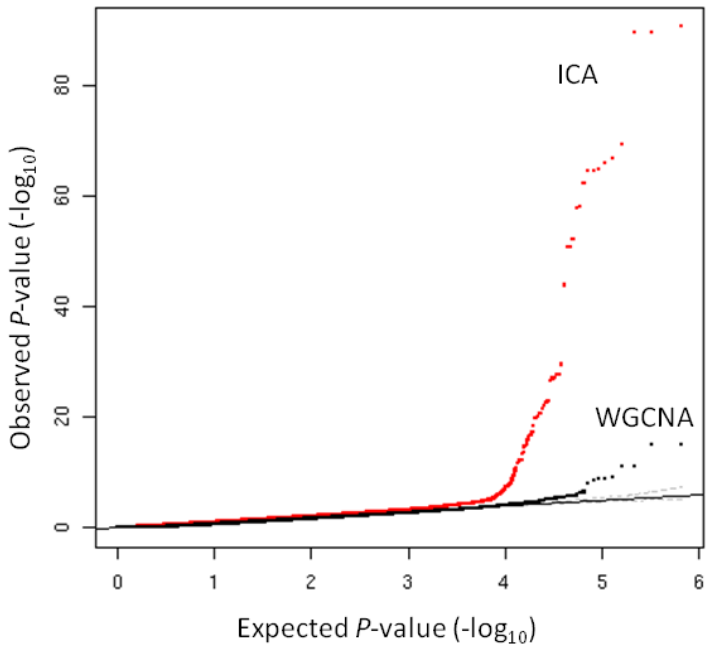


FIGURE 9.5 – **QQ-plot comparant les p-values d'association des 675 350 SNP avec les patterns extraits par l'ACI (rouge) et WGCNA (noir)** . Pour chaque SNP, la meilleure p-value obtenue sur les 26 modules de WGCNA et les 64 modules de l'ACI est tracée. Pour chaque méthode, une correction est appliquée pour tenir compte du nombre de patterns testés ainsi qu'indiqué par l'équation 9.2

situé sur le chromosome 15q25 autour du gène EBI2 (Epstein Barr virus Induced 2) était associé à une régulation en *trans* de l'expression des gènes du module. Ce gène code pour un récepteur à la protéine G contrôlant la migration des lymphocytes B dans les tissus et est très fortement exprimé dans les macrophages. Les auteurs de l'article ont montré que chez le rat, l'inhibition du gène EBI2 par des ARN interférents [18] augmentait l'expression du gène IRF7 dans les macrophages, suggérant que le gène EBI2 joue un rôle d'inhibiteur de la réponse immunitaire innée dans les macrophages.

Le locus 15q25 du rat est orthologue au locus 13q32 chez l'homme. Nous avons donc cherché à vérifier si le locus 13q32 était associé à l'expression des gènes du module dans le monocyte à partir des données de GHS et Cardiogenics. L'association du module avec le locus EBI2 a été répliquée dans Cardiogenics ($p = 10^{-4}$) mais pas dans GHS. Cette différence pourrait s'expliquer par la méthode de sélection des monocytes, qui diffère entre les deux études.

Les auteurs ont ensuite postulé que le module de gènes autour d'IRF7 pouvait être impliqué dans le diabète de type I. Le diabète de type I est en effet une maladie auto-immune dans laquelle les cellules β du pancréas (responsables de la production de l'insuline) sont attaquées par le système immunitaire [19] du patient, conduisant à une mauvaise régulation de la glycémie. Son apparition peut être déclenchée par un emballement de la réponse immunitaire suite à l'infection par certains virus [105]. Les auteurs ont montré que le module IRF7 était enrichi en gènes trouvés associé au diabète par GWAS ($p = 2.5\,10^{-10}$). Ils ont également trouvé qu'un SNP du locus 13q32 était associé ($p < 10^{-7}$) à un risque accru de développer un diabète de type I [106]. Ces résultats suggèrent donc un effet du gène EBI2 sur le diabète de type I passant par la régulation en *trans* dans les macrophages, du facteur de transcription IRF7 et du module des gènes liés à l'interféron (qui incluent majoritairement des cibles directes et indirectes du gène IRF7). Ce travail montre également l'intéret des grandes études transcriptomiques comme GHS et Cardiogenics pour tester chez l'homme des hypothèses physiopathologiques générées dans des modèles animaux.

18. Petites molécules d'ARN s'hybridant aux ARNm du gène pour inhiber sa traduction en protéine ou favoriser la dégradation des ARNm.

19. Lymphocytes T, B et macrophages.

CHAPITRE 10

Déconvolution de signaux dans un mélange de types cellulaires

Nous avons évoqué dans le chapitre précédent, le cas du locus ARHGEF3 dont l'effet sur le transcriptome a permis de mettre en évidence un problème lié à la contamination des ARN monocytaires par de l'ARN de plaquettes dans l'étude GHS. Suite à cette découverte, j'ai travaillé au développement de méthodes visant à estimer avec précision les niveaux de contamination par d'autres types cellulaires que le monocyte dans les données de GHS et à dissocier les signaux d'association propres aux monocytes, de signaux d'association provenant de types cellulaires contaminants. Nous développerons dans cette partie les résultats de ce travail.

Nous reviendrons tout d'abord sur le phénomène de contamination observé dans GHS, en décrivant les mécanismes à l'origine de ce phénomène et de quelle façon il influence les résultats. Nous détaillerons ensuite une méthode de "déconvolution"[1] présentée par Abbas [107] pour estimer les quantités dans un mélange composé de plusieurs types cellulaires distincts. Nous verrons comment l'application de cette méthode permet d'identifier une contamination des données et de s'en servir comme variable d'ajustement dans les analyses. Nous étudierons enfin la possibilité de corriger a posteriori les biais liés à la contamination et d'identifier la provenance des signaux dans un mélange cellulaire à partir de l'estimation faite des niveaux de contamination avant de discuter les limites de cette approche.

1. Le terme de déconvolution est ici emprunté à Abbas pour désigner la séparation d'éléments à partir de l'observation de leur somme. Son emploi se justifie par le fait que la densité d'une somme de variables aléatoires peut s'écrire comme le produit de convolution des densités de ces variables aléatoires. Il s'agit néanmoins d'un abus de langage puisque les méthodes utilisées ici, ne font pas appel directement à la modélisation de la densité des signaux issus des différents types cellulaires et cherchent seulement à estimer les proportions respectives de ces types cellulaires.

1 La contamination dans l'étude GHS

1.1 L'origine de la contamination

Afin de bien comprendre le phénomène de contamination observé dans GHS, il est essentiel de se pencher plus en détail sur la composition des échantillons à partir desquels est extrait l'ARN. Dans ce but, nous décrivons d'abord les différents types cellulaires présents dans le sang, avant de détailler les méthodes de purification cellulaire utilisées pour l'isolation des monocytes.

1.1.1 Composition du sang

Le sang est composé d'éléments cellulaires flottant dans un milieu liquide appelé plasma et composé d'eau, de solutés minéraux (O_2, CO_2, divers ions,...), de nutriments (lipides, acides aminés, glucides) et de protéines en suspension (albumine, fibrinogène, cholestérol, ...). Les élements cellulaires contenus dans le sang sont générés par le processus de l'hématopoïèse présenté dans la figure 10.1. Parmi ces éléments on trouve :

- des **globules rouges** (ou érythrocytes), cellules sans noyau responsables du transport de l'oxygène. On compte entre 4 et 5 millions de globules rouges par ml de sang.
- des **plaquettes** (ou thrombocytes), également sans noyau et impliquées dans le processus de coagulation et de réparation de la paroi artérielle. On compte près de 400 000 plaquettes par ml de sang.
- des **globules blancs** (ou leucocytes) assurant la réponse immunitaire contre les agents pathogènes. On dénombre entre 5000 et 7000 globules blancs par ml de sang parmi lesquels on trouve :
  - 50 à 75% de granulocytes. Ces cellules polynucléaires sont impliquées dans la réponse immunitaire non spécifique et participent à diriger celle-ci par l'exocytose de granules contenant des agents pro-inflammatoires (comme l'histamine) ou anti-inflammatoires.
  - 20 à 40% de lymphocytes répartis entre les lymphocytes T, les lymphocytes B et les cellules tueuses ("Natural Killer Cells" an anglais). Les lymphocytes sont impliqués dans la réponse immunitaire spécifique (lymphocytes T et B) et non spécifique (cellules tueuses et certains lymphocytes T), et agissent soit par présentation d'antigènes cytotoxiques conduisant à l'apoptose [2] des cellules infectées (lymphocytes T et cellules tueuses), soit par la libération d'anticorps (lymphocytes B) qui se fixent sur les antigènes et participent à la réponse immunitaire.

2. Mort cellulaire programmée.

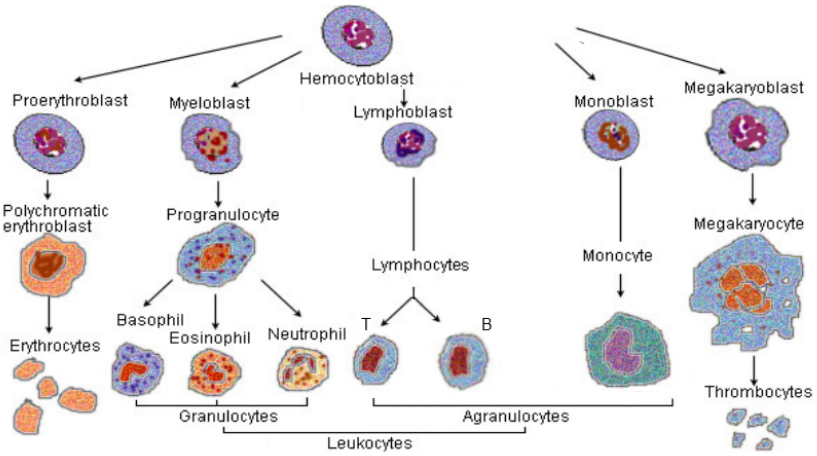


FIGURE 10.1 – **L'hématopoïèse.**

- 3 à 8% de monocytes, qui sont des cellules immunitaires et inflammatoires dont le rôle est détaillé dans le chapitre 2 (section 2.2). C'est le type cellulaire ciblé dans les études GHS et Cardiogenics.

Lorsque nous nous intéressons aux monocytes circulants comme c'est le cas dans GHS, il est donc nécessaire de séparer les différents types cellulaires présents dans le sang par des méthodes de purification. Lorsque cette séparation ne se fait pas parfaitement, les échantillons peuvent donc se trouver contaminés par des types cellulaires sanguins non désirés.

1.1.2 Méthodes de purification cellulaire

L'isolation d'un type cellulaire se fait généralement en deux étapes :

- Dans un premier temps on introduit dans les échantillons de sang complet un réactif nommé Ficol. Ce réactif permet de séparer les différents composants du sang en fonction de leur densité, lors de la centrifugation. Le sang se décompose alors en 4 phases distinctes
  - le plasma et les plaquettes,
  - le "buffy coat" qui contient les cellules mononucléaires (lymphocytes et monocytes),
  - le Ficol,
  - les granulocytes et globules rouges qui sont piégés par le Ficol.
- On effectue ensuite une purification (ou tri cellulaire) des cellules mononucléaires contenues dans le "buffy coat". Cette purification peut se faire selon deux techniques illustrées par la figure 10.2 :
  - Dans la sélection positive, on utilise des billes magnétiques auxquelles sont fixés

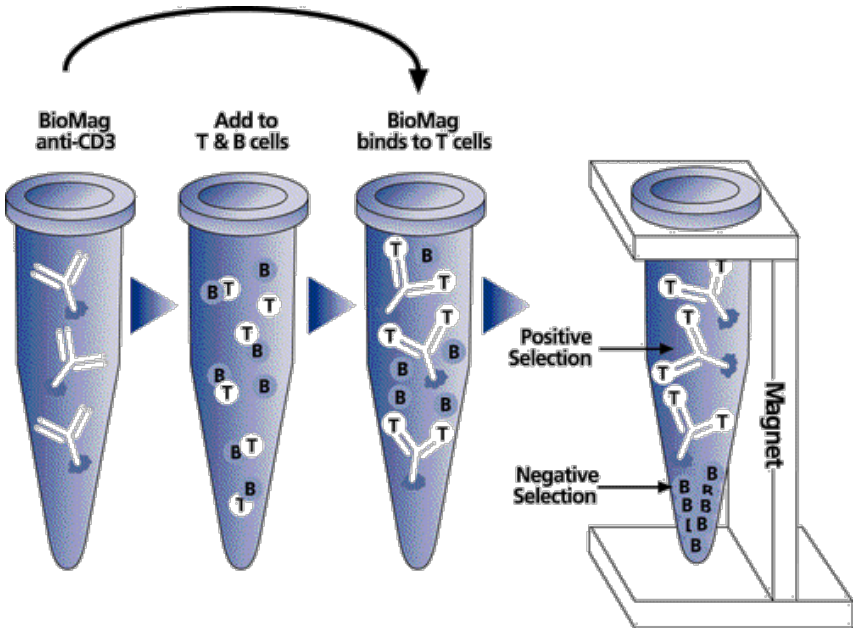


FIGURE 10.2 – **Principe simplifié du tri cellulaire par sélection positive et négative** (*Source :* The Scientist [108]).

des anticorps spécifiques des marqueurs membranaires spécifiques des cellules ciblées (dans le cas des monocytes on utilise des anticorps spécifiques du marqueur CD14). Les cellules présentant le marqueur membranaire ciblé sont ensuite extraites à l'aide d'un électro-aimant. L'inconvénient de cette méthode est que la liaison des marquers membranaires aux billes magnétique peut activer des voies de signalisation et par conséquent modifier le transcriptome de la cellule ciblée.

– Dans la sélection négative, on utilise des marqueurs ciblant les principaux types cellulaires présents à savoir des lymphocytes T (CD2, CD3, CD8), B (CD19) et NK (CD2, CD8, CD56), ainsi que des marqueurs spécifiques des globules rouges (Glycophorin A) et des granulocytes (CD66b). On retire ensuite les types cellulaires ciblés à l'aide d'un électro-aimant comme lors d'une sélection positive. Les résidus présents dans la solution après extraction des types cellulaires non désirés forment alors une solution enrichie en monocytes, mais dans laquelle peuvent néanmoins rester des traces des autres types cellulaires.

Le choix d'une de ces deux méthodes est un choix souvent discuté dans la littérature [109,110] et pour lequel il n'existe pas de consensus à ce jour. Dans GHS, c'est l'approche par sélection négative qui a été choisie afin d'éviter l'activation des monocytes par la liaison d'anticorps sur leurs marqueurs membranaires. Des contrôles préliminaires ont montré que les monocytes représentaient entre 94% et 99% des cellules présentes dans la solution enrichie, suggérant un bon fonctionnement de la méthode de tri utilisée, le reste étant principalement des plaquettes. Malgré cela, comme nous le verrons dans les

sections suivantes, l'analyse du transcriptome suggère une réalité assez différente et montre clairement que la présence de contaminants, même dans des proportions infimes, peut fortement influencer les résultats des analyses d'association [3]. Bien que le phénomène de contamination semble moins important dans des données sélectionnées par tri positif (en particulier la contamination par les plaquettes), ce phénomène ne peut pas être exclu comme nous le verrons plus loin.

1.2 Impact de la contamination

Dans la suite nous considérons les données d'expression comme provenant d'un mélange de plusieurs types cellulaires sanguins présents en des quantités variables d'un échantillon à l'autre. On note q_{kj} la quantité du k^{e} type cellulaire chez le j^{e} individu, avec $q_{1\cdot}$ le type cellulaire majoritaire (ici les monocytes). Soit une variable d'intérêt notée z dont on souhaite tester l'association avec les expressions (par exemple un SNP). La contamination peut biaiser les résultats par deux mécanismes différents (et non exclusifs) :

1. La variable z influence la proportion des différents types cellulaires présents dans le mélange : si un transcrit a des niveaux d'expression différents selon le type cellulaire, la variation des proportions due à la variable z induit une association entre l'expression du transcrit et la variable z. Dans ce cas les proportions des différents types cellulaires estimés agissent comme des facteurs confondants et leur effet peut être retiré par un ajustement simple sur les proportions du mélange $(q_{k\cdot})_{k>1}$.

2. La variable z influence le niveau du transcrit dans un type cellulaire autre que le type majoritaire : l'effet de la variable z sur l'expression peut être observé dans le mélange si le type cellulaire affecté par cette variable est présent en quantité suffisante. Dans ce cas, si z affecte le type cellulaire k, on s'attend à observer un effet de z d'autant plus fort que le type cellulaire k est présent en grande quantité dans le mélange, ce qui se traduira par une interaction entre la quantité du type cellulaire k et la variable étudiée z. Ainsi l'ajout dans le modèle de régression, des termes d'interactions entre z et les proportions du mélange $(q_{k\cdot})_{k>1}$ permet de contrôler pour un éventuel effet de z dans les types cellulaires contaminants, voire de déterminer les types cellulaires dans lesquels le niveau d'expression du transcrit est affecté par la variable z.

Cette approche rejoint celle proposée par Shen-Orr *et al.* [111] qui consiste à tirer parti de comptes cellulaires réalisés par cytométrie de flux pour identifier la provenance des signaux d'association dans un mélange de types cellulaires.

3. Et ce, même en présence d'un design d'expérience équilibré et randomisé.

2 Estimation des proportions des différents types cellulaires

Dans notre cas, les proportions du mélange ne sont pas connues a priori, il est donc nécessaire de les estimer afin de pouvoir inclure ces termes dans le modèle étudié. Dans cette section, nous présentons un algorithme de déconvolution décrit par Abbas [107] pour estimer les proportions des différents types cellulaires présents dans un mélange et évaluons l'efficacité de cette méthode.

2.1 Méthode utilisée

Soit un mélange de K types cellulaires différents. On suppose ici qu'on est capable d'isoler par sélection positive les divers types cellulaires présents dans le mélange pour en mesurer les profils moyens d'expression $\boldsymbol{\psi}_{.k}$. On peut alors utiliser ces profils pour estimer les quantités $q_{k.}$ des différents types cellulaires du mélange.

Une première approche que j'ai développée consiste à identifier pour chaque type cellulaire un sous-ensemble S_k de gènes caractéristiques de ce type cellulaire tel que le niveau des gènes appartenant à S_k dans les types cellulaires autres que le type k soit négligeable par rapport au niveau observé de ces gènes dans le type k. On considère ensuite le niveau moyen des gènes du sous-ensemble S_k dans l'échantillon i comme un indicateur de la quantité du type k dans cet échantillon. Dans GHS, j'ai dans un premier temps appliqué cette approche en utilisant des listes de gènes considérés dans HaemAtlas comme "spécifiques" des différents types cellulaires sanguins [112]. Bien que cette approche ait permis de mettre en évidence les effets de la contamination dans GHS [4], elle demeure imparfaite pour plusieurs raisons :

- Elle suppose qu'il est possible de trouver des transcrits qui soient spécifiques de chaque type cellulaire. Or, on observe généralement de très fortes corrélations des niveaux des transcrits entre les différents types cellulaires sanguins et la majeure partie des transcrits considérés ici comme spécifiques d'un type cellulaire sont en réalité exprimés à des niveaux parfois élevés dans d'autres types cellulaires. Cette approche a donc tendance à surestimer la corrélation entre les niveaux de contamination.
- Cette méthode ne fournit qu'un indicateur relatif de la quantité des différents types cellulaires entre les individus et ne permet pas d'estimer le niveau absolu de contamination des échantillons.

4. C'est cette approche qui a été retenue dans l'article traitant de l'extraction de modules co-régulés pour identifier les patterns liés à la contamination.

Afin de corriger ces défauts, j'ai donc adapté l'approche proposée par Abbas [107] à l'étude des données de GHS. Cette approche consiste à écrire la matrice d'expression X comme un produit matriciel

$$X \approx \Psi.Q$$

où Ψ est la matrice des profils d'expressions $\boldsymbol{\psi}_{.k}$ des différents types cellulaires et Q est la matrice des quantités $q_{k.}$. On applique une procédure en 2 étapes :

1. On estime les profils d'expression $\boldsymbol{\psi}_{.k}$ de chaque type cellulaire, à partir des atlas d'expression publics décrits dans le chapitre 2.
2. On utilise la matrice des profils estimés $\widehat{\Psi}$ pour estimer les quantités Q par des estimateurs des moindres carrés ordinaires du type :

$$\widehat{Q} = (\widehat{\Psi}'\widehat{\Psi})^{-1}\widehat{\Psi}X$$

Pour la première étape qui consiste à estimer les profils $\boldsymbol{\psi}_{.k}$, nous avons utilisé les données de HaemAtlas [112]. nous avons tout d'abord appliqué aux données une normalisation par quantile afin de faire coïncider leur distribution avec celle des données de GHS. Puis, les profils des différents échantillons ont été moyennés pour chacun des types cellulaires considérés (4 à 7 échantillons par type cellulaire dans HaemAtlas), afin d'obtenir les profils types $\widehat{\boldsymbol{\psi}_k}$. Ces profils peuvent être estimés soit sur l'ensemble des sondes, soit à partir d'un sous-ensemble représentatif de sondes de taille $p' < p$. On applique ensuite l'algorithme de déconvolution à la matrice $\widehat{\Psi}$ formée.

Notons que l'étape de normalisation par quantiles a pour conséquence de forcer les quantités $\widehat{q_{kj}}$ estimées pour chaque individu à sommer à 1. En effet, pour deux matrices Q et $Q' = \theta.Q$, on a

$$X' = \Psi.Q' = \theta\Psi.Q = \theta X\,.$$

d'où

$$\widehat{\Psi}' = \theta\widehat{\Psi}$$

du fait de la normalisation par quantile et

$$\widehat{Q} = \widehat{Q'}\,.$$

On parlera donc dans la suite indifféremment de proportions ou de quantités pour désigner les estimations des q_{kj}.

On estime dans un deuxième temps les proportions Q des différents types cellulaires à partir des profils $\widehat{\Psi}$. Dans ce but, on impose, ainsi que proposé, des contraintes de positivité sur les quantités estimées par un processus itératif :

1. Estimer pour chaque échantillon $x_{.j}$ les quantités des différents types cellulaires

$$\widehat{q_{.j}} = (\widehat{\Psi}'\widehat{\Psi})^{-1}\widehat{\Psi}x_{.j}$$

2. Si certaines quantités sont négatives :
   (a) fixer à zéro la quantité du type cellulaire dont la quantité estimée $\widehat{q_{kj}}$ est minimale
   (b) répéter l'estimation pour les autres quantités $\widehat{q_{kj}}$
3. Répéter l'opération précédente jusqu'à ce que toutes les quantités du vecteur $q_{.j}$ soient positives ou nulles

2.1.1 Choix du sous-ensemble de sondes utilisé pour la déconvolution

Lorsqu'on estime les quantités des différents types cellulaires du mélange, une question cruciale concerne le choix du sous-ensemble de sondes utilisé pour la déconvolution :

En effet, l'algorithme de déconvolution peut être appliqué de deux façons :

- En utilisant l'intégralité des sondes communes aux puces d'expression utilisées dans GHS (Illumina HT-12) et HaemAtlas (Illumina WG6), soit plus de 35 000 sondes.
- En se restreignant à un sous-ensemble S de sondes considérées comme spécifiques des différents types cellulaires.

Dans la pratique, la qualité de la déconvolution obtenue est fonction du conditionnement de la matrice $\Psi'\Psi$ [107].

Le conditionnement κ d'un système linéaire $Ax = b$, où x représente l'inconnue, A est une matrice inversible et b un vecteur est défini par

$$\kappa(A) = ||A^{-1}||_2||A||_2 = \left|\frac{\lambda_{max}}{\lambda_{min}}\right|$$

où $||.||_2$ désigne la norme $\ell 2$ et λ_{max} et λ_{min} désignent les valeurs propres maximales et minimales de la matrice A. Ce nombre fournit une mesure de la stabilité de la solution de ce système. En effet, lorsqu'on introduit une perturbation ΔA sur la matrice A, l'erreur relative théorique commise sur x, $\frac{||\Delta x||_2}{||x||_2}$ est majorée par

$$\frac{||\Delta x||_2}{||x||_2} \leq \kappa(A)\frac{||\Delta A||_2}{||A||_2}\,.$$

Un conditionnement élevé indique ainsi une forte instabilité de la solution du système linéaire tandis qu'un conditionnement proche de 1 indique que l'erreur relative commise sur x est du même ordre de grandeur que l'erreur relative commise sur A. Le conditionnement de la matrice $\Psi'\Psi$ reflète donc ici la robustesse de la méthode de déconvolution utilisée à une erreur sur les profils Ψ estimés.

Lorsqu'on calcule le conditionnement à partir de l'ensemble des sondes, on trouve une valeur de κ égale à 30. En revanche, l'utilisation d'un sous-ensemble S de sondes spécifiques, constitué de l'union sur les K types cellulaires des sous-ensembles S_k de

sondes spécifiques [5] de chaque type cellulaire, permet d'atteindre une conditionnement de 14, ce qui suggère une amélioration conséquente de la stabilité de la déconvolution.

2.1.2 Evaluation de la qualité de l'estimation

Afin d'évaluer la capacité de la méthode utilisée à distinguer les différents types cellulaires, nous avons effectué une étape de validation croisée de la méthode de type "leave-one-out" sur les échantillons de HaemAtlas.

Pour chaque échantillon de HaemAtlas, on calcule la matrice $\widehat{\Psi_{(-j)}}$ en retirant l'échantillon j des données d'HaemAtlas, avant le calcul des profils transcriptomiques $\widehat{\boldsymbol{\psi}_{.k}}$, et on estime la proportion du type cellulaire j par

$$\widehat{q_{.j}}^{VC} = (\widehat{\Psi_{(-j)}}' \widehat{\Psi_{(-j)}})^{-1} \widehat{\Psi_{(-j)}} x_{.j}$$

On peut alors calculer le risque quadratique moyen en prenant :

$$RQM = \frac{1}{nK} \sum_{j=1}^{n} \sum_{k=1}^{K} \left(\widehat{q_{kj}}^{VC} - q_{kj} \right)^2$$

avec q_{kj} la proportion du type cellulaire k dans l'échantillon j qui vaut 1 si l'échantillon j est un échantillon du type k et 0 sinon, et n le nombre total d'individus dans HaemAtlas.

On observe que le risque quadratique moyen obtenu par les deux méthodes est assez semblable et relativement faible, avec une valeur de $4.\,10^{-4}$ dans le cas où on se restreint aux gènes spécifiques d'un type cellulaire et $3.\,10^{-4}$ dans le cas où l'on utilise tous les gènes. Dans la suite, on choisit d'utiliser la méthode basée sur le sous-ensemble de gènes spécifiques qui présente une plus forte robustesse aux erreurs d'estimation de profils Ψ.

La figure 10.3 montre les résultats obtenus par la procédure de validation croisée pour les différents types cellulaires. On constate sur cette figure que la classification des échantillons marche globalement bien, à l'exception des lymphocytes T cytotoxiques (CD8+) pour lesquels on estime une proportion non nulle de lymphocytes CD4+ et de cellules tueuses CD56+.

En fait, ce résultat s'explique en partie par de très fortes similarités fonctionnelles entre les cellules tueuses et certains sous-types de lymphocytes T. De plus, la littérature mentionne la présence d'antigène CD8 sur certains sous-types de cellules tueuses [113,114], ce qui suggère que l'isolation de lymphocytes par marquage de l'antigène CD8 manque probablement de spécificité. Pour cette raison, nous avons dans la suite regroupé les différents sous-types de lymphocytes T et les cellules tueuses en une catégorie unique "Lymphocytes T".

5. Les sondes spécifiques du type k sont identifiées dans l'article original de Watkins *et al.* [112] en sélectionnant les sondes significativement sur-exprimées dans ce type cellulaire par rapport aux 5 autres types cellulaires par un critère de FDR à 5% et pour lesquelles le niveau d'expression dans le type cellulaire k est au moins le double du niveau dans les 5 autres types cellulaires.

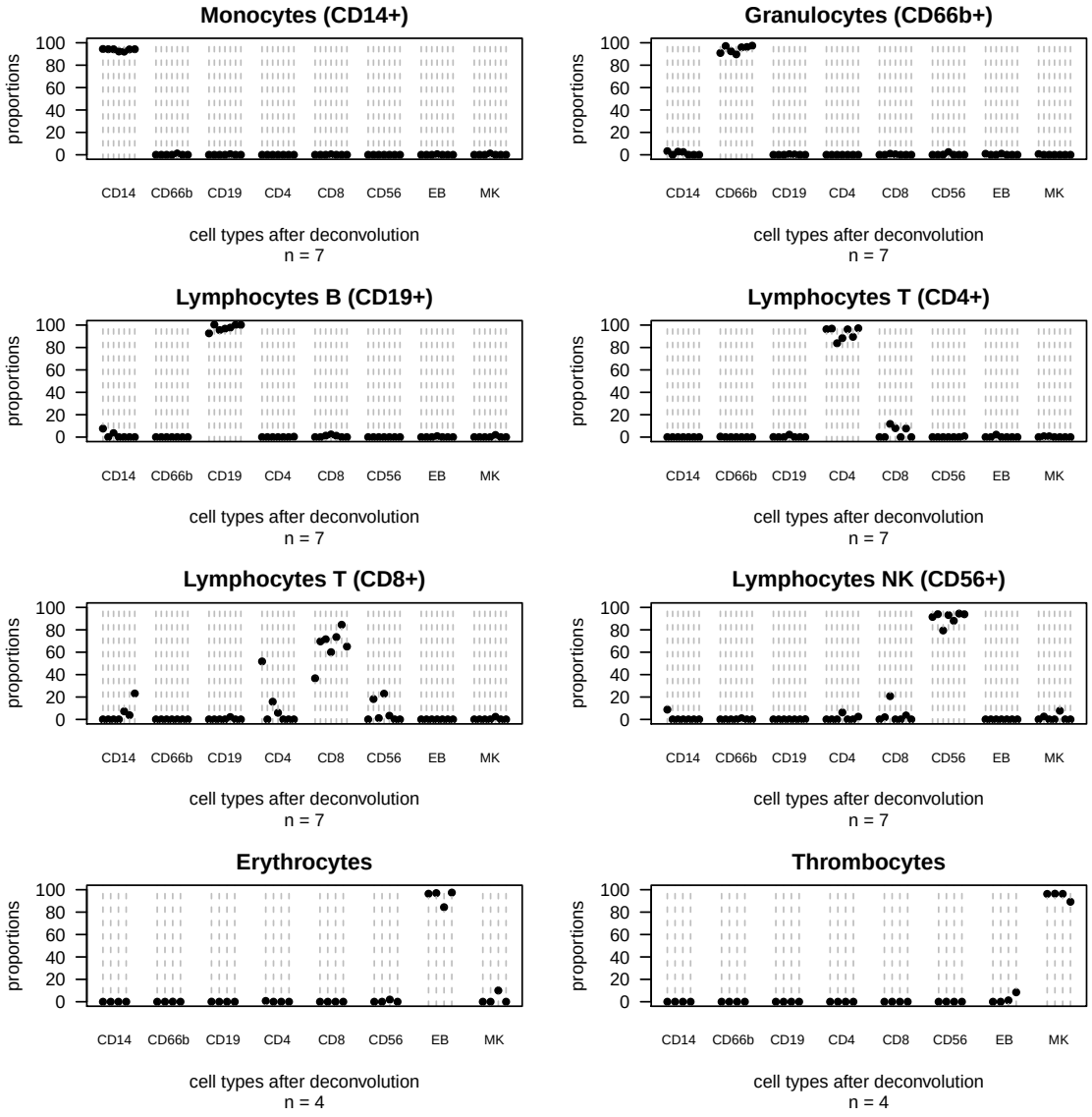


FIGURE 10.3 – **Proportions des différents types cellulaires estimées par la méthode de déconvolution dans les échantillons de HaemAtlas :** Chaque échantillon est supposé refléter un seul type cellulaire.

2.2 Application aux données de GHS

2.2.1 Estimation des proportions des différents types cellulaires

Lorsqu'on applique cette approche de déconvolution aux données de GHS, on trouve que l'ARN monocytaire représente en moyenne 81% de l'ARN présent dans les échantillons. On trouve également 8% d'ARN provenant des plaquettes, et 6% d'ARN issu de lymphocytes T et NK en moyenne. Le reste est réparti entre les granulocytes, les érythrocytes et les lymphocytes B (tableau 10.1).

	GHS	Cardiogenics
	Moyenne (écart-type)	Moyenne (écart-type)
Monocytes (CD14+)	81,1% (3,8%)	92,1% (4,6%)
Lymphocytes B (CD19+)	2,8% (2,7%)	0,1% (0,6%)
Granulocytes (CD66b+)	1,5% (2,1%)	5,9% (5,5%)
Erythrocytes	0,6% (1,3%)	1,2% (0,9%)
Thrombocytes	8,0% (3,8%)	0,2% (0,4%)
Lymphocytes T-NK (CD4+, CD8+, CD56+)	6,0% (4,6%)	0,4% (1,9%)

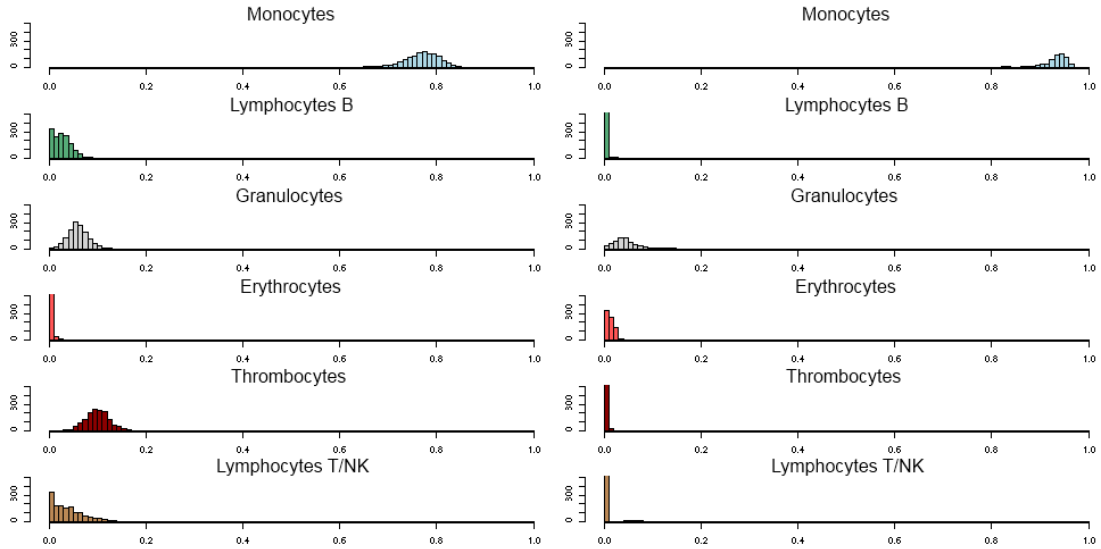
Tableau 10.1 – **Proportions estimées des différents types cellulaires dans GHS et Cardiogenics.**

En comparaison, dans l'étude Cardiogenics, où les monocyts ont été extraits par sélection positive, on estime à plus de 90% la part d'ARN monocytaire, et la majorité des échantillons présentent des niveaux plus faibles de contamination. Les figures 10.4a et 10.4b montrent les distributions des quantités des différents types cellulaires dans les études GHS et Cardiogenics. A l'exception de la contamination par les granulocytes qui semble être présente de façon systématique, la majeure partie des types cellulaires ne sont détectés qu'à l'état de traces (moins de 2%) dans la majorité des échantillons de Cardiogenics. L'apparente contamination des échantillons de Cardiogenics par les granulocytes peut sembler surprenante du fait de la méthode de tri utilisée. Cette observation peut s'expliquer soit par une fixation non spécifique des granulocytes aux anticorps CD14 (par présentation d'un antigène similaire, ou par l'existence d'une sous-population de granulocytes exhibant le marqueur CD14) soit par l'activation au sein des monocytes, sous certaines conditions, de pathways habituellement caractéristiques des granulocytes. Toutefois, la corrélation positive observées entre la quantité estimée de granulocytes et les comptes de neutrophiles [6] obtenus par cytométrie de flux ($\rho = 0,18$, $p = 3.0\,10^{-6}$) conduit à favoriser la première hypothèse. Dans GHS, on voit que les types cellulaires contami-

6. Qui représentent la grande majorité des granulocytes.

nants sont présents de façon relativement homogène dans les échantillons, ce qui est une conséquence du caractère moins spécifique de la méthode de tri cellulaire choisie.

On trouve donc comme attendu initialement, une pureté plus importante des monocytes extraits par sélection positive par rapport aux monocytes extraits par sélection négative, sans toutefois atteindre une pureté complète lorsqu'on procède par sélection positive. Ce gain de pureté est toutefois contre-balancé par de l'activation de la voie CD14 induite par la sélection positive.



(a) GHS (b) Cardiogenics

FIGURE 10.4 – **Distribution des quantités estimées des différents types cellulaires dans les données de GHS et Cardiogenics :** les quantités sont estimées par la méthode de déconvolution proposée par Abbas [107] en utilisant les profils d'expression de HaemAtlas.

2.2.2 Fonctions biologiques liées aux différents types cellulaires contaminants

Nous avons identifié pour chaque chaque type cellulaire les transcrits significativement corrélés à la contamination par ce type cellulaire. Nous avons ensuite testé l'enrichissement de ces transcrits en fonctions biologiques connues, afin de vérifier la cohérence des fonctions biologiques trouvées avec le type cellulaire considéré. On s'attend en effet à ce que les gènes dont l'expression est la plus corrélée à la présence d'un type cellulaire donné soient les gènes spécifiques de ce type cellulaire et donc que leur fonction biologique soit pertinente vis-à-vis de ce type cellulaire.

On sélectionne donc pour chaque type cellulaire, les gènes pour lesquels la contamination par ce type cellulaire explique plus de 10% de la variabilité totale de l'expression dans GHS et on calcule les enrichissements par les pathways Gene Ontology sur ces listes

de gènes. Cette analyse montre que la contamination par les plaquettes est significativement associée à des fonctions telles que la coagulation ($p = 1.8\,10^{-11}$), le nucléosome ($p = 3.3\,10^{-10}$), et l'activation des plaquettes ($p = 5.9\,10^{-9}$), ou que la contamination par les lymphocytes T est liée à la réponse immunitaire ($p = 4.0\,10^{-9}$), l'activité des récepteurs membranaires ($p = 1.1\,10^{-6}$) et la défense cellulaire ($p = 3.3\,10^{-7}$). On trouve également un enrichissement des gènes liés la régulation de la prolifération des lymphocytes B parmi les gènes corrélés à la contamination par les lymphocytes B ($p = 0.001$) et des enrichissements en fonctions liées à l'hémoglobine et aux érythrocytes ("différenciation des érythrocytes" : $p = 4.0\,10^{-5}$, "processus métabolique de l'hémoglobine" $p = 5.3\,10^{-5}$) parmi les gènes corrélés à la quantité estimée de globules rouges.

Ces résultats suggèrent donc que la méthode d'estimation des contaminations à partir des différents profils d'expression donne des résultats cohérents avec les connaissances biologiques a priori.

3 Impact de la contamination sur les associations du transcriptome avec le génotype

3.1 Influence des loci associés aux patterns d'expression sur la contamination

Pour chaque locus trouvé associé à un pattern d'expression dans le chapitre 9, nous avons testé si ce locus était associé au niveau de contamination par des types cellulaires autres que le monocyte. Pour 4 des 11 SNP initialement trouvés associés à un module de co-expression, une association avec un des types cellulaires contaminants a pu être mise en évidence (tableau 10.2). Nous détaillons ici chacune des associations trouvées et proposons des éléments d'interprétation. Pour chaque locus, seul le contaminant le plus associé est décrit.

3p21 Ainsi que déjà décrit dans la section 2.2.2 du chapitre 9, le locus 3p21 qui contient le gène ARHGEF3 a été trouvé associé dans des GWAS avec le volume plaquettaire moyen et la quantité de plaquettes [100, 115]. Dans GHS ce locus est associé à un module d'expression contenant un grand nombre de gènes sur-exprimés dans les plaquettes et fortement enrichi en fonctions liées aux plaquettes (coagulation : $p = 1,3\,10^{-12}$, hémostase : $p = 2,6\,10^{-11}$). Dans GHS, l'allèle mineur du SNP rs12485738 présent à ce locus, qui a été associé dans les GWAS à une augmentation du volume plaquettaire et une diminution du nombre de plaquettes est associé à une diminution de la contamination par les plaquettes ($p = 1.3\,10^{-19}, R^2 = 5,5\%$). Cette association pourrait s'expliquer par une meilleure séparation des plaquettes

SNP associé à un pattern	Locus	Gènes au locus	*P*-value de l'association SNP-pattern	Nombre de gènes dans le module associé	Fonctions biologiques associées au module	Association significative du SNP à un type cellulaire contaminant (type cellulaire)	Meilleure *p*-value d'association du SNP avec les quantités de types cellulaires contaminants.
rs2300573	1q24.2	*TBX19*	3.15 E-08	176	réponse aux virus	Non	0.08
rs13023213	2p12	*CD8A*	7.52 E-08	292	activation des lymphocytes T, défense cellulaire	Oui (Lymphocytes T)	1.3E-06
rs12485738	3p21	*ARHGEF3*	8.76 E-24	379	coagulation, hémostase	Oui (Plaquettes)	1.3 E-19
rs1344142	3p21	*ARHGEF3*	1.48 E-18	135	coagulation, hémostase	Oui (Plaquettes)	2.4 E-20
rs13196564	6q15	*MAP3K7*	5.24 E-08	311	mitose, cycle cellulaire	Oui (Lymphocytes B)	5.6 E-06
rs2842892	6q23.1	*STX7*	9.40 E-08	137	–	Non	.34
rs12705417	7q21	*MAGI2*	1.23 E-08	395	hémoglobine, transport de l'oxygène	Oui (Erythrocytes)	3.9 E-08
rs1058348	10p13	*CUGBP2*	3.49 E-08	189	–	Non	0.04
rs653178	12q24	*ATXN2, SH2B3*	2.36 E-09	62	–	Non	0.005
rs11177644	12q15	*LYZ, YEATS4*	1.14 E-92	45	–	Non	0.22
rs11171739	12q13	*RPS26, SUOX*	2.89 E-70	14	–	Non	0.17

Tableau 10.2 – **Contamination et loci associés à des modules de gènes co-régulés :** Pour chaque locus associé à un module de gènes co-régulés dans GHS, on rapporte le SNP le plus associé au pattern extrait par l'ACI à ce locus, les gènes situés à ce locus, la *p*-value d'association du meilleur SNP au pattern, le nombre de gènes identifiés comme faisant partie du module associé, et les fonctions biologiques associées à ce module. On indique ensuite si une des quantités estimées des différents types cellulaires est significativement associée à ce locus au seuil de Bonferroni pour le nombre de loci et de types cellulaires testés ($7.6\ 10^{-4}$). On rapporte à chaque fois le type cellulaire le plus fortement associé au locus et la *p*-value minimale observée.

lorsque leur volume augmente ainsi que par la diminution du nombre de plaquettes.

2p12 Sur le chromosome 2, le locus du gène CD8A est associé ($p = 1.33\,10^{-6}$) à la contamination par les lymphocytes T. Cette association, apparaît ici clairement comme un biais technique attribuable à la méthode de sélection. En effet, le gène CD8A code pour la protéine CD8 utilisée comme marqueur des lymphocytes T dans la procédure de sélection négative. Une modification de structure ou de quantité du marqueur CD8 est donc susceptible d'affecter la quantité de lymphocytes T CD8+ présents dans le mélange d'ARN.

7q21 Le locus du gène MAGI-2 (Membrane-Associated Guanylate kinase Inverted 2) est trouvé associé au niveau de contamination par les érythrocytes ($p = 3.9\,10^{-8}$), suggérant que ce gène pourrait être impliqué dans la modulation du nombre de globules rouges. Bien que ce gène n'ait pas été identifié formellement dans les GWAS portant sur les phénotypes hématologiques, des études ont montré une interaction de ce gène avec le récepteur à l'activine de type II qui est impliqué dans la formation des érythrocytes [116, 117].

6q15 Le locus du gène MAP3K7 (Mitogen Associated Proteine Kinase Kinase Kinase 7) est associé à un module de gènes enrichis en fonctions liées au cycle cellulaire et à la mitose. Le facteur latent associé à ce module montrait une forte corrélation avec la contamination par les lymphocyte B (r=-0.79). On trouve une association du SNP ($p = 5.6\,10^{-6}$) avec la contamination par les lymphocytes B. Une telle association pourrait s'expliquer par l'importante division cellulaire à laquelle sont sujets les lymphocytes B lors de la réaction immunitaire. Des études ont en effet montré le rôle du gène MAP3K7 dans la cascade de l'activation des lymphocytes B par la reconnaissance d'un antigène [118].

3.2 Association entre SNP et patterns de co-expression après ajustement sur les quantités des différents types cellulaires

Pour chaque SNP significativement lié à un pattern d'expression et trouvé associé à un type cellulaire contaminant, nous avons retesté l'effet du SNP sur le pattern après ajustement sur la quantité du type cellulaire k, selon le modèle

$$a_{l\cdot} = a + bSNP + \sum_{k=2}^{6} c_k \widehat{q_{k\cdot}} + \epsilon \tag{10.1}$$

On rapporte les résultats de cette analyse dans le tableau 10.3. Lorsqu'on teste l'effet du SNP sur les différents patterns de co-expression en ajustant sur les proportions des différents types cellulaires, on observe à chaque fois une diminution importante de l'effet du SNP.

			Association SNP-pattern					
			brute		ajustée sur un type cellulaire		ajustée sur tous les types cellulaires	
locus	SNP	type cellulaire associé	*p*-value d'association	différence entre homozygotes	*p*-value d'association	différence entre homozygotes	*p*-value d'association	différence entre homozygotes
2p12	rs13023213	T	1,33E-07	0,47	0,04	0,11	0,36	0,04
3p21	rs12485738	MK	2,83E-23	-0,70	0,0001	-0,14	0,0002	-0,12
3p21	rs1344142	MK	6,89E-19	0,57	8,47E-07	0,27	4,00E-07	0,26
6q15	rs13196564	CD19	6,70E-08	-0,92	0,004	-0,38	0,03	-0,25
7q21	rs12705417	EB	1,11E-08	0,97	0,002	0,42	0,11	0,18

Tableau 10.3 – **Effet sur l'association SNP-pattern de l'ajustement par les proportions du mélange :** Pour chaque SNP trouvé associé à un pattern, on rapporte la *p*-value d'association et la différence de niveau moyen d'expression entre les homozygotes rares et fréquents, dans le modèle non ajusté et les modèles ajustés (sur un ou cinq types cellulaires).

Cependant, dans la plupart des cas, l'effet du SNP sur le pattern reste significatif après ajustement. Seul l'effet du locus 2p12 est presque totalement dissipé par l'ajustement sur la proportion de lymphocytes T dans le mélange.

3.3 Analyse détaillée de l'effet du SNP sur les gènes les plus associés

Nous développons dans cette section le cas des loci CD8A et ARHGEF3 qui reflètent des situations contrastées. Pour chaque locus on étudie l'effet du génotype sur l'expression des gènes les plus associés avant et après ajustement sur les quantités estimées du type cellulaire le plus associé.

3.3.1 Locus CD8A

On s'intéresse tout d'abord aux gènes les plus fortement associés au génotype avant l'ajustement sur la contamination (tableau 10.4).

Parmi les gènes les plus associés au locus CD8A, on trouve des gènes spécifiques des celulles tueuses et des lymphocytes T (LCK, KIR2DL4, KRLD1,...). Lorsqu'on ajuste dans GHS sur la quantité estimée de lymphocytes T on voit que l'association des gènes avec le SNP est fortement réduite et n'est plus significative pour aucun des gènes au seuil de significativité de Bonferroni (10^{-6}). Dans cardiogenics aucune association n'est répliquée, soit parce que le transcrit n'est pas présent sur la puce utilisée dans cardiogenics, soit parce que l'effet du SNP n'est pas significatif. Cette absence de réplication dans Cardiogenics suggère que l'effet observé dans GHS est artefactuel. Après ajustement sur la contamination, on voit apparaître de nouvelles associations significatives. Ces associations

			GHS (avant ajustement)		GHS (ajusté sur la quantité de lymphocytes T)		Cardiogenics	
Identifiant de la sonde	gène ciblé par la sonde	Chromosome	p-value d'association	effet de l'allèle mineur	p-value d'association	effet de l'allèle mineur	p-value d'association	effet de l'allèle mineur
Sondes les plus associées dans GHS avant ajustement								
ILMN_1762207	SGSM1	22	1,07E-11	+	9,38E-06	+	ND	ND
ILMN_1699991	LCK	1	3,61E-10	+	5,3537E-05	+	ND	ND
ILMN_1762957	LOC648868	ND	5,74E-10	+	0,002	+	ND	ND
ILMN_1728298	SBK1	16	1,87E-09	+	0,003	+	0,24	-
ILMN_1792538	CD7	17	2,27E-09	+	0,007	+	0,11	-
ILMN_2196078	SLAMF6	1	5,90E-09	+	0,006	+	0,68	-
ILMN_1693207	KIR2DL4	19	8,83E-09	+	0,01	+	0,29	+
ILMN_1799134	KLRD1	12	9,46E-09	+	0,009	+	ND	ND
ILMN_2131828	KIR3DL1	19	1,20E-08	+	0,02	+	0,97	-
ILMN_1688279	PVRIG	7	1,21E-08	+	0,02	+	0,13	-
ILMN_2055781	KLRF1	12	1,24E-08	+	0,02	+	0,90	-
ILMN_1739756	KIR2DL4	19	1,40E-08	+	0,03	+	ND	ND
ILMN_2073184	S1PR5	19	1,85E-08	+	0,03	+	0,46	-
ILMN_1667232	KIR2DL1	19	2,26E-08	+	0,001	+	0,84	+
ILMN_1664828	APOBEC3H	22	2,32E-08	+	0,04	+	0,18	+
Sondes significativement associées dans GHS après ajustement								
ILMN_2353732	CD8A	2	0,001	-	2,90E-16	-	0,48	-
ILMN_1768482	CD8A	2	0,000	-	7,53E-14	-	0,21	-
ILMN_2094416	PLGLB1	2	3,10E-07	-	1,80E-07	-	0,0008	-

Tableau 10.4 – **Liste des 15 sondes les plus fortement associées au SNP rs13023213 (locus CD8A) avant ajustement sur la quantité de lymphocytes T et des 3 sondes significativement associées après ajustement** : On rapporte pour chaque sonde les résultats de l'association dans GHS avant et après ajustement sur la quantité de lymphopcytes T, ainsi que dans Cardiogenics (ajusté sur le centre). A chaque fois, on indique la p-value d'association et l'effet de l'allèle mineur sur l'expression de la sonde (augmentation (+) ou diminution (-) de l'expression).

impliquent des gènes situés en *cis* du locus (CD8A et PLGLB1) qui étaient déjà faiblement associés au SNP avant ajustement. On trouve ainsi après ajustement sur la quantité de lymphocytes T deux effets *cis* qui n'atteignaient pas le seuil de significativité dans les analyses non ajustées. La présence de SNP dans les sondes Illumina utilisées pour mesurer l'expression du gène CD8A suggère que l'effet observé sur le gène CD8A pourrait être un artefact. L'association en *cis* du gène PLGLB1 est en revanche répliquée dans Cardiogenics suggérant un effet *cis* dans les monocytes.

Il semble donc que l'association du locus CD8A avec le pattern de co-expression soit due à l'effet de la contamination par les lymphocytes T. Ici, l'ajustement a posteriori sur la contamination permet de supprimer cette association fallacieuse et de gagner en puissance pour détecter de nouvelles associations.

3.3.2 Locus ARHGEF3

Au locus ARHGEF3 on retrouve parmi les gènes les plus fortement associés au SNP, des gènes sur-exprimés dans les plaquettes (ITGB3, TSPAN9, COL6A3,...). Pour ces gènes, l'effet du SNP reste cependant très fortement significatif après ajustement sur la quantité de plaquettes, ce qui suggère que l'effet du SNP sur ces gènes n'est pas attribuable uniquement à son effet sur la quantité de plaquettes.

Lorsqu'on ajuste sur la quantité de plaquettes on voit apparaître de nouvelles associations avec des gènes situés en *trans* du locus du gène ARHGEF3. Parmi ces gènes on retrouve un important nombre de gènes codant pour des histones, protéines ubiquitaires influençant la conformation de l'ADN et dont certains liens avec l'activité des plaquettes ont été mis en évidence [119]. Cependant aucune de ces associations n'est répliquée dans Cardiogenics, ce qui semble indiquer que l'effet détecté ici ne reflète pas une modification du transcriptome des monocytes. Une interprétation possible à ces résultats pourrait être que le locus ARHGEF3, outre son effet sur la quantité de plaquettes pourrait également influencer le transcriptome des plaquettes. Comme énoncé dans la section 1.2 cette hypothèse pourrait en théorie être testée en introduisant dans le modèle des termes d'interactions entre le SNP et les proportions des différents types cellulaires. Toutefois, la faible variabilité des proportions des types cellulaires contaminants et l'imprécision de leur estimation ne permettent pas d'aller beaucoup plus loin dans l'interprétation. Dans le cas du locus ARHGEF3, si l'on peut donc exclure un effet majeur sur le transcriptome des monocytes, on ne peut pas conclure de manière évidente quant à l'orgine des effets observés, contrairement au locus CD8A.

Identifiant de la sonde	gène ciblé par la sonde	Chromosome	GHS (avant ajustement)		GHS (ajusté sur la quantité de lymphocytes T)		Cardiogenics	
			p-value d'association	effet de l'allèle mineur	p-value d'association	effet de l'allèle mineur	p-value d'association	effet de l'allèle mineur
Sondes les plus associées dans GHS avant ajustement								
ILMN_1760688	SAMD14	17	1,32E-36	-	5,64E-19	-	0,42	-
ILMN_1673548	HSPC159	2	1,31E-32	-	5,02E-15	-	0,62	-
ILMN_1733324	ITGB3	17	1,99E-30	-	4,34E-12	-	0,08	-
ILMN_1778991	NFIB	9	4,78E-30	-	5,58E-12	-	0,40	-
ILMN_1660114	MMRN1	4	1,21E-29	-	4,94E-11	-	0,27	-
ILMN_2412294	GNB5	15	3,10E-29	-	7,98E-11	-	0,40	-
ILMN_1729453	TSPAN9	12	6,82E-29	-	1,81E-10	-	0,59	-
ILMN_1706643	COL6A3	2	3,46E-28	-	9,68E-11	-	0,74	+
ILMN_1730487	CALD1	7	3,65E-28	-	1,07E-09	-	ND	ND
ILMN_1721888	ITGA2B	17	1,32E-27	-	2,82E-09	-	0,82	-
ILMN_1787919	PARVB	22	2,67E-27	-	2,31E-09	-	0,56	+
ILMN_1671928	PROS1	3	2,91E-27	-	9,58E-09	-	0,35	-
ILMN_1686373	C15orf26	15	7,07E-27	-	1,79E-08	-	0,08	-
ILMN_1671486	HOMER2	15	8,46E-27	-	1,63E-08	-	0,28	+
ILMN_1663519	SLC24A3	20	1,04E-26	-	1,32E-08	-	0,27	-
Sondes significativement associées dans GHS après ajustement								
ILMN_1760688	SAMD14	17	1,32E-36	-	5,64E-19	-	0,42	-
ILMN_1749368	HIST1H3H	6	0,03	+	1,45E-17	+	0,42	-
ILMN_1787567	TSC22D1	13	0,84	+	3,89E-17	+	ND	ND
ILMN_1788489	HIST1H3F	6	0,10	+	5,23E-16	+	0,70	-
ILMN_1693269	GNG8	19	0,15	+	7,61E-16	+	0,36	-
ILMN_1808907	HIST1H2BN	6	0,009	+	2,42E-15	+	ND	ND
ILMN_1673548	HSPC159	2	1,31E-32	-	5,02E-15	-	0,62	-
ILMN_1779373	HIST1H2BF	6	0,06	+	9,53E-15	+	0,04	-
ILMN_1705685	MEIS1	2	0,74	-	1,04E-14	+	ND	
ILMN_1699071	C21orf7	21	0,00	+	1,05E-14	+	0,01	+
ILMN_1747650	BMP6	6	0,05	-	1,63E-14	+	0,29	-
ILMN_1733937	MMD	17	0,01	-	2,63E-14	+	0,69	-
ILMN_1805643	RILPL1	12	0,86	+	6,83E-14	+	0,10	-
ILMN_1770338	TM4SF1	3	0,001	+	1,73E-13	+	0,70	-
ILMN_1756849	HIST1H2AE	6	0,07	+	1,85E-13	+	0,35	-

Tableau 10.5 – liste des 15 sondes les plus fortement associées au SNP rs12485738 avant et après ajustement sur la quantité de plaquettes. On rapporte pour chaque sonde les résultats de l'association dans GHS avant et après ajustement sur la quantité de plaquettes, ainsi que les résultats de l'association dans Cardiogenics (ajusté sur le centre). A chaque fois, on indique la p-value d'association et l'effet de l'allèle mineur sur l'expression de la sonde (augmentation (+) ou diminution (-) de l'expression). ND : non déterminé car sonde absente de la puce utilisée dans Cardiogenics.

4 Bilan et perspectives

Ce travail montre que la contamination par des types cellulaires non désirés peut fortement affecter le transcriptome et induire des associations artefactuelles lorsque la variable étudiée

- influence la quantité d'au moins un des types cellulaires présents
- et/ou influence le transcriptome d'un des types cellulaires contaminants

Nous avons montré que bien que ce type d'artefact soit plus important lorsqu'on recourt à des méthodes de sélection négative, il peut apparaître quelle que soit la méthode d'isolation utilisée.

Nous avons vu qu'il est possible en se basant sur des atlas d'expression de déceler ce type de biais liés à la contamination et de tenter d'estimer les proportions des différents types cellulaires présents. Cependant, comme nous l'avons vu avec le locus ARHGEF3 l'ajustement sur les quantités estimées ne permet pas de contrôler totalement les effets de la contamination. On peut peut proposer plusieurs raisons à cela :

- La présence d'erreurs de mesure sur les quantités estimées : dans ce cas, ce problème se ramène à celui d'un modèle avec erreur de mesure sur les covariables. Dans un tel modèle, l'erreur de mesure biaise l'estimation des paramètres et peut conduire à l'apparition de faux positifs. Si l'estimation des paramètres d'un tel modèle a été largement décrite dans la littérature [120] elle nécessite d'être capable de quantifier l'erreur de mesure réalisée sur les covariables. Ici deux cas de figure sont possibles :
  * L'imprécision des estimations des variables de contamination est liée à la méthode elle-même. On peut dans ce cas espérer la quantifier par des méthodes de bootstrap et intégrer cette erreur dans le modèle pour en tenir compte. La mise en œuvre des méthodes de correction des erreurs réalisées sur les covariables du modèle (SIMEX, Regression Calibration,...), constitue une des voies d'approfondissement de ce travail actuellement en cours d'exploration.
  * L'imprécision est liée à la mesure des profils utilisés pour la déconvolution. En effet, on peut imaginer que les données de la base HaemAtlas ne sont pas elles-même exemptes d'erreur et sont sujettes à des sources de variations mal contrôlées. Dans ce cas, en l'absence d'hypothèse précise sur la forme des erreurs de mesure auxquelles on est confronté, il apparaît plus difficile de tenir compte de l'imprécision des estimations pour améliorer le modèle.
- Outre des problèmes liés à l'imprécison de la mesure des quantités, une autre explication tient à l'existence d'effets des SNP sur le transcriptome des types cellulaires contaminants. Comme nous l'avons évoqué dans la section 1.2, si un facteur z affecte l'expression d'un des types cellulaires présents cela se traduira par une interaction entre la quantité de ce type cellulaire et le facteur z. Dans ce cas l'ajustement sur

les quantités ne suffit pas à éliminer les effets de la contamination. Un des objectifs initiaux de ce travail était de tenter de détecter la provenance des signaux d'associations par l'introduction de termes d'interactions dans le modèle. Cependant nous avons constaté sur des simulations :

- Que la présence d'erreurs de mesure sur les quantités peut introduire une augmentation de l'erreur de type I pour le test d'interaction.
- Que l'introduction du terme d'interaction dans le modèle réduit très fortement la puissance du test de l'effet principal du SNP dans les monocytes.
- Que la puissance du test d'interaction est généralement très faible puisque les quantités des différents types cellulaires sont peu variables et comprises dans un range de valeurs strictement positives entrainant une très forte corrélation des termes du modèle d'interaction.

Pour ces différentes raisons, la détection de la provenance des signaux d'associations n'apparaît pas réalisable de manière systématique [7] dans l'état actuel de ce travail.

7. Bien que l'augmentation de l'erreur de type I puisse être controlée par le recours à des méthodes de permutation, le manque de puissance reste problématique dès lors que les effets observés sont d'intensité modérée.

Conclusion et perspectives

En génomique humaine, l'intrégration du génome et du transcriptome est aujourd'hui considérée comme un axe de recherche majeur pour la compréhension de l'étiologie des maladies complexes. Cette intégration peut se faire de plusieurs manières, que nous avons tenté d'aborder dans cette thèse :

- par une recherche systématique des loci associés à des variations de l'expression (eQTL) ;
- par l'extraction de modules de co-expression, qui soient le reflet des processus biologiques à l'oeuvre dans la cellule et la recherche de polymorphismes affectant ces modules.

Dans mon travail de thèse, j'ai été en charge de l'analyse des données d'expression de l'étude GHS. J'ai ainsi participé à l'établissement d'une base de données recensant les eQTL présents dans les monocytes humains (`GHS_express`). Cette base de données, mise à disposition de la communauté scientifique, pourra être utilisée comme un outil facilitant l'interprétation des résultats fournis par les analyses pan-génomiques. L'analyse des eQTL identifiés dans GHS a établi que plus d'un tiers des gènes étaient génétiquement régulés dans les monocytes et que près de 70% des eQTL les plus forts pouvaient être répliqués entre tissus différents indiquant un fort partage des eQTLs d'un tissu à l'autre mais également une fraction non négligeable d'eQTL tissu-spécifiques.

Afin d'estimer l'information apportée par les eQTL pour l'interprétation des analyses génome entier, nous avons croisé les résultats de l'analyse des eQTL avec un catalogue recensant les loci de prédisposition identifiés dans des GWAS. Cette analyse a permis de retrouver l'enrichissement des loci de GWAS en *cis*-eQTL déjà observé à partir des données de HapMap. Nous avons cependant montré que la densité inégale des gènes sur le génome pouvait conduire à surestimer le lien entre les eQTL et les loci de prédisposition. Nous avons montré que dans les régions de forte densité, il arrive fréquemment que des loci de GWAS co-localisent avec des eQTL du seul fait du déséquilibre de liaison, ce qui peut compliquer l'identification des mécanismes causaux sous-jacents aux loci de prédisposition identifiés par les GWAS. Malgré cela nous avons pu montrer que dans un certain nombre de cas, l'étude des eQTL pouvait amener de nouvelles pistes à explorer.

C'est par exemple le cas au locus MTCH2, où un eQTL contrôlant l'expression du gène C1QTNF4, un paralogue de l'adiponectine, colocalise avec un locus associé à l'IMC. Nous avons en outre montré que la répartition des gènes sur le génome apportait à elle seule une information supérieure à l'étude des eQTL sur la localisation des loci de prédisposition, et qu'il était possible de tirer parti de cette information pour améliorer la puissance des GWAS.

Une deuxième partie de mon travail de thèse a consisté à développer une approche visant à identifier des modules de co-expression pour la recherche de loci affectant l'expression de groupes de gènes. Cette méthode est actuellement en cours d'intégration dans un package `R` d'analyse des données d'expression implémenté au sein du laboratoire. Dans GHS, cette approche a permis l'identification de 64 modules de gènes co-exprimés, dont les deux-tiers environ étaient enrichis en fonctions biologiques connues. Parmi ces modules, onze ont été associés à des polymorphismes, et trois de ces associations ont pu être répliquées dans l'étude Cardiogenics. Parmi les loci associés, deux avaient été trouvés associés au risque de diabète de type I dans des GWAS, suggérant que cette méthode est à même de capturer des variations du transcriptome reflétant des processus biologiques impliqués dans le développement des maladies complexes. De plus, la réplication chez l'homme d'un module lié à la réponse aux virus, trouvé associé chez le rat au gène EBI2, a permis de suggérer le rôle de ce gène dans le développement du diabète de type I. Pour chacun de ces modules, de plus amples expérimentations fonctionnelles, couplées à une modélisation fine[8] des réseaux de régulation à ces loci, sont à présent nécéssaires pour aller plus loin dans la compréhension des mécanismes liant ces loci aux maladies complexes.

Cette analyse a également permis de mettre en évidence l'existence de contamination par des types cellulaires non désirés dans l'étude GHS. Nous avons montré que ce genre d'artefact pouvait affecter les mesures du transcriptome lorsqu'on isole un type cellulaire par sélection négative. Cette contamination peut entraîner des associations fallacieuses lorsque les proportions des différents types cellulaires varient avec la variable étudiée. Nous avons montré qu'il était cependant possible d'estimer les proportions de contaminants présents dans les échantillons à l'aide d'informations provenant de bases de données publiques. Nous avons ainsi pu mettre en évidence des SNP affectant les quantités des différents types cellulaires et proposer une explication aux mécanismes biologiques sous-jacents. Toutefois, nous avons vu que l'ajustement sur les quantités estimées des types contaminants ne permettait pas de retirer tous les biais liés à la contamination. Ceci est vrai en particulier lorsque la variable étudiée affecte le transcriptome d'un des types cellulaires contaminants.

Plusieurs aspects de ce travail méritent à présent d'être approfondis :

8. Par exemple en recourant à des modèles structuraux.

- Un premier approfondissement consisterait à répéter l'analyse des eQTL en ajustant sur les quantités estimées des différents types cellulaires. Un tel ajustement contribuerait à augmenter la puissance pour la découverte des eQTL en retirant la variabilité induite par la contamination. Grâce au développement des moyens de calcul disponibles qui a eu lieu depuis le début de ce travail[9], il est à présent envisageable d'augmenter la puissance pour la recherche des eQTL en recourant à des méthodes de permutation pour tenir compte de la corrélation entre marqueurs dans l'estimation du FDR ou en ajustant progressivement les niveaux d'expression de chaque transcrit sur les eQTL déjà identifiés pour ce transcrit.
- Un second axe de recherche déjà en cours d'exploration est l'identification des eQTL potentiellement impliqués dans la prédisposition aux maladies complexes. Afin d'écarter les co-localisation fallacieuses attribuables à la densité de gènes évoquées dans le présent ouvrage, on peut recourir à un test formel de co-localisation développé par Plagnol *et al.* [75]. Ce test est en cours d'application à la recherche d'eQTL impliqués dans l'étiologie du diabète de type I.
- L'étude des modules de co-régulation pourrait également être améliorée :
  - Par le développement de méthodes d'analyse supervisées. Par exemple, une piste intéressante consisterait à rechercher des SNP associés à un "excès" de gènes en *trans* (par rapport au nombre d'associations *trans* attendues à un seuil de significativité donné).
  - Par l'application de méthodes d'association fondées sur la connaissance a priori de voies de signalisation telle que SigPathway [121] ou GSEA [122].
  - En testant l'association des patterns au génotype par des modèles multivariés de type lasso [123].
- Enfin, ainsi qu'évoqué dans la dernière section de ce document, le développement de méthodes capables d'identifier avec précision la provenance des signaux d'association détectés dans un mélange de types cellulaires constitue un domaine de recherche particulièrement intéressant, offrant de larges possibilités d'application. En effet, de telles méthodes pourraient non seulement permettre d'éliminer les biais dus à la contamination que l'on observe dans l'étude GHS, mais elles pourraient également trouver une application dans l'étude des mesures du transcriptome des échantillons de sang total ou dans l'étude du transcriptome de tissus complexes composés d'un mélange de types cellulaires distincts.

9. Du fait de la taille des données utilisées, les temps de calcul et les besoins en mémoire des algorithmes utilisées ont été une contrainte majeure au début de ce travail. Aujourd'hui, grâce à l'installation du cluster de calcul `iDataplex` à Jussieu en avril 2010, les moyen de calcul accessibles à l'unité ont été plus que décuplés ce qui autorise le recours à des approches plus gourmandes en mémoire et en temps de calcul.

Bibliographie

[1] Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nature Genetics *39*, 1217–1224. PMID : 17873874.

[2] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

[3] Göring, H. H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., Jowett, J. B. M., Abraham, L. J., Rainwater, D. L., Comuzzie, A. G., et al. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature Genetics *39*, 1208–1216. PMID : 17873875.

[4] Petretto, E., Mangion, J., Dickens, N. J., Cook, S. A., Kumaran, M. K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., et al. (2006). Heritability and tissue specificity of expression quantitative trait loci. PLoS Genet *2*, e172.

[5] Mahley, R. (1988). Apolipoprotein e : cholesterol transport protein with expanding role in cell biology. Science *240*, 622 –630.

[6] Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., et al. (2010). Genetics and beyond – the transcriptome of human monocytes and disease susceptibility. PLoS ONE *5*, e10693.

[7] Kannel, W. B., McGee, D., and Gordon, T. (1976). A general cardiovascular risk profile : The framingham study. The American Journal of Cardiology *38*, 46–51.

[8] Hubert, H., Feinleib, M., McNamara, P., and Castelli, W. (1983). Obesity as an independent risk factor for cardiovascular disease : a 26- year follow-up of participants in the framingham heart study. Circulation *67*, 968–977.

[9] Kannel, W. B. (1996). Blood pressure as a cardiovascular risk factor. JAMA : The Journal of the American Medical Association *275*, 1571 –1576.

[10] Luft, F. C. (2001). Twins in cardiovascular genetic research. Hypertension *37*, 350–356.

[11] Ding, K. and Kullo, I. J. (2009). Genome-wide association studies for atherosclerotic vascular disease and its risk factors. Circulation. Cardiovascular genetics *2*, 63–72. PMID : 19750184 PMCID : 2740629.

[12] Tunstall-Pedoe, H., Kuulasmaa, K., Amouyel, P., Arveiler, D., Rajakangas, A. M., and Pajak, A. (1994). Myocardial infarction and coronary deaths in the world health organization MONICA project. registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. Circulation *90*, 583–612. PMID : 8026046.

[13] Stevens, A. and Lowe, J. (1997). Human histology. (Mosby Toronto).

[14] Toussaint, J. (2003). L'athérosclérose : physiopathologie, diagnòstics, thérapeutiques. (Elsevier Masson).

[15] Cohen, A., Tzourio, C., Bertrand, B., Chauvel, C., Bousser, M., and Amarenco, P. (1997). Aortic plaque morphology and vascular events : A follow-up study in patients with ischemic stroke. Circulation *96*, 3838–3841.

[16] Meuwissen, M., van der Wal, A., Siebes, M., Koch, K., Chamuleau, S., van der Loos, C., Teeling, P., de Winter, R., Niessen, H., Tijssen, J., et al. (2004). Role of plaque inflammation in acute and recurrent coronary syndromes. Netherlands Heart Journal *12*, 106–109. undefinedPMCID : 2497049.

[17] Frostegård, J., Ulfgren, A. K., Nyberg, P., Hedin, U., Swedenborg, J., Andersson, U., and Hansson, G. K. (1999). Cytokine expression in advanced human atherosclerotic plaques : dominance of pro-inflammatory (Th1) and macrophage-stimulating cytokines. Atherosclerosis *145*, 33–43. PMID : 10428293.

[18] Hansson, G. K., Libby, P., Schönbeck, U., and Yan, Z. (2002). Innate and adaptive immunity in the pathogenesis of atherosclerosis. Circulation Research *91*, 281–291. PMID : 12193460.

[19] Arnett, D. K., Baird, A. E., Barkley, R. A., Basson, C. T., Boerwinkle, E., Ganesh, S. K., Herrington, D. M., Hong, Y., Jaquish, C., McDermott, D. A., et al. (2007). Relevance of genetics and genomics for prevention and treatment of cardiovascular disease : A scientific statement from the american heart association council on epidemiology and prevention, the stroke council, and the functional genomics and translational biology interdisciplinary working group. Circulation *115*, 2878–2901.

[20] Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., et al. (2011). ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Research *39*, D1002–1004. PMID : 21071405.

[21] Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., et al. (2011).

NCBI GEO : archive for functional genomics data sets–10 years on. Nucleic Acids Research *39*, D1005–1010. PMID : 21097893.

[22] Bhowmick, D., Davison, A. C., Goldstein, D. R., and Ruffieux, Y. (2006). A laplace mixture model for identification of differential expression in microarray experiments. Biostatistics (Oxford, England) *7*, 630–641. PMID : 16565148.

[23] Dunning, M. J., Thorne, N. P., Camilier, I., Smith, M. L., and Tavaré, S. (2006). Quality control and low-level statistical analysis of illumina beadarrays. Revstat Statistical Journal *4*, 1–30.

[24] Affymetrix (2002). Affymetrix white papers - statistical algorithms description documents.

[25] Illumina (2010). BeadStudio data analysis software from illumina, inc. - biocompare buyer's guide for life scientists. http ://www.biocompare.com/Articles/ProductReview/719/BeadStudio-Data-Analysis-Software-From-Illumina-Inc.html.

[26] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England) *4*, 249–264. PMID : 12925520.

[27] Lin, S. M., Du, P., Huber, W., and Kibbe, W. A. (2008). Model-based variance-stabilizing transformation for illumina microarray data. Nucl. Acids Res. *36*, e11.

[28] Bolstad, B. (2001). Probe level quantile normalization of high density oligonucleotide array data. Unpublished manuscript.

[29] Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) *19*, 185–193. PMID : 12538238.

[30] Qiu, X., Brooks, A. I., Klebanov, L., and Yakovlev, A. (2005). The effects of normalization on the correlation structure of microarray data. BMC bioinformatics *6*, 120.

[31] Lim, W. K., Wang, K., Lefebvre, C., and Califano, A. (2007). Comparative analysis of microarray normalization procedures : effects on reverse engineering gene networks. Bioinformatics (Oxford, England) *23*, i282–288. PMID : 17646307.

[32] Saviozzi, S. and Calogero, R. A. (2003). Microarray probe expression measures, data normalization and statistical validation. Comparative and Functional Genomics *4*, 442–446. PMID : 18629084 PMCID : 2447370.

[33] Ballman, K. V., Grill, D. E., Oberg, A. L., and Therneau, T. M. (2004). Faster cyclic loess : normalizing RNA arrays via linear models. Bioinformatics *20*, 2778 –2786.

[34] Berger, J. A., Hautaniemi, S., J\ärvinen, A. K., Edgren, H., Mitra, S. K., and Astola, J. (2004). Optimized LOWESS normalization parameter selection for DNA microarray data. BMC bioinformatics *5*, 194.

[35] Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., Saxild, H. H., Nielsen, C., Brunak, S., and Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol *3*, 1–16.

[36] Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi : a pipeline for processing illumina microarray. Bioinformatics (Oxford, England) *24*, 1547–1548. PMID : 18467348.

[37] Asea, A., Kraeft, S., Kurt-Jones, E. A., Stevenson, M. A., Chen, L. B., Finberg, R. W., Koo, G. C., and Calderwood, S. K. (2000). HSP70 stimulates cytokine production through a CD14-dependant pathway, demonstrating its dual role as a chaperone and cytokine. Nat Med *6*, 435–442.

[38] Barbosa-Morais, N. L., Dunning, M. J., Samarajiwa, S. A., Darot, J. F. J., Ritchie, M. E., Lynch, A. G., and Tavaré, S. (2010). A re-annotation pipeline for illumina BeadArrays : improving the interpretation of gene expression data. Nucleic Acids Research *38*, e17.

[39] Margulies, E. H., Kardia, S. L., and Innis, J. W. (2001). Identification and prevention of a GC content bias in SAGE libraries. Nucleic Acids Research *29*, E60–60. PMID : 11410683.

[40] Carvalho, B., Bengtsson, H., Speed, T. P., and Irizarry, R. A. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostat *8*, 485–499.

[41] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., et al. (2007). PLINK : a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics *81*, 559–575. PMID : 17701901.

[42] Aulchenko, Y. S., Ripke, S., Isaacs, A., and van Duijn, C. M. (2007). GenABEL : an r library for genome-wide association analysis. Bioinformatics (Oxford, England) *23*, 1294–1296. PMID : 17384015.

[43] Cox, M. A. and Cox, T. F. (2008). Multidimensional scaling. Handbook of data visualization pp. 315–347.

[44] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character *231*, 289 –337.

[45] Abdi, H. and Abdi, H. (2007). Encyclopedia of measurement and statistics. In Bonferroni and Šidák corrections for multiple comparisons In Bonferroni and Šidák corrections for multiple comparisons. (Thousand Oaks, CA : Sage).

[46] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289–300. ArticleType : research-article / Full publication date : 1995 / Copyright © 1995 Royal Statistical Society.

[47] Storey, J. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B (Statistical Methodology) *64*, 479–498.

[48] Efron, B., Storey, J. D., and Tibshirani, R. (2001). Microarrays, empirical bayes methods, and false discovery rates. GENET. EPIDEMIOL *23*, 70—86.

[49] Aubert, J., Bar-Hen, A., Daudin, J., and Robin, S. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR. BMC Bioinformatics *5*, 125.

[50] Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. Journal of the Royal Statistical Society. Series B (Statistical Methodology) *67*, 555–572. ArticleType : research-article / Full publication date : 2005 / Copyright © 2005 Royal Statistical Society.

[51] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association *102*, 93–103.

[52] Friguet, C. (2009). A factor model approach to multiple testing under dependence. Journal of the American Statistical Association *104*, 1406–1415.

[53] Strimmer, K. (2008). A unified approach to false discovery rate estimation. BMC Bioinformatics *9*, 303.

[54] Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956 –960.

[55] Hsu, Y., Zillikens, M. C., Wilson, S. G., Farber, C. R., Demissie, S., Soranzo, N., Bianchi, E. N., Grundberg, E., Liang, L., Richards, J. B., et al. (2010). An integration of genome-wide association study and gene expression profiling to prioritize the

discovery of novel susceptibility loci for osteoporosis-related traits. PLoS Genetics *6*, e1000977. PMID : 20548944.

[56] Doss, C. G. P., Rajasekaran, R., Arjun, P., and Sethumadhavan, R. (2010). Prioritization of candidate SNPs in colon cancer using bioinformatics tools : An alternative approach for a cancer biologist. Interdisciplinary Sciences, Computational Life Sciences *2*, 320–346. PMID : 21153778.

[57] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG : kyoto encyclopedia of genes and genomes. Nucleic Acids Research *27*, 29–34. PMID : 9847135.

[58] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology : tool for the unification of biology. the gene ontology consortium. Nature Genetics *25*, 25–29. PMID : 10802651.

[59] Sun, W., Yu, T., and Li, K. (2007). Detection of eQTL modules mediated by activity levels of transcription factors. Bioinformatics (Oxford, England) *23*, 2290–2297. PMID : 17599927.

[60] Bao, L., Xia, X., and Cui, Y. (2010). Expression QTL modules as functional components underlying higher-order phenotypes. PloS One *5*, e14313. PMID : 21179437.

[61] Veyrieras, J., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-Resolution mapping of Expression-QTLs yields insight into human gene regulation. PLoS Genet *4*, e1000214.

[62] Myers, A. J., Gibbs, J. R., Webster, J. A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P., et al. (2007). A survey of genetic human cortical gene expression. Nature Genetics *39*, 1494–1499. PMID : 17982457.

[63] Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biology *6*, e107. PMID : 18462017.

[64] Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., et al. (2007). A genome-wide association study of global gene expression. Nature Genetics *39*, 1202–1207. PMID : 17873877.

[65] Chan, E., Quon, G., Chua, G., Babak, T., Trochesset, M., Zirngibl, R., Aubin, J., Ratcliffe, M., Wilde, A., Brudno, M., et al. (2009). Conservation of core gene expression in vertebrate tissues. Journal of Biology *8*, 33.

[66] Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Arcelus, M. G., Sekowska, M., et al. (2009).

Common regulatory variation impacts gene expression in a cell type dependent manner. Science (New York, N.Y.) *325*, 1246–1250. PMID : 19644074 PMCID : 2867218.

[67] Wallace, C., Newhouse, S. J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R. J., Marçano, A. C. B., Hajat, C., et al. (2008). Genome-wide association study identifies genes for biomarkers of cardiovascular disease : serum urate and dyslipidemia. American Journal of Human Genetics *82*, 139–149. PMID : 18179892.

[68] Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nature Genetics *40*, 161–169. PMID : 18193043.

[69] Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719. PMID : 20686566.

[70] Franke, L. and Jansen, R. C. (2009). eQTL analysis in humans. Methods in Molecular Biology (Clifton, N.J.) *573*, 311–328. PMID : 19763935.

[71] Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., and Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genetics *6*, e1000895. PMID : 20369022.

[72] Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs : annotation to enhance discovery from GWAS. PLoS Genetics *6*, e1000888. PMID : 20369019.

[73] Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E., et al. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature *448*, 470–473.

[74] Ganesh, S. K., Zakai, N. A., van Rooij, F. J. A., Soranzo, N., Smith, A. V., Nalls, M. A., Chen, M., Kottgen, A., Glazer, N. L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE consortium. Nature genetics *41*, 1191–1198. PMID : 19862010 PMCID : 2778265.

[75] Plagnol, V., Smyth, D. J., Todd, J. A., and Clayton, D. G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics (Oxford, England) *10*, 327–334. PMID : 19039033.

[76] Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nature Genetics *41*, 25–34. PMID : 19079261.

[77] Matsuzawa, Y. (2010). Adiponectin : a key player in obesity related disorders. Current Pharmaceutical Design *16*, 1896–1901. PMID : 20370675.

[78] Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. Biometrika *93*, 509 –524.

[79] Kardia, S. L., Greene, M. T., Boerwinkle, E., Turner, S. T., and Kullo, I. J. (2008). Investigating the complex genetic architecture of ankle-brachial index, a measure of peripheral arterial disease, in non-Hispanic whites. BMC Medical Genomics *1*, 16. PMID : 18482449.

[80] Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. Nature Genetics *35*, 57–64. PMID : 12897782.

[81] Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E., and Schadt, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature genetics *40*, 854–861.

[82] Ayroles, J. F., Carbone, M. A., Stone, E. A., Jordan, K. W., Lyman, R. F., Magwire, M. M., Rollmann, S. M., Duncan, L. H., Lawrence, F., Anholt, R. R., et al. (2009). Systems genetics of complex traits in drosophila melanogaster. Nature genetics *41*, 299–307.

[83] Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics *37*, 710–717. PMID : 15965475.

[84] Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E. E., Drake, T. A., Lusis, A. J., et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet *2*, e130.

[85] Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428. PMID : 18344981.

[86] Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. Bioinformatics (Oxford, England) *18*, 51–60. PMID : 11836211.

[87] Biswas, S., Storey, J. D., and Akey, J. M. (2008). Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. BMC Bioinformatics *9*, 244. PMID : 18492285.

[88] Engreitz, J. M., Daigle, B. J., Marshall, J. J., and Altman, R. B. (2010). Independent component analysis : mining microarray data for fundamental human gene expression modules. Journal of Biomedical Informatics *43*, 932–944. PMID : 20619355.

[89] Cover, T. M. and Thomas, J. A. (2005). Elements of Information Theory. (Hoboken, NJ, USA : John Wiley & Sons, Inc.).

[90] Hyvarinen, A. and Oja, E. (2000). Independent component analysis : algorithms and applications. Neural Netw *13*, 411–430.

[91] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika *30*, 179–185.

[92] Zwick, W. R. and Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin *99*, 432–442.

[93] Holmes, R. B. (1991). On random correlation matrices. SIAM journal on matrix analysis and applications *12*, 239.

[94] Maio, A. D. (1999). Heat shock proteins : facts, thoughts, and dreams. Shock (Augusta, Ga.) *11*, 1–12. PMID : 9921710.

[95] Jakob, U., Gaestel, M., Engel, K., and Buchner, J. (1993). Small heat shock proteins are molecular chaperones. Journal of Biological Chemistry *268*, 1517 –1520.

[96] McLemore, E. C., Tessier, D. J., Thresher, J., Komalavilas, P., and Brophy, C. M. (2005). Role of the small heat shock proteins in regulating vascular smooth muscle tone. Journal of the American College of Surgeons *201*, 30–36. PMID : 15978441.

[97] Dhillon, O. S., Khan, S. Q., Narayan, H. K., Ng, K. H., Struck, J., Quinn, P. A., Morgenthaler, N. G., Squire, I. B., Davies, J. E., Bergmann, A., et al. (2010). Prognostic value of Mid-Regional Pro-Adrenomedullin levels taken on admission and discharge in Non-ST-Elevation myocardial infarction : The LAMP (Leicester acute myocardial infarction peptide) II study. J Am Coll Cardiol *56*, 125–133.

[98] Liuzzo, G., Biasucci, L. M., Gallimore, J. R., Grillo, R. L., Rebuzzi, A. G., Pepys, M. B., and Maseri, A. (1994). The prognostic value of c-reactive protein and serum amyloid a protein in severe unstable angina. The New England Journal of Medicine *331*, 417–424. PMID : 7880233.

[99] Patti, G., Sciascio, G. D., D'Ambrosio, A., Dicuonzo, G., Abbate, A., and Dobrina, A. (2002). Prognostic value of interleukin-1 receptor antagonist in patients under-

going percutaneous coronary intervention. The American Journal of Cardiology *89*, 372–376. PMID : 11835913.

[100] Meisinger, C., Prokisch, H., Gieger, C., Soranzo, N., Mehta, D., Rosskopf, D., Lichtner, P., Klopp, N., Stephens, J., Watkins, N. A., et al. (2009). A genome-wide association study identifies three loci associated with mean platelet volume. American Journal of Human Genetics *84*, 66–71. PMID : 19110211.

[101] Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S. F., Payne, F., et al. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nature Genetics *39*, 857–864. PMID : 17554260.

[102] Horvath, S. and Langfelder, P. (2008). WGCNA : an r package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

[103] Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree : the dynamic tree cut package for r. Bioinformatics *24*, 719 –720.

[104] Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., Li, Y., Sarwar, R., Langley, S. R., Bauerfeind, A., et al. (2010). A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature *467*, 460–464. PMID : 20827270.

[105] Horwitz, M. S., Bradley, L. M., Harbertson, J., Krahl, T., Lee, J., and Sarvennick, N. (1998). Diabetes induced by coxsackie virus : Initiation by bystander damage and not molecular mimicry. Nat Med *4*, 781–785.

[106] Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature Genetics *41*, 703–707. PMID : 19430480.

[107] Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PloS One *4*, e6098. PMID : 19568420.

[108] Vettese-Dadey, M. (1999). Going their separate ways : A profile of products for cell separation. The Scientist *13*, 21.

[109] Semple, J. W., Allen, D., Chang, W., Castaldi, P., and Freedman, J. (1993). Rapid separation of CD4+ and CD19+ lymphocyte populations from human peripheral blood by a magnetic activated cell sorter (MACS). Cytometry *14*, 955–960. PMID : 7507026.

[110] Stanciu, L. A., Shute, J., Holgate, S. T., and Djukanović, R. (1996). Production of IL-8 and IL-4 by positively and negatively selected CD4+ and CD8+ human t cells following a four-step cell separation method including magnetic cell sorting (MACS). Journal of Immunological Methods *189*, 107–115. PMID : 8576572.

[111] Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell type-specific gene expression differences in complex tissues. Nature Methods *7*, 287–289. PMID : 20208531.

[112] Watkins, N. A., Gusnanto, A., de Bono, B., De, S., Miranda-Saavedra, D., Hardie, D. L., Angenent, W. G. J., Attwood, A. P., Ellis, P. D., Erber, W., et al. (2009). A HaemAtlas : characterizing gene expression in differentiated human blood cells. Blood *113*, e1–9.

[113] Kieffer, L. J., Bennett, J. A., Cunningham, A. C., Gladue, R. P., McNeish, J., Kavathas, P. B., and Hanke, J. H. (1996). Human CD8 ? expression in NK cells but not cytotoxic t cells of transgenic mice. International Immunology *8*, 1617 –1626.

[114] Baume, D. M., Caligiuri, M. A., Manley, T. J., Daley, J. F., and Ritz, J. (1990). Differential expression of CD8[alpha] and CD8[beta] associated with MHC-restricted and non-MHC-restricted cytolytic effector cells. Cellular Immunology *131*, 352–365.

[115] Soranzo, N., Spector, T. D., Mangino, M., Kuhnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet *41*, 1182–1190.

[116] Shoji, H., Tsuchida, K., Kishi, H., Yamakawa, N., Matsuzaki, T., Liu, Z., Nakamura, T., and Sugino, H. (2000). Identification and characterization of a PDZ protein that interacts with activin type II receptors. The Journal of Biological Chemistry *275*, 5485–5492. PMID : 10681527.

[117] Liu, F., Shao, L. E., and Yu, J. (2000). Truncated activin type II receptor inhibits erythroid differentiation in k562 cells. Journal of Cellular Biochemistry *78*, 24–33. PMID : 10797563.

[118] Shinohara, H. and Kurosaki, T. (2009). Comprehending the complex connection between PKCbeta, TAK1, and IKK in BCR signaling. Immunological Reviews *232*, 300–318. PMID : 19909372.

[119] Soslau, G., Prest, P. J., Class, R., Jost, M., and Mathews, L. (2009). Inhibition of gamma-thrombin-induced human platelet aggregation by histone h1subtypes and h1.3 fragments. Platelets *20*, 349–356. PMID : 19637099.

[120] Fuller, W. A. (1987). Measurement error models. (John Wiley and Sons).

[121] Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. Proceedings of the National Academy of Sciences of the United States of America *102*, 13544.

[122] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545.

[123] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) *58*, 267–288.

Annexes

Liste des articles :

Travaux Publiés

- Heinig M*, Petretto E*, Wallace C, Bottolo L, Rotival M, et al. (2010). *A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk.* Nature 467, 460-464

- Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) *Genetics and Beyond - The Transcriptome of Human Monocytes and Disease Susceptibility.* PLoS ONE 5(5) : e10693. doi :10.1371/journal.pone.0010693

En préparation

- Rotival M, Zeller T, Wild P, Maouche S, Szymczak S et al., *Integrating Genome-Wide Genetic Variations and monocyte expression Data Reveals Trans-Regulated Gene Modules in Humans.*. En preparation pour une re-soumission à PloS Genetics.

- Rotival M, Zeller T, Truong V, Szymczak S, Castagné R et al. *Gene density influences GWAS discoveries and confounds the colocalization of GWAS loci with expression QTLs.* Soumis à l'European Journal of Human Genetics.

- Castagné R*, Rotival M*, Zeller T, Wild P, Truong V, et al. *Lack of evidence for dosage compensation of the active X-chromosome in human monocyte.* En préparation pour une soumission à genome research.

Article 1:

Lack of Evidence for Dosage Compensation of the active X-Chromosome in Human Monocytes

Lack of Evidence for Dosage Compensation of the Active X-Chromosome in Human Monocytes

Raphaële Castagné[1†], Maxime Rotival[1†], Tanja Zeller[2,3], Philipp S. Wild[2,3], Vinh Truong[1], David-Alexandre Trégouët[1], Thomas Münzel[2], Andreas Ziegler[4], François Cambien[1], Stefan Blankenberg[2,3], Laurence Tiret[1]

Abstract

Background: The hypothesis of dosage compensation of genes of the X chromosome, supported by previous microarray studies, was recently challenged by RNA-sequencing data. It was suggested that microarray studies were biased towards an over-estimation of X-linked expression levels as a consequence of the filtering of genes below the detection threshold of microarrays.

Methodology/Principal findings: To investigate this hypothesis, we used microarray expression data from circulating monocytes in 1,467 individuals. In total, 25,349 and 1,156 probes were unambiguously assigned to autosomes and the X chromosome, respectively. Globally, there was a clear shift of X-linked expressions towards lower levels than autosomes. We compared the ratio of expression levels of X-linked to autosomal transcripts (X:AA) under two different models: model 1 in which gene expressions were filtered using a statistical detection threshold irrespective of gene chromosomal location (equivalent to what is usually done in microarrays); model 2 in which the same proportions of genes were filtered separately on the X and on autosomes. For a wide range of filtering proportions, the X:AA ratio estimated under model 1 was not significantly different from 1, the value expected if dosage compensation was achieved, whereas it was significantly lower than 1 under model 2. The difference between the two models increased as the filtering became more stringent because of a greater truncation of lowly expressed genes on the X chromosome than on autosomes in model 1. These results were confirmed on simulated data. A similar pattern was observed for imprinted genes which are expressed in a single copy.

Conclusion/significance: This study leads to reject the hypothesis of dosage compensation of the X chromosome. It also shows that the method used for filtering lowly expressed genes may have a major impact on the results of microarray studies according to the hypothesis investigated.

[1] INSERM UMRS 937, Pierre and Marie Curie University (UPMC, Paris 6) and Medical School, Paris, F-75634, France
[2] II. Medizinische Klinik und Poliklinik, Johannes-Gutenberg Universität Mainz, Universitätsmedizin, Mainz, D-55131, Germany
[3] Present address: Department of General and Interventional Cardiology, University Heart Center, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
[4] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Lübeck, D-23538, Germany
[†] These authors contributed equally to this work
Corresponding Author: Laurence Tiret, INSERM UMRS 937, Faculté de Médecine Pitié-Salpêtrière, 91 bd de l'Hôpital, 75634 PARIS cedex 13, France
Email: laurence.tiret@upmc.fr

Keywords: Dosage compensation, Gene expression, X chromosome, Microarray

Introduction

It is widely admitted that in mammals, X-linked genes are upregulated to ensure balanced expression between the X chromosome, present in a single active copy per cell, and autosomes, present in two copies (Payer and Lee 2008). The hypothesis of dosage compensation first proposed by Ohno in 1967 (Ohno 1967) was supported by recent microarray studies showing that X-linked genes were expressed at similar levels to autosomal genes in mice and humans (Gupta et al. 2006; Nguyen and Disteche 2006; Johnston et al. 2008). However, the molecular mechanism responsible for this compensation is still unknown. Recently, Ohno's hypothesis was challenged by a study using RNA sequencing (RNA-Seq) data showing that the ratio of the median expression level of X-linked genes to that of autosomal genes (X:AA) was significantly lower than 1 in different human and mouse tissues (Xiong et al. 2010). The authors attributed the difference between their findings and previous ones to the fact that RNA-Seq is much more sensitive than microarray to detect small expression differences (Xiong et al. 2010; Sultan et al. 2008; Marioni et al. 2008) and that microarray studies are likely to be biased towards an over-estimation of X-linked expression levels as a consequence of the filtering of genes considered to be under the detection threshold of microarray (Xiong et al. 2010). This controversial finding led us to question the method conventionally used for the analysis of microarray-based expression data. Using data from a large-scale expression study in human monocytes (Zeller et al. 2010), we show that when appropriately correcting for the bias associated with the filtering of lowly expressed genes in microarrays, expression levels of X-linked genes are significantly lower than those of autosomal genes, leading to reject the hypothesis of dosage compensation.

Results

Filtering transcripts considered as undetected by microarrays may discard genes that show biologically relevant associations supporting cellular expression

In microarray studies, genes whose expression is not significantly different from the background signal are conventionally filtered out prior to analysis. These genes are often inappropriately considered as unexpressed in the cell type under study although they are only undetected. Recent studies based on RNA-seq have shown that a fraction of the genes undetected by microarrays were actually expressed at low levels in the cells investigated (Xiong et al. 2010; Sultan et al. 2008). This filtering based on a statistical detection criterion was justified in former small microarray studies which were mainly designed for discovering large expression differences between contrasted experimental conditions and therefore focused on highly expressed genes. However, it may be less founded in current large-scale transcriptomic studies which are more interested in characterizing the natural sources of variability of gene expression, such as genetic variations, environmental exposures, metabolic conditions, ageing or gender (Zeller et al. 2010). In a microarray experiment, the expression of a gene which is considered undetected according to a detection threshold may found related to a SNP or another relevant factor that provides biological evidence that the gene is expressed. This is a problem known as signal-to-noise in biology (Ideker et al. 2011).

To illustrate this issue, we re-analyzed the data of a previous study in which gene expression was simultaneously measured by microarray and RNA-Seq in two different cell lines, HEK and B cells (Sultan et al. 2008). This study reported 25% more genes detected by RNA-Seq than by microarray due to the greater sensitivity of RNA-Seq. When the authors focused on genes detected by both platforms and in both cell lines (n=7,043), they showed that the differences of gene expressions between HEK and B cells (measured by the log ratio of expression) strongly correlated across the two platforms ($r = 0.88$) in spite of a compression effect resulting in smaller ratios in microarrays. This result indicated that true biological differences between cell types were

reproducibly found across platforms. We performed the same analysis on the genes that were detected by RNA-Seq (at least five reads) but were undetected by microarray (detection score < 0.95) (1,640 genes). As shown in Figure 1 plotting the log ratio of expression (B versus HEK cell) measured by the two platforms, there was a subset of genes lying along the diagonal in which differential expression between HEK and B cells strongly correlated across the two platforms. For these genes which are likely to be truly differentially expressed between cells, the difference could be detected by microarrays even though their expression level was considered not different from the background noise. This demonstrates that for genes below the detection level of microarrays, biologically relevant signals can be found that indicate that the gene is truly expressed.

When appropriately correcting for the artifact due to the filtering of genes in microarrays, the hypothesis of dosage compensation of the active X chromosome in human monocytes is not supported

To investigate the hypothesis of dosage compensation of X-linked genes, we used expression data from the Gutenberg Heart Study (GHS), a population-based study in which the transcriptome of circulating monocytes was assessed in 1,467 unrelated subjects (51.1% of men) using the *Illumina* HT-12 v3 BeadChip

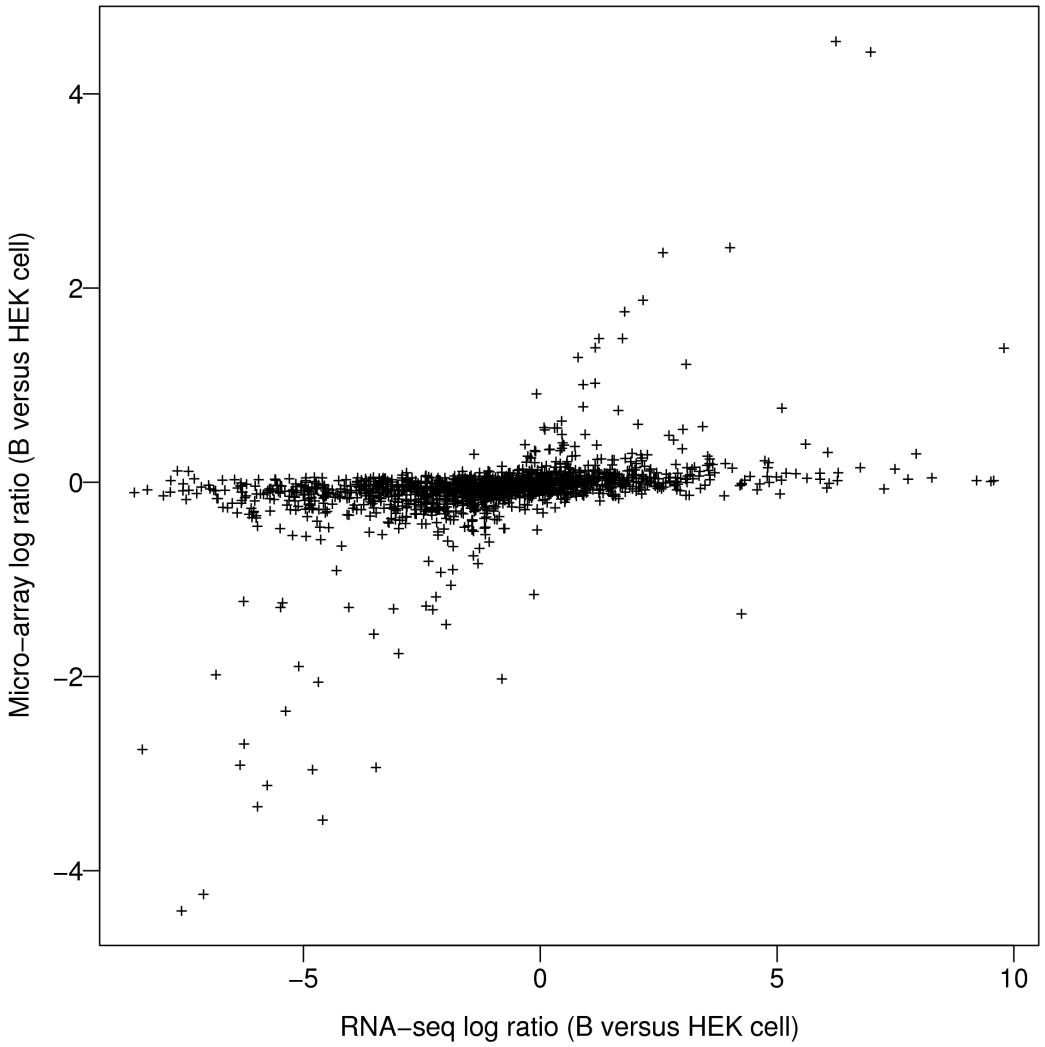


Figure 1. Comparison of differentially expressed genes (B versus HEK cells) by RNA sequencing and microarrays. Data are drawn from Sultan et al. (Sultan et al. 2008) and genes detected by RNA-Seq (at least five reads) but not detected by microarrays (detection score < 0.95) were selected (1,640 genes in total). The plot shows log2 ratios of expression in RNA-Seq (x axis) and microarrays (y axis).

(Zeller et al. 2010). After removing probes with a bad quality score according to ReMOAT (Barbosa-Morais et al. 2010) 25,349 and 1,156 probes were unambiguously assigned to the autosomes and the X chromosome, respectively. Analyses were performed at the probe level and for simplicity the term of "transcript" was used to denote a unique probe-hybridization product (although in few cases the same transcript could be targeted by several probes or conversely, a same probe could target several transcripts). Analyses were performed in males and females separately.

As usually performed in microarray experiments, we first selected the transcripts whose expression was detected in ≥ 95% of female samples or ≥ 95% of male samples. This filtering resulted in the selection of 10,896 autosomal and 360 X-linked transcripts, the vast majority of them being detected in both genders. The proportion of transcripts filtered prior to analysis was 57.5% as a whole, but it was much higher on the X chromosome than on autosomes (68.9% vs 57.0%, $P < 10^{-6}$). In the subset of selected transcripts, the median level of expression of X-linked transcripts was not significantly different from that of autosomal transcripts in both sexes (7.68 vs 7.87, $P = 0.09$ in males; 7.71 vs 7.85, $P = 0.12$ in females). As previously reported (Johnston et al. 2008), expression levels of X-linked genes fell within the global range of autosomes, suggesting that dosage compensation of X-linked genes was globally achieved in both sexes (Figures 2A and 2B).

However, as shown by the quantile functions of expression levels plotted separately for the X chromosome and for autosomes, there was a clear shift of X-linked expressions towards lower levels than autosomes in both sexes (Figures 3A and 3B). As a consequence, taking a common detection threshold for the X and for autosomal chromosomes led to a greater truncation of lowly expressed transcripts on the X chromosome than on autosomes (as shown by the horizontal plain red lines in Figures 3A and 3B) resulting in an over-estimation of X-linked expressions in the subset of filtered genes. In order to circumvent this artefact, we compared expression levels between X-linked and autosomal transcripts after excluding equal proportions of the less expressed transcripts on the X and on autosomes separately (as shown by the vertical green lines in Figures 3A and 3B). When excluding the same proportion of transcripts as above (57.5%), but considering this time the X chromosome and autosomes separately, the difference of median expression levels between X-linked and autosomal transcripts became highly significant in both sexes (6.82 vs 7.91, $P < 10^{-31}$ in males; 6.85 vs 7.91, $P < 10^{-30}$ in females).

We then compared the X:AA ratio of expression level of X-linked transcripts to autosomal transcripts when filtering either a global proportion of the lowest gene expressions irrespective of their chromosome location (model 1 which is equivalent to considering a common detection threshold) or the same proportion of genes on the X and on autosomes separately (model 2). In order to investigate the impact of the filtering threshold on the estimation of the X:AA ratio, we varied the proportion of filtered transcripts, which corresponded to moving the horizontal red lines from the bottom to the top (model 1) or the vertical green lines from the left to the right (model 2) in Figures 3A and 3B. As shown in Figures 4A and 4B, the X:AA ratio was always higher when using a common filtering threshold not depending on the gene chromosomal location (model 1, red triangles) than when using a chromosome-specific threshold (model 2, green circles). The difference between the two X:AA estimates increased as the filtering became more stringent as a result of a greater truncation in model 1 of lowly expressed genes on the X chromosome than on autosomes. When the proportion of genes filtered prior to analysis was greater or equal to 50%, the X:AA ratio estimated when taking a common filtering threshold was no longer significantly different from 1, the value expected if there was dosage compensation, whereas it was significantly lower than 1 when taking a chromosome-specific threshold. Results were very similar in males and females (Figure 4A and 4B).

We estimated the X:AA ratio for each autosome

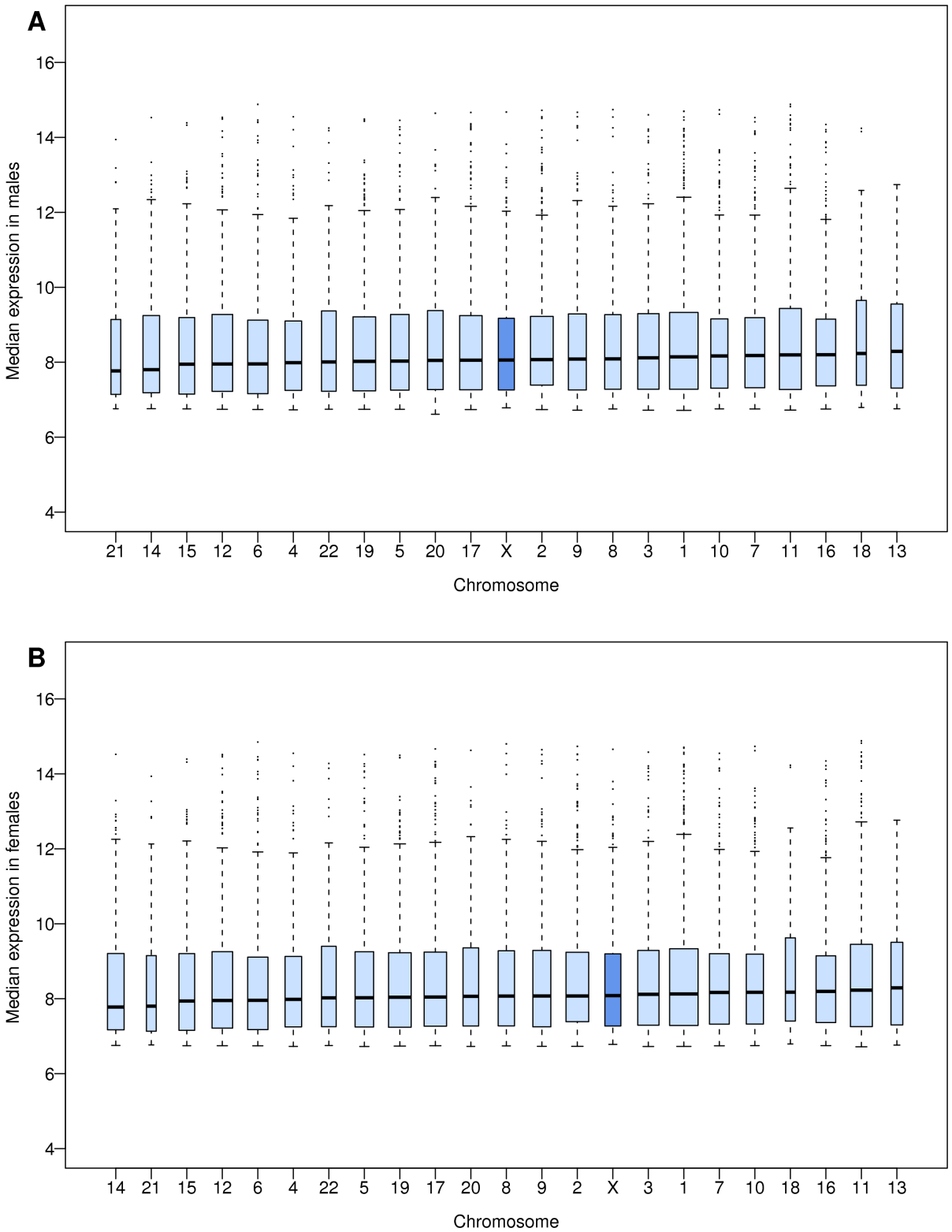


Figure 2. Box plots of the median expression levels in human monocytes according to chromosome when selecting the transcripts detected in at least 95 % of individuals. (A) Males and (B) Females.

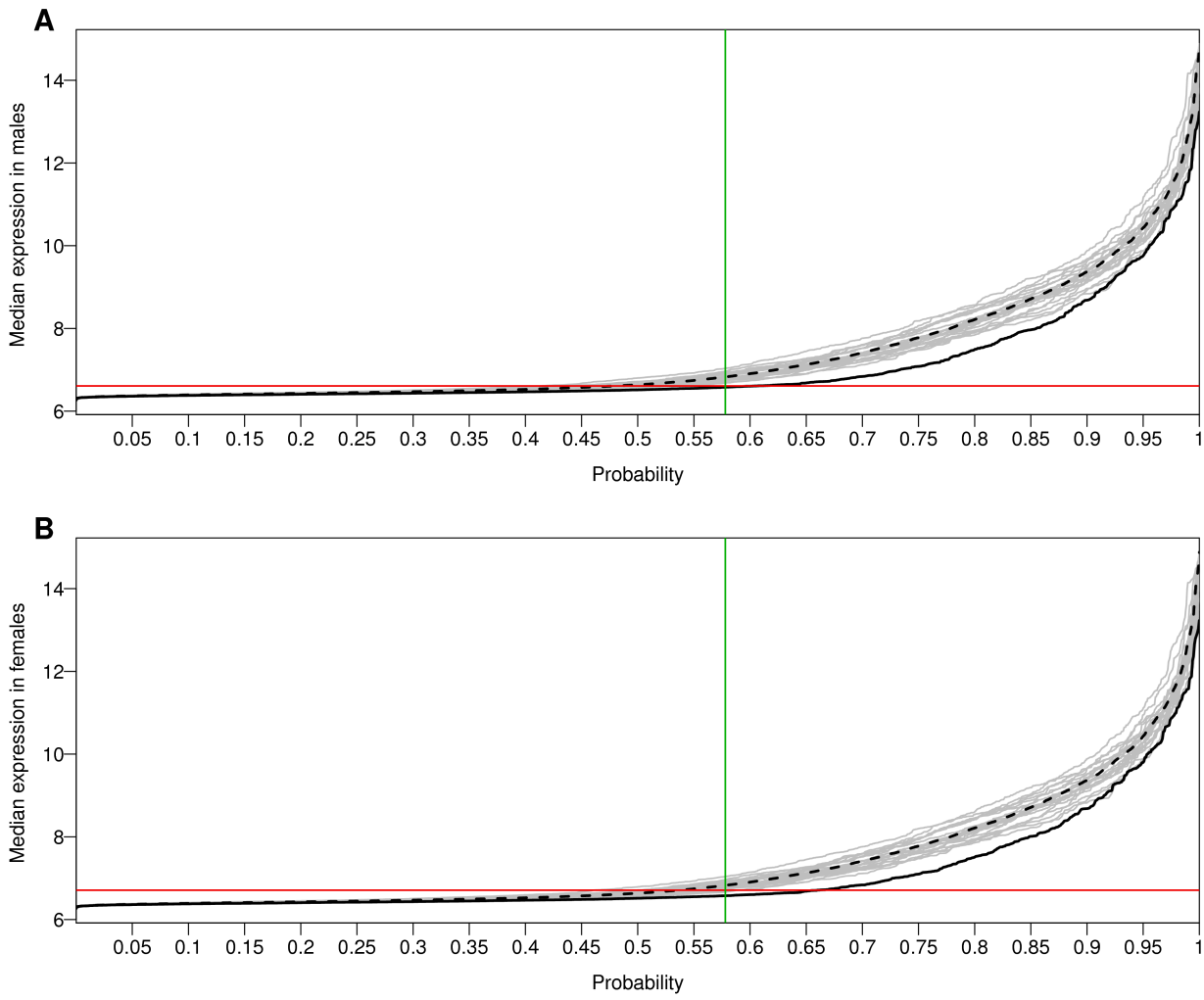


Figure 3. Quantile functions of median expression levels of X-linked and autosomal transcripts in human monocytes. (A) Males and (B) Females: X-linked transcripts are shown by the plain black curve, autosomal transcripts by the dashed black curve and each autosome by an individual grey curve. For a probability p (x-axis), the y-axis shows the median expression level below which px100% of transcripts fall. The quantile function of the X chromosome is always below that of autosomes, indicating lower expression levels for X-linked than for autosomal transcripts. The horizontal red line corresponds to the filtering performed when selecting transcripts detected in ≥ 95% of individuals. The vertical green line corresponds to excluding equal proportions (57.5%) of the less expressed genes on the X chromosome and on autosomes separately.

individually after filtering the same proportion (50%) of the lowest gene expressions on each chromosome. The X:AA ratio was different from 1 for all autosomes (Figure 4C and 4D). Worthy of note, the X:AA ratio associated to chromosome 21, which was the highest in human liver RNA-seq data (Xiong et al. 2010), was also the highest in human monocytes.

To validate these results, we simulated expression data under two different models assuming full dosage compensation (hypothesis 1, X:AA ratio=1) or no dosage compensation (hypothesis 2, X:AA ratio=0.5). Transcript levels were simulated according to the model proposed by Lin et al. (Lin et al. 2008) with parameters based on the empirical values observed in the GHS dataset (see Methods). Under hypothesis 1, the distribution of X-linked transcripts was taken not different from that of autosomal genes whereas under hypothesis 2, mean levels of X-linked transcripts were divided by 2 to mimic the inactivation of one X copy. Data were normalized and transformed using VST. The X:AA ratio was then estimated under the two contrasted hypotheses of dosage compensation and using the two different methods for filtering gene expressions. The simulation was repeated 10,000 times to generate confidence intervals.

Under the hypothesis of full dosage

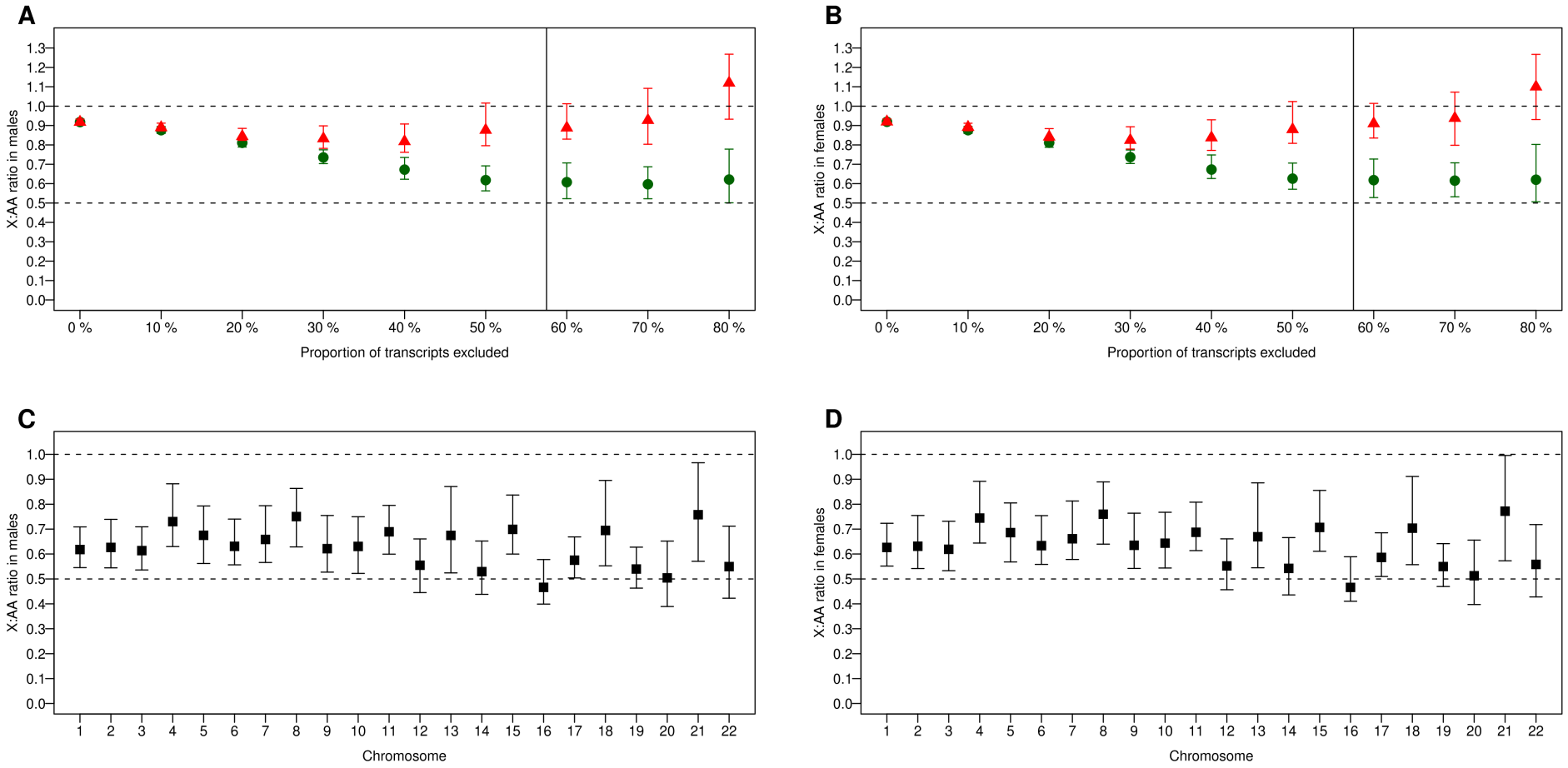


Figure 4. Comparison of expression levels between the X chromosome and autosomes in human monocytes. (A) Males and (B) Females: The graph plots the X:AA ratio of median expression of X-linked genes to autosomal genes according to the proportion of transcripts filtered prior to analysis, using either a common threshold (red triangles) or individual thresholds on the X and on autosomes (green circles). Error bars show the 95% bootstrap confidence intervals. The horizontal dashed lines show the ratios expected if there was no dosage compensation (X:AA = 0.5) or full compensation (X:AA = 1). The vertical line corresponds to the proportion of genes filtered when using a P-detection ≥ 95%. (C) Males and (D) Females: X:AA ratios when the X is compared to individual autosomes and the same proportion of transcripts (50%) is filtered on the X and on each autosome.

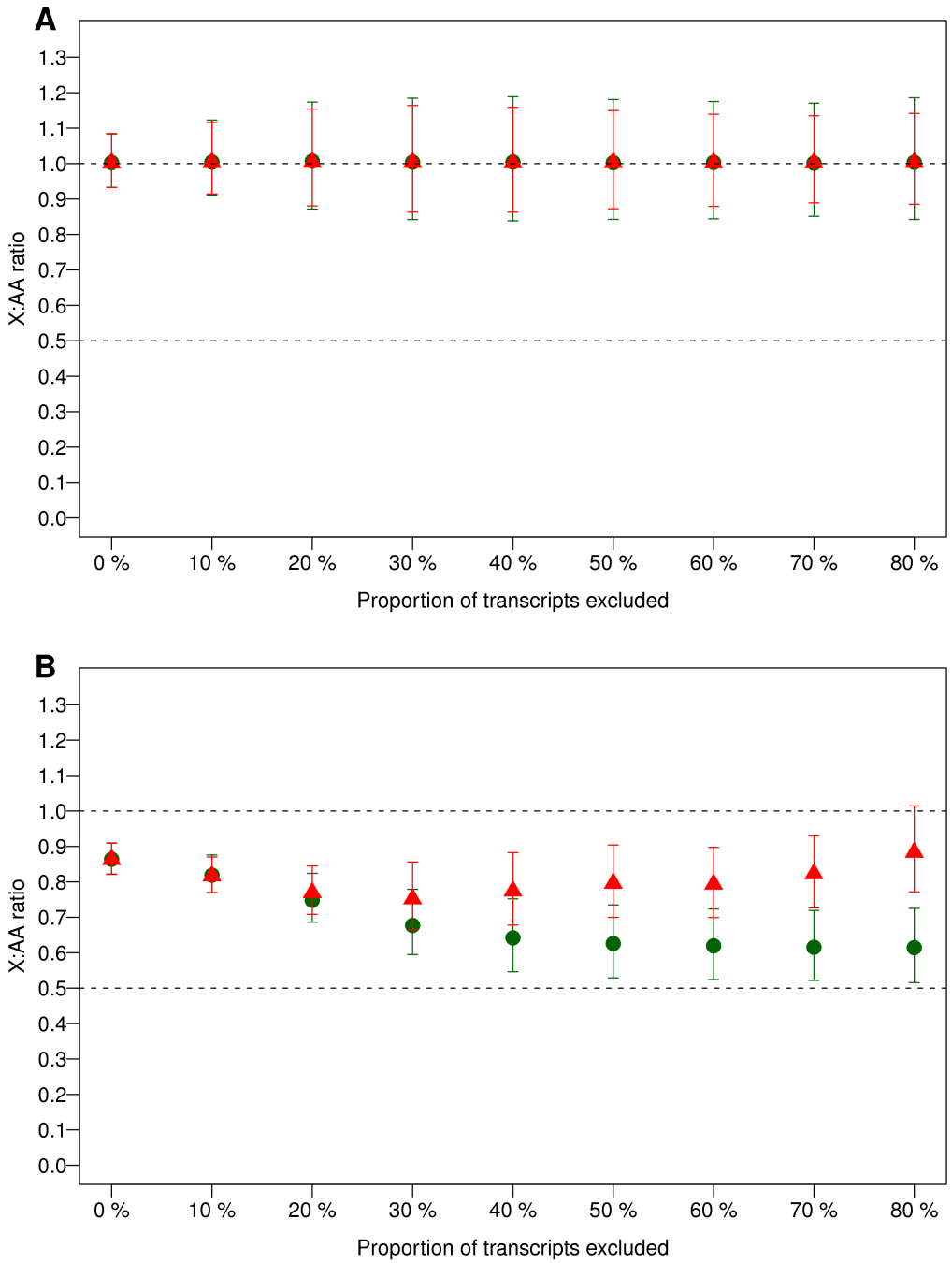


Figure 5. Comparison of expression levels between the X chromosome and autosomes on data simulated under a model of full dosage compensation (A) or lack of dosage compensation (B). See legend of figures 4A and 4B.

compensation, the X:AA ratio did not differ from 1 whatever the method used for filtering and the proportion of genes excluded (Figure 5A). By contrast, under the hypothesis of lack of dosage compensation, we observed a pattern similar to the one found in the real data, that is an X:AA ratio always lower and closer to 0.5 when using chromosome-specific filtering proportions than when using a common filtering (Figure 5B). The fact that the true ratio of 0.5 was not reached is likely explained by the compression of low expression values by the VST transformation. This result on simulated data demonstrates that filtering genes irrespective of their X/autosomal location results in an underestimation of the true differences of expression between the X chromosome and autosomes.

Dosage compensation of X-linked genes in other human tissues

To check whether the same conclusions could be drawn from other tissues, we analyzed publicly available microarray expression data from three different human tissues, the meibomian glands (access number GSE17822), the muscle (access number GSE20319) and the colon (access number GSE26305). For each tissue dataset, we selected either the transcripts for which the detection score was greater than 95% (filtering not depending of chromosome), or the same proportion (arbitrarily taken to 50%) of the most highly expressed transcripts separately on the X chromosome and on autosomes. Whatever the tissue under consideration, the differences of expression levels between X-linked and autosomal genes were always more pronounced when filtering separately genes on the X and on autosomes, indicating that the conclusion drawn from monocytes was not restricted to that specific tissue (Figure 6).

Comparison with genes submitted to genomic imprinting

In mammals, genomic imprinting affects a small proportion (< 1%) of autosomal genes and results in the expression of only one allele inherited from the father or the mother. In terms of expressed alleles, imprinted genes are thus comparable to X-linked genes which are submitted to inactivation of one of the two alleles. We hypothesized that imprinted genes may exhibit a similar pattern of expression to the one observed in X-linked genes.

To test this hypothesis, we compared the levels of imprinted transcripts to those of non-imprinted transcripts in the GHS dataset. Among the 10,806 well-annotated autosomal probes, 97 probes (list in Table S1) corresponded to genes that were reported to be submitted to imprinting in two databases (http://igc.otago.ac.nz/home.html and http://www.geneimprint.com/). When first selecting the transcripts whose expression was detected in ≥ 95% of samples, the proportion of transcripts filtered prior to analysis was much higher for imprinted than for non-imprinted transcripts (72.2% vs 57.3%, $P < 10^{-6}$). In the subset of selected transcripts, the median level of expression of imprinted transcripts was not significantly different from that of non-imprinted transcripts (8.32 vs 8.08, $P = 0.88$). By contrast, when selecting a similar proportion (arbitrarily taken to 50%) of the most highly expressed transcripts separately among imprinted and non-imprinted genes, expression levels were significantly lower in imprinted transcripts than in non-imprinted ones (6.89 vs 7.78, $P < 0.001$) (Figure 7).

Discussion

The present study raised a critical issue related to the way of defining that a gene is expressed in a given cell type. In microarray studies, it is generally advocated to select only the genes whose expression is detected in the majority of samples, the remaining genes being considered as not expressed in the cell type under study. However, with the advent of more sensitive techniques like RNA-Seq as well as the greater power of contemporary transcriptomic studies, it is realized that many genes considered as non expressed in microarray experiments are actually expressed at low levels, or only in a fraction of the population, for example when expression is modulated by a genetic or an environmental factor.

This issue is particularly well illustrated by the analysis of X-linked gene expression. Analyzing data by conventional filtering methods used in microarrays would lead to the conclusion that expression levels of X-linked genes do not differ from those of autosomal genes in human monocytes as previously reported by microarray studies (Gupta et al. 2006; Nguyen and Disteche 2006; Johnston et al. 2008). This result is the consequence of a higher proportion of filtered genes on the X chromosome than on autosomes. When correcting for this bias, we showed that X-linked genes are expressed at lower levels than autosomal genes, in keeping with recent findings based on RNA-sequencing (Xiong et al. 2010). This result was not specific of monocytes and could be extended to other human cell types. Moreover, a similar pattern of

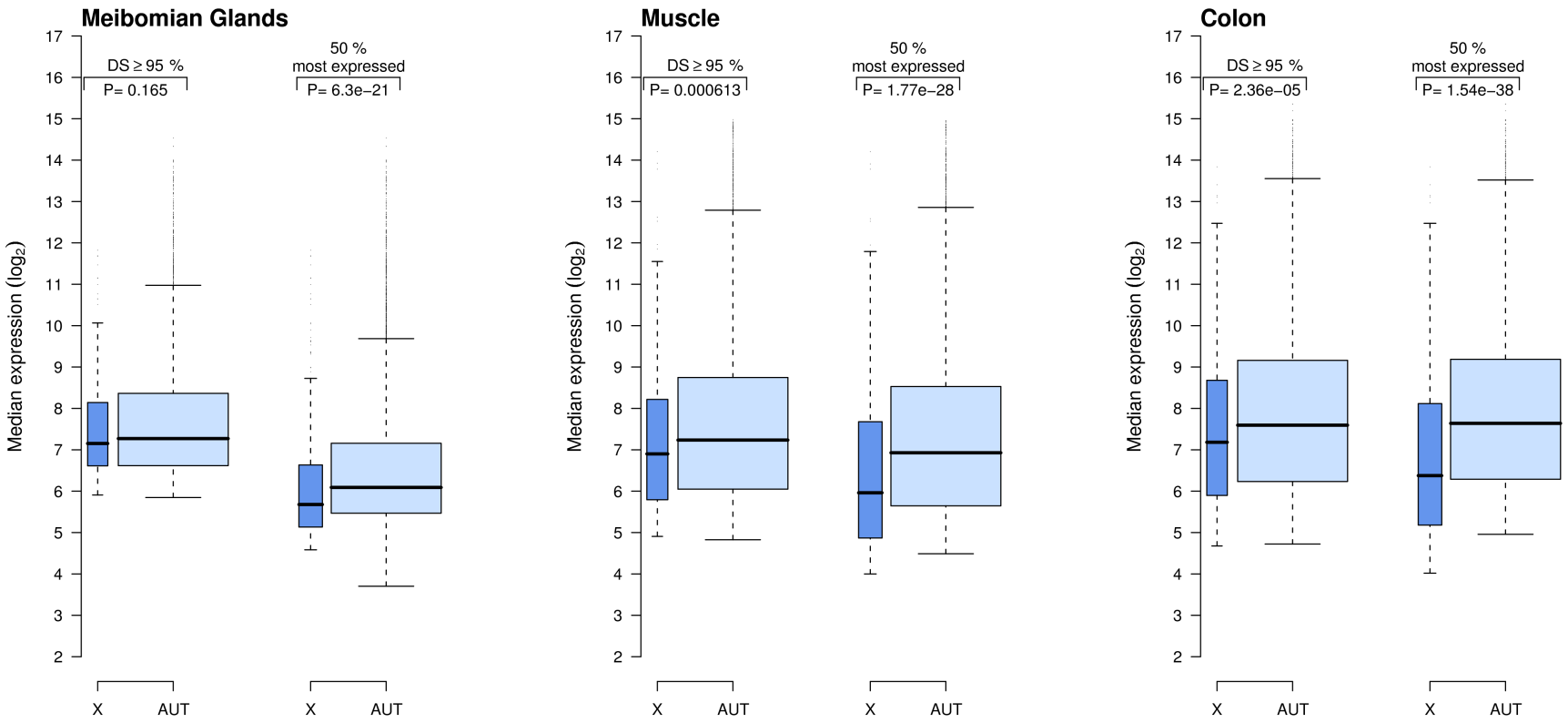


Figure 6. Expression levels of X-linked and autosomal transcripts in different human tissues. The graph shows boxplots of expression levels either when filtering the genes according to a common detection threshold (detection score ≥ 0.95) or when excluding the 50% lowest gene expressions separately on the X chromosome and on autosomes.

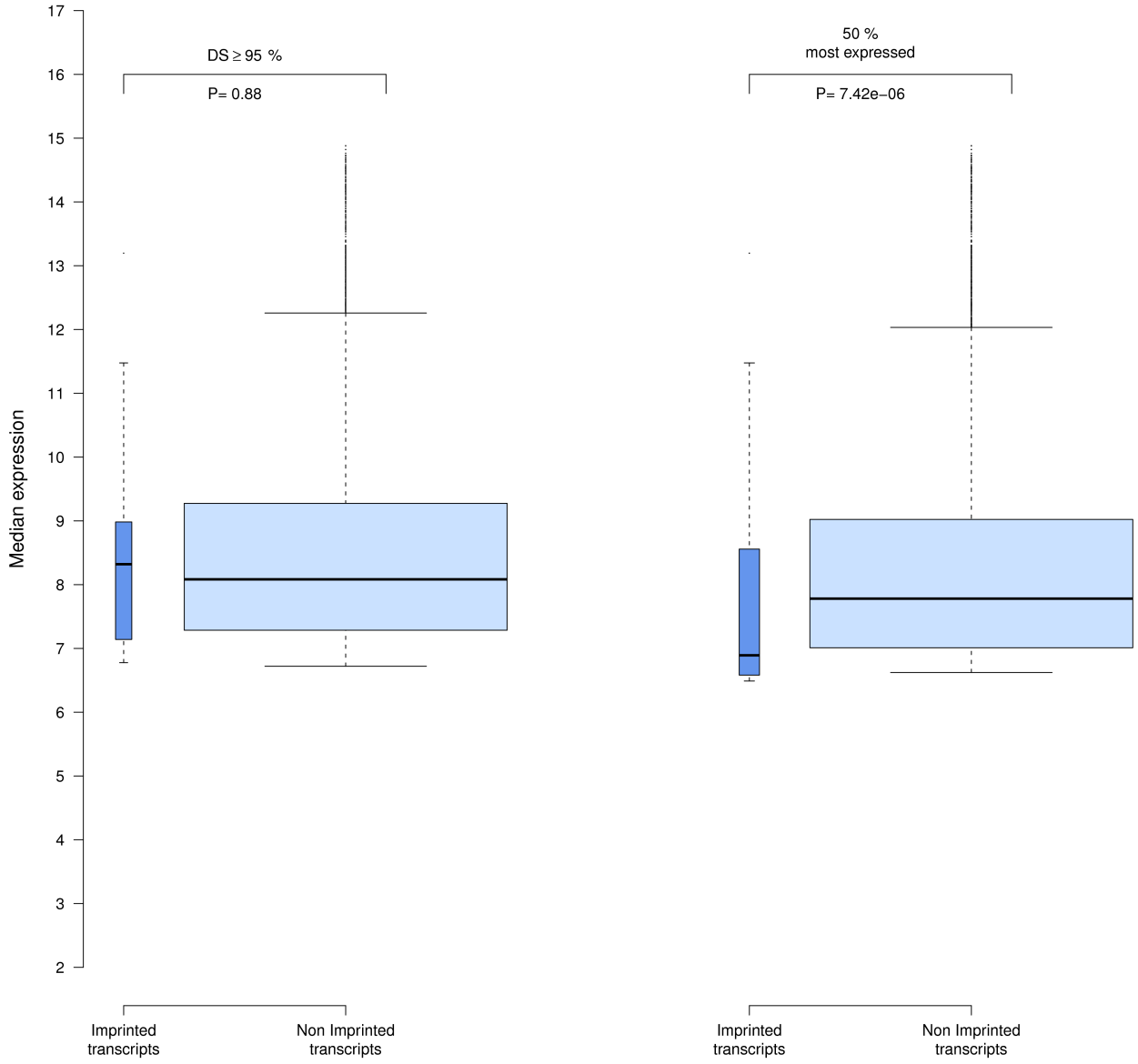


Figure 7. Expression levels of imprinted and non-imprinted autosomal transcripts in human moncoytes. The graph shows boxplots of expression levels either when filtering the genes according to a common detection threshold (detection score ≥ 0.95) or when excluding the 50% lowest gene expressions separately in imprinted and non-imprinted genes.

under-expression was observed for imprinted genes, another category of genes that are expressed in a single copy.

It has been suggested that up to 15% of X-linked genes escape inactivation and are expressed from the two X chromosomes in females(Carrel and Willard 2005). This phenomenon might result in different patterns of expression in males and females. However, for those genes that escape inactivation, expression from the inactive X chromosome is generally partial and/or occurs only in a fraction of females, resulting in relatively small differences of expression between sexes(Carrel and Willard 2005). This is probably why we did not observe major differences of patterns between males and females in our analyses.

In conclusion, the present analysis leads to reject the hypothesis of dosage compensation of the X chromosome in humans. It also shows that confounding the true cellular expression with the statistical detection level may be highly misleading according to the hypothesis investigated.

Materials and Methods

Ethic statement

The study protocol and drawing of the blood sample have been approved by the local ethics committee and by the local and federal data safety commissioners (Ethik-Kommission der Landesärztekammer Rheinland-Pfalz). All subjects included signed an informed consent.

Study Population

The study has been described in details elsewhere (Zeller et al. 2010). Study participants of both sexes aged 35-74 yr, were successively enrolled into the Gutenberg Heart Study (GHS), a community-based single centre cohort study conducted in the Rhein-Main region in western mid-Germany. All subjects were of European descent. Individuals for whom we found a discrepancy between the phenotypic gender and the sex inferred from expression of Y-linked transcripts were excluded, leaving 1,467 individuals for analysis (750 men and 717 women).

Genome-wide expression

Genome-wide expression profiles were assessed from peripheral blood monocytes. Separation of monocytes was conducted within 60 min after blood collection by negative selection using RosetteSep Monocyte Enrichment Cocktail (StemCell Technologies, Vancouver, Canada). Total RNA was extracted the same day using Trizol extraction and purification by silica-based columns. Expression profiles were assessed using the Illumina HT-12 v3 BeadChip. The pre-processing of data was performed using Beadstudio. Values from probes with ≤ 1 bead were re-imputed using the SVD impute from the pcaMethods R package. Data were normalized using quantile normalization and VST transformation (Lin et al. 2008) as implemented in the lumi R package (Du et al. 2008). After removing probes with a bad quality score according to ReMOAT (http://remoat.sysbiol.cam.ac.uk), 25,349 and 1,156 probes were unambiguously assigned to the autosomes and the X chromosome, respectively.

Statistical Analysis

Analyses were performed at the probe level. To select the transcripts that were detected in ≥ 95% of individuals, we used the detection *P*-values provided by the *Illumina* software and considered that a transcript was detected in a sample when the detection *P*-value for that sample was < 0.05. Median expression levels were compared between the X chromosome and autosomes using a Mann-Whitney U test. The X:AA ratio was the ratio between median expression levels of X-linked genes to median expression levels of autosomal genes. The 95% bootstrap confidence interval was estimated by resampling 1000 times the datasets of X-linked and autosomal transcripts.

We estimated the X:AA ratio using either a common filtering proportion of genes irrespective of the chromosomal location, or a proportion specific of the X/autosomal location. In the former case, we excluded the *k*% lowest transcripts among all transcripts, whereas in the latter case, we excluded the *k*% lowest X-linked transcripts and the *k*% lowest autosomal transcripts (*k* varying from 0% to 80%). For the comparison of the X chromosome to individual autosome, this proportion was taken chromosome-specific.

Simulations

We simulated expression levels in 25,349 autosomal transcripts and 1,156 X-linked transcripts under two contrasted models assuming either full dosage compensation (model 1) or lack of dosage compensation (model 2). For each transcript, expression level was simulated using the model proposed by Lin et al. (Lin et al. 2008):

$$\hat{\mu} = B + \mu e^{\eta}$$

where $\hat{\mu}$ is the observed transcript level, μ is the noise-free expression level, B is the background error following a Gaussian distribution (μ_B,σ_B) and η the multiplicative error following a Gaussian distribution (0,σ_η). To mimic real expression data, μ was sampled from the

empirical distribution of untransformed, background-corrected, expression levels of autosomal genes observed in GHS data. In model 2, the value of μ for X-linked transcripts was multiplied by 0.5 to mimic the inactivation of one X copy whereas in model 1, it was left unchanged. Values for μ_B and σ_B were taken from the empirical distribution of the negative controls provided by *Illumina* while σ_η was estimated from the relation between the bead average expression and the bead standard error as in Lin et al. (Lin et al. 2008). Simulated data were then transformed using the VST transformation and the X:AA ratio was computed under the two different models. The simulation was repeated 10,000 times to generate confidence intervals.

All analyses were performed in R v. 2.10.1.

Acknowledgments

The Gutenberg Heart Study is funded through the government of Rheinland-Pfalz ("Stiftung Rheinland Pfalz für Innovation", contract AZ 961-386261/733), the research programs "Wissen schafft Zukunft" and "Schwerpunkt Vaskuläre Prävention" of the Johannes Gutenberg-University of Mainz and its contract with Boehringer Ingelheim and PHILIPS Medical Systems including an unrestricted grant for the Gutenberg Heart Study. The present study was supported by the National Genome Network "NGFNplus" (contract A3 01GS0833 and 01GS0831) and by a joint funding from the Federal Ministry of Education and Research, Germany (contract BMBF 01KU0908A) and from the Agence Nationale de la Recherche, France (contract ANR 09 GENO 106 01) for the project CARDomics. MR is supported by a grant from the Fondation pour la Recherche Médicale (FDT20101220928).

Author Contributions

Conceived and designed the experiments: TZ, TM, AZ, FC, SB, LT.
Performed the experiments: TZ, PSW.
Analyzed the data: RC, MR, VT, DAT.
Wrote the paper: RC, MR, LT.

Supporting Information

Supplemental data include one supplementary table (Table S1).

References

Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JFJ, Ritchie ME, Lynch AG, and Tavaré S. 2010. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res.* **38**: e17.

Carrel L, and Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**: 400-404.

Du P, Kibbe WA, and Lin SM. 2008. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**: 1547-1548.

Gupta V, Parisi M, Sturgill D, Nuttall R, Doctolero M, Dudko OK, Malley JD, Eastman PS, and Oliver B. 2006. Global analysis of X-chromosome dosage compensation. *J. Biol.* **5**: 3.

Ideker T, Dutkowski J, and Hood L. 2011. Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power. *Cell* **144**: 860-863.

Johnston CM, Lovell FL, Leongamornlert DA, Stranger BE, Dermitzakis ET, and Ross MT. 2008. Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS Genet.* **4**: e9.

Lin SM, Du P, Huber W, and Kibbe WA. 2008. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.* **36**: e11.

Marioni JC, Mason CE, Mane SM, Stephens M, and Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**: 1509-1517.

Nguyen DK, and Disteche CM. 2006. Dosage compensation of the active X chromosome in mammals. *Nat. Genet.* **38**: 47-53.

Ohno S. 1967. *Sex Chromosomes and Sex-linked Genes*. Springer,Berlin. Springer,Berlin.

Payer B, and Lee JT. 2008. X chromosome dosage compensation: how mammals keep the balance. *Annu. Rev. Genet* **42**: 733-772.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**: 956-960.

Xiong Y, Chen X, Chen Z, Wang Xunzhang, Shi S, Wang Xueqin, Zhang J, and He X. 2010. RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat. Genet* **42**: 1043-1047.

Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, et al. 2010. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5**: e10693.

Article 2:

Genetics and Beyond - The Transcriptome of Human Monocytes and Disease Susceptibility

OPEN ACCESS Freely available online

PLoS one

Genetics and Beyond – The Transcriptome of Human Monocytes and Disease Susceptibility

Tanja Zeller[1], Philipp Wild[1], Silke Szymczak[2], Maxime Rotival[3], Arne Schillert[2], Raphaele Castagne[3], Seraya Maouche[3], Marine Germain[3], Karl Lackner[4], Heidi Rossmann[4], Medea Eleftheriadis[1], Christoph R. Sinning[1], Renate B. Schnabel[1], Edith Lubos[1], Detlev Mennerich[5], Werner Rust[5], Claire Perret[3], Carole Proust[3], Viviane Nicaud[3], Joseph Loscalzo[6], Norbert Hübner[7], David Tregouet[3], Thomas Münzel[1], Andreas Ziegler[2], Laurence Tiret[3], Stefan Blankenberg[1]*[9], François Cambien[3]*[9]

1 Medizinische Klinik und Poliklinik, Johannes-Gutenberg Universität Mainz, Mainz, Germany, **2** Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Lübeck, Germany, **3** INSERM UMRS 937, Pierre and Marie Curie University and Medical School, Paris, France, **4** Institut für Klinische Chemie und Laboratoriumsmediizin, Johannes-Gutenberg Universität Mainz, Mainz, Germany, **5** Boehringer Ingelheim Pharma GmbH and Co. KG, Biberach, Germany, **6** Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, **7** Max Delbrück Center for Molecular Medicine, Berlin, Germany

Abstract

Background: Variability of gene expression in human may link gene sequence variability and phenotypes; however, non-genetic variations, alone or in combination with genetics, may also influence expression traits and have a critical role in physiological and disease processes.

Methodology/Principal Findings: To get better insight into the overall variability of gene expression, we assessed the transcriptome of circulating monocytes, a key cell involved in immunity-related diseases and atherosclerosis, in 1,490 unrelated individuals and investigated its association with >675,000 SNPs and 10 common cardiovascular risk factors. Out of 12,808 expressed genes, 2,745 expression quantitative trait loci were detected ($P<5.78\times10^{-12}$), most of them (90%) being *cis*-modulated. Extensive analyses showed that associations identified by genome-wide association studies of lipids, body mass index or blood pressure were rarely compatible with a mediation by monocyte expression level at the locus. At a study-wide level ($P<3.9\times10^{-7}$), 1,662 expression traits (13.0%) were significantly associated with at least one risk factor. Genome-wide interaction analyses suggested that genetic variability and risk factors mostly acted additively on gene expression. Because of the structure of correlation among expression traits, the variability of risk factors could be characterized by a limited set of independent gene expressions which may have biological and clinical relevance. For example expression traits associated with cigarette smoking were more strongly associated with carotid atherosclerosis than smoking itself.

Conclusions/Significance: This study demonstrates that the monocyte transcriptome is a potent integrator of genetic and non-genetic influences of relevance for disease pathophysiology and risk assessment.

Citation: Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) Genetics and Beyond – The Transcriptome of Human Monocytes and Disease Susceptibility. PLoS ONE 5(5): e10693. doi:10.1371/journal.pone.0010693

Editor: Zoltán Bochdanovits, VU University Medical Center and Center for Neurogenomics and Cognitive Research, VU University, Netherlands

Received March 15, 2010; **Accepted** April 26, 2010; **Published** May 18, 2010

Funding: The Gutenberg Heart Study is funded through the government of Rheinland-Pfalz (Stiftung Rheinland Pfalz für Innovation, contract number AZ 961-386261/733), the research programs Wissen schafft Zukunft and Schwerpunkt Vaskuläre Prävention of the Johannes Gutenberg-University of Mainz and its contract with Boehringer Ingelheim and PHILIPS Medical Systems including an unrestricted grant for the Gutenberg Heart Study. Specifically, the research reported in this article was supported by the National Genome Network NGFNplus (contract number project A3 01GS0833) by the Federal Ministry of Education and Research, Germany. For this particular research paper, Boehringer Ingelheim provided payment of two employees for expression microarray analyses and array purchase. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: In this study, Boehringer Ingelheim provided payment for two employees for expression microarray analyses and array purchase and PHILIPS Medical Systems provided instruments for ultrasound studies. However, this does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials, as detailed online in the guide for authors of the Journal.

* E-mail: blankenberg@2-med.klinik.uni-mainz.de (SB); francois.cambien@upmc.fr (FC)

[9] These authors contributed equally to this work.

Introduction

The transcriptome, i.e. the whole set of RNA transcripts in a cell, is generally conceived as a system whose major function is to pass information encoded in the genome sequence to the realm of phenotypes that underlie physiological and pathological traits. This messenger paradigm justifies the current interest for the genetics of gene expression [1–6] which has been further enhanced by the numerous associations between genetic markers and diseases reported in recent genome-wide association studies (GWAS) and the expected relevance of genome wide expression (GWE) to characterize the biological basis of these associations [7–10]. However the variability of gene expression not only reflects genetic variation but depends on other factors as well such

as environmental exposures [11,12], metabolic conditions [13], ageing [14,15] or gender [16–17].

Based on these premises we reasoned that if the state of the transcriptome and its changes are important determinants of cell functions, differences in transcript abundance whatever their origin, genetic or non genetic, may contribute to disease pathogenesis. Moreover, the transcriptome might integrate information from numerous sources and inform on the current pathophysiological state of the organism. To assess these possibilities, a global characterization of the variability of the transcriptome, integrating genetic and non genetic influences was undertaken. The study focused on peripheral blood monocytes because these cells may be isolated from an easily accessible tissue and play a key role in the pathogenesis of immune disorders and atherosclerosis-related diseases [18]. In addition, working on a single cell type reduces the complexity of transcriptome data and may avoid possible biases resulting from the heterogeneous cell-types distribution in different samples as it is the case when using whole blood or leucocytes RNAs.

Results

The genome-wide expression of circulating monocytes

To reduce potential artefacts, fresh samples were collected and processed in a short period of time according to a very strict protocol. Monocytes were obtained from 1,490 unrelated individuals, 730 women and 760 men, aged 35 to 74 years, recruited in the Gutenberg Heart Study (GHS), a community-based project conducted in a single centre in the region of Mainz (Germany) (Table 1). GWE profiles were generated using *Illumina* Human HT-12 expression BeadChips, and after normalization and filtering out genes undetected in monocytes or non-well characterized (see Materials and Methods), 12,808 expressions traits (averaged over probes) remained for analysis.

Identification of eSNPs and eQTLs

Genotyping was performed using *Affymetrix* 6.0 arrays. After filtering out SNPs poorly performing or having a minor allele frequency <0.01, 675,350 SNPs were kept for further analyses. All associations between SNPs and expression traits with a *P*-value $<10^{-5}$ ($n>225{,}000$) were stored in the "GHS_Express" database ("GHS_Express" is available online, see Methods S1). At a study-wise threshold of significance correcting for the number of SNPs and expressions ($P<5.78\times10^{-12}$), 37,403 associations, involving 29,912 SNPs and 2,745 expression traits (referred to as eSNPs and eQTLs, respectively), were identified (Table 2). The median number of eSNPs by eQTL was 11 with an interquartile range of 4 to 26. Owing to its large sample size, the study had an 80% power to detect a SNP effect accounting for 4% or more of the variability (R^2) of any expression trait. Among the 2,745 significant eQTLs, the R^2 observed for the best eSNP varied from 3.1% to >80% with a median of 7.7%. For 290 eQTLs, the R^2 was greater than 25%.

Table 1. Description of the GHS study population.

	Men	Women	*P*-value
N	760	730	
Age (years)	56.4 (10.6)	53.9 (11.2)	2.4×10^{-5}
BMI (kg/m^2)	27.6 (3.9)	26.2 (5.1)	1.2×10^{-8}
HDL cholesterol (mg/dL)	54.4 (14.9)	69.2 (17.8)	2.2×10^{-16}
LDL cholesterol (mg/dL)	133.7 (36.1)	133.0 (36.8)	NS
Triglycerides (mg/dl)	143.2 (97.5)	114.4 (56.8)	2.1×10^{-12}
Systolic blood pressure (mmHg)	135.8 (16.7)	128.5 (18.2)	2.3×10^{-16}
Diastolic blood pressure (mmHg)	85.2 (9.6)	81.2 (9.5)	5.4×10^{-16}
Current smoker	128 (16.8%)	113 (15.5%)	NS
Plasma CRP (mg/L) (sqrt)	1.509 (0.818)	1.545 (0.743)	NS
Plasma glucose (mg/dL)	97.8 (18.5)	91.8 (15.4)	1.1×10^{-4}

Values are means (SD) or numbers (%).

doi:10.1371/journal.pone.0010693.t001

Cis versus trans associations

Associations involving SNPs located within 1 Mb of either the 5′ or 3′ end of the associated gene were considered *cis* (File S1) and other associations were considered *trans*. In accordance with previous results [2–6], most of the genetic variability affecting the transcriptome was of *cis* origin. At study-wise significance, the number of *cis* and *trans* eQTLs were 2,477 and 349, respectively, yielding a *cis*/*trans* ratio of 7.1 (81 eQTLs were both *cis*- and *trans*-modulated). At less stringent levels of significance the number of *trans* associations considerably increased, as expected by chance, whereas the number of *cis* eQTLs only modestly increased, indicating that the high stringency used for *cis* eQTLs identification did not result in an important under-estimation of the true number of *cis* eQTLs (Figure 1).

Comparison with previous GWAS of gene expression

We examined the overlap between the *cis* eQTLs identified in the present study and those found in three previous association studies in which gene expression was explored in LCLs [1,2] and hepatic cells [3]. For this comparison, a significance threshold of 3.9×10^{-6} (Bonferroni-corrected for 12,808 genes) was used for the analysis of eQTLs in GHS data, corresponding to a single hypothesis tested per gene. Among the *cis* eQTLs considered significant in each of the studies and involving expression traits detected in GHS, 66.7%, 56.5% and 54.1%, respectively, were significant in our data (Table 3, Files S2–S4). The proportion of *cis* eQTLs that replicated in GHS increased with the increasing level of significance reported in each study, consistent with the fact that stronger associations are more robust and more likely to be shared by different types of cells. These comparisons revealed a relatively high rate of replication of the previous findings in GHS. However, as a consequence of its greater power, *P*-values observed in GHS were considerably lower than those previously reported (Figure 2).

We also examined the overlap between *cis* eQTLs in GHS and *cis*-heritable eQTLs found by expression profiling of lymphocyte RNA in the San Antonio Family Heart Study (SAFHS) [4]. Among the eQTLs with a *cis* heritability ≥0.1 in SAFHS, 62% were significantly *cis* modulated in GHS, and this proportion reached 89% for heritabilities ≥0.6 (Figure 3, File S5).

Trans associations showed much weaker consistency across studies. Among the 50 eQTLs having a *trans* lod score >4.0 in SAFHS [4] with corresponding expression detected in GHS, only one, *MAPK8IP1*, was replicated in GHS ($P<10^{-300}$). Replication of the *trans* associations in studies of similar power as the present one would be of interest.

Identifying eQTLs that may result from the presence of SNPs in probe sequences

For all probes present on the *Illumina* HT12 array, a systematic search for sequence polymorphisms was undertaken, using the HapMap database as reference (Release 27; Phase II+III, Feb09,

Table 2. Number of gene expression-by-SNP associations at various levels of significance.

Significance level	Minimum R^2 [$]	Total number of associations	*cis*/*trans* ratio for associations	Total number of associated expressions (eQTLs)	*cis*/*trans* ratio for eQTLs	Total number of associated SNPs (eSNPs)	*cis*/*trans* ratio for eSNPs
$<10^{-6}$	0.016	93491	2.1	8575	0.5	67190	2.4
$<10^{-8}$	0.022	54749	7.3	3857	3.0	41425	11.2
$<10^{-10}$	0.028	42421	9.8	2998	6.0	33339	16.3
$<5.78\times10^{-12}$	**0.031**	**37403**	**10.7**	**2745**	**7.1**	**29912**	**17.1**
$<10^{-15}$	0.042	27330	12.7	2180	9.5	22591	17.8
$<10^{-20}$	0.057	19655	14.7	1725	12.8	16883	19.2
$<10^{-25}$	0.071	15015	16.4	1429	16.2	13045	21.5
$<10^{-35}$	0.099	9673	17.1	1031	21.6	8516	22.9
$<10^{-50}$	0.140	5873	14.0	712	28.8	5224	21.7
$<10^{-100}$	0.263	1790	10.5	290	28.1	1598	11.1
$<10^{-150}$	0.371	922	5.5	156	21.4	772	5.9
$<10^{-200}$	0.463	635	3.7	97	15.3	504	3.9
$<10^{-300}$	0.606	321	1.7	38	11.7	213	1.7

[$]*Minimum R^2 (proportion of gene expression variability explained by a SNP) observed for a given significance level. Numbers corresponding to study-wise significance are shown in bold. For investigating cis associations or performing any other hypothesis-based test, lower levels of significance may be considered.*

doi:10.1371/journal.pone.0010693.t002

on NCBI B36 assembly and dbSNP b126). Among the 2,477 genes whose expression was associated with *cis* eSNPs, 173 (7%) were probed by one or several polymorphic sequences (180 probes) (Table S1). For 32 of these probes, the HapMap SNP was present on the Affymetrix array used in this study and for 41 other probes, the HapMap SNP had one or several perfect proxies on the array. For those eQTLs, we cannot exclude the possibility of an artefactual association due to a differential binding of the probe to its target sequence.

Gene expression, a link between DNA sequence variability and clinical phenotypes?

A link between genetic variability and clinical phenotypes is supported in human studies by several observations relating variants

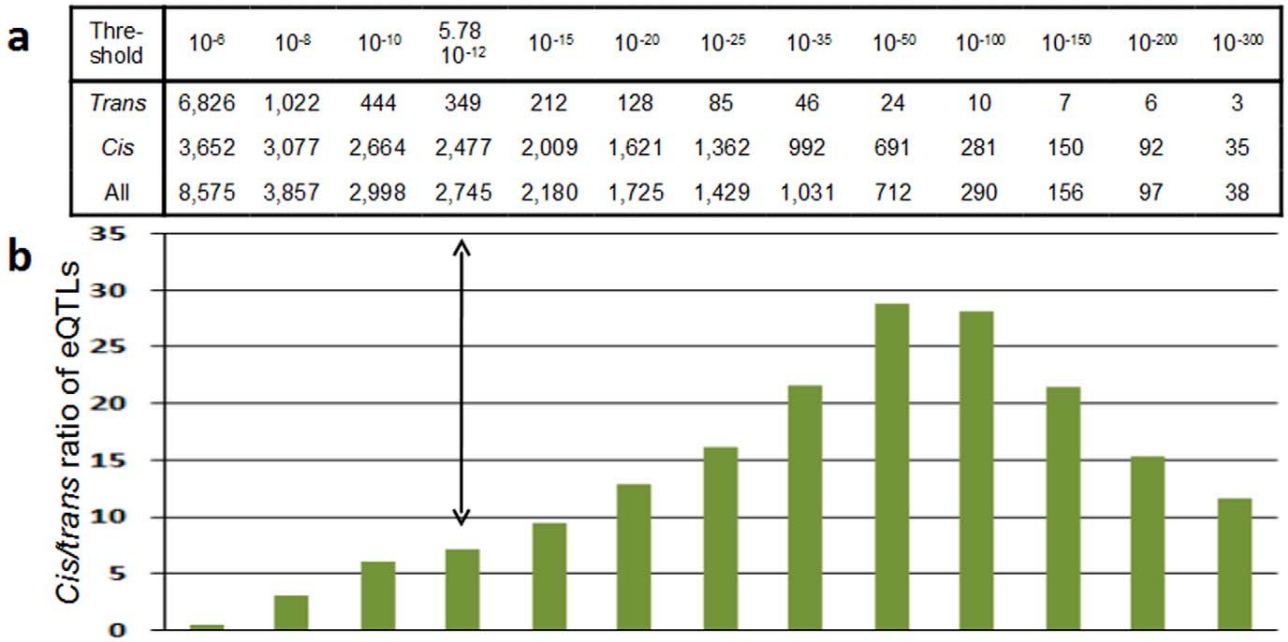


Figure 1. Number of eQTLs according to the significance threshold adopted and corresponding *cis/trans* eQTL ratio. The vertical arrow indicates the study-wise level of significance correcting for the number of hypotheses tested. Some eQTLs being associated with both *cis* and *trans*-acting eSNPs, the sum of *cis* and *trans* eQTLs is greater than the total number of eQTLs.

doi:10.1371/journal.pone.0010693.g001

Table 3. Number of *cis* eQTLs identified in previous studies and replicated in GHS.

	Stranger et al.		Dixon et al.		Schadt et al.	
Level of signifi-cance	**Number of eQTLs at level of significance**	**Percent signifi-cant in GHS***	**Number of eQTLs at level of significance**	**Percent signif-icant in GHS***	**Number of eQTLs at level of significance**	**Percent signifi-cant in GHS***
$>10^{-8}$	86	55.8	110	50.9	928	47.9
10^{-8}–10^{-10}	63	69.8	162	50.0	168	57.7
10^{-10}–10^{-15}	144	63.2	237	54.8	211	57.3
10^{-15}–10^{-20}	60	70.0	102	60.7	120	66.7
10^{-20}–10^{-25}	38	89.5	73	65.7	73	67.1
$\leq 10^{-25}$	48	70.8	89	67.4	103	73.8
All	439	66.7	773	56.5	1603	54.1

** Comparisons were based on sets of gene expressions overlapping between each study and GHS and were restricted to autosomal cis eQTLs. All cis eQTLs considered significant in each study were retrieved and replication was assessed in GHS ($P<3.9\times10^{-6}$ correcting for 12,808 gene expressions).*
For Stranger et al [1]*, data were extracted from Table S2. We considered as significant the associations found in at least 3 HAPMAP populations. For Dixon et al* [2]*, data were extracted from Table S1 and trans eQTLs were excluded. Matching of probes was done using a table provided by the authors on their web site. For Schadt et al* [3]*, cis eQTLs considered significant (First.Pass.Indicator set to 1) were extracted from Table S3. For each eQTL, we selected in GHS the P-value of the best cis eSNP. The full data used to generate this table are provided in Files S2–S4.*
doi:10.1371/journal.pone.0010693.t003

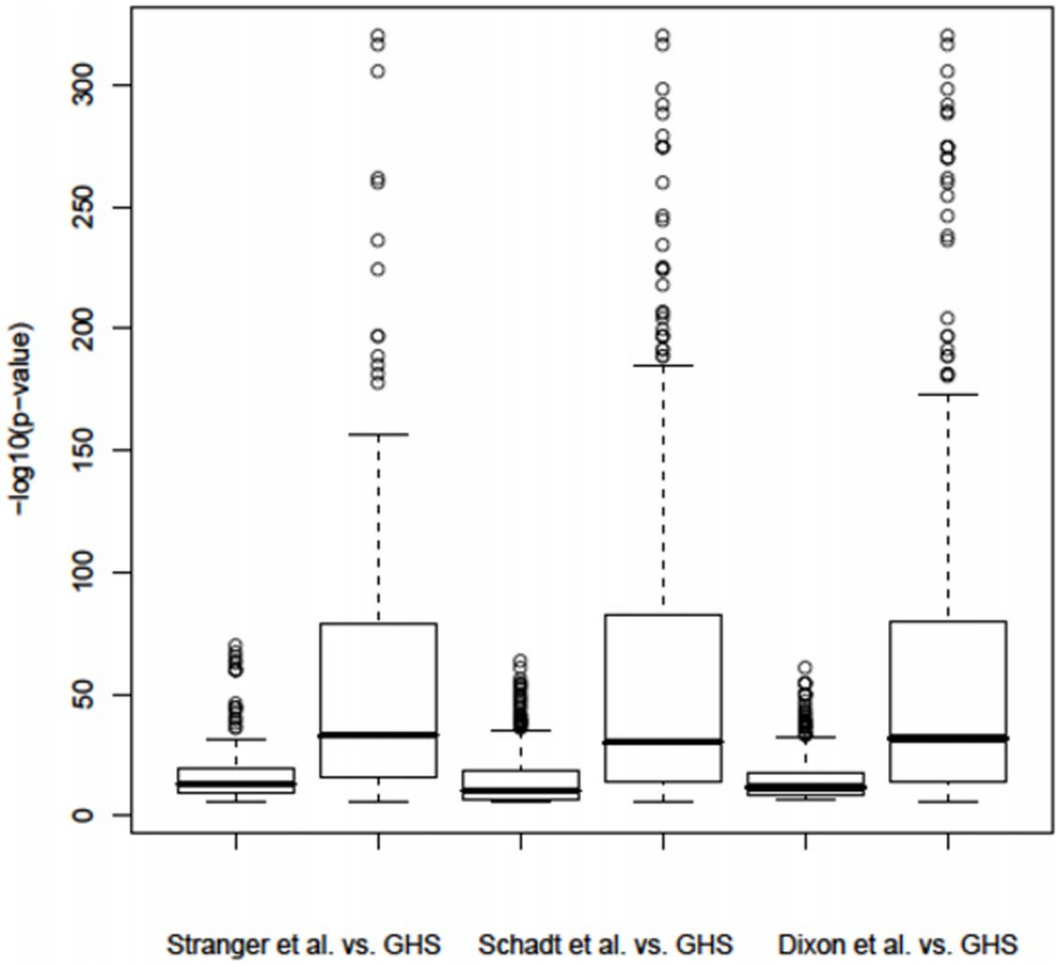


Figure 2. Comparison of the distributions of *P*-values of *cis* eQTLs reported as significant in three previous association studies with *P*-values observed in GHS for the same eQTLs. For each of the 3 comparisons, we selected in GHS the subset of gene expressions claimed as significant in the study of comparison. Only autosomal genes were considered in these comparisons. The data used to generate this figure are provided in Files S2–S4. See also footnote of Table S3 for details.
doi:10.1371/journal.pone.0010693.g002

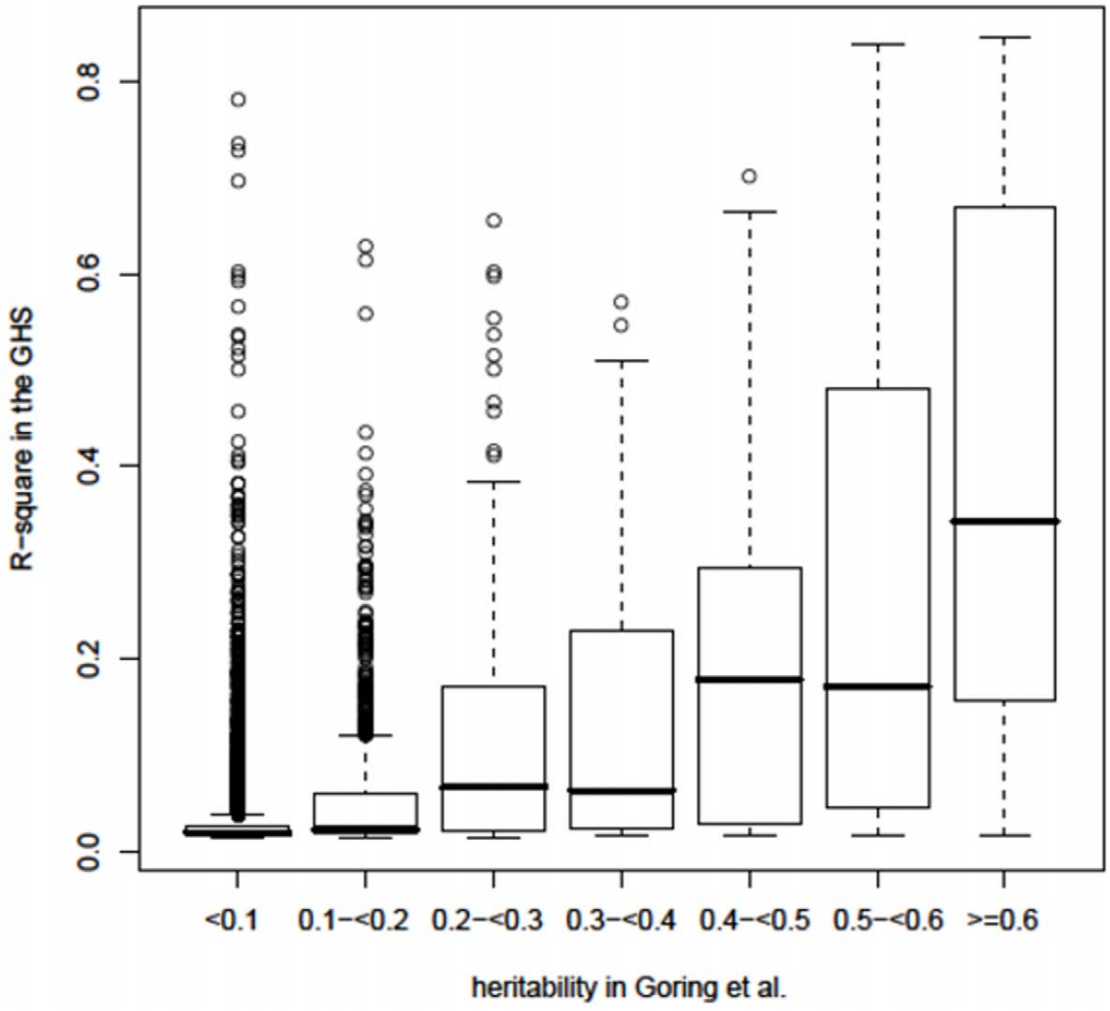


Figure 3. Comparison of the heritability of *cis* eQTLs estimated in the SAFHS study with the R^2 of the corresponding *cis* eQTLs in GHS. Data were extracted from Supplementary Table 4 in Göring et al. [4] and comparisons were restricted to genes having a corresponding gene symbol in GHS. Heritability in the SAFHS was estimated by linkage analysis and accounts for the whole variability at a locus while R^2 refers to a single eSNP (the best eSNP) and therefore underestimates the global variability affecting gene expression at a locus. The data used to generate this figure are provided in File S5. The median R^2 was globally lower than the heritability, consistent with the fact that the R^2 is referring to a single SNP whereas heritability reflects the whole genetic variation at a locus.
doi:10.1371/journal.pone.0010693.g003

in regulatory gene regions to protein phenotypes or diseases [19,20]. However, this message-passing paradigm has never been evaluated on a genome-wide basis. We therefore tested whether monocytes gene expression might mediate the effects of loci recently identified by GWAS of cardiovascular risk factors. For each locus identified in GWAS of lipid variables [21], blood pressure (BP) [22] and body mass index (BMI) [23], we selected the lead SNP or a tag SNP having an $r^2 \geq 0.8$ with the lead SNP in GHS data. Associations between lead/tag SNPs and corresponding risk factors were checked in all GHS subjects for whom genome-wide data was available ($n = 3{,}175$). Most previous GWAS loci for circulating lipids were replicated in our data (Table 4) but only few of the findings of GWAS of BMI and BP were replicated (Table S2). This low replication is probably due to a lack of power, as the maximum R^2 observed in the GWAS of BP [22] was 0.09% and the power of GHS to replicate such an association was only 38%.

For each GWAS locus, we examined whether the lead/tag SNP correlated with any expression trait in GHS data and when a significant association was found, we checked whether the expression trait was significantly associated with the risk factor under consideration. This analysis revealed that very few GWAS results were compatible with an effect mediated by gene expression at the locus (Table 4 and Table S2). There were, however, two exceptions: the first one concerned the *LPL* locus, where the minor allele of rs17489282 was associated with higher HDL-cholesterol ($P = 5.91 \times 10^{-5}$) and *LPL* expression ($P = 2.18 \times 10^{-6}$), while HDL-cholesterol and *LPL* expression were positively correlated ($P = 6 \times 10^{-4}$), consistent with an effect mediated by LPL; the second one concerned the association between the 1p13.3 locus and LDL-cholesterol. This locus encompasses three potential candidate genes, *CELSR2*, *PSRC1* and *SORT1*, and it has been suggested that *CELSR2* or *SORT1* could be responsible for the reported associations of this locus with LDL [3,24,25]. In our data, the minor allele of rs629301 (a perfect tag of the lead SNP identified by GWAS), was associated with lower LDL-cholesterol ($P = 2.6 \times 10^{-4}$) and higher *PRSC1* expression ($P = 2.3 \times 10^{-56}$)

Table 4. Loci identified in GWAS of circulating lipids – associations of lead/tag SNPs with phenotypes and expression, and of expression with phenotype in GHS.

Lead SNP in GWAS	Phenotype	Chr	Position (Mb)	Genes in region	Tag SNP in Affy 6.0 with $r^2>0.8$	r^2 between lead SNP and tag SNP	Association between tag SNP and phenotype (*P*-value)	eQTL associated with tag SNP	Association between tag SNP and eQTL (*P*-value)	Association between eQTL and phenotype (*P*-value)
rs10889353	TG	1	62.83	*DOCK7*	rs10889353	1.00	1.86E-04	*DOCK7*	7.09E-52	0.3970
rs646776*	**LDL**	**1**	**109.53**	***CELSR2/PSRC1/ SORT1***	**rs629301**	**1.00**	**2.62E-04**	***PSRC1***	**2.34E-56**	**0.0190**
								CELSR2	7.56E-06	0.6195
rs693	LDL	2	21.14	*APOB*	rs693	1.00	1.61E-04	none		
rs6754295	HDL	2	21.12	*APOB*	rs673548	0.86	0.0435	none		
rs673548	TG	2	21.15	*APOB*	rs673548	1.00	4.10E-05	none		
rs780094	TG	2	27.65	*GCKR*	rs780094	1.00	3.15E-08	none		
rs6756629	LDL	2	43.98	ABCG5	rs4953023	1.00	7.87E-05	none		
rs3846662	LDL	5	74.69	*HMGCR*	rs12654264	0.84	9.26E-06	none		
rs12670798	LDL	7	21.38	*DNAH11*	none					
rs2240466	TG	7	72.3	*MLXIPL*	rs2074755	1.00	0.0015	none		
rs2083637	**HDL**	**8**	**19.91**	***LPL***	**rs17489282**	**1.00**	**5.91E-05**	***LPL***	**2.18E-06**	**0.0006**
rs2083637	TG	8	19.91	*LPL*	rs17489282	1.00	3.31E-07	*LPL*	2.18E-06	0.3520
rs3905000	HDL	9	104.74	*ABCA1*	rs3890182	1.00	0.33	none		
rs7395662	HDL	11	48.48	*MADD-FOLH1*	rs7395662	1.00	0.17	*MYBPC3*	1.17E-09	0.1435
								SPI1	4.42E-06	0.0222
rs174570	LDL	11	61.35	*FADS2/3*	rs174570	1.00	0.0386	none		
rs12272004	TG	11	116.11	*APO(A1/A4/A5/C3)*	rs10488699	1.00	1.63E-06	none		
rs1532085	HDL	15	56.47	*LIPC*	none					
rs1532624	HDL	16	55.56	*CETP*	none					
rs2271293	HDL	16	66.46	*CTCF-PRMT8*	rs2271293	1.00	0.0145	*DPEP3*	5.09E-17	0.8275
								DUS2L	5.15E-42	0.1410
								GFOD2	1.48E-17	0.6400
								LCAT	6.00E-06	0.3347
								PARD6A	7.88E-07	0.8772
								PRMT7	2.03E-06	0.1690
rs4939883	HDL	18	45.42	*LIPG*	rs7240405	1.00	0.0233	none		
rs2228671	LDL	19	11.07	*LDLR*	none					
rs157580	LDL	19	50.09	*TOMM40-APOE*	none					

GWAS loci were taken from Table 2 in ref. 21.

**This SNP was also found in GWAS of CAD. The association between tag SNP and phenotype was tested in the 3,175 GHS subjects having GWV data. Association between eQTL and tag SNP or phenotype was tested in the 1,490 GHS subjects having GWE data. In bold are shown the loci for which the SNP-phenotype association found in GWAS is compatible with mediation by gene expression. Similar analyses for BMI and BP are given in Table S2.*

doi:10.1371/journal.pone.0010693.t004

while *PRSC1* expression and LDL-cholesterol were negatively correlated ($P=0.019$). Results for *CELSR2* were much less consistent and *SORT1*, the third gene at the locus, was not *cis*-modulated in monocytes.

Several loci associated with coronary artery disease (CAD) have been identified by GWAS [26–29]. The strongest association involves SNPs in the 9p21 region. Recently it was reported that deletion in mice of the region orthologous to the 9p21 CAD interval in human affects the expression of the nearby *cdkn2a* and *cdkn2b* genes as well as the properties of proliferation of vascular cells [30]. The Cyclin-dependent kinase inhibitor coding genes, *CDNK2A* and *CDNK2B*, are also located close to the CAD locus in humans. *CDKN2A* expression in monocytes was not detected in our study, we therefore focused our analysis on *CDKN2B*. All SNPs available in GHS in the region encompassing the CAD locus were tested for association with the expression of *CDKN2B*. Figure 4 shows that *CDKN2B* expression was strongly associated with several SNPs located in a region upstream of the gene sequence ($P<10^{-60}$). However, these SNPs were not associated with CAD (this result was obtained in a yet unpublished GWAS comparing GHS individuals to a cohort of CAD patients), whereas proxies of the CAD-associated SNPs were unrelated with *CDKN2B* expression (see legend of Figure 4 for more details). The SNPs associated with *CDKN2B* expression are located within the sequence of the non-coding alternatively spliced gene *ANRIL* (also named *CDKN2BAS*) whose implication in the association with CAD has been hypothesized [31]. Although our results are limited by the fact that neither *CDKN2A* nor *ANRIL* expressions could be

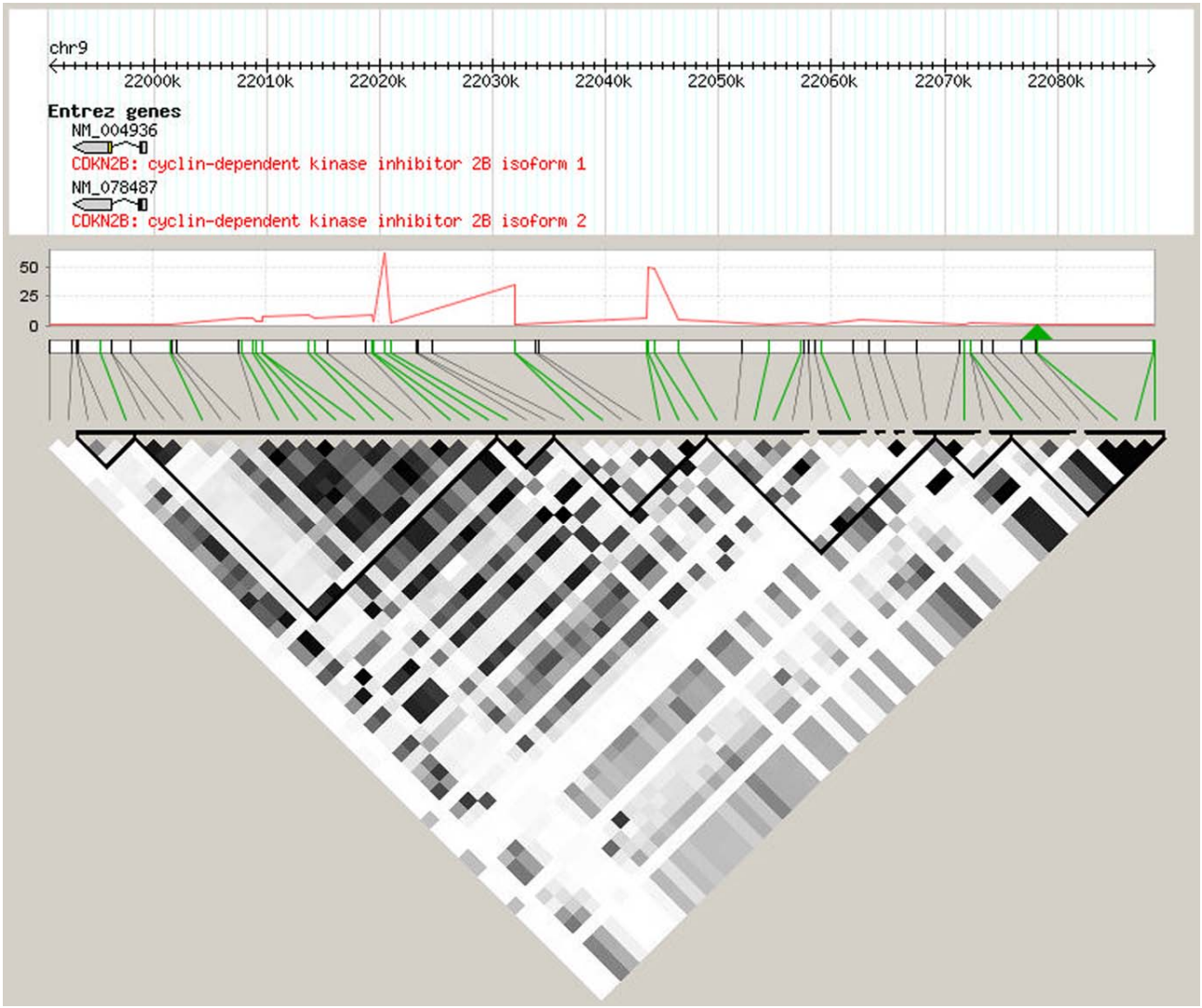


Figure 4. The loci affecting CDKN2B expression and CAD on chromosome 9p21 are independent. The lead SNP rs1333049 generally reported at the CAD locus was not present on the Affymetrix 6.0 array, we therefore selected its best proxy, rs10757272 (position 22078260, $r^2 = 0.9$ with rs1333049), using SNAP (https://www.broadinstitute.org/mpg/snap). Positions of genotyped SNPs are shown using a green link and position of the proxy SNP, rs10757272, is represented by a green triangle. The red curve reflect the –log10(*P*-value) for the association between SNPs and *CDKN2B* expression. The LD (r^2) between pairs of SNPs is shown at the bottom of the figure using a range of colors between white ($r^2 = 0$) and black ($r^2 = 1$). The *CDKN2B* and CAD-associated SNPs are located in different blocks of LD strongly suggesting that the genetic effects on *CDKN2B* expression and CAD are independent.
doi:10.1371/journal.pone.0010693.g004

evaluated, they reveal that in humans, SNPs that affect *CDKN2B* expression are different from those that are known to affect CAD risk (Figure 4).

Expression traits associated with risk factors

To investigate gene expression in relation to risk factors (age, gender, BMI, HDL and LDL cholesterol, triglycerides, Systolic and Diastolic Blood pressure, smoking and plasma CRP), the study-wise significance threshold was set at 3.9×10^{-7} to correct for the number of risk factors ($n = 10$) and expressions ($n = 12{,}808$) tested. Overall, 1,662 expression traits (13.0%) were associated with at least one risk factor (Table 5 and File S6). Gender and age were the two major factors influencing expression levels (807 gene expressions were affected by gender and 396 by age). BMI, smoking and C-reactive protein (CRP) levels were also correlated with numerous expression traits (230, 294 and 328, respectively). Conversely, few associations with BP and lipids were observed (Table 5).

Genetic and non-genetic factors act additively on gene expression

Cis eQTLs were over-represented among expression traits that were also affected by gender, age, BMI, smoking and CRP, with odds ratios as high as 3.24 for cigarette smoking (Table 5). This suggests that some genes are more responsive than others to the influence of multiple factors. For expression traits that were simultaneously associated with *cis* eSNPs and risk factors ($n = 465$), we determined the joint effects of the two sources of variability on expression level. For this purpose, each eQTL was modelled as a function of the best *cis*-acting eSNP, the associated risk factor and

Table 5. Number of expression traits associated with risk factors and *cis* eSNPs.

Risk factor	Number of expression traits associated with the specified risk factor	Number (%) of expression traits also associated with *cis* eSNPs	Odds ratio (95% CI)*
Gender	807	230 (28.5%)	1.73 (1.47–2.03)
Age	396	94 (23.7%)	1.31 (1.03–1.66)
BMI	230	72 (31.3%)	1.92 (1.45–2.56)
HDL	9	1 (11.1%)	ND
LDL	1	0	ND
Triglycerides	9	3 (33.3%)	ND
SBP	48	6 (12.5%)	0.59 (0.25–1.40)
DBP	18	2 (11.1%)	ND
Smoking	294	126 (42.9%)	3.24 (2.56–4.10)
CRP	328	116 (35.4%)	2.34 (1.86–2.95)
All (irrespective of any association with risk factor)	12,808	2,477 (19.3%)	

Study-wise levels of significance were considered for associations of expression traits with risk factors and SNPs (3.9×10^{-7} and 5.78×10^{-12}, respectively). Associations of expression traits with BMI, CRP and smoking were adjusted for age and sex, and association with HDL, LDL, triglycerides, SBP and DBP were additionally adjusted for BMI.

*Odds ratio (OR) of being influenced by a *cis* eSNP for an expression trait associated with a given risk factor. For example, gender-related expression traits have an OR of 1.73 of being influenced by *cis* eSNPs by comparison to expression traits unrelated to gender. ND: not determined because of small numbers.

doi:10.1371/journal.pone.0010693.t005

the interaction between the two. When several risk factors were associated with the same eQTL, each of them was tested separately for interaction with the corresponding SNP. The best FDR-corrected *P*-values for interaction (File S7) were 0.014 (*ISCU* expression, gender and rs4830487) and the second one was 0.042 (*HIST1H2AE* expression, smoking and rs16891378). This first genome-wide exploration of interaction between *cis* eSNPs and risk factors on gene expression therefore suggests that the two sources of variability mostly act additively on expression. This is illustrated for eQTLs affected by smoking in Figure 5. It must be noted however that despite the large size of this study, its power may nevertheless be insufficient to assess weak interaction.

Expressions influenced by genetic and risk factors are enriched in immunity and defense genes

An ontology analysis using the Panther system demonstrated that, by reference to the list of 12,808 genes expressed in monocytes, the set of 465 expression traits affected by multiple sources of variability was enriched in "Immunity and defense" genes (69 observed/35.8 expected, $P=4.5\times10^{-6}$), especially in the sub-categories of "Macrophage-mediated immunity" (18/3.6, $P=7.5\times10^{-6}$).

The variability of each risk factor can be characterized by a limited set of independent gene expressions

The preceding analyses revealed that each risk factor was associated with a large number of expression traits, thus emphasizing the multiple inter-relations existing between the transcriptomic and risk factor profiles of an individual. While it is important from a biological and mechanistic perspective to characterize at best all the genes that are influenced by a given condition, from a clinical perspective, it might be more relevant to identify a limited set of gene expressions that could efficiently discriminate individuals with different risk profiles. Indeed, because of the tight co-regulation of genes within biological systems, numerous gene expressions are inter-correlated and consequently, their association with risk factors are not independent. To account for this inter-dependency, we conducted a multivariate analysis to identify expression traits that were independently associated with each risk factor.

To obtain reliable results, we randomly divided the study population into two sub-samples of equal size which were used for screening and validation purposes respectively (see Materials and Methods). In these analyses, each risk factor was considered separately whereas all expression traits were envisaged jointly for their potential association with the risk factor, considered here as the dependent variable. The screening/validation procedure was repeated 250 times and for each risk factor, we report expression traits associated ($P<0.01$) with the risk factor in more than 25% of the replicates. This stringent approach led to the identification of 106 independent expression correlates for the ten risk factors (Table 6), a much reduced number compared to the 1,662 expression traits previously identified by the one-to-one association analysis presented in Table 5.

Gender and age. Even after exclusion of sex-linked genes, gender was independently associated with the largest number of expression traits ($n=31$) which, considered all together, contributed to a highly significant discrimination between males and females ($P<10^{-100}$). By contrast the number of expression traits independently associated with age was more limited ($n=12$).

BMI and CRP. Both factors were independently associated with 12 and 14 expression traits respectively. Inspection of the genes listed in Table 6 shows that several of them, including *CX3CR1*, *CD209*, *CLEC10A*, *FCER1A*, *FCGBP*, *C1RL*, *C1QB*, *CD36*, *ADM* and *VSIG4*, encode proteins involved in the differentiation or maturation of immunity-related cells and in host defence [32–38]. We may speculate that the variability of expression of these genes is the consequence of an already present heterogeneity of monocytes [39,40] or reflects a particular transcription pattern that prefigures future functional changes. The example of *CX3CR1* which was positively associated with both BMI and CRP is particularly interesting as this gene encodes the fractalkine receptor whose role is essential in the migration of

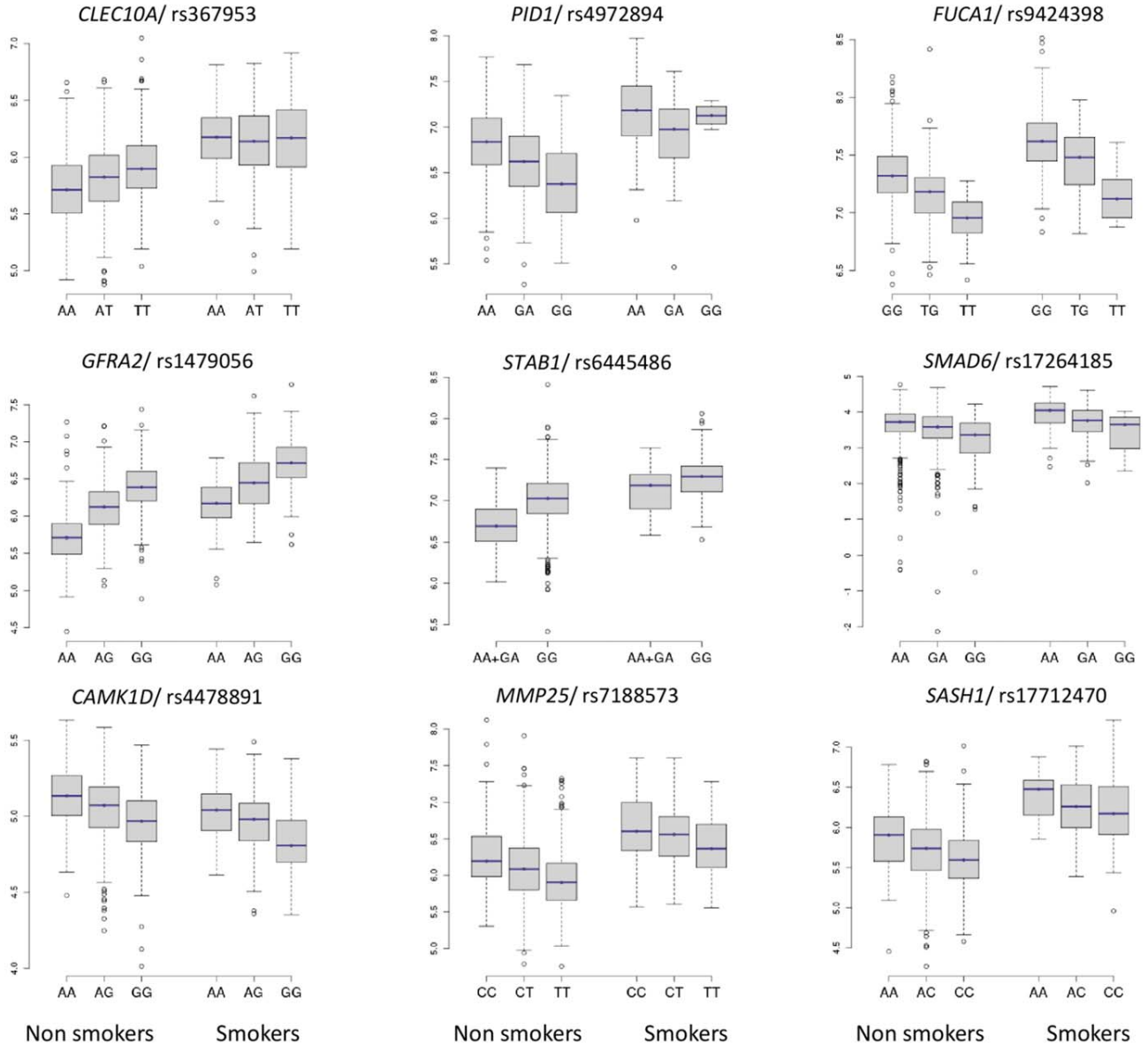


Figure 5. Effect of the best *cis* eSNP and smoking on expression of smoking-related eQTLs. The proportion of variability of expression explained by the best *cis* eSNP varied from 3.1% for *CLEC10A* to 27.2% for *GFRA2* while the proportion explained by smoking varied from 2.8% for *SMAD6* to 21.6% for *SASH1*. The lowest *P*-value for interaction between SNP and smoking was 0.02 for *STAB1*.
doi:10.1371/journal.pone.0010693.g005

monocytes to sites of inflammation and injury, especially in atherosclerotic lesions [41].

Lipids. *ABCA1* and *ABCG1* gene expressions were both associated with circulating lipids. The proteins encoded by these genes are key players in reverse cholesterol transport and the regulation of lipid-trafficking mechanisms in macrophages respectively [42,43]. MYLIP (Idol) is a ligase involved in the ubiquitination and degradation of LDL receptors [44,45] and SCD is a stearoyl-CoA desaturase involved in the conversion of saturated into monounsaturated fatty acids that regulates lipid metabolism and may be modulated by dietary intake [46].

Blood pressure. One of the main independent correlates of SBP was *ARID5B* (*MRF2*), whose relevance in the physiology of BP regulation is supported by its role as a regulator of smooth muscle differentiation and proliferation [47]. *GFOD1*, another expression correlate of SBP and DBP is a gene of unknown function which has been associated with attention deficit hyperactivity disorder [48].

Cigarette smoking has a major impact on gene expression and atherosclerosis

Smoking was independently associated with 18 expression traits (Table 6) which, considered all together, contributed to a highly significant discrimination between smokers and non smokers ($P<10^{-107}$, $R^2>50\%$). Nine of these genes were modulated by *cis* eSNPs (Table 7) and as already mentioned above, the genetic and smoking effects on these gene expressions were additive (Figure 5).

Among the 18 expression traits associated with smoking, *SASH1*, *P2RY6* and *PTGDS* were systematically retrieved in all

Table 6. Subsets of expression traits showing robust independent association with the different risk factors in the validation sample.

Risk factor	Median (range) of the global *P*-values across replicates*	List of gene expressions associated with risk factor after adjustment on covariates[$]
Gender[#]	All $P<10^{-100}$	CCDC106, MMEL1, ANKRD57, CXCR7, FCGR2B, SOX15, FCGBP, LPXN, CD24, HOXA9, PTK6, DDX43, RAB11FIP1, PTK2, CLEC4G, ADARB1, PROK2, MYBPH, PER3, TPPP3, MPO, FAM24B, EMR3, ENOSF1, TPM2, PTTG1IP, CELSR3, CD1A, FOLR2, BOLA3, OPLAH
Age	1.1×10^{-50} (2.9×10^{-70}–3.6×10^{-37})	PARP3, PDGFRB, NEFH, P2RY2, SPINK2, GPER, NFKBIZ, ZSCAN18, IGLL1, BLK, ITM2C, C1RL
BMI	1.5×10^{-37} (2.6×10^{-50}–2.7×10^{-26})	CX3CR1, MAP3K6, FCGBP, CD209, LYPD2, VSIG4, RPGRIP1, PACAP, LGALS3BP, ELA2, CD36, ABCA1
HDL-chol	5.5×10^{-10} (2.2×10^{-15}–2.7×10^{-3})	PRDM1, SCD, DPEP2,TMEM43
LDL-chol	2.3×10^{-3} (3.8×10^{-7}–0.5)	BYSL, ABCA1
Triglycerides	4.2×10^{-6} (2.1×10^{-11}–0.08)	MYLIP, PHGDH, ABCA1, ELA2, ABCG1, SASH1
SBP	5.0×10^{-20} (4.6×10^{-27}–2.6×10^{-12})	CRIP1, GFOD1, DHRS9, NR4A2, TSC22D3, ARID5B, PAPSS2, HVCN1
DBP	1.8×10^{-10} (3.6×10^{-16}–1.2×10^{-4})	GFOD1, CRIP1, TPPP3, NR4A2,EMP1
CRP	6.1×10^{-51} (7.3×10^{-67}–2.0×10^{-30})	FAM20A, CETP, FCGBP, COL9A2, C1RL, ADM, CREB5, APBB1IP, CX3CR1, C1QB, MS4A4A, FCER1A, ALDH1A1, FLVCR2
Smoking	9.7×10^{-108} (1.9×10^{-128}–3.7×10^{-84})	SASH1, P2RY6, PTGDS, PID1, CYP4F22, MMP25, WWC3, FUCA1, PDE4B, STAB1, GFRA2, CLEC10A, CAMK1D, DHRS9, CNTNAP2, IQCK, ITGB7, SMAD6,

*The global *P*-value is the *P*-value obtained by comparing the model with all significant expression traits and covariates to the model with covariates only. Expressions that are underlined are associated negatively to the risk factor (or to male gender), others are associated positively (or to female gender).
[$]Covariates: age and gender for BMI, CRP and smoking; age, gender and BMI for lipids and BP.
[#]Gender-associated traits were selected from autosomal genes only.
doi:10.1371/journal.pone.0010693.t006

screening/validation replicates (Table S3). *SASH1* is a tumor suppressor gene [49], *P2RY6* encodes a G-protein-coupled receptor involved in the proinflammatory response to UDP in monocytes [50] and *PTGDS* encodes a prostaglandin D synthase involved in smooth muscle contraction/relaxation and inhibition of platelet aggregation, two functions known to be modified by tobacco consumption. Recently, in a GWE study of leukocytes RNA, *PTGDS* and *SASH1* expressions were found associated with cotinine, a metabolite of nicotine used as a marker of tobacco exposure [51].

Cigarette smoking is a major risk factor for atherosclerosis [52]. In GHS participants, the prevalence of atherosclerotic plaques in the right and left carotid arteries, assessed by echography, was strongly increased in smokers ($P=9.1\times10^{-7}$ after adjustment for age and gender). Among the 18 gene expressions independently associated with smoking, four were individually correlated with the number of carotid plaques, *PTGDS* ($P=1.8\times10^{-7}$) negatively and *MMP25* ($P=3.5\times10^{-4}$), *SASH1* ($P=1.4\times10^{-5}$) and *WWC3* ($P=6.3\times10^{-4}$) positively (Table 7). In a multivariate model including the four expression traits and smoking, as well as age and gender, *PTGDS* ($P=5.7\times10^{-4}$) and *SASH1 (P = 0.012)* remained significantly associated with the number of plaques whereas *MMP25* ($P=0.09$), *WWC3* ($P=0.10$) and smoking ($P=0.5$) were no longer significant, suggesting that the association between smoking and atherosclerosis was mostly reflected (or mediated) by its effect on the expression of these four genes. The fact that *PTGDS* and *SASH1* expression remained associated with carotid plaques after adjustment on smoking status may indicate a broader implication of these genes in atherosclerosis than the sole effect induced by smoking. However it is also possible that the expression of these two genes more faithfully reflects tobacco consumption than the dichotomous variable used to define smoking. This illustrates the dual aspect of the transcriptome which may be viewed either as an element in a causal chain or as reflecting ongoing processes with no implied causation. Because genetics may help to dissect causal pathways, we examined whether the best *cis* eSNPs associated with expression of the smoking-related genes were also associated with carotid atherosclerosis, but no such association was detected (Table 7).

Discussion

This large-scale investigation of the transcriptome of monocytes in healthy individuals provides new biological insights into the mechanisms by which gene expression might contribute to disease pathogenesis. In the line of previous studies [1–5], we could build a detailed map of *cis*-regulated eQTLs in monocytes. Even if cell-specific eQTLs exist [53], a large fraction of them are likely to be common to other cell types, and the eQTL map provided here constitutes the most extensive one so far.

Despite the large number of eQTLs identified, the transcriptome of circulating monocytes, contrary to initial expectations [7–10], appeared of modest help to dissect the relationship between genome variability and complex human traits such as cardiovascular risk factors. One explanation for this finding might be that monocytes are not the most relevant cells for unravelling links between genome variation and the risk factors investigated. With regard to circulating lipids for example, only 28 of the 45 genes located in regions harbouring SNPs associated with circulating lipids in GWAS [24] were expressed in monocytes. The difficulty to corroborate the messenger paradigm in human clinical studies may also relate to the fact that the linear model of

Table 7. Smoking-related expression traits: association of expression with *cis*-acting eSNP, smoking and extent of atherosclerosis.

Smoking-related gene expressions	Association of the best *cis* eSNP with expression		Association of the best *cis* eSNP with the extent of atherosclerosis		Association of expression with smoking		Association of expression with the extent of atherosclerosis	
	SNP number	*P*-value	t- value	*P*-value	t- value	*P*-value	t- value	*P*-value
CAMK1D	**rs4478891**	**1.1×10^{-26}**	0.7	0.51	−7.4	2.3×10^{-13}	−3.1	0.002
CLEC10A	**rs367953**	**3.6×10^{-12}**	0.5	0.65	13.0	5.4×10^{-37}	−0.4	0.68
CNTNAP2	rs1110144	8.6×10^{-12}	−1.1	0.28	6.5	7.3×10^{-11}	1.7	0.082
CYP4F22	rs11253478	1.5×10^{-6}	0.0	1.00	−17.5	1.8×10^{-62}	−2.2	0.027
DHRS9	rs1386426	4.3×10^{-7}	2.5	0.012	−6.3	3.5×10^{-10}	0.2	0.82
FUCA1	**rs9424398**	**6.8×10^{-42}**	1.0	0.34	13.7	1.4×10^{-40}	2.1	0.035
GFRA2	**rs1479056**	**1.6×10^{-108}**	0.5	0.62	13.3	3.1×10^{-38}	0.8	0.44
IQCK	rs1879894	1.1×10^{-7}	0.6	0.52	−7.6	2.7×10^{-14}	−1.7	0.089
ITGB7	rs17080239	1.0×10^{-6}	0.4	0.68	6.3	3.9×10^{-10}	0.3	0.73
MMP25	**rs7188573**	**4.7×10^{-18}**	0.0	1.00	15.6	5.4×10^{-51}	**3.6**	**3.51×10^{-4}**
P2RY6	rs3781305	9.0×10^{-7}	−1.0	0.34	16.0	1.5×10^{-53}	0.5	0.58
PDE4B	rs4352802	4.5×10^{-7}	−1.2	0.24	−7.7	1.2×10^{-14}	−1.8	0.077
PID1	**rs4972894**	**1.1×10^{-25}**	−0.1	0.95	10.5	6.2×10^{-25}	1.4	0.16
PTGDS	rs10870158	6.0×10^{-9}	0.7	0.49	−14.0	3.5×10^{-42}	**−5.2**	**1.77×10^{-7}**
SASH1	**rs17712470**	**5.6×10^{-14}**	0.5	0.63	20.5	5.1×10^{-83}	**4.4**	**1.38×10^{-5}**
SMAD6	**rs17264185**	**2.7×10^{-15}**	−1.5	0.13	6.8	9.6×10^{-12}	1.6	0.11
STAB1	**rs9867823**	**2.8×10^{-28}**	0.4	0.68	12.3	2.9×10^{-33}	0.7	0.50
WWC3	rs1013478	8.0×10^{-6}	−1.1	0.26	10.6	1.8×10^{-25}	**3.4**	**6.32×10^{-4}**

Atherosclerosis was assessed by the number of carotid plaques.

Associations of expressions with smoking and number of carotid plaques were adjusted on gender and age. *Cis* eSNPs significant at a study-wise level ($P<5.78\times10^{-12}$) and associations with carotid plaques significant after correction for multiple testing ($n=18$ tests) are shown in bold characters.

doi:10.1371/journal.pone.0010693.t007

causality generally assumed to reflect the relation between genome variability, expression and phenotype may be too simplistic to account for a much more complex biological reality. The effects of genetic variants may be too weak to allow detection even in a study of this size. It is also important to keep in mind that most reported eSNPs are acting in *cis*, whereas *trans* eSNPs may actually be those that mainly drive the changes in gene expression that affect disease risk.

Most importantly, the present study highlighted for the first time the strong link existing between the transcriptome of an individual and his (her) clinical and epidemiological profile. The fact that the transcriptome tightly mirrors the variability of risk factors at a cellular level may have profound implications from a biological and clinical perspective. Until now, the traditional way of viewing the role of genes in the susceptibility to human diseases was through the effect of their variability of sequence. The present findings suggest that another important, if not greater, impact of genes on human phenotypes relates to the variability of their expression, whatever the origin of this variability. The global association observed between most cardiovascular risk factors and the transcriptome and the fact that each risk factor could be characterized by a limited and specific set of independent gene expressions further suggests that this relationship might be clinically relevant. This was particularly well illustrated by the response of the transcriptome to cigarette smoking. We showed that less than 20 genes among the 12,000 expressed in monocytes could highly discriminate smokers and non-smokers, and among them, four genes were sufficient to account for the strong association existing between smoking and atherosclerosis. Whether these genes are causally involved in the mechanisms linking smoking to the development of atherosclerotic plaques or whether they are only markers of ongoing pathological processes remains to be elucidated.

In conclusion, the variability of the transcriptome of monocytes can be viewed from two perspectives. On one hand it reflects the accumulation of effects originating from the genome and the environment and may inform on a number of ongoing processes relevant to disease. On the other hand, it may reflect or anticipate differences in monocytes biology that could have pathophysiological implications. This dual perspective suggests that a better understanding of the sources of variability of the transcriptome of monocytes and other easily accessible cells, will contribute in an important way to our understanding of complex diseases.

Materials and Methods

Ethic statement

The study protocol and drawing of the blood sample have been approved by the local ethics committee and by the local and federal data safety commissioners (Ethik-Kommission der Landesärztekammer Rheinland-Pfalz). All subjects included signed an informed consent.

Study population

The Gutenberg Heart Study (GHS) is designed as a community-based, prospective, observational single-center cohort study in the Rhein-Main region in western mid-Germany. The primary aim of the study is to improve the individual cardiovascular risk

prediction by identifying genetic and non genetic risk factors contributing to cardiovascular diseases, with a strong emphasis on atherosclerosis.

A sample of eligible participants was randomly drawn from the registers of the local registry offices in the city of Mainz and the district of Mainz-Bingen. This sample was stratified in a ratio of 1:1 for gender and residence, and in equal numbers for decades of age. Inclusion criteria were an age between 35 and 74 years and a written consent; exclusion criteria were insufficient knowledge of the German language to understand explanations and instructions, and physical or psychic inability to participate in the examinations in the study center. Individuals were invited for a 5-hour baseline-examination to the study center where clinical examinations and collection of blood samples were performed. The present analysis was based on an initial sample of 3,336 subjects successively enrolled into the GHS from April 2007 to April 2008. Genomic DNA was isolated from all participants. Monocyte RNA was isolated from half of the participants recruited each day to ensure rapid sample processing and isolation of total RNA. For approximately 1,500 study participants, both DNA and RNA were available.

Measurement and definition of cardiovascular risk factors

Blood pressure measurements were performed by an automated sphygmomanometer blood pressure meter (Omron 705CP-II, OMRON Medizintechnik Handesgesellschaft GmbH, Germany) after 5, 8 and 11 minutes of rest. The mean from the 2nd and 3rd standardized measurement was calculated for the systolic and diastolic blood pressure. For the anthropometric measurements, calibrated, digital scales (Seca 862, Seca Germany), a measuring stick (Seca 220, Seca, Germany) and a waist measuring tape were used. The blood sampling was carried out under fasting conditions. HDL-cholesterol, LDL-cholesterol, triglycerides and C-reactive protein (CRP) measurements were performed on an Architect c8000 by commercially available tests (CRP, Ultra HDL, Direct LDL and Triglycerides) from Abbott (www.abbottdiagnostics.de). All tests were measured under standardized conditions in an accredited laboratory of the institute of clinical chemistry and laboratory medicine at the University of Mainz. Smoking was defined by dichotomizing the population into non-smokers (never smokers and former smokers) and smokers (occasional smoker, i.e. <1 cigarette/day, and smoker, i.e. >1 cigarette/day).

Ultrasound of the Carotid Arteries and evaluation of the number of atherosclerotic plaques

IMT was assessed with an ie33 ultrasound system (Philips, NL) using an 11 to 3 MHz linear array transducer. Experienced technologists blinded to participants' clinical data made all ultrasound measurements. The IMT was visualized bilaterally at the far wall of the CCA. In brief, a cursor representing the region of interest (10 mm) was positioned 1 cm in front of the beginning before the carotid bulb. Evaluation was performed using an automatic computerized system (Philips, NL Qlab software) and triggering was performed according to the Q wave of the ECG to enable measurement in complete relaxation of the ventricle. IMT was recorded 1 cm before the carotid bulb in a part without plaque on the left and right side. As mean IMT, the CCA was reported with the sum of IMT of the left and right side and afterwards divided by two. Plaques were defined as thickening of the IMT of at least 1.5 mm and presence was checked in all measured arteries. The number of plaques from both sides was recorded and subjects being classified as plaque positive when at least one plaque was measured on either side or plaque negative, when no plaque was recorded.

Genotyping

For each participant genomic DNA was extracted from buffy-coats prepared from EDTA blood samples (9 mL) using the method of Miller [54]. Genotyping was performed using the *Affymetrix* Genome-Wide Human SNP Array 6.0 (http://www.affymetrix.com), as described by the *Affymetrix* user manual. Genotypes were called using the *Affymetrix* Birdseed-V2 calling algorithm and quality control was performed using GenABEL (http://mga.bionet.nsc.ru/nlru/GenABEL/). Individuals with a call rate below 97% or a too high autosomal heterozygosity (False Discovery Rate $<1\%$) were excluded. After applying standard quality criteria (minor allele frequency $>1\%$, genotype call rate $>98\%$ and *P*-value of deviation from Hardy-Weinberg equilibrium $>10^{-4}$), 675,350 out of 900,392 SNPs remained for analysis.

Separation of monocytes

Separation of monocytes was conducted within 60 min after blood collection and RNA was extracted the same day. Total RNA was isolated from monocytes using Trizol extraction and purification by silica-based columns. To separate monocytes, 8 mL blood was collected using the Vacutainer CPT Cell Preparation Tube System (BD, Heidelberg, Germany) and 400 µL Rosette Sep Monocyte Enrichment Cocktail (StemCell Technologies, Vancouver, Canada) was added immediately after blood collection. Monocytes, not labeled by antibodies, are collected as a highly enriched fraction at the interface between plasma and the density medium in the tube. After separation, cells were washed twice in ice cold PBS buffer containing 2 mM EDTA. Success of monocyte separation was controlled using an ADVIA 2120 Analyser (Siemens Healthcare Diagnostics, Eschborn, Germany) for part of the samples.

Preparation of RNA

After separation, cells were resuspended in 1.5 mL Trizol Reagent (Invitrogen, Karlsruhe, Germany) immediately and frozen at -20°C until isolation of RNA at the same day (maximal storage time 5 h). After thawing, samples were transferred into Phase Lock Gel Tubes (Eppendorf, Hamburg, Germany), 200 mL chloroform was added and phases were separated by centrifugation at 4600 rpm for 15 min. Purification of total monocytes RNA was performed using the RNeasy Mini kit (Qiagen, Hilden, Germany) according to the manufactures' Animal Cell Spin and RNA Cleanup protocols including an additional DNase digestion step. Total RNA was eluted in 20 µL RNase-free water. Yield of RNA was checked spectrophotometrically by NanoDrop N-1000 measuring the OD260 as well as the ratio OD260 and OD280. The integrity of the total RNA was assessed through analysis on an Agilent Bioanalyzer 2100 (Agilent Technologies, Boeblingen, Germany).

Genome-Wide Expression analysis

GWE analysis was performed on monocytes RNA samples using the *Illumina* HT-12 v3 BeadChip (http://www.Illumina.com). RNA samples were processed in batches of 96 samples. Two hundred ng of total RNA was reverse transcribed, amplified and biotinylated using the Illumina TotalPrep-96 RNA Amplification Kit (Ambion/Applied Biosystems, Darmstadt, Germany). 700 ng of each biotinylated cRNA was hybridized to a single BeadChip at

58°C for 16–18 hours. BeadChips were scanned using the *Illumina* Bead Array Reader.

Pre-processing of expression data

The summary probe-level data delivered by the *Illumina* scanner (mean and SD computed over all beads for a particular probe) was loaded in *Beadstudio*. The pre-processing done by the *Illumina* software, at the level of the scanner and by *Beadstudio* included: correction for local background effects, removal of outlier beads, computation of average bead signal and SD for each probe and gene, calculation of detection *P*-values using negative controls present on the array, quantile normalization across arrays, check of outlier samples using a clustering algorithm, check of positive controls. Analyses were carried out on the mean level for all probes in each gene. To stabilise variance across expression levels, we applied an arcsinh transformation to the expression data [55]. Compared to a log transformation, this transformation has the advantage not to discard negative expression values which can occur in *Illumina* data.

The *Illumina* HT-12 BeadChip included 37,804 genes (some probes being not assigned to RefSeq genes). A gene was declared significantly expressed in the dataset, i.e. expressed above background (as measured by the negative controls present on each array), when the detection *P*-value calculated by *Beadstudio* was <0.05 in more that 5% of the samples. This resulted in 22,305 genes considered as being significantly expressed in our dataset. After removing 8,058 putative and/or non well characterized genes, i.e. gene names starting by KIAA ($n = 165$), FLJ ($n = 214$), HS ($n = 4{,}262$), C*x*orf ($n = 842$), MGC ($n = 72$), LOC ($n = 2{,}503$), 12,808 well characterized detected genes remained for analysis.

Genome wide association analysis

To test all associations between SNPs and expressions in a reasonable amount of time, a C script calling the GNU Scientific Library (GSL) "TAMU ANOVA" (www.stat.tamu.edu/~aredd/tamuanova/) was written. For all significant associations, results were checked against the R-lm library [56]. When the numbers of homozygotes for the minor allele of a SNP was lower than 30, they were grouped with heterozygotes. We used a family-wise error rate of 0.05 corrected for the number of tested SNP x expression associations, which corresponds to declare significant any association with a *P*-value $<5.78\times10^{-12}$. To increase robustness, associations significant by ANOVA were further checked by a Kruskall-Wallis (KW) test and only associations with a *P*-value $<10^{-10}$ by the KW test were retained. *P*-values given in the results and in the GHS-Express database are those obtained by ANOVA. For SNPs on chromosome X, associations with gene expression were assessed separately in women and men and the *P*-values were combined using the Fisher method [57].

Association of gene expressions with CVD risk factors

The relationship between gene expression and each CHD risk factor was tested by a linear regression model using R-lm, with gene expression as the dependent variable. Association with age was adjusted for gender while association with other risk factors were adjusted for gender, age and, if specified, BMI. A square root transformation was applied to CRP levels to remove positive skewness. A study-wise statistical significance threshold of 3.9×10^{-7} was used to correct for the number of tests (10 risk factors $\times$12,808 gene expressions). For each expression trait that was associated with a risk factor and also affected by cis eSNPs, we tested the interaction between the risk factor and the best cis eSNP on expression in a regression model.

Global assessment of associations between the monocyte transcriptome and CVD risk factors

The goal of this analysis was to identify subsets of expression traits independently associated with each risk factor. To increase the robustness of the analysis the population was randomly divided into 2 sub-samples of equal size which were used for screening and validation purpose respectively. The screening step was focused on the subsets of expression traits that were associated with each covariate-adjusted risk factor in univariate analysis at $P<3.9\times10^{-6}$ (Bonferroni correction for 12,808 expressions). Each risk factor and corresponding subset of expression traits were included as dependent and predictor variables respectively in a forward stepwise regression model to identify expression traits that were independently associated with the risk factor ($P<0.01$). Gene expressions selected at the screening step were then jointly tested in the validation sample for association with the risk factor by multiple regression analysis. This screening/validation procedure was repeated 250 times and for each risk factor, expression traits associated ($P<0.01$) with the risk factor in more than 25% of the replicates are reported.

Power of the SNP-expression association analysis

Power was calculated using the program Quanto (http://hydra.usc.edu/GxE/). Assuming a quantitative expression trait with mean 0 and SD 1, a sample size of 1,490 subjects, a type I error of 5.78×10^{-12} and an additive allele effect, the study had a 82% power to detect the effect of a SNP explaining 4% of gene expression.

Functional classification of genes

An ontology analysis was performed using the Panther database (http://www.pantherdb.org/). Lists or sublists of genes involved in associations with eSNPs or risk factors were compared to the background list of the 12,808 genes. The *P*-value calculated by the binomial statistic and Bonferroni-corrected was used.

Quality checking and exclusion of outliers

Population stratification and quality of genotypes and expression data were tested extensively and outliers were excluded on the basis of multidimensional scaling analysis (see Methods S2)

GHS_Express

A downloadable SQL database compiling the results of the various associations tested is available online (http://genecanvas.ecgene.net/uploads/ForReview/), see also Methods S1. This database can be used to test specific hypotheses.

Supporting Information

Table S1 Characterization of polymorphic probes in eQTLs.
Found at: doi:10.1371/journal.pone.0010693.s001 (0.33 MB DOC)

Table S2 Loci identified in GWAS of BMI and BP - associations of lead/tag SNPs with phenotypes and expressions, and of expressions with phenotype in GHS.
Found at: doi:10.1371/journal.pone.0010693.s002 (0.07 MB DOC)

Table S3 Sets of gene expressions robustly and independently associated with each risk factor in the validation samples.
Found at: doi:10.1371/journal.pone.0010693.s003 (0.21 MB DOC)

Methods S1 GHS-Express database.

Found at: doi:10.1371/journal.pone.0010693.s004 (0.18 MB DOC)

Methods S2 Data quality checking.
Found at: doi:10.1371/journal.pone.0010693.s005 (0.21 MB DOC)

File S1 GHS - Characteristics of cis eQTLs.
Found at: doi:10.1371/journal.pone.0010693.s006 (0.40 MB XLS)

File S2 Comparison cis eQTL in GHS and the study of Stranger et al.
Found at: doi:10.1371/journal.pone.0010693.s007 (0.10 MB XLS)

File S3 Comparison cis eQTL in GHS and the study of Dixon et al.
Found at: doi:10.1371/journal.pone.0010693.s008 (0.14 MB XLS)

File S4 Comparison cis eQTL in GHS and the study of Schadt et al.
Found at: doi:10.1371/journal.pone.0010693.s009 (0.28 MB XLS)

File S5 Comparison cis eQTL in GHS and the study of Goring et al.
Found at: doi:10.1371/journal.pone.0010693.s010 (0.83 MB XLS)

File S6 GHS - Expression traits associated with risk factors.
Found at: doi:10.1371/journal.pone.0010693.s011 (0.29 MB XLS)

File S7 GHS - Interaction cis eSNPs with risk factors.
Found at: doi:10.1371/journal.pone.0010693.s012 (0.12 MB XLS)

Acknowledgments

We thank John Todd and Chris Wallace for very helpful comments on our manuscript.
We acknowledge Carolin Neukirch, Fatma Karaman, Jutta Bähr and Stefanie Müller for technical assistance, Andreas Weith for help during technical performance of GWV and GWE experiments and Alexandru Munteanu for assistance in informatics.

Author Contributions

Conceived and designed the experiments: TZ KJL JL NH TM SB FC. Performed the experiments: TZ PSW CRS RBS DM WR CP CP SB. Analyzed the data: SS MR AS RC SM MG ME EL VN DT AZ LT FC. Contributed reagents/materials/analysis tools: TZ PSW KJL HR ME CRS RBS EL DM WR AZ. Wrote the paper: TZ LT SB FC.

References

1. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–24.
2. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nat Genet 39: 1202–7.
3. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6: e107.
4. Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39: 1208–16.
5. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452: 423–8.
6. Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, et al. (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat Genet 42: 62–7.
7. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10: 184–94.
8. Nica AC, Dermitzakis ET (2008) Using gene expression to investigate the genetic basis of complex disorders. Hum Mol Genet 17: R129–134.
9. Cheung VG, Spielman RS (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet 10: 595–604.
10. Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet 24: 408–415.
11. Crujeiras AB, Parra D, Milagro FI, Goyenechea E, Larrarte E, et al. (2008) Differential expression of oxidative stress and inflammation related genes in peripheral blood mononuclear cells in response to a low-calorie diet: a nutrigenomics study. OMICS 12: 251–261.
12. Büttner P, Mosig S, Funke H (2007) Gene expression profiles of T lymphocytes are sensitive to the influence of heavy smoking: A pilot study. Immunogenetics 59: 37–43.
13. Capel F, Klimcáková E, Viguerie N, Roussel B, Vítková M, et al. (2009) Macrophages and adipocytes in human obesity: adipose tissue gene expression and insulin sensitivity during calorie restriction and weight stabilization. Diabetes 58: 1558–1567.
14. de Magalhães JP, Curado J, Church GM (2009) Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics 25: 875–881.
15. Tan Q, Zhao J, Li S, Christiansen L, Kruse TA, et al. (2008) Differential and correlation analyses of microarray gene expression data in the CEPH Utah families. Genomics 92: 94–100.
16. Yang X, Schadt EE, Wang S, Wang H, Arnold AP, et al. (2006) Tissue-specific expression and regulation of sexually dimorphic genes in mice. Genome Res 16: 995–1004.
17. Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet 8: 689–698.
18. Weber C, Zernecke A, Libby P (2008) The multifaceted contributions of leukocyte subsets to atherosclerosis: lessons from mouse models. Nat Rev Immunol 8: 802–15.
19. Sigurdsson S, Nordmark G, Garnier S, Grundberg E, Kwan T, et al. (2008) A risk haplotype of STAT4 for systemic lupus erythematosus is over-expressed, correlates with anti-dsDNA and shows additive effects with two risk alleles of IRF5. Hum Mol Genet 17: 2868–2876.
20. Handunnetthi L, Ramagopalan SV, Ebers GC, Knight JC (2010) Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. Genes Immun 11: 99–112.
21. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41: 47–55.
22. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, et al. (2009) Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet 41: 666–76.
23. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41: 18–24.
24. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41: 56–65.
25. Linsel-Nitschke P, Heeren J, Aherrahrou Z, Bruse P, Gieger C, et al. (2009) Genetic variation at chromosome 1p13.3 affects sortilin mRNA expression, cellular LDL-uptake and serum LDL levels which translates to the risk of coronary artery disease. Atherosclerosis 208: 183–9.
26. Samani NJ, Braund PS, Erdmann J, Götz A, Tomaszewski M, et al. (2008) The novel genetic variant predisposing to coronary artery disease in the region of the PSRC1 and CELSR2 genes on chromosome 1 associates with serum cholesterol. J Mol Med 86: 1233–1241.
27. Trégouët D, König IR, Erdmann J, Munteanu A, Braund PS, et al. (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat Genet 41: 283–5.
28. Erdmann J, Grosshennig A, Braund PS, König IR, Hengstenberg C, et al. (2009) New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat Genet 41: 280–282.
29. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, et al. (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 41: 334–341.
30. Visel A, Zhu Y, May D, Afzal V, Gong E, et al. (2010) Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature 464: 409–12.
31. Jarinova O, Stewart AFR, Roberts R, Wells G, Lau P, et al. (2009) Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. Arterioscler Thromb Vasc Biol 29: 1671–1677.
32. Auffray C, Fogg DK, Narni-Mancinelli E, Senechal B, Trouillet C, et al. (2009) CX3CR1+ CD115+ CD135+ common macrophage/DC precursors and the role of CX3CR1 in their response to inflammation. J Exp Med 206: 595–606.
33. Gordon S, Taylor PR (2005) Monocyte and macrophage heterogeneity. Nat Rev Immunol 5: 953–64.

34. Zhou T, Chen Y, Hao L, Zhang Y (2006) DC-SIGN and immunoregulation. Cell Mol Immunol 3: 279–83.
35. Collot-Teixeira S, Martin J, McDermott-Roe C, Poston R, McGregor JL (2007) CD36 and macrophages in atherosclerosis. Cardiovasc Res 75: 468–77.
36. Zudaire E, Portal-Núñez S, Cuttitta F (2006) The central role of adrenomedullin in host defense. J Leukoc Biol 80: 237–44.
37. Chan Y, Chiang M, Tsai Y, Su S, Chen M, et al. (2009) Absence of the Transcriptional Repressor Blimp-1 in Hematopoietic Lineages Reveals Its Role in Dendritic Cell Homeostatic Development and Function. J Immunol 183: 7039–7046.
38. He JQ, Wiesmann C, van Lookeren Campagne M (2008) A role of macrophage complement receptor CRIg in immune clearance and inflammation. Mol Immunol 45: 4041–4047.
39. Auffray C, Sieweke MH, Geissmann F (2009) Blood Monocytes: Development, Heterogeneity, and Relationship with Dendritic Cells. Annu Rev Immunol 27: 669–692.
40. Swirski FK, Weissleder R, Pittet MJ (2009) Heterogeneous In Vivo Behavior of Monocyte Subsets in Atherosclerosis. Arterioscler Thromb Vasc Biol 29: 1424–32.
41. Combadière C, Potteaux S, Gao J, Esposito B, Casanova S, et al. (2003) Decreased atherosclerotic lesion formation in CX3CR1/apolipoprotein E double knockout mice. Circulation 107: 1009–16.
42. Oram JF, Vaughan AM (2006) ATP-Binding cassette cholesterol transporters and cardiovascular disease. Circ Res 99: 1031–43.
43. Schmitz G, Langmann T, Heimerl S (2001) Role of ABCG1 and other ABCG family members in lipid metabolism. J Lipid Res 42: 1513–1520.
44. Zelcer N, Hong C, Boyadjian R, Tontonoz P (2009) LXR regulates cholesterol uptake through Idol-dependent ubiquitination of the LDL receptor. Science 325: 100–104.
45. Lindholm D, Bornhauser BC, Korhonen L (2009) Mylip makes an Idol turn into regulation of LDL receptor. Cell Mol Life Sci 66: 3399–3402.
46. Flowers MT, Ntambi JM (2008) Role of stearoyl-coenzyme A desaturase in regulating lipid metabolism. Curr Opin Lipidol 19: 248–256.
47. Watanabe M, Layne MD, Hsieh C, Maemura K, Gray S, et al. (2002) Regulation of smooth muscle cell differentiation by AT-rich interaction domain transcription factors Mrf2alpha and Mrf2beta. Circ Res 91: 382–389.
48. Lasky-Su J, Neale BM, Franke B, Anney RJL, Zhou K, et al. (2008) Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. Am J Med Genet B Neuropsychiatr Genet 147B: 1345–1354.
49. Rimkus C, Martini M, Friederichs J, Rosenberg R, Doll D, et al. (2006) Prognostic significance of downregulated expression of the candidate tumour suppressor gene SASH1 in colon cancer. Br J Cancer 95: 1419–1423.
50. Cox MA, Gomes B, Palmer K, Du K, Wiekowski M, et al. (2005) The pyrimidinergic P2Y6 receptor mediates a novel release of proinflammatory cytokines and chemokines in monocytic cells stimulated with UDP. Biochem Biophys Res Commun 330: 467–73.
51. Charles PC, Alder BD, Hilliard EG, Schisler JC, Lineberger RE, et al. (2008) Tobacco use induces anti-apoptotic, proliferative patterns of gene expression in circulating leukocytes of Caucasian males. BMC Med Genomics 1: 38.
52. Ambrose JA, Barua RS (2004) The pathophysiology of cigarette smoking and cardiovascular disease: an update. J Am Coll Cardiol 43: 1731–1737.
53. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246–1250.
54. Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16: 1215.
55. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1: S96–104.
56. R Development Core Team (2008) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing Available at: http://www.R-project.org.
57. Fisher R (1932) Statistical Methods for Research Workers. London: Oliver and Boyd.

Article 3:

Gene density influences GWAS discoveries and confounds the colocalization of GWAS loci with expression QTLs

Gene density influences GWAS discoveries and confounds the colocalization of GWAS loci with expression QTLs

Maxime Rotival 1, Tanja Zeller 2, Vinh Truong 1, Silke Szymczak 3, Raphaele Castagné 1, Philipp S. Wild 2, Arne Schillert 3, Karl J. Lackner 4, Thomas F. Münzel 2, Andreas Ziegler 3, François Cambien 1, Stefan Blankenberg 2, Laurence Tiret 1

Abstract

We studied the impact of cis acting expression quantitative trait loci (cis eQTLs) on the occurrence of hits in genome-wide association studies (GWAS) by confronting findings reported in the NHGRI GWAS catalog with cis eQTLs reported in the GHS_Express database derived from a study where gene expression was assessed from circulating monocytes in 1,467 individuals. We observed a clear enrichment of GWAS susceptibility loci in cis eQTLs (OR = 2.48, CI: 2.13 – 2.87, $p = 5.86 \times 10^{-33}$) which increased with higher significance threshold for defining cis eQTLs. However, we also found that the probability of detection of a GWAS locus was strongly dependent on the gene density at the locus. When accounting for this confounding factor, the association of GWAS loci with cis eQTLs was greatly weakened (OR = 1.64, CI: 1.39 - 1.93, $p = 2.42 \times 10^{-9}$). We found several examples in the GWAS catalog where the cis eQTL information may prove misleading for the identification of the underlying causal mechanism. Overall, gene density had a higher predictive value than cis eQTL information for the detection of GWAS susceptibility loci (pseudo R^2 = 2.6% vs 1.3%). To enhance power of GWAS, we propose to use prior knowledge on the gene density around SNPs for weighting SNPs in analysis. We illustrate the gain of power provided by such a weighted procedure with a GWAS of circulating homocysteine levels.

Introduction

Over the past 5 years, genome-wide association studies (GWAS) have met indisputable successes in the discovery of new loci involved in the susceptibility to common diseases. By January 2011, more than 4800 associations were listed in the GWAS catalog established by the National Human Genome Research Institute (NHGRI)[1]. However, despite even larger meta-analyses, only a small fraction of heritability of complex traits is accounted for by GWAS loci[2-5] and the strategies for identifying the so-called "missing heritability" are a matter of intense debate[6,7].

The concurrent availability of large-scale studies of gene expression in a variety of human tissues[8-12] has raised the prospect of leveraging expression information to enhance discovery of trait-associated loci in GWAS [13-17]. The presence of an expression quantitative trait locus (eQTL) in a region detected by GWAS may indeed provide a biological interpretation to the association and help prioritize among many equally plausible causative genes the one(s) which deserve deeper investigations[18-20]. In support to the concept of integrating information from expression in GWAS, two recent studies using eQTL data from HapMap

[1] INSERM UMRS 937, Pierre and Marie Curie University (UPMC, Paris 6) and Medical School[2] II. Medizinische Klinik und Poliklinik, Universitätsmedizin der Johannes-Gutenberg Universität [3] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, [4] Institut für Klinische Chemie und Laboratoriumsmedizin, Johannes-Gutenberg Universität.

Correspondence should be addressed to L.T. (laurence.tiret @upmc.fr)

samples have shown that GWAS associated loci were significantly enriched in *cis* eQTLs[18,21]. However, the coincidental occurrence of a GWAS locus and an eQTL in the same region does not necessarily imply a shared common mechanism but may only result from the strong linkage disequilibrium (LD) structure existing in the genome. Based on the observations that GWAS loci frequently encompass several genes and that eQTLs are particularly abundant throughout the genome[8-12], we hypothesized that gene density might be a confounding factor in the association between GWAS loci and *cis* eQTLs. We investigated this hypothesis using data from the Gutenberg Heart Study (GHS), a population-based study where expression was assessed from peripheral blood monocytes[11]. Due to the large sample size (n = 1,467), this eQTL database constitutes the most extensive one so far. Moreover, the key role played by monocytes in the pathogenesis of several disorders, such as immune, infectious and atherosclerosis-related diseases, makes this eQTL database relevant for a wide variety of complex traits.

While we confirmed the enrichment of GWAS loci in *cis* eQTLs, we showed that accounting for the gene density around SNPs considerably weakened this association. Although every case is unique, on a global scale gene density proved to be a better predictor of GWAS loci than *cis* eQTL information. This finding has practical implications since gene density is readily available from public databases and could be easily used in GWAS for weighting SNPs and increasing power to detect trait-associated loci. We provide an illustration of such a gain of power with a GWAS on circulating homocysteine levels.

Material and Methods

GWAS database

We downloaded the GWAS catalog[1] (version 07-27-10) from the NHGRI website (http://www.genome.gov/gwastudies). GWAS associations involving copy number variation polymorphisms (n=5) or haplotypes (n=14) and those for which no SNP at the locus was reported (n=103) were discarded. For each SNP reported in the GWAS catalog, we checked whether it was present among the 675 934 SNPs of the *Affymetrix* 6.0 chip used in the GHS_Express eQTL database described below. If the SNP was not present on the *Affymetrix* chip, we sought to identify the best proxy within a 500 kb distance using SNAP (SNP Annotation and Proxy Search , http://www.broadinstitute.org/mpg/snap) on the CEU Hapmap 3 panel[22]. If there was no good proxy ($r^2 > 0.8$), the SNP was discarded from analysis. When several GWAS associations reported from different studies for the same trait or a proxy trait (e.g. weight and body mass index, [BMI]) pointed to a likely common GWAS locus, we restricted the analysis to the SNP showing the best p-value for association. A GWAS locus was defined as a set of adjacent SNPs separated by less than 500 kb from each other. After these filtering steps, 1 712 SNP-phenotype associations were retained, involving 1 587 distinct SNPs (a same SNP being possibly associated to different phenotypes). The filtering steps of SNPs are summarized in supplementary figure 1.

GHS_Express eQTL database

The eQTL database used in the present study was derived from the GHS_Express database (available at http://genecanvas.ecgene.net/news.php) which reports genome-wide association between SNPs and expression levels measured in circulating monocytes from 1 467 subjects included in GHS[11]. Genotyping was performed using the *Affymetrix* Genome-Wide Human SNP Array 6.0 (Affymetrix, CA, USA). After pre-processing of data and filtering of SNPs with minor allele frequency (MAF) ≤ 0.01, call rate $\leq$ 0.98 or that escaped Hardy-Weinberg equilibrium ($p<0.0001$), 675 934 SNPs were kept for analysis.

Genome-wide expression profiles were assessed using the *Illumina* HT-12 v3 BeadChip (Illumina, CA, USA). Data were normalized using quantile normalization and VST transformation

as implemented in the lumi R package[23,24]. Probes that had no target sequence on the NCBI build 36.1 of the genome, matched to non RefSeq sequence or targeted repeat sequences that are prone to cross-hybridization problems were filtered out using ReMOAT[25] (Reannotation and Mapping of Oligonucleotide Arrays Technologies, http://remoat.sysbiol.cam.ac.uk). Probes matching to SNP-containing sequence according to ReMOAT were also removed to avoid spurious association due to a difference of binding between the probe and its transcript. The final dataset included 14 261 probes corresponding to 11 109 well annotated genes (supplementary figure 1).

We focused the present analysis on *cis* eQTLs because *trans* eQTLs are more prone to false positive findings. A c*is* association was defined by a maximum interval of 250 kb between the SNP and the 5' or 3' end of the associated transcript, as most *cis* eQTLs are found in this interval[26]. Genotype-expression association was tested at the probe level by analysis of variance with 2 df as implemented in the *gsl* C library TAMU_ANOVA. Homozygous and heterozygous genotypes were grouped for SNPs with less than 30 rare homozygotes. Unless noted otherwise, we used a Bonferroni-corrected p-value threshold of 3.2×10^{-8} to identify significant SNP-expression associations. To ensure robustness, associations significant by ANOVA were further checked by a Kruskall-Wallis (KW) test and only associations with a p-value < 10^{-6} (Bonferroni threshold for the number of significant associations with ANOVA) by the KW test were retained.

Modelling the relation between SNPs of the two databases

Our analysis was based on the 675 934 SNPs of the *Affymetrix* chip included in the GHS_Express database. Each SNP was characterized by a vector of five variables: 1. a binary variable *X1* coded 1 if the SNP was the best reported SNP for an association of the GWAS catalog (or was the best proxy for this SNP) and 0 otherwise; 2. a binary variable *X2* coded 1 if the SNP was significantly associated in *cis* with expression in the GHS_Express database and 0 otherwise; 3. the MAF of the SNP; 4. the gene density around the SNP defined as the number of RefSeq genes present on the *Illumina* HT-12 Beadchip which fell within a 250 kb interval on each side of the SNP (i.e. within a 500 kb window) ; 5. the SNP density around the SNP defined as the number of SNPs of the *Affymetrix* 6.0 chip falling within a 250 kb interval on each side of the SNP (500 kb window).

In all analyses, *X1* was considered as the dependent variable. Odds ratios (ORs) and 95% confidence intervals (CIs) were estimated using logistic regression analysis. The strength of association between *X1* and covariates was measured by the pseudo-R^2 statistics[27] :

$$R^2 = \frac{Var(y^*)}{Var(y^*) + V(e)}$$

where *y** is the continuous latent variable underlying the observed binary outcome from the logit model. The quantity *Var(y*)* is estimated directly from the log odds of the predicted probabilities and *V(e)* is the variance of the logistic probability density function equal to $\frac{\pi^2}{3}$.

RSAE score

For each SNP associated to expression (referred to as *cis* eSNP), we defined a score of relative strength of association with expression (RSAE) as the ratio of the R^2 of gene expression explained by that SNP to the maximal R^2 explained by the best *cis* eSNP at the eQTL. A SNP with an RSAE score of 1 indicates that among all the SNPs in the *cis* interval, this SNP has the strongest effect. Note that it does not imply that the SNP has a causal role in expression regulation since there might exist SNPs with a stronger impact on expression that are not assayed, or the SNP may be a neutral marker in complete association with a functional variant.

GWAS of circulating homocysteine levels

Because gene density was found to be highly

predictive of GWAS loci, we investigated whether using this information for weighting GWAS SNPs increased the power of discovering trait-associated loci. For this purpose, we performed a GWAS on circulating homocysteine levels in 3,306 individuals from GHS genotyped with the *Affymetrix* 6.0 chip (see above). Homocysteine was measured by standards protocols in fresh EDA-plasma samples using a commercially available assay (Abbot diagnostics, Delkenhein, Germany) on the Architect system. Homocysteine levels were log transformed to remove positive skewness and outliers (deviation from the mean > 4 SDs) were excluded from analysis. Levels were adjusted for age and gender prior to analysis. The GWAS was carried out using PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) with an additive model[28]. Significance of results was ascertained using the Benjamini-Hochberg false discovery rate (FDR) procedure[29].

We first performed a classical GWAS and then carried out a GWAS using a weighted FDR procedure[30] based on gene density. Weights were taken proportional to the gene density as defined above, to which we added 1 to avoid division by 0. Weights were scaled to have a mean equal to 1, as requested by the weighted FDR procedure. A FDR < 0.05 was considered significant.

Analyses were performed in R v. 2.12.0.

Results

Relation between Gwas SNPs and *cis* eSNPs

Among the 675 934 SNPs selected from the *Affymetrix* 6.0 chip, 1 587 (0.23%) were retained as directly involved, or proxies for SNPs involved, in one or several associations reported in the GWAS catalog. From the GHS_Express database, 37 307 (5.5%) of the 675 934 SNPs were associated with expression in *cis* at a study-wise level ($p < 3.2 \times 10^{-8}$). Combining information from the GWAS catalog and the GHS-Express database revealed an increased proportion of *cis* eSNPs among the Gwas-associated SNPs (12.6% vs 5.5%, OR = 2.48, CI: 2.13 – 2.87, $p = 5.86 \times 10^{-33}$) (Table 1).

The association was little affected by adjustment on the SNP MAF (OR = 2.36, CI: 2.03 – 2.74, $p = 2.15 \times 10^{-29}$). Considering more stringent significance thresholds for defining *cis* eQTLs strengthened the association (OR = 2.76, CI: 2.35 – 3.22, $p = 1.82 \times 10^{-36}$ for a threshold of 10^{-10} and OR = 2.91, CI: 2.46-3.43, $p = 9.43 \times 10^{-37}$ for a threshold of 10^{-12}). As already observed in studies on lymphoblastoid cell lines (LCLs) [18,21], the enrichment of Gwas SNPs in *cis* eSNPs was the strongest for diseases where monocytes have a key role in pathophysiology, such as infectious and autoimmune diseases (Figure 1).

			GWAS-associated SNP		
			Yes	No	All
Cis eSNP	Yes	All	200	37 107	37 307
		RSAE score	82	12 278	12 360
	No		1 387	637 240	638 627
	All		1 587	674 347	675 934

Table 1. Number of SNPs reported in the GWAS catalog and associated in cis to monocyte gene expression
RSAE: relative strength of association with expression

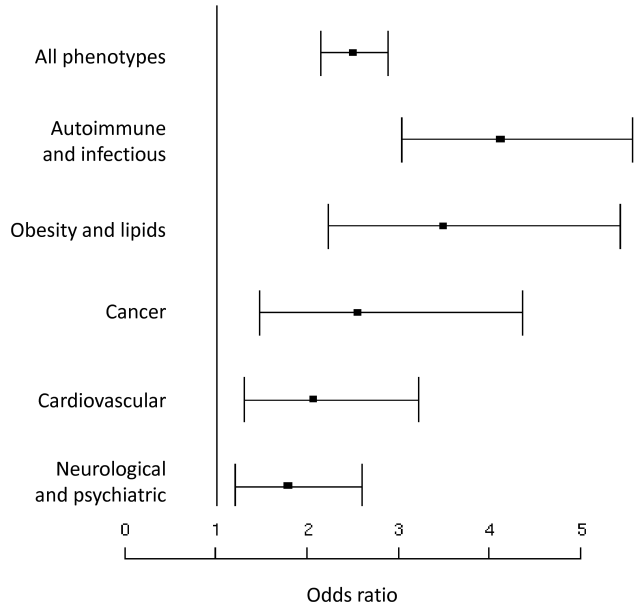


Figure 1. Association (OR and 95% CI) between GWAS SNPs and cis eSNPs according to broad categories of phenotypes

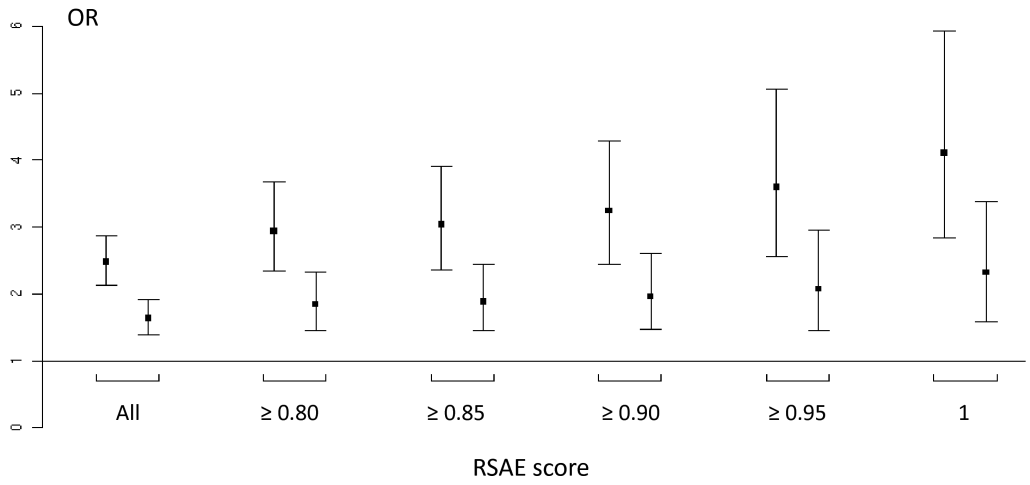


Figure 2. Association (OR and 95% CI) between GWAS SNPs and cis eSNPs according to the relative strength of association with expression (RSAE) score of the cis eSNP.
For each class of RSAE score, the ORs before (left bar) and after (right bar) adjustment on the gene density are shown.

The association was stronger when selecting the *cis* eSNPs with an RSAE score higher than 0.8 (OR = 2.94, CI: 2.35 – 3.67, p = 2.71 x 10^{-21}) (Table 1) and increased with more stringent RSAE scores (Figure 2, left bars). For a *cis* eSNP with an RSAE score of 1 (n = 3,073), the likelihood of being a GWAS SNP was multiplied by more than 4 (OR = 4.11, CI: 2.84 – 5.93, p = 6.38 x 10^{-14}).

Because the co-localization of Gwas SNPs and *cis* eSNPs seemed to preferentially occur in regions of the genome characterized by a high density of genes (Supplementary figure 2), we examined in more details the influence of the gene density on the relation between Gwas-associated SNPs and *cis* eSNPs. The number of genes around Gwas-associated SNPs varied from 0 to 40 with a median of 2 and an interquartile range of 1 to 4. The probability for any SNP of being a Gwas-associated SNP was positively related to the gene density in a 500 kb window around the SNP (OR = 1.10 per gene

in the region, CI=1.09 - 1.11, p = 1.48 x 10^{-82}), whereas it was modestly, negatively, associated to the SNP density (OR = 0.996 per SNP in the region, CI: 0.995 – 0.997, p = 4.20 x 10^{-18}). This negative association might be explained by the strong LD generally observed between neighbouring SNPs, implying that GWAS loci often encompass several SNPs with similar magnitude of association[31]. The probability that a particular SNP is the most significant one in a GWAS therefore decreases with the number of SNPs in the region.

The SNP density and the gene density were negatively correlated (r = -0.21). Introducing both variables in the logistic model little affected their respective effect (OR = 1.09, CI: 1.08 – 1.10, p = 4.67 x 10^{-63} for the gene density and OR = 0.997, CI: 0.996 – 0.998, p = 5.14 x 10^{-7} for the SNP density). Because of the strong influence of the gene density on the probability of being a Gwas SNP, we re-assessed the association between Gwas SNPs and *cis* eSNPs after adjusting for this covariate and found that is was considerably weakened (OR = 1.64, CI: 1.39 - 1.93, p = 2.42 x 10^{-9}). This weaker effect was observed whatever the RSAE score of the *cis* eSNP was (Figure 2, right bars). Exclusion of the HLA region did not modify the results (data not shown).

The predictive power of gene density and cis eQTLs on the probability of being a Gwas SNP was assessed by the pseudo-R^2. The fact of being a *cis* eSNP had a lower predictive value than the gene density (R^2 = 1.3% vs 2.6%).

Some examples of co-localizations of GWAS loci and monocytes *cis* eQTLs

All SNPs from the GWAS catalog that co-localized with a *cis* eSNP (or a proxy) from the GHS_Express database are shown in Supplementary Table 1. We focused on *cis* eSNPs with an RSAE score ≥ 0.8 to increase the likelihood that the co-localization reflects a shared causal mechanism. Supporting the notion that gene density is a major factor influencing the co-localization of GWAS loci and *cis* eQTLs, the median density of genes around the Gwas SNPs was 8, as compared to a median of 2 in the whole set of Gwas SNPs.

Supplementary Table 1 illustrates diverse situations encountered when seeking to use the *cis* eQTL information for prioritizing candidate genes at a GWAS locus. For example, a GWAS on BMI reported an association at a locus on chromosome 11 and indicated *MTCH2* (mitochondrial carrier homolog 2) as the most likely gene responsible for BMI variation[32]. The Gwas SNP rs10838738 has a prefect proxy rs4752856 on the *Affymetrix* 6.0 chip which strongly correlates to *C1QTNF4* (C1q and tumor necrosis factor
related protein 4) expression in monocytes (R^2 = 0.22, p = 1.6 x 10^{-78}, RSAE score = 1). The minor allele is associated with lower BMI and higher *C1QTNF4* expression. The *C1QTNF4* gene encodes a protein that belongs to a family of paralogs of adiponectin, a molecule known to be negatively associated to adiposity[33]. The *cis* eQTL information suggests that *C1QTNF4* might be an alternative candidate to *MTCH2* for explaining the association with BMI. Another example where the *cis* eQTL information is clearly meaningful is the GWAS association at locus 8p21 with HDL-cholesterol and triglycerides where the causal gene reported is *LPL* (lipoprotein lipase), a gene well-known to be involved in lipid metabolism[34] and whose expression is strongly *cis*-modulated in monocytes.

However, there are several counter-examples where the co-localization of a GWAS locus and a *cis* eQTL might be misleading. For instance, in the GWAS of red blood cell counts[35], although there is a strong *cis* eQTL at the 7q22 locus with an RSAE score of 1 (*GIGYF1*), the most likely candidate gene is the *EPO* (erythropoietin [MIM 133170]) gene which is not *cis* regulated in monocytes. Similarly, in the GWAS of the liver enzyme alkaline phosphatase (ALP)[36], the candidate gene at the 1p36 locus is the *ALPL* gene which encodes the ALP protein, although there is a *cis* eQTL with an RSAE score of 1 (*NBPF3*).

These examples illustrate the fact that each situation is unique and that the presence of a *cis* eQTL may not be a sufficient argument for

prioritizing genes at a GWAS locus in the absence of additional biological relevance.

GWAS of circulating homocysteine levels using a weighted FDR procedure based on gene density

Because gene density was a strong predictor of GWAS loci, we next investigated whether using this prior knowledge for weighting SNPs in a GWAS analysis increased the power to detect signals. For this purpose, we performed a GWAS of circulating homocysteine levels in 3,306 subjects of the GHS cohort. To interpret the results, we used eQTL data from the GHS_Express database.

Homocysteine is a risk factor for coronary artery disease, stroke, thrombosis and neurological disease[37]. Its circulating levels are partly genetically determined and the most consistent association reported so far is with the non-synonymous C677T polymorphism (rs1801133) located in exon 5 of the 5,10-methylenetetrahydrofolate reductase (*MTHFR*) gene which encodes a key enzyme for homocysteine metabolism[38]. In a previous GWAS of circulating homocysteine concentrations, the *MTHFR* gene was the most significant locus, whereas a second locus was detected on chromosome 11 (*SYT9*) but was not replicated in an independent study[39].

In the first GWAS that we performed without any weighting, no SNP reached the pre-specified FDR of 0.05. The *MTHFR*/C677T SNP (rs1801133) was not on the *Affymetrix* 6.0 chip and had no good proxy on the chip (best proxy rs12404124, r^2 = 0.36), which probably explained that the *MTHFR* locus was not detected. We then used the weighted FDR procedure based on gene density for prioritizing SNPs. This allowed the detection of two regions on chromosomes 1 and 6, respectively (Table 2).

The chromosome 1 region was centred on the *MTHFR* gene and included 7 significant SNPs (Table 2, Figures 3 and 4A). Each of the significant SNPs was strongly associated in *cis* with *MTHFR* expression (p-value ranging from 4.1×10^{-14} to 5.9×10^{-32}) (Table 2, Figure 4B). The SNPs were also associated to a lesser extent with *CLCN6* (chloride channel 6) expression. Although the *MTHFR* gene is the most likely candidate in the region, the RSAE score of the significant SNPs with *MTHFR* expression were moderate, not exceeding 0.4, suggesting dissociation between the variants affecting *MTHFR* expression and those associated with plasma homocysteine, as also shown in figures 4A and 4B. Actually, the SNP the most strongly associated to expression (rs1023252, $p = 2.7 \times 10^{-84}$) is located upstream of the *MTHFR* gene whereas the C677T polymorphism (rs1801133), considered to be responsible for genetic determination of plasma homocysteine, is located in exon 5 and alters an amino acid, leading to impaired enzyme function[38]. The LD between rs1023252 and rs1801133 is low (r^2 = 0.17). Another SNP (rs12121543) adjacent to the C677T polymorphism was also strongly associated to *MTHFR* expression ($p = 3.7 \times 10^{-72}$) but was also in weak LD with rs1801133 (r^2 = 0.14) (Figure 4C).

The chromosome 6 region included 12 significant SNPs with a variable number of surrounding genes (Table 2). The densest part of the region encompassed 34 genes, among which was a cluster of 4 genes belonging to the solute carrier family 17 *(SLC17A1*/A2/A3/*A4*).. One of these genes, *SLC17A2* have been previously reported to be associated with homocysteine levels using a robust candidate gene approach with cross-validation[40]. Worthy of note, The *SLC17A1*/A2/A3/*A4* gene cluster has also been reported to be associated with uric acid concentrations in several GWAS[41-43]. This finding might suggest that homocysteine and uric acid share a common determinism. However, this co-localization more likely illustrates the fact that the higher the gene density, the greater the probability of coincidental occurrence of GWAS signals is. Several SNPs at the locus were weakly associated to expression traits (Table 2). However those SNPs had a low RSAE score suggesting that the association with homocysteine and the eQTLs involved distinct causal mechanisms.

SNP ID	Chrom	Position	Unweighted FDR	Nb of surrounding genes[1]	FDR weighted by gene density[2]	Genes associated in *cis*[3]	Best p-value for *cis* association[4]	RSAE scores[5]
rs17037429	1	11 796 374	0.102	15	0.022	*MTHFR/CLCN6*	4.10E-14	0.17/0.65
rs12404124	1	11 796 456	0.093	15	0.012	*MTHFR/CLCN6*	5.25E-27	0.33/0.58
rs198391	1	11 799 004	0.093	15	0.012	*MTHFR/CLCN6*	9.17E-27	0.33/0.59
rs12567136	1	11 806 318	0.093	15	0.012	*MTHFR/CLCN6*	6.09E-15	0.18/0.62
rs7537765	1	11 809 890	0.115	15	0.024	*MTHFR/CLCN6*	1.89E-14	0.18/0.64
rs198401	1	11 810 971	0.093	15	0.011	*MTHFR/CLCN6*	2.75E-26	0.32/0.61
rs535107	1	11 812 055	0.093	15	0.012	*MTHFR/CLCN6*	3.02E-26	0.32/0.59
rs4712969	6	25 872 171	0.093	7	0.012	-	-	-
rs4712972	6	25 880 026	0.093	9	0.012	-	-	-
rs1892252	6	25 880 618	0.124	10	0.045	-	-	-
rs1892253	6	25 890 293	0.093	12	0.012	-	-	-
rs6940698	6	25 929 559	0.145	16	0.040	-	-	-
rs9461219	6	25 944 906	0.121	16	0.034	-	-	-
rs10484433	6	26 138 471	0.191	34	0.040	*HIST1H2BD*	1.42E-08	0.07
rs10484435	6	26 139 790	0.093	34	0.012	*HIST1H2BD*	6.08E-08	0.06
rs13220395	6	26 163 347	0.093	34	0.011	*HIST1H2BD*	1.05E-07	0.06
rs16891264	6	26 180 424	0.093	34	0.009	*HIST1H2BD*	3.06E-08	0.06
rs9467704	6	26 427 465	0.207	30	0.045	*BTN3A2/HIST1H2BD/BTN2A2*	6.74E-36	0.40/0.10/0.12
rs13207082	6	27 359 358	0.093	9	0.012	-	-	-

Table 2. GWAS loci associated with plasma homocysteine levels in the GHS study (n = 3,306) using a weighted FDR procedure based on gene density. The table shows all the SNPs significant using either procedure (FDR < 0.05)

[1] Number of genes within a 500 kb window around the SNP.

[2] FDR weighted by the number of genes in a 500 kb window.

[3] Genes within a 500 kb window whose expression is associated with the SNP. Genes are ranked by increasing p-value.

[4] P-value of the best cis association in the GHS_Express database between the SNP and the transcripts within a 500kb window.

[5] RSAE score of the SNP for the expression traits associated in cis (ordered as in the column 7).

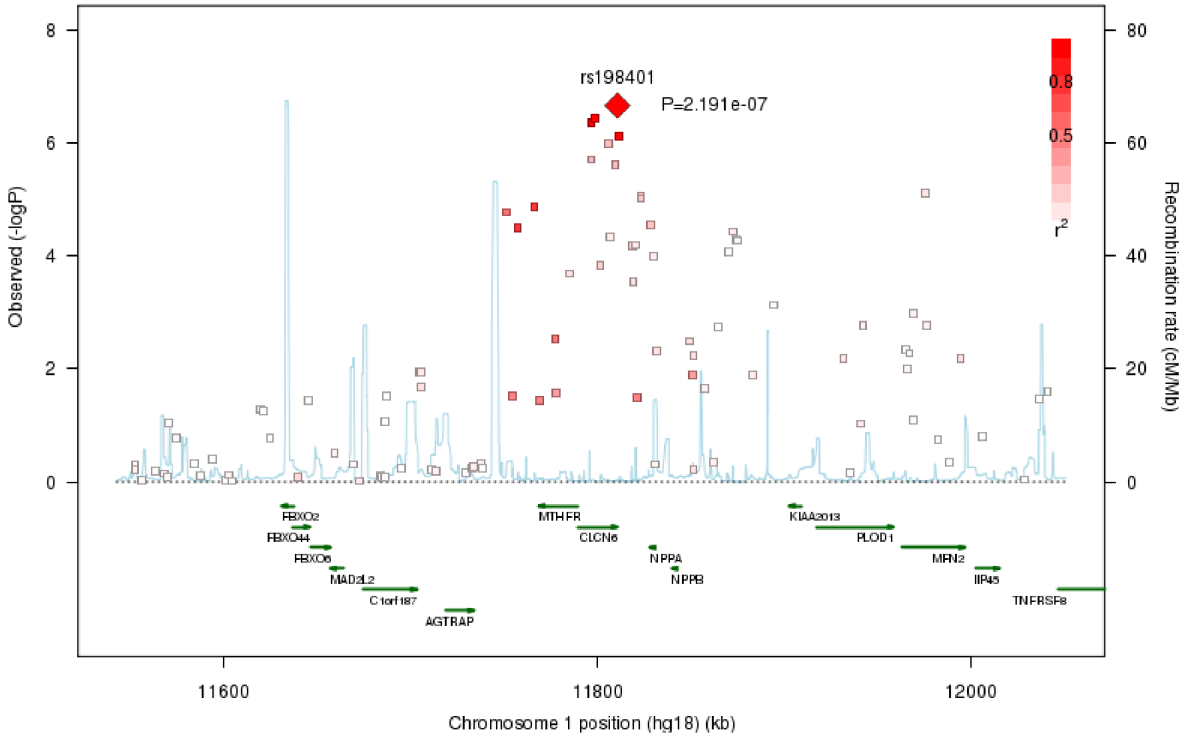


Figure 3. Region of chromosome 1 associated to circulating homocysteine (HCY) levels in a GWAS of 3,306 subjects.
Association results with HCY levels in a 600kb region (SNAP version 2.2). The most strongly associated SNP (rs198401) is shown by a red diamond. Other SNPs are shown by a square filled in shaded red according to their LD (r2) to rs198401

Discussion

Using a comprehensive database of *cis* eQTLs in human monocytes, we confirmed in the present study that Gwas SNPs were strongly enriched in SNPs influencing gene expression. This finding extends the results previously reported in other types of cells or tissues such as LCLs, liver or adipose tissue [5,44-46,18,21]. The presence of *cis* eQTLs is often used as an argument for prioritizing the genes found at a GWAS locus, upon the rationale that expression may provide a biological link to interpret the statistical association between SNPs and phenotype. In the present study, we found several examples where the GWAS locus and the *cis* eQTL pointed to a common plausible causal gene (e.g. the *LPL* gene and triglycerides and HDL-cholesterol,the *VKORC1* gene and the response to warfarin). However, even when the likely GWAS gene is also a *cis* eQTL, it does not necessarily imply that the effect on phenotype is mediated by expression, as illustrated by the *MTHFR* gene for which there is dissociation between the non-synonymous SNP influencing homocysteine levels and the SNP modulating *MTHFR* expression.

We also found that in several cases, despite the presence of a strong *cis* eQTL at the locus, the most likely candidate was not the *cis* eQTL but another gene (e.g. the *EPO* gene with red blood cell counts, the *ALPL* gene with plasma ALP levels). Similarly, in the GWAS of homocysteine levels, although there were several *cis* eQTLs at the chromosome 6 locus (*BTN3A2, HIST1H2BD* and *BTN2A2*), the most likely candidate was not a *cis* regulated gene (*SLC17A2*).

The usefulness of *cis* eQTL databases in guiding the interpretation of GWAS signals depends on several criteria, including the power of the *cis* eQTL study used for reference, the tissue in which gene expression is assessed and the sensitivity of the method used for

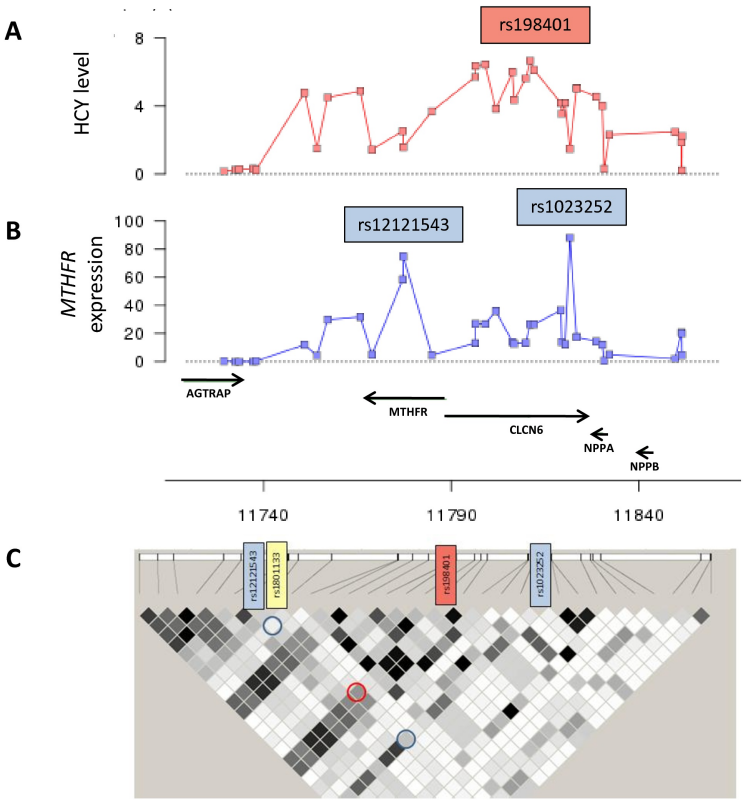


Figure 4. Colocalization of GWAS circulating homocysteine (HCY) levels and eQTL signal on chromosome 1

(A) Zoom of the association of SNPs with HCY levels in the 150 kb region around the MTHFR gene.

(B) Association of the same SNPs as in the (B) panel with MTHFR expression. The peak of association with expression is with rs1023252, the second peak is with rs12121543.

(C) Linkage disequilibrium plot (Haploview 4.2, r2 color scheme) between SNPs shown in (B) and (C) panels. The functional MTHFR/C677T polymorphism (rs1801133, yellow color), not present on the Affymetrix chip, was added in the plot. The top SNP for association with HCY is shown in pink color and the top SNPs for association with MTHFR expression are shown in blue color. The LD coefficients of these SNPs to rs1801133 are surrounded by red and blue circles, respectively

quantifying expression. An under-powered study may miss some potentially interesting *cis* eQTLs associated with modest influence on expression. In the GHS_Express database, we had an 80% power to detect a SNP effect accounting for more than 3% of the variability of expression. With the advent of more sensitive techniques for quantifying gene expression, such as RNA-sequencing, and the easier access to a larger diversity of tissues and cells, the catalog of eQTLs is expected to rapidly expand. Ideally, gene expression should be measured in a tissue relevant to the pathophysiology of the trait under study. Although recent studies support a substantial overlap of *cis* eQTLs across different tissues[11,14,44,47], genetic *cis* regulation may be in some cases restricted to a particular tissue. Such a tissue-specific regulation was recently shown for *SORT1* (sortilin). The *SORT1* gene is located at the 1p13 lipid-related locus which encompasses two other genes, *CELSR2* and *PSRC1* [48]. In monocytes, *PSCR1* has a strong *cis* eQTL whereas *CELSR2* is weakly, and *SORT1* is not, *cis* regulated[11]. According to these observations in monocytes, *PSCR1* would appear as the most attractive candidate for mediating the association with LDL-cholesterol, an interpretation supported by the correlation between *PSCR1* expression and LDL-cholesterol in GHS participants[11]. However, it was recently shown by functional experiments that the causal gene is more likely to be *SORT1* and that the effect on LDL metabolism would be due to

an alteration of *SORT1* expression that was liver-specific and not observed in adipose tissues[49].

As outlined above, the variety of situations encountered suggest that using the global information yielded by *cis* eQTLs for interpreting GWAS signals might often be misleading. To improve the method, a Regulatory Trait Concordance (RTC) score that accounts for local LD structure has been proposed to prioritize the GWAS SNP-eQTL couples that have a higher likelihood of tagging the same functional SNP[18]. This RTC score, although different from our RSAE score, was developed with the same goal of giving more weight to SNPs that have a higher impact on expression (an RTC score of 1 is equivalent to an RSAE score of 1). However, we found some examples where, in spite of an RSAE score of 1, the *cis* eQTL was clearly not the causal gene. To get better insight into the underlying biological mechanisms, a statistical method based on a fine modelling of the LD structure between SNPs has been proposed to disentangle true causal mechanisms from coincidental associations but such a method relies on a detailed knowledge of the LD structure in the region and cannot be applied on a global scale[50].

Limitations related to the tissue investigated, combined to the fact that numerous causal variants have phenotypic effects that are not mediated by gene expression (e.g. non-synonymous coding polymorphisms or splicing variants) probably explain why *cis* eQTLs are a weaker predictor of GWAS loci than gene density. Actually, gene density is a very crude and naïve indicator which does not assume any underlying biological mechanism, but just reflect the fact that causal genetic variants are more likely to be found within or close to genes. Even though there are some examples of GWAS loci falling in a gene desert, as the 10q11 locus, associated to Parkinson disease[51], most of the GWAS loci actually fall in regions encompassing several genes (Supplementary Table 1).

Because gene density is readily available from public databases, this information can be profitably used to enhance the detection of GWAS signals. The weighted FDR procedure we applied in the present study has been shown to improve power when the assignment of weights is positively associated with the null hypotheses being false[37]. This gain of power was illustrated by the GWAS on homocysteine levels where the previously known *MTHFR* locus was detected by the weighted FDR procedure whereas it was missed by an unweighted procedure. The second signal on chromosome 6 needs further confirmation.

In conclusion, our study showed that gene density is a major confounding factor in the co-localization of GWAS loci and *cis* eQTLs and that using *cis* eQTLs for prioritizing candidate genes at GWAS loci might be misleading in the absence of further biological arguments. We also showed that using a weighted FDR procedure based on gene density can increase the power for detecting susceptibility loci in GWAS.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

Acknowledgments

The Gutenberg Heart Study is funded through the government of Rheinland-Pfalz (“Stiftung Rheinland Pfalz für Innovation”, contract AZ 961-386261/733), the research programs “Wissen schafft Zukunft” and “Schwerpunkt Vaskuläre Prävention” of the Johannes Gutenberg-University of Mainz and its contract with Boehringer Ingelheim and PHILIPS Medical Systems including an unrestricted grant for the Gutenberg Heart Study. The present study was supported by the National Genome Network “NGFNplus” (contract A3 01GS0833 and 01GS0831) and by a joint funding from the Federal Ministry of Education and Research, Germany (contract BMBF 01KU0908A) and from the Agence Nationale de la Recherche, France (contract ANR 09 GENO 106 01) for the project CARDomics. MR is supported by a grant from the Fondation pour la Recherche Médicale (FDT20101220928).

References

1. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A*. 2009;**106**:9362-9367.
2. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. 2010;**466**:707-713.
3. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet*. 2010;**42**:937-948.
4. Heid IM, Jackson AU, Randall JC, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet*. 2010;**42**:949-960.
5. Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 2010;**467**:832-838.
6. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet*. 2010;**11**:446-450.
7. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;**461**:747-753.
8. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. *Nat. Genet*. 2007;**39**:1202-1207.
9. Myers AJ, Gibbs JR, Webster JA, et al. A survey of genetic human cortical gene expression. *Nat. Genet*. 2007;**39**:1494-1499.
10. Göring HHH, Curran JE, Johnson MP, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet*. 2007;**39**:1208-1216.
11. Zeller T, Wild P, Szymczak S, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS ONE*. 2010;**5**:e10693.
12. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet*. 2007;**39**:1217-1224.
13. Charlesworth JC, Peralta JM, Drigalenko E, et al. Toward the identification of causal genes in complex diseases: a gene-centric joint test of significance combining genomic and transcriptomic data. *BMC Proc*. 2009;**3 Suppl 7**:S92.
14. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature*. 2008;**452**:423-428.
15. Naukkarinen J, Surakka I, Pietiläinen KH, et al. Use of genome-wide expression data to mine the "Gray Zone" of GWA studies leads to novel candidate obesity genes. *PLoS Genet*. 2010;**6**:e1000976.
16. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet*. 2010;**86**:581-591.
17. Schadt EE, Lamb J, Yang X, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet*. 2005;**37**:710-717.
18. Nica AC, Montgomery SB, Dimas AS, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet*. 2010;**6**:e1000895.
19. Hsu Y, Zillikens MC, Wilson SG, et al. An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits. *PLoS Genet*. 2010;**6**:e1000977.
20. Parikh H, Lyssenko V, Groop LC. Prioritizing genes for follow-up from genome wide association studies using information on

gene expression in tissues relevant for type 2 diabetes mellitus. *BMC Med Genomics*. **2**:72-72.

21. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*. 2010;**6**:e1000888.
22. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008;**24**:2938-2939.
23. Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucl. Acids Res.* 2008;**36**:e11.
24. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*. 2008;**24**:1547-1548.
25. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Research*. 2010;**38**:e17.
26. Veyrieras J, Kudaravalli S, Kim SY, et al. High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet*. 2008;**4**:e1000214.
27. McKelvey RD, Zavoina W. A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*. 1975;**4**:103.
28. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 2007;**81**:559-575.
29. Hochberg Y, Benjamini. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*. **57**:289-300.
30. Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. *Biometrika*. 2006;**93**:509 -524.
31. Zhang J, Rowe WL, Clark AG, Buetow KH. Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. *Am. J. Hum. Genet*. 2003;**73**:1073-1081.
32. Willer CJ, Speliotes EK, Loos RJF, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet*. 2009;**41**:25-34.
33. Matsuzawa Y. Adiponectin: a key player in obesity related disorders. *Curr. Pharm. Des*. 2010;**16**:1896-1901.
34. Fisher RM, Humphries SE, Talmud PJ. Common variation in the lipoprotein lipase gene: effects on plasma lipids and risk of atherosclerosis. *Atherosclerosis*. 1997;**135**:145-159.
35. Ganesh SK, Zakai NA, van Rooij FJA, et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat Genet*. 2009;**41**:1191-1198.
36. Yuan X, Waterworth D, Perry JR, et al. Population-Based Genome-wide Association Studies Reveal Six Loci Influencing Plasma Levels of Liver Enzymes. *The American Journal of Human Genetics*. 2008;**83**:520-528.
37. Graham IM. Homocysteine as a risk factor for cardiovascular disease. *Trends Cardiovasc. Med*. 1991;**1**:244-249.
38. Frosst P, Blom HJ, Milos R, et al. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nat. Genet*. 1995;**10**:111-113.
39. Tanaka T, Scheet P, Giusti B, et al. Genome-wide Association Study of Vitamin B6, Vitamin B12, Folate, and Homocysteine Blood Concentrations. *Am J Hum Genet*. 2009;**84**:477-482.
40. Kardia SL, Greene MT, Boerwinkle E, Turner ST, Kullo IJ. Investigating the complex

genetic architecture of ankle-brachial index, a measure of peripheral arterial disease, in non-Hispanic whites. *BMC Med Genomics*. 2008;**1**:16.

41. Kolz M, Johnson T, Sanna S, et al. Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet*. 2009;**5**:e1000504.

42. Yang Q, Köttgen A, Dehghan A, et al. Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. *Circ Cardiovasc Genet*. 2010;**3**:523-530.

43. Dehghan A, Köttgen A, Yang Q, et al. Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *Lancet*. 2008;**372**:1953-1961.

44. Ding J, Gudjonsson JE, Liang L, et al. Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *Am. J. Hum. Genet*. 2010;**87**:779-789.

45. Dubois PCA, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet*. 2010;**42**:295-302.

46. Zhong H, Beaulaurier J, Lum PY, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet*. 2010;**6**:e1000932.

47. Bullaughey K, Chavarria CI, Coop G, Gilad Y. Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Hum. Mol. Genet*. 2009;**18**:4296-4303.

48. Kathiresan S, Melander O, Guiducci C, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet*. 2008;**40**:189-197.

49. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010;**466**:714-719.

50. Plagnol V, Smyth DJ, Todd JA, Clayton DG. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics*. 2009;**10**:327-334.

51. Fung H, Scholz S, Matarin M, et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol*. 2006;**5**:911-916.

Article 4:

A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk

LETTER

doi:10.1038/nature09386

A *trans*-acting locus regulates an anti-viral expression network and type 1 diabetes risk

Matthias Heinig[1,2]*, Enrico Petretto[3,4]*, Chris Wallace[5], Leonardo Bottolo[3,4], Maxime Rotival[6], Han Lu[3], Yoyo Li[3], Rizwan Sarwar[3], Sarah R. Langley[3], Anja Bauerfeind[1], Oliver Hummel[1], Young-Ae Lee[1,7], Svetlana Paskas[1], Carola Rintisch[1], Kathrin Saar[1], Jason Cooper[5], Rachel Buchan[3], Elizabeth E. Gray[8], Jason G. Cyster[8], Cardiogenics Consortium†, Jeanette Erdmann[9], Christian Hengstenberg[10], Seraya Maouche[6], Willem H. Ouwehand[11,12], Catherine M. Rice[12], Nilesh J. Samani[13], Heribert Schunkert[9], Alison H. Goodall[13], Herbert Schulz[1], Helge G. Roider[2], Martin Vingron[2], Stefan Blankenberg[14], Thomas Münzel[14], Tanja Zeller[14], Silke Szymczak[15], Andreas Ziegler[15], Laurence Tiret[6], Deborah J. Smyth[5], Michal Pravenec[16], Timothy J. Aitman[3], Francois Cambien[6], David Clayton[5], John A. Todd[5], Norbert Hubner[1,17] & Stuart A. Cook[3,18]

Combined analyses of gene networks and DNA sequence variation can provide new insights into the aetiology of common diseases that may not be apparent from genome-wide association studies alone. Recent advances in rat genomics are facilitating systems-genetics approaches[1,2]. Here we report the use of integrated genome-wide approaches across seven rat tissues to identify gene networks and the loci underlying their regulation. We defined an interferon regulatory factor 7 (IRF7[3])-driven inflammatory network (IDIN) enriched for viral response genes, which represents a molecular biomarker for macrophages and which was regulated in multiple tissues by a locus on rat chromosome 15q25. We show that Epstein–Barr virus induced gene 2 (*Ebi2*, also known as *Gpr183*), which lies at this locus and controls B lymphocyte migration[4,5], is expressed in macrophages and regulates the IDIN. The human orthologous locus on chromosome 13q32 controlled the human equivalent of the IDIN, which was conserved in monocytes. IDIN genes were more likely to associate with susceptibility to type 1 diabetes (T1D)—a macrophage-associated autoimmune disease—than randomly selected immune response genes ($P = 8.85 \times 10^{-6}$). The human locus controlling the IDIN was associated with the risk of T1D at single nucleotide polymorphism rs9585056 ($P = 7.0 \times 10^{-10}$; odds ratio, 1.15), which was one of five single nucleotide polymorphisms in this region associated with *EBI2* (*GPR183*) expression. These data implicate *IRF7* network genes and their regulatory locus in the pathogenesis of T1D.

Although genome-wide association studies (GWASs) have uncovered many common genetic variants associated with human diseases, the molecular mechanisms by which DNA variation affects disease risk remain poorly characterized[6]. To translate genetic association into biological function, DNA variation has been correlated with gene expression to identify the genetic drivers of gene networks, which are coordinately regulated by transcription factors and represent important determinants of disease aetiology[7–9]. Here we used a panel of recombinant inbred rat strains[1] to study transcription-factor-driven gene networks and their regulatory loci and integrated these data with human gene expression and GWAS data to identify genes, networks and pathways for human disease (Supplementary Fig. 1).

We combined expression quantitative trait loci (eQTLs) from fat, kidney and heart[1,2] with new eQTL data in aorta, skeletal muscle, adrenal gland and liver to create genome-wide eQTL data sets across seven rat tissues. We used a two-step procedure to integrate eQTLs and transcription factor target genes to identify transcription-factor-driven gene networks (Supplementary Information). In the first step, we identified 147 transcription factors whose expression mapped to 587 eQTLs across seven tissues, which were mostly (>90%) under *trans*-regulatory genetic control, in keeping with previous studies in yeast[10,11]. In the second step, we tested for enrichment of transcription factor binding sites (TFBSs)[12] in the putative promoter sequences of genes whose expression mapped to *trans*-eQTLs. Out of the 13 transcription-factor-driven gene networks identified (Supplementary Table 1) we observed the strongest TFBS enrichment for interferon regulatory transcription factor *Irf7* ($P < 1 \times 10^{-6}$; false discovery rate (FDR), $< 5 \times 10^{-5}$). *Irf7* TFBSs were predicted in the promoters of 23 genes, including *Irf7* itself, that all mapped to a single *trans*-eQTL on rat chromosome 15q25 in adrenal gland, kidney, heart and liver. We confirmed a subset of the predicted *Irf7* targets by chromatin immunoprecipitation and quantitative PCR that established direct interaction of *Irf7* with the promoters of these genes (Fig. 1a–c). Taken together, this provides evidence for a transcription-factor-driven regulatory cascade in which genetic variation on chromosome 15q25 modulates the expression of *Irf7* and *Irf7* target genes.

Irf7 is a master regulator of the type 1 interferon response[3], and genes directly regulated by *Irf7* may comprise the core components of a larger network, which we identified by genome-wide co-expression analysis of *Irf7* target genes across tissues (Supplementary Information). This revealed a network of 247 genes across seven tissues, which was expanded to 305 genes in four of the seven tissues where additional gene expression data were available (FDR < 0.1%) (Supplementary Table 2). Gene Ontology analysis of the network showed enrichment for specific biological processes, including 'immune response'

[1]Max-Delbrück-Center for Molecular Medicine (MDC), Robert-Rössle-Straße 10, 13125 Berlin, Germany. [2]Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany. [3]Medical Research Council Clinical Sciences Centre, Faculty of Medicine, Imperial College London, Hammersmith Hospital, Du Cane Road, London W12 ONN, UK. [4]Department of Epidemiology and Biostatistics, Faculty of Medicine, Imperial College London, Praed Street, London W2 1PG, UK. [5]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 OXY, UK. [6]INSERM UMRS 937, Pierre and Marie Curie University (UPMC, Paris 6) and Medical School, 91 Boulevard de l'Hôpital, Paris 75013, France. [7]Pediatric Pneumology and Immunology, Charité - Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany. [8]Howard Hughes Medical Institute and Department of Microbiology and Immunology, University of California San Francisco, California 94143, USA. [9]Universität zu Lübeck, Medizinische Klinik II, 23538 Lübeck, Germany. [10]Klinik und Poliklinik für Innere Medizin II, Universität Regensburg, 93053 Regensburg, Germany. [11]Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge CB2 0PT, UK. [12]Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton CB10 1SA, UK. [13]Department of Cardiovascular Sciences, University of Leicester and Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester LE3 9QP, UK. [14]Medizinische Klinik und Poliklinik, Johannes-Gutenberg Universität Mainz, Universitätsmedizin, Langenbeckstrasse 1, 55131 Mainz, Germany. [15]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Straße 1, 23562 Lübeck, Germany. [16]Institute of Physiology, Czech Academy of Sciences and Centre for Applied Genomics, Videnska 1083, 14220 Prague 4, Czech Republic. [17]CC4, Campus Charité Mitte, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany. [18]National Heart and Lung Institute, Imperial College, Dovehouse Street, London SW3 6LY, UK.
*These authors contributed equally to this work.
†A list of participants and their affiliations appears at the end of the paper.

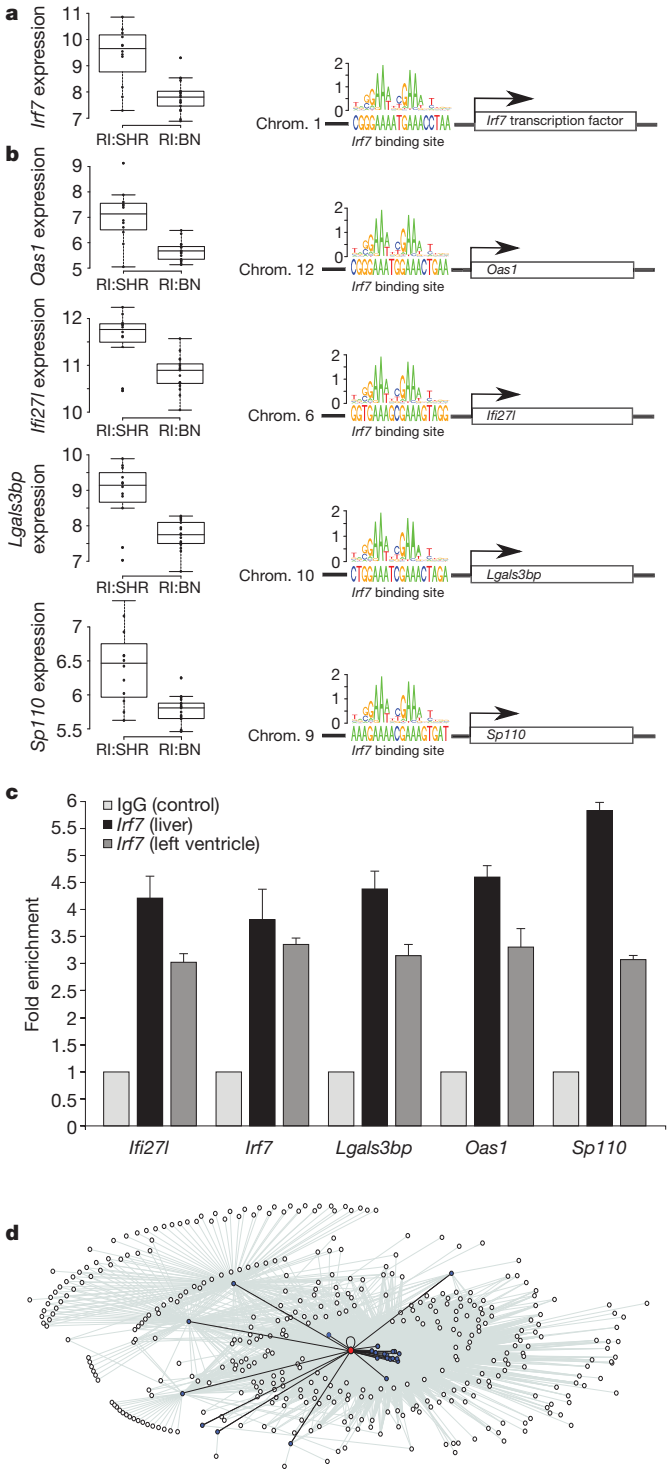


Figure 1 | The rat *Irf7*-driven inflammatory gene network. **a**, **b**, *Trans*-regulated expression of *Irf7* (**a**) and genes containing *Irf7* TFBSs (**b**) by rat chromosome 15q25. Left panels, gene expression in the left ventricle is shown in the recombinant inbred rat strains grouped by SHR or Brown Norway genotype at SNP J666808 (SHR allele, RI:SHR; Brown Norway allele, RI:BN). Data are presented as box plots (box, 25th–75th percentiles; solid bar, median; whiskers, 10th–90th percentiles; total $n = 29$). Right panels, TFBS predictions are represented for the five (out of 23 predicted) *Irf7* target genes. The chromosome encoding the *Irf7* target is shown to the left of the predicted *Irf7* binding sites. These data provide evidence for a regulatory cascade in which a locus on chromosome 15q25 regulates the expression of *Irf7* on chromosome 1 in an allele-dependent manner with consequent effects on *Irf7* target genes mediated through *Irf7* TFBSs. **c**, Quantitative chromatin immunoprecipitation of predicted *Irf7* target genes. Direct binding of *Irf7* to the promoters of the predicted targets *Ifi27l* (*Ifi27*), *Irf7*, *Lgals3bp*, *Oas1* (*Oas1a*) and *Sp110* was confirmed in liver and heart tissues. Fold enrichments are shown relative to non-immune immunoglobulin-G (IgG) control. Error bars, s.d. ($n = 5$). **d**, The expanded IDIN comprising 305 genes. Nodes represent genes; the node representing *Irf7* is coloured red and its predicted targets are coloured blue (Supplementary Table 2). Edges connect genes that are either predicted *Irf7* targets (black) or show significant Pearson correlation (FDR < 0.1%) to one of the predicted targets (grey).

($P = 7.5 \times 10^{-29}$), and 'response to virus' ($P = 9.6 \times 10^{-7}$) (Supplementary Table 3). We designated the network the *Irf7*-driven inflammatory gene network (IDIN) (Fig. 1d), which was most enriched for expression in mouse bone marrow macrophages ($P = 1.6 \times 10^{-159}$) and human monocytes ($P = 6.0 \times 10^{-177}$), with high levels of expression in other immune cells, including B lymphocytes (Supplementary Fig. 2).

Although a core of 23 *Irf7* target genes mapped to the same *trans*-eQTL on rat chromosome 15, the overall genetic control of the IDIN was unknown. We used sparse Bayesian regression models[13] to determine the association between expression levels of IDIN genes across seven tissues with genome-wide single nucleotide polymorphisms (SNPs) and identify regulatory 'hot spots'[14]. The same rat 15q25 locus, which controlled *Irf7* and its targets in *trans*, was associated with IDIN expression in all tissues (FDR < 1%) and showed the strongest evidence for common regulation in five out of seven tissues with increased expression of IDIN genes associated with the spontaneously hypertensive rat (SHR) allele (Fig. 2). The IDIN, which is highly expressed in immune cells, may represent a molecular signature of macrophages that are associated with risk of common inflammatory diseases[15] and autoimmune disease T1D[16]. Hence, we characterized expression of *Cd68*, an established marker of macrophages[17], in SHR and Brown Norway hearts and the recombinant inbred strains. *Cd68* messenger RNA levels were elevated in SHR relative to Brown Norway heart ($P = 0.01$), which reflected increased numbers of macrophages ($P = 2 \times 10^{-22}$). In the recombinant inbred strains, *Cd68* was under *trans*-acting genetic control at the 15q25 locus that regulates the IDIN (Supplementary Fig. 3).

We then analysed genetic variation in the recombinant inbred strains using SNPs[18] from the 15q25 region, which contains seven annotated protein-coding genes, and determined the expression of IDIN genes in seven inbred rat strains of known genotype that refined the locus to a 700-kilobase region (Supplementary Fig. 4). Using the SHR genome sequence[19], only *Dock9*, *Ebi2* and *Tm9sf2* showed DNA variation within the region, which was synonymous for *Dock9*, non-synonymous but not predicted to be functional for *Tm9sf2*, and a 5′ untranslated region SNP for *Ebi2* (Supplementary Table 4). *Ebi2* was the only differentially expressed gene between parental strains within the region and was *cis*-regulated in heart and kidney and highly expressed in myeloid cell types (Supplementary Figs 4 and 5). We assessed the effect of the *Ebi2* 5′ untranslated region SNP by luciferase assay; the SHR allele resulted in reduced luciferase activity relative to the Brown Norway allele (Supplementary Fig. 5).

Ebi2 encodes an orphan G-protein-coupled receptor that controls B-cell migration[4,5] and is a candidate for the regulation of the IDIN at the chromosome 15q25 region. We localized *Ebi2* expression to $Cd68^{+}$ macrophages within the rat heart (Supplementary Fig. 6), an observation that we confirmed and extended across tissues (pancreas, liver, kidney and heart) in the $Ebi2^{GFP/+}$ mouse[4] (Supplementary Fig. 7). Short interfering RNA knockdown of *Ebi2* in primary cultures of rat macrophages (Supplementary Fig. 8a) increased expression of *Irf7*, the central hub of the IDIN, and of IDIN genes (Supplementary Fig. 8b). This suggests that *Ebi2* is a negative regulator of the innate immune response in macrophages, which would be consistent with lower *Ebi2* expression in the SHR, which has more macrophages than the Brown Norway rat (Supplementary Fig. 3).

To translate our findings to humans, we tested whether the IDIN was recapitulated in human immune cells using genome-wide expression

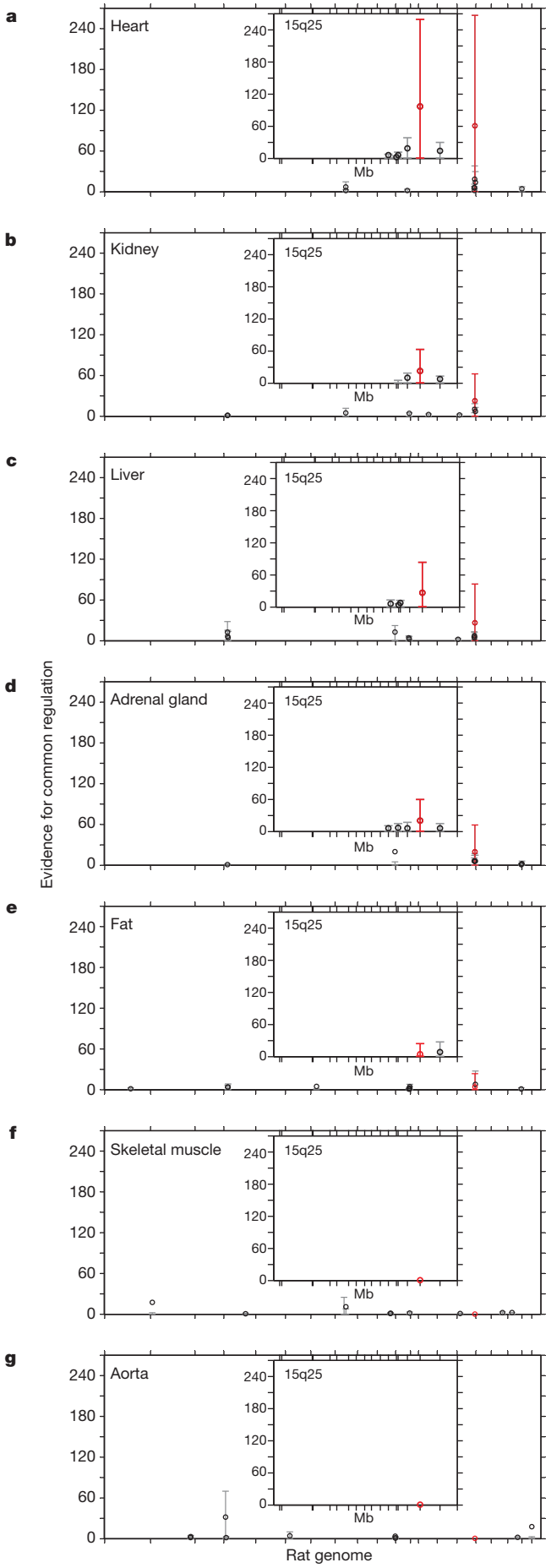


Figure 2 | Genetic mapping of regulatory hot spots for the IDIN. a–g, For each rat autosomal chromosome (horizontal axes), the strength of evidence for a SNP being a regulatory hot spot for controlling the network expression in seven tissues is measured by the average Bayes factor (vertical axes). Controlling the FDR at the 1% level for each eQTL, the average Bayes factor indicates the evidence in favour of common genetic regulation versus no genetic control, and is reported as a ratio between the strengths of these models (Supplementary Information). For the ten largest regulatory hot spots, the average Bayes factors (circles) and their 90% range (5th–95th percentiles, bars) are reported; a single SNP (J666808) that is consistently and most strongly associated with the network in five out of seven tissues is highlighted in red. Insets, average Bayes factors and 90% range for the SNPs on rat chromosome 15q25 (base pairs 87,479,238 to 108,949,015). SNP positions in the region are indicated by tick marks. Mb, megabase.

data from monocytes isolated from 1,490 individuals from the Gutenberg Heart Study[20] (GHS). We performed TFBS enrichment and co-expression analysis, analogous to that performed in the rat, and identified the human *IRF7*-driven network (Supplementary Table 5), which had strong overlap with the rat IDIN ($P = 9.1 \times 10^{-20}$) and was most significantly annotated by the Gene Ontology term 'response to virus' ($P = 1.9 \times 10^{-13}$) (Supplementary Table 6). Using monocyte gene expression data from a distinct cohort of 758 subjects from the Cardiogenics Study (Supplementary Information), we found the same set of co-regulated *IRF7* target genes (Supplementary Table 5) and significant overlap with the expanded *IRF7*-driven network indentified in the GHS ($P = 8.3 \times 10^{-23}$).

We determined whether the human chromosome 13q32 locus (spanning ~1 Mb; Supplementary Table 7), which is orthologous to the critical rat chromosome 15q25 region, was associated with expression of IDIN genes in humans. Multivariate analysis of the Cardiogenics Study monocyte expression and genotype data revealed that six SNPs in the 13q32 region (including rs9557217 ($P = 5.0 \times 10^{-5}$) and rs9585056 ($P = 1.1 \times 10^{-3}$)) were associated with *trans*-regulated expression of *IRF7* and *IRF7* target genes (Supplementary Fig. 9). We did not, however, detect a signal for *trans*-regulation of *IRF7* or *IRF7* target genes at the 13q32 locus in the GHS cohort. This may reflect differences between the monocyte selection protocols used in the two studies (Supplementary Information and data not shown).

In both the GHS and Cardiogenics Study cohorts, *EBI2* expression in monocytes was *cis*-regulated at the 13q32 locus, but the peak SNPs differed between the two cohorts (most-associated SNPs: Cardiogenics Study, rs9585056 ($P = 2.2 \times 10^{-8}$); GHS, rs9517725 ($P = 6.8 \times 10^{-13}$)) (Fig. 3). However, a formal hypothesis test[21] of a common causal genetic variant was not rejected ($P = 0.14$). Two of the five SNPs contained in the model explaining *EBI2* expression, rs9557217 and rs9585056, also had a significant *trans*-effect on IDIN gene expression in the Cardiogenics Study cohort (Supplementary Fig. 9), suggesting common regulatory control by this locus on the *IRF7* network and *EBI2* expression.

Monocyte-derived macrophages are critical determinants of inflammatory processes important for common diseases[15], including autoimmune T1D[22]. The IDIN is expressed in macrophages, enriched for immune response genes, and contained *IFIH1*, a well-characterized T1D susceptibility gene[23,24]. We evaluated the association of the human orthologues of rat IDIN genes and genes in the human IDIN (Fig. 3) with T1D (Supplementary Information). SNPs close to (≤1 Mb from) any IDIN genes were significantly more likely to associate with T1D in large-scale GWASs than SNPs close to genes not in the network ($P = 2.4 \times 10^{-10}$) (Supplementary Table 8). We also tested the IDIN association with T1D against all genes annotated by the Gene Ontology term 'immune response' and established an over-representation of T1D-associated genes ($P = 8.85 \times 10^{-6}$), indicating that the IDIN more specifically categorizes T1D genes than the Gene Ontology term 'immune response'. The association of the IDIN with T1D genes remained when the human leukocyte antigen locus was removed from the analysis ($P = 8.57 \times 10^{-4}$; Supplementary Table 8).

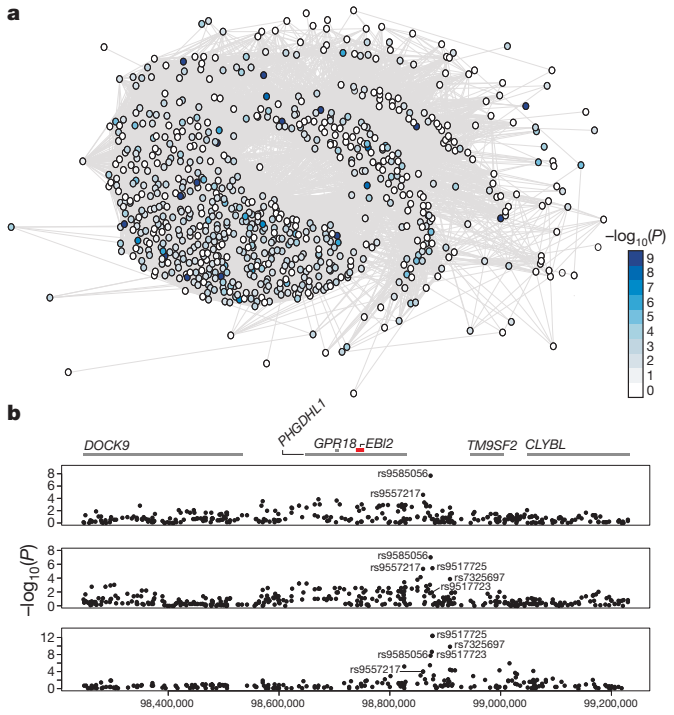


Figure 3 | A gene network and locus for T1D risk. **a**, Schematic of the union of *IRF7*-driven gene networks ($n = 697$) that was created using the set of human orthologues of rat IDIN genes ($n = 252$) and human IDIN genes ($n = 495$). A Wilcoxon rank test showed SNPs close to IDIN genes to be significantly more likely to associate with T1D in large-scale GWASs than SNPs close to randomly selected genes ($P = 2.5 \times 10^{-10}$) and randomly selected immune response genes ($P = 8.8 \times 10^{-6}$). Nodes represent IDIN genes and the node colour indicates the *P* value (negative log scale) of the association of SNPs within 1 Mb of any IDIN gene with T1D (see Methods and http://www.t1dbase.org). **b**, Results of T1D association (top) and *EBI2* eQTL analysis in the Cardiogenics Study (middle) and of *EBI2* eQTL analysis in the GHS (bottom) at the human chromosome 13q32 locus that is orthologous to the 700-kilobase rat chromosome 15q25 region. The top panel shows the $-\log_{10}(P)$ values of T1D association with SNPs in the region. SNP rs9585056 showed the strongest association with T1D ($P = 7.0 \times 10^{-10}$) among the genotyped markers. The middle and bottom panels show the nominal $-\log_{10}(P)$ values of marker regression against gene expression of *EBI2* for all SNPs in the region. We defined *EBI2* eQTL models by selecting SNPs using lasso regression (Supplementary Information) in the GHS (rs9585056, rs9517723, rs7325697). When adding imputed SNPs, rs9517725 explains most of the variation of the *EBI2* expression ($P = 6.8 \times 10^{-13}$) at this locus. Lasso model selection in the Cardiogenics Study yielded an overlapping set of three SNPs (rs9557217, rs9585056, rs9517725). The locations of genes in the region are depicted on the upper horizontal axis. On the lower horizontal axis, the genomic position on chromosome 13 is shown in base pairs.

In a GWAS meta-analysis of T1D in 7,514 cases and 9,045 controls[25], we found evidence for association of the chromosome 13q32 region at SNP rs9585056 ($P = 1.3 \times 10^{-7}$) that had not been reported before (Fig. 3b). We genotyped this SNP in two independent large cohorts and increased the strength of the T1D association (combined $P = 7.0 \times 10^{-10}$; odds ratio (95% confidence interval), 1.15 (1.09–1.21); Supplementary Table 9). The minor C allele of SNP rs9585056 was associated with T1D risk, lower *EBI2* expression in both GHS and Cardiogenics Study cohorts, and, on average, increased expression levels of IDIN genes in the Cardiogenics Study cohort. Although we cannot discriminate between single and multiple causal variants, overall these results show an overlap of association signals in the same region on human chromosome 13q32 for IDIN genes, *EBI2 cis*-regulation and T1D. We also noted that the *EBI1* (*CCR7*) and *EBI3* genes are also associated with T1D susceptibility: *EBI1* is in the confirmed T1D region 17q21.2[25], and *EBI3* encodes the β-subunit of the IL-27 cytokine, for which the α-subunit gene, *IL27*, is in the T1D region 16p11.2[25], suggesting a link between Epstein–Barr virus infection and T1D.

The immunopathology of autoimmune T1D is characterized by infiltration of the pancreas with B and T lymphocytes and macrophages[16]. We have shown that IDIN genes contribute to T1D risk and implicate the innate viral response pathway and macrophages in the aetiology of T1D. Loci that perturb gene networks can be important for disease risk[8] and the new T1D susceptibility locus that we identified may regulate innate immune response genes in macrophages, as we demonstrated in the rat. *Ebi2*, which controls *Irf7*[3], represents a candidate for *trans*-regulation of the human IDIN and for T1D risk. A role for *IRF7* in the pathogenesis of T1D is supported by functional studies[26] and by other T1D genes, namely *TLR7*, *TLR8*[27] and *IFIH1*[23,24], which are regulated by or act through *IRF7*[28]. Our study shows that co-expression networks across species provide functional annotation of genes in biological processes that can be used to reveal the signal of common genetic variation of small effect that is not detected by GWASs.

METHODS SUMMARY

We generated genome-wide expression data in the rat from seven tissues (adrenal gland, aorta, fat, kidney, left ventricle, liver and skeletal muscle) using Affymetrix RAE 230a and RAE 230_2 chips. eQTL mapping was carried out using the genetic map of the BXH/HXB recombinant inbred strains generated in a previous large-scale effort by the STAR consortium[18], as previously described[1,2]. In humans, expression data from isolated monocytes were obtained from 1,490 population-based individuals from the GHS[20] and from 758 individuals from the Cardiogenics Study. eQTL data were analysed in conjunction with TFBS enrichment analysis using PASTAA[12] to identify core gene networks centred on transcription factors. The core networks were expanded to include genes showing co-expression (FDR < 0.1%) with any of the core network genes in seven rat tissues and isolated human monocytes. We determined association between expression levels of the network genes and genome-wide SNPs in the rat using sparse Bayesian regression models[13], and identified the major regulatory control points (hot spots)[14] for the entire network. Genes at the locus associated with the rat network were characterized by DNA sequencing, RNA sequencing, quantitative PCR analyses, luciferase assay and combined *in situ* hybridization and immunohistochemistry. A combined network, comprising the union or intersection of the rat and human networks, was constructed and analysed for association with T1D by means of a stratified Wilcoxon rank test to compare SNPs genotyped in T1D GWASs[25,29] close to (≤1 Mb from) any network gene or to those close to any gene not in the network (see http://www.t1dbase.org for all T1D SNP association data). SNPs across the human locus, that is, orthologous to rat chromosome 15q25 controlling the network, were tested for association with T1D as described elsewhere[25]. Supplementary Fig. 1 provides an overview of the study design. Full methods are provided in Supplementary Information.

Received 18 December 2009; accepted 28 July 2010.
Published online 8 September 2010.

1. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genet.* **37,** 243–253 (2005).
2. Petretto, E. *et al.* Integrated genomic approaches implicate osteoglycin (*Ogn*) in the regulation of left ventricular mass. *Nature Genet.* **40,** 546–552 (2008).
3. Honda, K. *et al.* IRF-7 is the master regulator of type-I interferon-dependent immune responses. *Nature* **434,** 772–777 (2005).
4. Pereira, J. P., Kelly, L. M., Xu, Y. & Cyster, J. G. EBI2 mediates B cell segregation between the outer and centre follicle. *Nature* **460,** 1122–1126 (2009).
5. Gatto, D., Paus, D., Basten, A., Mackay, C. R. & Brink, R. Guidance of B cells by the orphan G protein-coupled receptor EBI2 shapes humoral immune responses. *Immunity* **31,** 259–269 (2009).
6. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).
7. Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461,** 218–223 (2009).
8. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–435 (2008).
9. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325,** 1246–1250 (2009).
10. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752–755 (2002).
11. Yvert, G. *et al.* *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **35,** 57–64 (2003).

12. Roider, H. G., Manke, T., O'Keeffe, S., Vingron, M. & Haas, S. A. PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics* **25,** 435–442 (2009).
13. Petretto, E. *et al.* New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol.* **6,** e1000737 (2010).
14. Breitling, R. *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* **4,** e1000232 (2008).
15. Nathan, C. & Ding, A. Nonresolving inflammation. *Cell* **140,** 871–882 (2010).
16. Eizirik, D. L., Colli, M. L. & Ortis, F. The role of inflammation in insulitis and β-cell loss in type 1 diabetes. *Nature Rev. Endocrinol.* **5,** 219–226 (2009).
17. Holness, C. L. & Simmons, D. L. Molecular cloning of CD68, a human macrophage marker related to lysosomal glycoproteins. *Blood* **81,** 1607–1613 (1993).
18. Saar, K. *et al.* SNP and haplotype mapping for genetic analysis in the rat. *Nature Genet.* **40,** 560–566 (2008).
19. Atanur, S. S. *et al.* The genome sequence of the spontaneously hypertensive rat: analysis and functional significance. *Genome Res.* **20,** 791–803 (2010).
20. Zeller, T. *et al.* Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5,** e10693 (2010).
21. Plagnol, V., Smyth, D. J., Todd, J. A. & Clayton, D. G. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* **10,** 327–334 (2009).
22. von Herrath, M. Diabetes: a virus–gene collaboration. *Nature* **459,** 518–519 (2009).
23. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324,** 387–389 (2009).
24. Smyth, D. J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nature Genet.* **38,** 617–619 (2006).
25. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genet.* **41,** 703–707 (2009).
26. Li, Q. *et al.* Interferon-α initiates type 1 diabetes in nonobese diabetic mice. *Proc. Natl Acad. Sci. USA* **105,** 12439–12444 (2008).
27. Cooper, J. D. *et al.* Follow-up of 1715 SNPs from the Wellcome Trust Case Control Consortium genome-wide association study in type I diabetes families. *Genes Immun.* **10** (suppl. 1), S85–S94 (2009).
28. Kawai, T. *et al.* Interferon-α induction through Toll-like receptors involves a direct interaction of IRF7 with MyD88 and TRAF6. *Nature Immunol.* **5,** 1061–1068 (2004).
29. The. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge funding from the German National Genome Research Network (NGFN-Plus 'Genetics of Heart Failure'), the Helmholtz Association Alliance on Systems Biology (MSBN), EURATools (LSHG-CT-2005-019015), European Union FP6 (LSHM-CT-2006-037593), PHC ALLIANCE 2009 (19419PH), UK National Institute for Health Research Biomedical Research Unit (Royal Brompton and Harefield NHS Trusts, University Hospitals of Leicester NHS Trusts) and Biomedical Research Centre (Imperial College NHS Trust) awards, the British Heart Foundation, grant P301/10/0290 from the Grant Agency of the Czech Republic, grant 1M6837805002 from the Ministry of Education of the Czech Republic, the Fondation Leducq, the Medical Research Council UK, Research Councils UK, the Juvenile Diabetes Research Foundation International, National Institute for Health Research (UK), National Institute of Diabetes and Digestive and Kidney Diseases (USA), and the Wellcome Trust. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. HEALTH-F4-2010-241504 (EURATRANS). O. Burren performed T1DBase analyses.

Author Contributions S.A.C., N.H. and E.P. initiated the study. M.H., E.P., N.H. and S.A.C. participated in the conception, design and coordination of the study. H.L., Y.L., R.S., Y.A.L., S.P., C.R., K.S. and R.B. performed genetic, biochemical and functional analyses in rats. E.E.G. and J.G.C. provided *Ebi2*$^{GFP/+}$ mouse data. M.P. and T.J.A. contributed materials and discussion of the manuscript. M.H., E.P., C.W., D.J.S., D.C., A.B., S.R.L., L.B., M.R. and L.T. designed and applied the modelling methodology and statistical analyses. M.H., E.P. and H. Schulz performed eQTL analysis in the rat. L.B. designed and performed the Bayesian analysis. C.W., D.J.S. and D.C. performed association analyses in humans. M.H., O.H., H.R. and M.V. designed and performed bioinformatics analyses in rats. J.E., C.H., S.M., W.H.O., C.M.R., N.J.S., H. Schunkert, A.H.G., S.B., T.M., T.Z., S.S., A.Z., M.R., L.T. and F.C. provided the human monocyte expression data and contributed to the transcriptomic analyses in the Cardiogenics Study and Gutenberg Heart Study cohorts. M.H., E.P., N.H. and S.A.C. wrote the paper with significant contributions from C.W. and J.A.T. All authors discussed the results and commented on the manuscript.

Author Information Microarray expression data in the rat have been deposited at ArrayExpress with the following identity codes: skeletal muscle, E-TABM-458; aorta, E-MTAB-322; liver, E-MTAB-323. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to addressed to N.H. (nhuebner@mdc-berlin.de) or S.A.C. (stuart.cook@csc.mrc.ac.uk).

Cardiogenics Consortium

Peter Braund[1], Jay Gracey[1], Unni Krishnan[1], Jasbir S. Moore[1], Chris P. Nelson[1], Helen Pollard[1], Tony Attwood[2], Abi Crisp-Hihn[2], Nicola Foad[2], Jennifer Jolley[2], Heather Lloyd-Jones[2], David Muir[2], Elizabeth Murray[2], Karen O'Leary[2], Angela Rankin[2], Jennifer Sambrook[2], Tiphaine Godfroy[3], Jessy Brocheton[3], Carole Proust[3], Gerd Schmitz[4], Susanne Heimerl[5], Ingrid Lugauer[5], Stephanie Belz[6], Stefanie Gulde[6], Patrick Linsel-Nitschke[6], Hendrik Sager[6], Laura Schroeder[6], Per Lundmark[7], Ann-Christine Syvannen[7], Jessica Neudert[8], Michael Scholz[8], Panos Deloukas[9], Emma Gray[9], Rhian Gwilliams[9] & David Niblett[9].

[1]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester LE3 9QP, UK. [2]Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge CB2 2PT, UK. [3]INSERM UMRS 937, Pierre and Marie Curie University (UPMC, Paris 6) and Medical School, 91 Boulevard de l'Hôpital, Paris 75013, France. [4]Institut für Klinische Chemie und Laboratoriumsmedizin, Universität Regensburg, 93053 Regensburg, Germany. [5]Klinik und Poliklinik für Innere Medizin II, Universität Regensburg, 93053 Regensburg, Germany. [6]Universität zu Lübeck, Medizinische Klinik II, 23538 Lübeck, Germany. [7]Molecular Medicine, Department of Medical Sciences, Uppsala University, SE-751 85 Uppsala, Sweden. [8]Trium, Analysis Online GmbH, Hohenlindenerstraße 1, 81677 München, Germany. [9]Wellcome Trust Sanger Institute, Genome Campus, Hinxton CB10 1SA, Cambridge.

Article 5:

Integrating Genome-Wide Genetic Variations and Monocyte Expression Data Reveals Reproducible Trans-Regulated Gene Modules in Humans

Integrating Genome-Wide Genetic Variations and Monocyte Expression Data Reveals Reproducible Trans-Regulated Gene Modules in Humans

Maxime Rotival [1,12] , Tanja Zeller [2,12], Philipp S. Wild [2,12] , Seraya Maouche [1], Silke Szymczak [3], Arne Schillert [3], Raphaele Castagné [1], Arne Deiseroth [2], Carole Proust [1], Jessy Brocheton [1] , Tiphaine Godefroy [1], Claire Perret [1], Marine Germain [1], Medea Eleftheriadis [2], Christoph R. Sinning [2], Renate B. Schnabel [2], Edith Lubos [2], Viviane Nicaud [1], Karl Lackner [4], Heidi Rossmann [4], Thomas F. Münzel [2], Augusto Rendon [5,6], Cardiogenics Consortium [7] , Panos Deloukas [8], Christian Hengstenberg [9], Patrick Linsel-Nitschke [10], Gilles Montalescot [1], Willem H. Ouwehand [5,8], Nilesh J. Samani [11], Heribert Schunkert [10], David-Alexandre Tregouet [1], Andreas Ziegler [3], Alison H. Goodall [11], François Cambien [1] , Laurence Tiret [1,13], Stefan Blankenberg [2,13]

Abstract

One major expectation from the transcriptome in humans is to help characterize the biological basis of associations identified by genome-wide association studies. So far, few *cis* expression quantitative trait loci (eQTLs) have been reliably related to disease susceptibility. *Trans*-regulating mechanisms may play a more prominent role in disease susceptibility. We analyzed 12,808 well-characterized genes expressed in circulating monocytes in a population-based sample of 1,490 European unrelated subjects. We applied first a method of extraction of expression patterns – independent component analysis – to identify sets of co-regulated genes. These patterns were then related to 675,350 SNPs to identify major *trans*-acting regulators. We detected three genomic regions significantly associated with patterns which replicated in Cardiogenics, an independent study in which expression profiles of monocytes were available in 758 subjects. The locus 12q13 (lead SNP rs11171739), previously identified as a type 1 diabetes locus, was associated with a pattern including two *cis* eQTLs, *RPS26* and *SUOX*, and 5 *trans* eQTLs, one of which, *MADCAM1*, being a plausible candidate for mediating T1D susceptibility. The locus 12q24 (lead SNP rs653178) which has demonstrated extensive disease pleiotropy, including type 1 diabetes, hypertension and celiac disease, was associated to a pattern strongly correlating to blood pressure level. The strongest *trans* eQTL in this pattern was *CRIP1*, a known marker of cellular proliferation in cancer. The locus 12q15 (lead SNP rs11177644) was associated with a pattern driven by two *cis* eQTLs, *LYZ* and *YEATS4*, and including 34 *trans* eQTLs, several of them being tumor-related genes. The present study shows that a method exploiting the structure of co-expressions among genes can help identify genomic regions involved in *trans* regulation of sets of genes and provide clues for understanding the mechanisms linking genome-wide association loci to disease.

Introduction

Owing to the development of genome-wide association studies (GWAS), the last two years have witnessed spectacular successes in the identification of new loci involved in the susceptibility to complex diseases [1]. However, most of these associations have yet to be translated into a full understanding of the genetic mechanisms that are mediating disease susceptibility. The possibility of assaying genome-wide expression (GWE) and genome-wide variability (GWV) simultaneously in large-scale studies opens new perspectives for

1 INSERM UMRS 937, Pierre and Marie Curie University (UPMC, Paris 6) and Medical School, 91 Bd de l'Hôpital 75013, Paris, France; 2 II. Medizinische Klinik und Poliklinik, Universitätsmedizin der Johannes-Gutenberg Universität Mainz, Langenbeckstraße 1, 55131 Mainz, Germany; 3 Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Str. ; 1, 23562 Lübeck, Germany; 4 Institut für Klinische Chemie und Laboratoriumsmedizin, Johannes-Gutenberg Universität Mainz, Universitätsmedizin, Langenbeckstraße 1, 55131 ; Mainz, Germany; 5 Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge, UK; 6 MRC Biostatistics Unit, Robinson Way, Cambridge CB2 0SR, UK; 7 http://www.cardiogenics.eu/web/; 8 Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, UK; 9 Klinik und Poliklinik für Innere Medizin II, Universität Regensburg, Regensburg, Germany; 10 Medizinische Klinik II, Universität zu Lübeck, Lübeck, German; 11 Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK; 12 These authors contributed equally to this work; 13 These authors contributed equally to this work

unravelling these mechanisms [2].

Several studies on the genetics of expression have shown that a considerable number of genes are regulated by expression SNPs and that *cis* expression quantitative loci (eQTLs) largely outnumber *trans* eQTLs [3-8]. A reason for this imbalance might be that *trans* eQTLs are beneath the level of detection of most studies because, unlike *cis* eQTLs, they do not directly influence gene expression. Moreover, *trans* associations are more sensitive to confounding factors including technical experimental effects and stratification of the cell population [9].

Identification of *trans*-acting SNPs might enhance our understanding of the molecular mechanisms that control transcriptional modules, i.e. sets of genes highly co-regulated, which are thought to be involved in pathophysiological processes [10]. Such transcriptional modules have been described in yeast [11, 12], *Drosophila* [13], mice [14-16] and humans [6]. Since *trans*-acting SNPs are expected to have pleiotropic effects on a large number of genes, each being modestly affected, statistical mapping of these SNPs may be facilitated by prior recognition of subsets of co-regulated genes [17].

In the present study, we have analyzed 12,808 well-characterized genes expressed in circulating monocytes in relation to GWV in a population-based sample of 1,490 unrelated subjects participating in the Gutenberg Heart Study (GHS). We applied first a method of extraction of expression patterns – independent component analysis (ICA) [18, 19] – in order to identify sets of co-regulated genes. These patterns were then related to 675,350 SNPs to identify major *trans*-acting regulators. We identified three genomic regions associated to expression patterns that were centred on the *ERBB3, SH2B3* and *LYZ-YEATS4* genes, respectively. Connecting these results with recent GWAS findings provided potential clues for better understanding the genetic basis of complex diseases.

Results

The study was conducted in 1,490 individuals of European origin (730 women and 760 men) aged 35 to 74 years that were recruited in the GHS, a community-based project conducted in a single centre in the region of Mainz (Germany) [20]. Monocytes were freshly isolated from peripheral blood by negative separation using a cocktail of antibodies directed against non-monocytic cells (CD2, CD3, CD8, CD19, CD56 and CD66b). GWE profiles were generated using Illumina Human HT12 BeadChip expression arrays, and after normalization and filtering out of genes whose expression was under the detection level and genes not well characterized, 12,808 expressions traits (averaged over probes) remained for analysis (see Methods).

Description of ICA method

The goal of ICA [18, 19] is to find hidden variables, called "independent components", which represent underlying processes that influence gene expression. The expression of each gene is written as a linear function of these components, where the influences of different components show minimal statistical dependencies. Each component defines groups of co-induced and/or co-repressed genes. These components may be viewed as reflecting distinct biological causes influencing gene expression, such as activation of signaling pathways, binding of transcription factors, posttranscriptional regulation…

We consider an expression data matrix X whose rows correspond to genes and columns to individuals. The ICA model splits the matrix into a matrix product $X \sim SA$ (see Figure 1), subject to the condition that the statistical dependence between the K columns of S be minimized. The expression level of gene i in individual j is

$$x_{ij} = \sum_k s_{ik}\, a_{kj}$$

where s_{ik} is the contribution of component k on gene expression i and a_{kj} is the level of "activation" of that component in individual j. Note that the components can be interpreted in

a dual view. First, each column of *S* is a vector of the linear contributions of the component on each gene expression which can be interpreted as the "signature" of the underlying biological process. To minimize the dependence between the columns of *S*, ICA identifies components that exhibit approximately sparse signatures, showing an increased proportion of contributions close to zero. Each component can then be characterized by a set of genes for which its contributions are "significantly" different from zero (see the definition of modules below). Importantly, different components can be characterized by overlapping sets of genes. For this reason, ICA is likely to better reflect biological reality than methods that partition genes into distinct clusters.

Alternatively, each component can be characterized by its pattern of expression in individuals (rows of *A*) which reflects the level of "activation" of the underlying biological process. Pattern levels are estimated by linear combinations of gene expression levels obtained by inverting the equation $X \sim SA$. Patterns can be correlated with each other in the population. This is an advantage of ICA over classical methods of dimensionality reduction relying on orthogonality of factors like principal component analysis (PCA) [21, 22].

In the following, we used the term of "signature" or that of "pattern" for a component according to whether it referred to columns of *S* (genes) or rows of *A* (individuals). Figure 1 shows an illustration of ICA for $K = 2$.

Analysis workflow

Figure 2 shows the analysis workflow. After normalization of raw expression data, filtering of undetected probes and removal of outlier samples by multi-dimensional scaling (MDS) analysis, singular value decomposition (SVD) was used prior to ICA to reduce the dimensionality of data and determine the

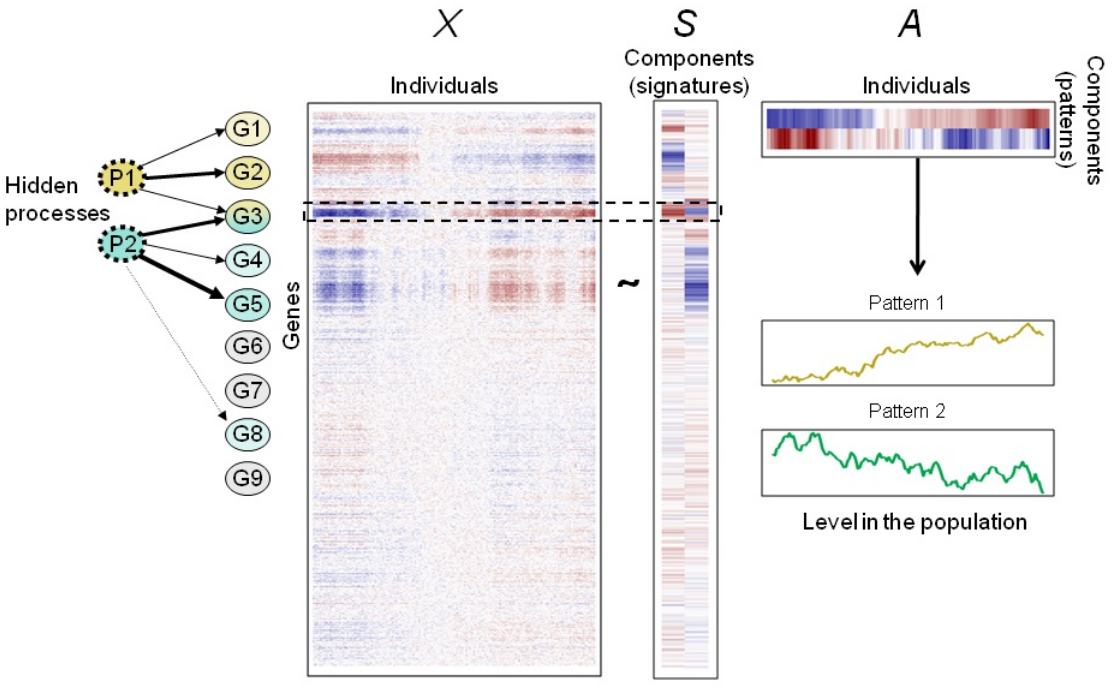


Figure 1. An example of the ICA method with K = 2 (K, number of independent components).
Data are represented using a heat color map, from dark blue (minimum) to dark red (maximum). ICA splits the gene expression matrix X into a matrix product X = SA, introducing two new components ("signatures", contained in the columns of S) with minimal statistical dependencies between them. These components may be viewed as reflecting hidden underlying processes influencing gene expressions (P1 and P2). In the example, P1 influences 3 genes and P2 influences 4 genes. Gene G3 is influenced by both processes, which is reflected by the dark red and dark blue colors in the row corresponding to G3 in matrix S. The rows of matrix A represent the levels of the two components in individuals ("patterns"). The same data are shown as continuous profiles below. Individuals have been ordered to show that when levels of pattern 1 increase, levels of pattern 2 decrease, resulting in a negative correlation between the two patterns.

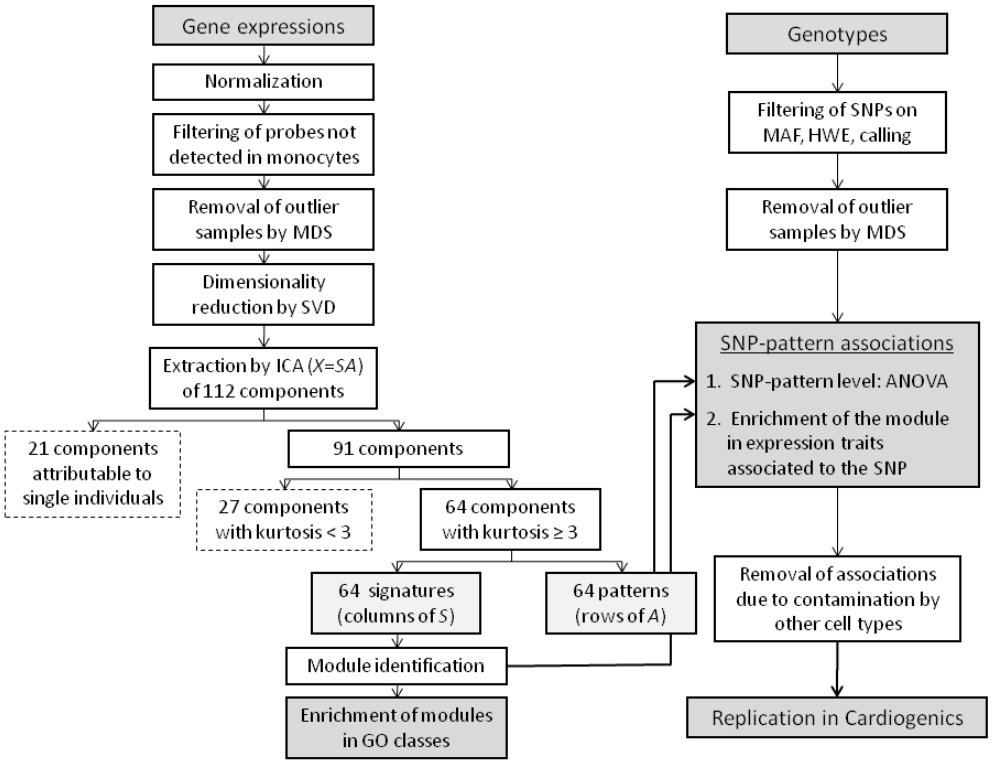


Figure 2. Analysis workflow.
The graph shows the workflows used in parallel for expression data and genotype data. MDS: multidimensional scaling; SVD: singular value decomposition; ICA: independent component analysis; GO: gene ontology; MAF: minor allele frequency; HWE: Hardy-Weinberg equilibrium.

optimal number of components to extract by ICA [18, 19] (see Text S1). As shown by the SVD screeplot (Figure S5), 30 orthogonal components were able to capture 50% of the global variability of the transcriptome. However, as we were interested in components potentially explaining small, but meaningful, variations of the transcriptome, we extended the number of components up to the limit beyond which variability appeared mostly attributable to random noise. According to the SVD screeplot, this limit was 112 (Text S1). The FastICA algorithm was then run with this fixed number of components.

Twenty-one of the 112 components identified by ICA were characterized by a single individual who explained more than 10% of the variability of the pattern in the population. Actually, we found that most of these individuals, although not having been initially identified as outliers, were at the periphery of the main cluster of individuals obtained from the MDS analysis of expression data performed prior to ICA (Text S1). These 21 "individual-specific" components were no longer considered, leaving 91 components for further analysis.

Modules of genes characterizing signatures

The fundamental principle of ICA estimation is that the columns of *S* (signatures) must be as non-gaussian as possible, typically a peaked distribution with few genes at the tails to which the signature strongly contributes, and the majority of genes in the center being weakly or not influenced [18, 19]. To determine the most non-gaussian signatures, hence the most informative components, we used the kurtosis which measures the peakedness of the distribution [18] and focused on signatures showing a kurtosis ≥ 3. This criterion led to the selection of 64 signatures. As explained above, the 64 signatures correspond to 64 patterns of expression in the population. Some of these patterns exhibited strong pairwise correlations (see the correlation matrix in the GHS_ICA_Modules database at http://genecanvas.ecgene.net/uploads/ForReview/).

For each of these 64 signatures, we defined the

"module" as the subset of genes the most strongly influenced, i.e. genes at both extremes of the distribution. For this purpose, we used a method proposed for false discovery rate (FDR) estimation [23]. Genes associated with an FDR < 10^{-3} were considered as belonging to the module characterizing the signature. The size of the modules varied from 14 to 670 genes (median 179).

Gene Ontology enrichment analysis of modules

A Gene Ontology (GO) analysis was performed to identify modules that were associated with specific biological processes. For 42 of the 64 modules (66%), we found a significant enrichment of GO classes from genes of the module (Table S1). Over-represented biological categories included a large number of categories related to immune and inflammatory response (reponse to virus, T-cell activation, response to bacteria/fungus, cytokine activity, acute inflammatory response, humoral immune response …) and several low level biological process categories such as the nucleotide metabolic process, mRNA metabolic process, ribosome biogenesis, regulation of cell proliferation, nucleosome assembly (histone genes), and cell-cycle.

Association between patterns and genome-wide variability (GWV)

We next investigated whether the level of expression of patterns in the population was influenced by SNPs. GWV genotyping was carried out using *Affymetrix* 6.0 arrays. After quality control filters, 675,350 SNPs were available for testing association with the 64 patterns.

Association between patterns and SNPs was tested in a 2-step approach (Figure 2). First, we applied a filtering to select SNP-pattern associations that were significant at $P < 10^{-7}$ (suggestive associations). The significance threshold used in this first step was taken not too stringent in order to increase the sensitivity. The second step was aimed at discarding the SNP-pattern associations that were almost entirely explained by a single or very few genes of the module whose expression strongly correlated to the SNP. This would be the case, for example, for a SNP having a strong effect on a *cis* eQTL belonging to the module, but not associated with any other expression trait of the module. To exclude these cases of less interest for the present study, a SNP-pattern association was retained at step 2 if the corresponding module was significantly enriched in expression traits individually associated to the SNP by reference to the whole set of expression traits. Since the goal here was to detect associations not necessarily very strong but clustering within modules, a threshold of $P < 10^{-5}$ was taken for associations between SNP and individual expression traits (the threshold adopted for results reported in the publicly available GHS_Express database of SNP-expression associations http://genecanvas.ecgene.net/uploads/ForReview/). Enrichment of the module in significant associations was tested using a hypergeometric test with a threshold of significance of 1.15 x 10^{-9} (Bonferroni-corrected for 64 modules x 675,350 SNPs).

This 2-step approach led to the detection of 11 patterns associated with one or several SNPs at the same locus. The proportion of variability of the pattern explained by the lead SNP at the locus varied from 1.9% to 24.8% (Table 1). Because the method of monocyte enrichment did not yield a 100% purity and even modest heterogeneity of cell content may induce artefactual correlations among expressions [24], we checked whether contamination by non-monocyte cells might affect the associations observed. For this purpose, we generated surrogate variables of contamination corresponding to each blood cell type reported in the HaemAtlas [25]. We created 7 variables corresponding to the different cell types (CD4+, CD8+, CD19+, CD56+, CD66b+, erythroblasts and megakaryocytes) by averaging in each individual his (her) levels of expression for the transcripts reported to be specific of that cell type. When re-testing the 11 SNP-pattern associations by multiple regression analysis simultaneously adjusting for the 7 contamination variables, 5

Pattern	Lead SNP associated to the pattern at the locus	Chr	Position (bp)	Genes nearby (genes *cis* regulated by the SNP are underlined)	*P* -value for the SNP-pattern association[a]	Variance of the pattern explained by the SNP (R^2)	Number of genes within the module	Number of expression traits associated to the SNP within the module[b]	*P* -value for enrichment in expression traits associated to the SNP within the module[c]	SNP-pattern a potentially du contaminatior incriminated)
33	rs2300573	1	166560874	*TBX19*	3.15 E-08	0.023	176	18	1.25 E-34	No
21	rs13023213	2	86875454	*CD8A*	7.52 E-08	0.022	292	36	3.08 E-59	Yes (T cells)
12	rs12485738	3	56840816	*ARHGEF3*	8.76 E-24	0.069	379	288	< 1.0 E 250	Yes (MKs)
93	rs1344142	3	56832473	*ARHGEF3*	1.48 E-18	0.054	135	61	1.05 E-60	Yes (MKs)
48	rs13196564	6	91563760	*MAP3K7*	5.24 E-08	0.022	311	7	5.24 E-10	Yes (B cells)
66	rs2842892	6	132856076	*STX7*	9.40 E-08	0.019	137	5	1.30 E-10	No
35	rs12705417	7	77856777	*MAGI2*	1.23 E-08	0.024	395	10	6.96 E-16	Yes (erythrobl
7	rs1058348	10	11342351	*CUGBP2*	3.49 E-08	0.02	189	49	1.45 E-46	No
62	rs653178	12	110492139	*ATXN2, SH2B3*	2.36 E-09	0.026	62	5	5.49 E-10	No
98	rs11177644	12	68072015	*LYZ , YEATS4*	1.14 E-92	0.248	45	36	1.22 E-86	No
102	rs11171739	12	54756892	*RPS26 , SUOX*	2.89 E-70	0.194	14	7	2.45 E-21	No

Table 1. Genome-wide association of SNPs with patterns

[a]P-values < 10^{-7} were considered for SNP-pattern associations; [b]P-values < 10^{-5} were considered for associations between the SNP and expression traits within the module; [c]P-values < 1.15 x 10^{-9} were considered for the enrichment in expression traits associated to the SNP within the module. MKs: megakaryocytes.

associations lost significance (Table 1). Worthy of note, the corresponding modules were enriched in GO categories relevant for the incriminated cell types (Table S1). Moreover, in several cases the best associated SNP was located in a gene highly relevant to the type of cell: the *ARHGEF3* gene, which has been reported to influence mean platelet volume [26], was involved in potential contamination by platelets; the *CD8A* gene, encoding the alpha chain of the CD8 antigen found on T cells, was involved in the level of likely contamination by T cells; the *MAP3K7* gene, a gene involved in B-cell specific immune response [27], was involved in the level of contamination by B cells. Following the same reasoning, we might anticipate a biological link between the *MAGI2* gene and potential contamination by erythroblast-derived cells (Table 1).

For associations that were not affected by potential contamination, we checked whether they replicated in the Cardiogenics Study in which monocyte GWE profiles and GWV genotypes were available in 758 subjects (see Methods). Replication in Cardiogenics was assessed by examining the association between the lead SNP (or a proxy when it was not available) and each expression trait of the module. For three of the SNP-pattern associations (rs1058348-pattern7, rs2300573-pattern33 and rs2842892-pattern66), replication was not achieved in Cardiogenics as none of the expression traits in the module was significantly associated to the SNP. Detailed results of these associations are available in the GHS_ICA_modules database (http://genecanvas.ecgene.net/uploads/ForReview/). Worthy of note, module of pattern 33 was strongly enriched in genes involved in the immune response and largely overlapped with the recently identified rat network centered on the transcription factor IRF7, a master regulator of the type-1 interferon response [28].

Association of locus 12q13 with type 1 diabetes (T1D) might be mediated by *MADCAM1* expression

The association between pattern 102 and rs11171739 on chromosome 12q13 ($P = 2.9 \times 10^{-70}$ for association, $P = 2.5 \times 10^{-21}$ for enrichment) is of particular interest as rs11171739 has been identified by GWAS as a marker for T1D susceptibility [29, 30]. The locus 12q13 encompasses two genes, *ERBB3* coding for a receptor tyrosine kinase and *RPS26* coding for a ribosomal protein. *Cis* regulation of *RPS26* in diverse tissues, in particular the pancreas, has been used to argue that this gene was a more likely candidate than *ERBB3* for T1D association although this is a matter of controversy [7, 30, 31].

Module 102 contained two *cis* eQTLs associated to rs11171739, *RPS26* and *SUOX* ($P < 10^{-300}$ and 3.1×10^{-18}, respectively). The *cis* regulation of *RPS26* in monocytes confirms that reported in other cell types [4, 7, 8]. Module 102 also contained several paralogs of *RPS26* (*RPS26L*, *RPS26L1* and *RPS26P10*) whose association with rs11171739 was probably due to cross-hybridization artifacts (Table S2). Among the other significant genes was *CCDC4* (also known as *BEND4*) on chromosome 4, a gene of unknown function whose expression was also found associated to the 12q13 locus in leukocytes [8]. *CCDC4* expression strongly correlated to *RPS26L* expression ($r = 0.83$) and its association with rs11171739 completely vanished after adjustment for *RPS26L*. The last significant gene, *MADCAM1* on chromosome 19, moderately correlated with *RPS26* ($r = 0.31$) and its association with rs11171739 ($P = 7.6 \times 10^{-20}$) lost significance after accounting for *RPS26* level ($P = 0.38$) whereas it was not modified by adjustment for *SUOX*. *MADCAM1* was expressed at low level, albeit sufficient to be considered as present according to our detection criteria. In monocytes from Cardiogenics, despite similarly low levels of *MADCAM1* that would have led to consider the gene as undetected according to standard criteria (see Discussion), the association strongly replicated ($P = 3.9 \times 10^{-13}$) with rs10876864, a proxy of rs11171739 (LD $r^2 = 0.91$). The association was in the same direction in the two studies and the SNPs were associated with similar R^2 (5.7% in GHS and

7.2% in Cardiogenics) (Figure 3). The minor allele of rs11171739 (*C*) which was associated with increased risk of type 1 diabetes [7, 30, 31] was associated with higher *MADCAM1* expression.

MADCAM1 (mucosal addressin cell adhesion molecule-1) is preferentially expressed in the intestinal tract where it participates in lymphocyte homing [32]. Its extra-intestinal expression is suspected to promote the development of chronic inflammation at other sites such as the liver or the pancreas [33]. In nonobese diabetic mice, *MAdCAM1* has been shown to be involved in the lymphocytic infiltration of pancreatic islets [34] and the development of neonatal diabetes [35]. All these elements concur to suggest that *MADCAM1* might be responsible for T1D susceptibility through a modulation of its expression by *RPS26*. Although this finding was observed in monocytes where *MADCAM1* was weakly expressed, the same mechanism of regulation of *MADCAM1* by *RPS26* may be active in other tissues that are directly engaged in T1D etiology such as the pancreas.

The *SH2B3* locus is associated with a pattern related to blood pressure and *CRIP1* expression

The association between pattern 62 and rs653178 at locus 12q24 (P = 2.4x10^{-9} for association, P = 5.5x10^{-10} for enrichment) deserved attention for several reasons. First, the locus 12q24 has been reported in GWAS to be involved in pleiotropic phenotypes including celiac disease [36], T1D [30], asthma [37], myocardial infarction and coronary artery disease [26, 37], blood pressure (BP) [38-40], platelets counts [26], eosinophil number [37] and hematocrit [39]. The locus encompasses two genes, *SH2B3* and *ATXN2*, *SH2B3* being generally considered as the most likely candidate for disease susceptibility. Second, pattern 62 strongly correlated with systolic (P = 2.7x10^{-20}) and diastolic (P = 5.7x10^{-15}) BP in GHS subjects. Third, the most significant gene expression within module 62 was *CRIP1* (P = 2.8x10^{-7}) which, in a previous analysis of GHS data, emerged as the strongest correlate of systolic BP [20]. The association of rs653178 with *CRIP1* expression replicated in Cardiogenics (P = 2.2x10^{-5}). The association was in the same direction in the two studies and the SNP was associated with comparable R^2 (2.0% in GHS and 2.6% in Cardiogenics) (Table S3, Figure 4).

SH2B3, also known as *LNK,* is a member of the family of adaptor proteins mediating the interaction between the extracellular receptors and intracellular signaling pathways. It is expressed in hematopoietic precursor cells and endothelial cells and acts as a broad inhibitor of growth factor and cytokine signaling

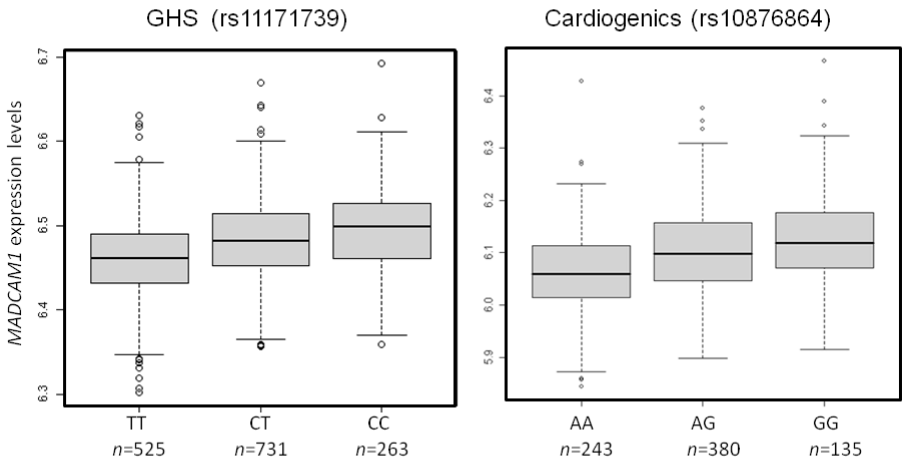


Figure 3. Box plots showing the association of locus 12q13 with MADCAM1 expression in GHS and Cardiogenics.
Because rs11171739 was not available in Cardiogenics, rs10876864 was used as a proxy (r^2 = 0.91 in HapMap3 CEU samples).

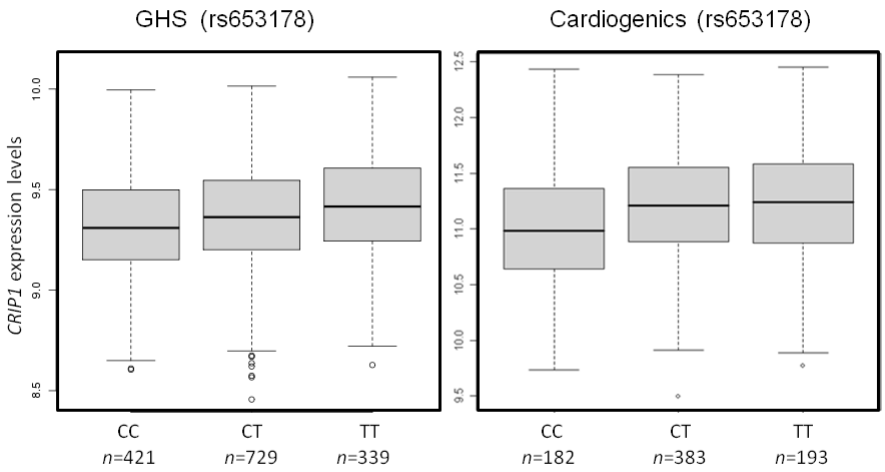


Figure 4. Box plots showing the association of rs653178 at locus 12q24 (SH2B3) with CRIP1 expression in GHS and Cardiogenics.

pathways [41]. The association of rs653178 with pattern 62 was not mediated by a *cis* effect on *SH2B3* or by any other *cis* eQTL. In addition to *CRIP1*, module 62 included four expression traits significantly associated in *trans* with rs653178 (*RAB11FIP1*, *MYADM*, *TIPARP* and *TREM1*), among which *RAB11FIP1* showed a borderline association in Cardiogenics ($P = 0.01$) (Table S3).

Rs653178 belongs to a long-range haplotype which also carries rs3184504, a non-synonymous polymorphism (*R262W*) of the *SH2B3* gene which is located in a pleckstrin homology domain involved in intracellular signaling. This haplotype has probably arisen from a selective sweep specific to Europeans since it is not observed in African and Asian populations [26]. The *C* allele of rs653178, which is the allele associated with increased BP and higher risk of disease in GWAS, was associated with decreased expression of *CRIP1* (Figure 4). However, in the GHS population, *CRIP1* expression was positively related to SBP ($r = 0.28$) and DBP ($r = 0.18$), suggesting a complex relationship between genetic variation, gene expression and disease.

CRIP1 (cysteine-rich intestinal protein) belongs to a family of proteins with a LIM domain. LIM domains are protein interaction domains functioning in the regulation of gene expression, cell adhesion and signal transduction [42]. *CRIP1* is highly expressed in immune cells and overexpression of *CRIP1* in transgenic mice has been shown to alter the immune response [43]. *CRIP1* has also been identified as a marker of cellular proliferation in several types of cancer [44]. Consistent with this role, module 62 included several genes involved in cellular growth and/or tumorigenicity (*MYADM*, *SGMS2*, *EMP1*, *ITGA5*, *KLF6, FOXO1*). The present results suggest that *CRIP1* might play a central role in the pleiotropic effects of *SH2B3* in several diseases.

Association at locus 12q15 involves a large number of *trans* effects mediated by two *cis* eQTLs, *LYZ* and *YEATS4*

The strongest association was between rs11177644 at locus 12q15 and pattern 98 ($P = 1.1\text{x}10^{-92}$ for association, $P = 1.2\text{x}10^{-86}$ for enrichment) (Table 1). The block of association included 40 SNPs and the lead SNP explained 24.8% of the pattern variance. The module included two *cis* eQTLs, *LYZ* and *YEATS4* (48.6% and 37.7% of expression variability explained by the lead SNP, respectively) as well as 34 genes associated in *trans,* 17 of which with a *P*-value $< 10^{-12}$. Almost all associations were confirmed in Cardiogenics (Table S4). Most expression traits of module 98 negatively correlated to *LYZ* and *YEATS4* (Figure S6). When including expression levels of *LYZ* and *YEATS4* as

covariates in the linear regression model relating each *trans* eQTL to rs11177644, all *trans* associations considerably decreased (median R^2 decreasing from 3.2% to 0.5%), suggesting that these *trans* associations were mediated by *cis* regulation at the locus. *LYZ* encodes human lyzozyme which is secreted by monocytes and has a bacteriolytic function. *YEATS4* (also known as *GAS41*) is a member of a large family of domain proteins which form complexes involved in chromatin modification and transcriptional regulation and has a strong link to cancer [45]. It was not possible from the present data to infer whether pattern 98 reflects a unique pathway involving *LYZ* and *YEATS4* or whether it was a mixture of two independent pathways that showed coincidental correlation because of the physical proximity of *LYZ* and *YEATS4* on chromosome 12.

Comparison of ICA and weighted gene co-expression network analysis (WGCNA)

To validate the approach used in this study, we compared the results obtained by ICA to those obtained by WGCNA, a method recently proposed to identify sets of co-expressed genes [46-48]. The WGCNA method clusters genes into non overlapping classes, called "modules", based on their profiles of co-expression, and each module is characterized by its first principal component referred to as the module eigengene (ME).

Applied to our data, the WGCNA method clustered the 12,808 gene expression traits into 26 modules. We computed the correlations between the 26 MEs and the 64 patterns obtained by ICA (Figure 5). Twenty-three MEs (88%) exhibited a correlation > 0.8 with at least one ICA pattern. Conversely, only 20 ICA patterns (31%) correlated to a ME with the same intensity, suggesting that ICA was able to identify patterns that were not represented by WGCNA modules, such as patterns 62 and 102 described above. Eleven MEs (42%) were found enriched in GO categories against 42 (66%) for ICA using the same significance threshold, suggesting that ICA was able to recover more biologically relevant patterns than WGCNA.

We next compared the power of the two methods for identifying SNPs associated with sets of co-expressed genes. Figure 6 compares the quantile-quantile plots of the Sidak-corrected *P*-values obtained when testing the 675,350 SNPs against the 26 WGCNA MEs on

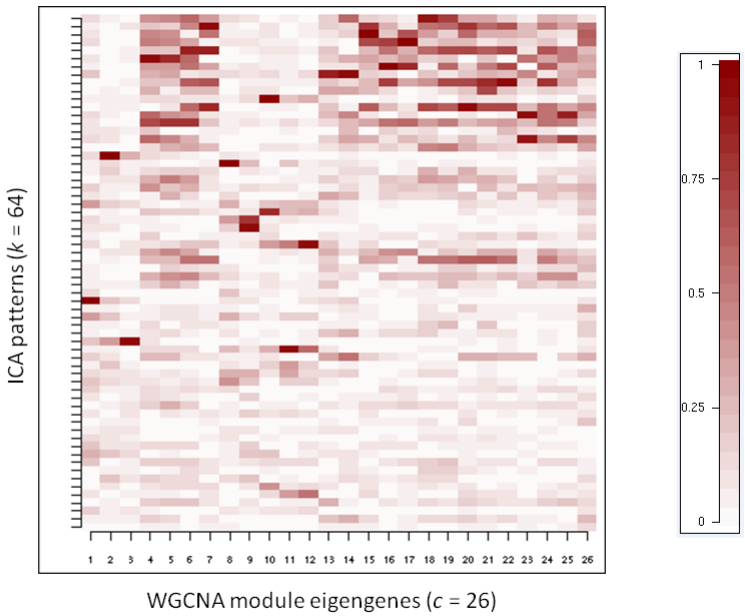


Figure 5. Matrix of absolute Pearson correlation coefficients between expression patterns obtained by ICA and module eigengenes obtained by WGCNA. ICA patterns (rows) are ordered by decreasing explained variance.

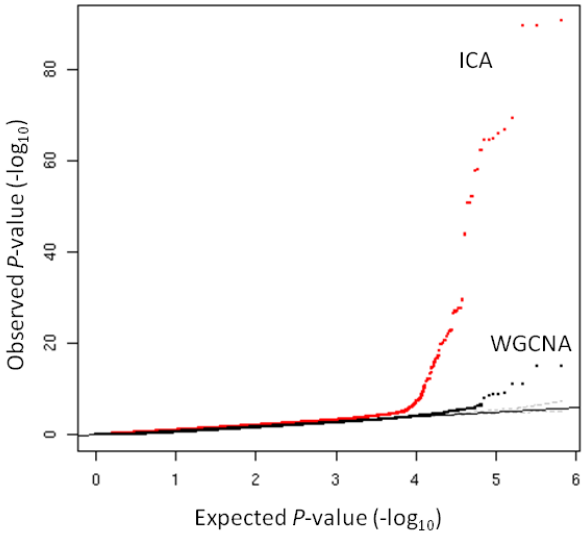


Figure 6. Quantile-quantile plot comparing the associations of 675,350 SNPs with module eigengenes (MEs) obtained by WGCNA (black) and patterns obtained by ICA (red). For each SNP, the best P-value over the 26 MEs (64 ICA patterns, respectively) is shown. A Sidak correction was applied to correct for the number of MEs (patterns, resp.) tested.

one hand, and the 64 ICA patterns on the other hand. Much stronger associations were found with ICA patterns than with WGCNA MEs. The strongest associations ($P < 10^{-9}$) detected with the WGCNA MEs involved SNPs of the *ARHGEF3* locus, which turned out to be explained by contamination by platelet RNA (see above).

Discussion

Various methods have been proposed to detect sets of co-expressed genes, including nonnegative matrix factorization [49], connectivity-based approaches such as WGCNA [6, 46-48, 50] or Bayesian networks [7, 10, 12, 51]. The ICA method [18, 19] used in the present study is based on the assumption that the co-expression of genes may be described by a small number of latent features exerting independent influences on expression. Ideally, these features may be related to distinct biological causes of variation, like regulators of gene expression, cellular functions or response to environment [19]. ICA has the advantages over co-expression network methods like WGCNA to allow for overlapping modules of genes, and over methods based on orthogonality of factors like PCA, to allow for correlation between patterns. For these reasons it might better reflect the complexity of biological systems. ICA has been applied to different types of microarray data, in particular to identify expression signatures in cancer [52-54].

Most of the components extracted by ICA could be characterized by a specific module of genes. A GO analysis indicated that two thirds of these modules were enriched in genes belonging to GO categories, thus highlighting the ability of ICA to recover biologically meaningful covariation. As pointed out by others [13], modules can help in functional annotation of genes of unknown function based on known annotations of other genes in the module, such as *CCDC4* in the *RPS26*-associated module.

Patterns of co-expression might be confounded by systematic variations introduced during sample processing or microarray measurements and by heterogeneity of the cell population [9, 24]. In particular, patterns observed in unseparated peripheral blood mononuclear cells or whole tissues are more likely to reflect variations in the tissue composition rather than

true cell-specific co-expression. In the present study, monocytes were isolated by negative selection. The choice of the method for separation of leukocytes is a matter of debate. Negative selection results in lower cell purity, while positive selection may induce cellular activation and altered transcription due to cross-linking cell surface antigens. A comparison of the two methods in 6 subjects suggested that positive selection did not induce important changes in gene expression [24]. However, the study had little power to detect modest variations such as those involved in *trans* associations. Thanks to the recent advances in the characterization of genes specific of the different blood cell lineages [25], it is now possible to better control *in silico* for potential heterogeneity of the cell population under study. We used this information to test the robustness of the SNP-pattern associations after adjustment for surrogate variables of contamination. It was not possible, however, to adjust expression data prior to ICA since we observed that such adjustment could induce other artifactual correlations, probably because genes supposed to be specific of a given cell type may also be expressed, although at lower levels, in the monocyte. Since none of the presently available methods yields an 100% purity, *in silico* adjustment appears as a solution for *post hoc* controlling the robustness of associations as recently proposed [55, 56].

Using this robust approach, we detected three genomic regions associated in *trans* with clusters of co-expressed genes, associations that were replicated in an independent study, Cardiogenics. For two of these regions, the *trans* effects appeared to be mediated by one (or two) *cis* eQTLs (*RPS26* and *LYZ/YEATS4*) while in the third case (*SH2B3*), the *trans* associations were likely to be explained by an alteration of the intracellular signaling. Most importantly, this approach provided some potential clues for deciphering the mechanisms underlying the relationship between GWAS loci and disease susceptibility, as illustrated by *MADCAM1* and *CRIP1*. The presence of *MADCAM1* in a module of genes co-expressed with *RPS26* provides a potential interpretation for its role in T1D susceptibility via a regulation by *RPS26*. The mechanism of regulation has to be confirmed in other cells or tissues where *MADCAM1* is more specifically expressed. This result emphasizes the fact that peripheral circulating blood may offer an easily accessible window to decipher genetic mechanisms that are ubiquitous but may have a pathophysiological relevance only in specific disease-related tissues. The example of *MADCAM1* also raises an important issue related to the criteria usually adopted for considering that a gene is expressed or not in a tissue or cell type. Conventionally, a gene is considered as expressed if its level is significantly higher than the background level. But even when its level is below the statistical detection threshold, a transcript may be of tremendous interest if, as *MADCAM1*, it proves to be related to a SNP or any other relevant factor. The increasing power of most contemporary transcriptomic studies should facilitate the detection of such effects that were missed in earlier less-powered studies.

While the *trans* association between *CRIP1* expression and SNPs at the *SH2B3* locus would not have reached the conventional level of significance required in a genome-wide analysis, its presence in a module related to BP and enriched in *SH2B3*-associated expressions makes this gene a plausible candidate for mediating the relationship between *SH2B3* genetic variations and various complex traits, in particular blood pressure. This hypothesis will require further experimental validation.

In conclusion, the present study shows that a method exploiting the structure of co-expressions among genes such as ICA can help identify genomic regions involved in *trans* regulation of sets of genes and provide clues for understanding the mechanisms linking GWAS loci to disease. It also suggests that *trans* associations involving large sets of gene expressions may reflect stratification of the cell population that can be controlled for by *in silico* adjustment.

Methods

More details are provided in Text S1.

Subjects. Study participants of both sexes aged 35-74 yr, were successively enrolled into the GHS, a community-based, prospective, observational single-center cohort study in the Rhein-Main region in western mid-Germany. The majority of participants were of European origin. A few non-European individuals detected by MDS analysis of genetic data (see below) were excluded prior to analysis, leaving 1,490 subjects for further analysis.

Ethics statement. All subjects gave written informed consent. Ethical approval was given by the local ethics committee and by the local and federal data safety commissioners.

Genotyping. GWV genotyping was performed using the *Affymetrix* Genome-Wide Human SNP Array 6.0 and the Genome-Wide Human SNP *Nsp*I/*Sty*I 5.0 Assay kit. Genotypes were called using the *Affymetrix* Birdseed-V2 calling algorithm and quality control was performed using GenABEL [57] (http://mga.bionet.nsc.ru/nlru/GenABEL/).

Separation of monocytes. Separation of monocytes was conducted within 60 min after blood collection. 8 mL blood was collected using the Vacutainer CPT Cell Preparation Tube System (BD, Heidelberg, Germany) and 400 µL RosetteSep Monocyte Enrichment Cocktail (StemCell Technologies, Vancouver, Canada) was added immediately after blood collection. This cocktail contains antibodies directed against cell surface antigens on human hematopoietic cells (CD2, CD3, CD8, CD19, CD56, CD66b) and glycophorin A on red blood cells. Total RNA was extracted the same day using Trizol extraction and purification by silica-based columns.

Microarray hybridization and data pre-processing. GWE assessment was performed using the *Illumina* HT-12 v3 BeadChip**.** Pre-processing of data and quantile normalization was performed using *Beadstudio*. Analysis was performed on the mean levels of probes of genes. To stabilize variance across gene expression levels, data were arcsinh-transformed. The *Illumina* HT-12 chip included 37,804 genes (including probes not assigned to RefSeq transcripts). A gene was declared expressed when the fraction of samples with a detection *P*-value < 0.05 for that gene was significantly higher than 5% (Text S1). After removing putative and/or non well characterized genes (i.e. gene names starting by KIAA, FLJ, HS., C*x*orf, MGC, LOC, NT_, ENSG), 12,808 genes remained for analysis.

Outliers. Multi-dimensional scaling (MDS) was performed on GWE and GWV datasets and outliers in either dataset were excluded from analyses (Text S1).

Independent component analysis (ICA). After normalization, the distribution of each expression trait across individuals was centred and standardized. The R function *svd* was used prior to ICA to reduce the dimensionality of data and determine the optimal number of patterns to be extracted by ICA (Text S1). ICA was performed with the R *fastICA* algorithm which uses negentropy to minimize the dependency between components. The algorithm was configured using parallel extraction method and *logcosh* approximation of negentropy with α=1. To avoid trapping in a local maximum, 10 runs of the algorithm were performed and the run with the maximal negentropy was kept.

Definition of modules and enrichment analyses. The *fdrtool* R package [23] was used to define the subset of genes characterizing each signature ("module"). The statistics to which the method was applied was the entry s_{ik} of matrix S, considered as a normal score. The signature of each component was modeled as a mixture of two distributions (null and alternative). The method fits a null (Gaussian) distribution around the median of the signature distribution. A gene i was considered as

belonging to the module of the signature k if s_{ik} had a probability < 10^{-3} of being drawn under the null (FDR < 10^{-3}).

Functional annotations were made using the Gene Ontology database. Module enrichment was tested using a hypergeometric test. A threshold of 5.45 x 10^{-6} correcting for the number of categories tested was taken to declare that a category was significantly enriched in genes from a module.

Testing association of patterns with genotype. Association of the 64 patterns with the 675,350 SNPs was first tested by ANOVA with 2 d.f. using the C variance program of the GNU library TAMU_ANOVA (www.stat.tamu.edu/~aredd/tamuanova/). In this first step, a P-value < 10-7 was considered as suggestive. For suggestive SNP-pattern associations, we tested in a second step the enrichment of the module in expressions individually associated to the SNP by ANOVA at P < 10-5. For this second test, we used a hypergeometric test with a study-wise threshold of significance of 1.15 x 10-9 (Bonferroni-corrected for 64 modules x 675,350 SNPs).

Adjustment for potential contamination by non-monocytic cells. We applied an approach described elsewhere (Maouche et al, personal data). For each blood cell type (CD4+, CD8+, CD19+, CD56+, CD66b+, erythroblasts and megakaryocytes), we listed from the HaemAtlas [25] the genes reported as specific of that lineage (Table S5 from [25]). Genes were considered specific from one lineage when they were over-expressed with a fold change higher than 2 in the considered lineage compared to all others [25]. In every GHS sample, the degree of contamination by a given cell type was assessed by averaging the expression levels of the subset of the cell-specific genes in that sample. This resulted in 7 surrogate variables for contamination. All significant SNP-pattern associations were re-tested by simultaneously including these 7 variables as covariates in the regression linear model.

Replication in Cardiogenics. The population study included 363 patients with coronary artery disease recruited in Lübeck and Regensburg (Germany), Leicester (UK) and Paris (France) and 395 healthy individuals recruited in Cambridge (UK) within the Cardiogenics Consortium (http://www.cardiogenics.eu). All subjects were of European descent (Text S1). Genome-wide genotyping was carried out using the *Illumina* Sentrix Human Custom 1.2M array and the Human 610 Quad Custom array. Monocytes were isolated from whole blood using CD14 micro beads (*Miltenyi*). Gene expression profiling was performed using *Illumina* Human Ref-8 Sentrix Bead Chip arrays. Pre-processing of data and statistical analysis were performed in the R statistical environment. For the genes to be replicated, we did not apply any filtering on the level of detection. Association of gene expression with genotype was tested by analysis of variance with adjustment on age, gender and center. Analysis was performed at the probe level and the probe showing the strongest association was selected.

Weighted gene co-expression network analysis (WGCNA). WGCNA was performed on normalized expression data using the *blockwiseModules* function from the WGCNA R package (v0.92). The TOM matrix was computed from the whole set of 12,808 gene expressions (maxblocksize was set to 12,808) and all other tuning parameters were set to their default value (including dynamic tree cutting and automated merging of close modules). Module eigengenes (MEs) were computed by the *blockwiseModules* function as the first principal component of each module.

Comparison of ICA and WGCNA methods. Pairwise Pearson correlation coefficients were computed between the 64 patterns obtained by ICA and the 26 MEs obtained by WGCNA. To compare the power of the two methods to detect associations with SNPs, we tested the association of the 675,350 SNPs with the 26 WGCNA MEs by linear regression analysis

assuming an additive allele effect. For each SNP, we retained the best *P*-value over the 26 MEs and applied a Sidak correction for the number of WGCNA MEs tested. The same analysis was performed with the 64 ICA patterns. QQ plots of the 675,350 corrected *P*-values were displayed for the two methods.

GHS_Express. A downloadable SQL database compiling the results of the various associations between SNPs and expression traits is available online (http://genecanvas.ecgene.net/uploads/ForReview/). For using this database, see Methods S1 in [56].

GHS_ICA_modules. More detailed results of the analyses performed in the present study are compiled in an HTML database that is available online (http://genecanvas.ecgene.net/uploads/ForReview/). These results include correlations between patterns, module composition and enrichment, associations between SNPs and individual expression traits within modules in GHS and Cardiogenics.

Acknowledgments

We appreciate the contribution of participants of the Gutenberg Heart Study. We gratefully acknowledge the excellent medical and technical assistance of all technicians, study nurses, and coworkers involved in the Gutenberg Heart Study. We acknowledge Andreas Weith, Detlev Mennerich and Werner Rust for help during technical performance of GWV and GWE experiments and Alexandru Munteanu for assistance in informatics.

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461:747-753.
2. Cookson W, Liang L, Abecasis G, Moffatt M ,Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10:184-194.
3. Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39:1208-1216.
4. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nat Genet 39:1202-1207.
5. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39:1217-1224.
6. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452:423-428.
7. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6:e107.
8. Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, et al. (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat Genet 42:62-67.
9. Kang HM, Ye C ,Eskin E (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics 180:1909-1925.
10. Schadt EE, Zhang B ,Zhu J (2009) Advances in systems biology are enhancing our understanding of disease and moving us closer to novel disease treatments. Genetica 136:259-269.
11. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, et al. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35:57-64.

12. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, et al. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40:854-861.
13. Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, et al. (2009) Systems genetics of complex traits in Drosophila melanogaster. Nat Genet 41:299-307.
14. Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, et al. (2005) Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. Nat Genet 37:1224-1233.
15. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710-717.
16. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, et al. (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet 2:e130.
17. Biswas S, Storey JD ,Akey JM (2008) Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. BMC Bioinformatics 9:244.
18. Hyvarinen A ,Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw 13:411-430.
19. Liebermeister W (2002) Linear modes of gene expression determined by independent component analysis. Bioinformatics 18:51-60.
20. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) Genetics and beyond - The transcriptome of human monocytes and disease susceptibility. PLoS ONE 5:e10693.
21. Lee SI ,Batzoglou S (2003) Application of independent component analysis to microarrays. Genome Biol 4:R76.
22. Carpentier AS, Riva A, Tisseur P, Didier G ,Henaut A (2004) The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. Comput Biol Chem 28:3-10.
23. Strimmer K (2008) A unified approach to false discovery rate estimation. BMC Bioinformatics 9:303.
24. Lyons PA, Koukoulaki M, Hatton A, Doggett K, Woffendin HB, et al. (2007) Microarray analysis of human leucocyte subsets: the advantages of positive selection and rapid purification. BMC Genomics 8:64.
25. Watkins NA, Gusnanto A, de Bono B, De S, Miranda-Saavedra D, et al. (2009) A HaemAtlas: characterizing gene expression in differentiated human blood cells. Blood 113:e1-9.
26. Soranzo N, Spector TD, Mangino M, Kuhnel B, Rendon A, et al. (2009) A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet 41:1182-1190.
27. Sato S, Sanjo H, Takeda K, Ninomiya-Tsuji J, Yamamoto M, et al. (2005) Essential function for the kinase TAK1 in innate and adaptive immune responses. Nat Immunol 6:1087-1095.
28. Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M, et al. (2010) A conserved trans-acting regulatory locus underlies a proinflammatory gene expression network and susceptibility to autoimmune type 1 diabetes. Nature (in press).
29. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.
30. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of

type 1 diabetes. Nat Genet 39:857-864.
31. Plagnol V, Smyth DJ, Todd JA, Clayton DG (2009) Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics 10:327-334.
32. Briskin M, Winsor-Hines D, Shyjan A, Cochran N, Bloom S, et al. (1997) Human mucosal addressin cell adhesion molecule-1 is preferentially expressed in intestinal tract and associated lymphoid tissue. Am J Pathol 151:97-110.
33. Adams DH ,Eksteen B (2006) Aberrant homing of mucosal T cells and extra-intestinal manifestations of inflammatory bowel disease. Nat Rev Immunol 6:244-251.
34. Hanninen A, Taylor C, Streeter PR, Stark LS, Sarte JM, et al. (1993) Vascular addressins are induced on islet vessels during insulitis in nonobese diabetic mice and are involved in lymphoid cell binding to islet endothelium. J Clin Invest 92:2509-2515.
35. Hanninen A, Jaakkola I ,Jalkanen S (1998) Mucosal addressin is required for the development of diabetes in nonobese diabetic mice. J Immunol 160:6018-6025.
36. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 40:395-402.
37. Gudbjartsson DF, Bjornsdottir US, Halapi E, Helgadottir A, Sulem P, et al. (2009) Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat Genet 41:342-347.
38. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, et al. (2009) Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet
39. Ganesh SK, Zakai NA, van Rooij FJ, Soranzo N, Smith AV, et al. (2009) Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet 41:1191-1198.
40. Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, et al. (2009) Genome-wide association study of blood pressure and hypertension. Nat Genet
41. Velazquez L, Cheng AM, Fleming HE, Furlonger C, Vesely S, et al. (2002) Cytokine signaling and hematopoietic homeostasis are disrupted in Lnk-deficient mice. J Exp Med 195:1599-1611.
42. Kadrmas JL ,Beckerle MC (2004) The LIM domain: from the cytoskeleton to the nucleus. Nat Rev Mol Cell Biol 5:920-931.
43. Lanningham-Foster L, Green CL, Langkamp-Henken B, Davis BA, Nguyen KT, et al. (2002) Overexpression of CRIP in transgenic mice alters cytokine patterns and the immune response. Am J Physiol Endocrinol Metab 282:E1197-1203.
44. Hao J, Serohijos AW, Newton G, Tassone G, Wang Z, et al. (2008) Identification and rational redesign of peptide ligands to CRIP1, a novel biomarker for cancers. PLoS Comput Biol 4:e1000138.
45. Schulze JM, Wang AY ,Kobor MS (2009) YEATS domain proteins: a diverse family with many links to chromatin modification and transcription. Biochem Cell Biol 87:65-75.
46. Keller MP, Choi Y, Wang P, Davis DB, Rabaglia ME, et al. (2008) A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Res 18:706-716.
47. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, et al. (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452:429-435.
48. Plaisier CL, Horvath S, Huertas-Vazquez A, Cruz-Bautista I, Herrera MF, et al. (2009) A systems genetics

approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. PLoS Genet 5:e1000642.

49. Brunet JP, Tamayo P, Golub TR ,Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 101:4164-4169.
50. Grieve IC, Dickens NJ, Pravenec M, Kren V, Hubner N, et al. (2008) Genome-wide co-expression analysis in multiple tissues. PLoS One 3:e4033.
51. Friedman N, Linial M, Nachman I ,Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7:601-620.
52. Zhang XW, Yap YL, Wei D, Chen F ,Danchin A (2005) Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. Eur J Hum Genet 13:1303-1311.
53. Teschendorff AE, Journee M, Absil PA, Sepulchre R ,Caldas C (2007) Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput Biol 3:e161.
54. Lutter D, Ugocsai P, Grandl M, Orso E, Theis F, et al. (2008) Analyzing M-CSF dependent monocyte/macrophage differentiation: expression modes and meta-modes derived from an independent component analysis. BMC Bioinformatics 9:100.
55. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, et al. (2010) Cell type-specific gene expression differences in complex tissues. Nat Methods 7:287-289.
56. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z ,Clark HF (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 4:e6098.
57. Aulchenko YS, Ripke S, Isaacs A ,van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23:1294-1296.

Titre: Approches intégrées du génome et du transcriptome dans les maladies complexes humaines

Résumé: Cette thèse a pour objet l'étude du lien génotype-transcriptome et de son influence sur le développement des maladies multifactorielles. Les apports de ce travail sont à la fois méthodologiques et appliqués.

Nous étudions d'abord le lien génotype-transcriptome en établissant la liste des eQTL (expression Quantitative Trait Loci) dans le monocyte et nous évaluons l'apport de l'observation des eQTL pour l'interprétation des analyses d'association génome entier (GWAS). Nous proposons ensuite une méthode pour l'identification de variants génétiques affectant des modules de gènes co-régulés que nous appliquons à l'étude des données d'expression de monocytes issus d'une large étude populationnelle (GHS). Nous mettons ainsi en évidence plusieurs loci affectant l'expression de modules de gènes co-régulés, dont plusieurs sont impliqués dans la prédisposition au diabète de type I. Nous montrons également que le processus d'isolation des cellules monocytaires peut engendrer des biais liés à la contamination par des types cellulaires non désirés et nous proposons une approche pour contrôler ce type de biais dans les analyses.

Mot-clefs: Génomique, Transcriptome, Maladies Complexes, eQTL, Statistique, Cardiovasculaire, GWAS

Title: Integrated approaches of genome and transcriptome in the study of human complex diseases.

Abstract: This thesis deals with the study of the relation between genotype and expression and its influence on the development of complex diseases. This work brings both methodological and applied results.

First, we study the relation between genoype and transcriptome by establishing a database of eQTL (expression Quantitative Trait Loci) in monocytes and we evaluate the contribution of eQTL for the interpretation of results from Genome Wide Association Studies (GWAS).We then provide a methodology for the identification of genetic polymorphisms regulating modules of co-expressed genes that we apply to a large scale populationnal study of the monocyte transcriptome. We thus identify several loci associated with modules of co-regulated genes, several of which are involved in the susceptibility to type I diabetes. We also show that the isolation of monocytes can induce complex bias through contamination from unwanted cell types and we provide a method to control for such bias in the analysis.

Keywords: Genomics, Transcriptome, Complex Diseases, eQTL, Statistics, Cardiovascular, GWAS

ADDRESSE : INSERM UMRS 937: "Génomique cardiovasculaire"
Faculté de Médecine, 6^{eme} étage
91, Boulevard de l'hopital, 75013 PARIS

ISBN : | | | | | | | | | | |