Inserm
UMRS 937
Génomique Cardiovasculaire

UNIVERSITÉ PARIS SUD
ÉCOLE DOCTORALE 420
SANTÉ PUBLIQUE

THÈSE de doctorat
de l'Université Paris-Sud 11
Spécialité : Génétique Statistique

Présentée et soutenue publiquement par
Nicolas GRELICHE
le 18 février 2013

Stratégies de recherches de phénomènes d'interactions dans les maladies multifactorielles

Sous la direction de :
David-Alexandre TRÉGOUËT

Membres du jury :

Mme Gaëlle LELANDAIS	Institut Jacques Monod, Paris	(Rapporteur)
M David COX	Centre de recherche en Cancérologie, Lyon	(Rapporteur)
Mme Nadine ANDRIEU	Institut Curie, Paris	(Examinateur)
M Hervé SEITZ	Institut de Génétique Humaine, Montpellier	(Examinateur)
M David-Alexandre TRÉGOUËT	Inserm UMRS 937, Paris	(Directeur de Thèse)

AVANT OUVERTURE : *Conserver à température ambiante.*

APRES OUVERTURE : *Conserver au frais et consommer dans les 48h.*

Soupe délice de légumes d'antan au beaufort A.O.C., Knorr

nature
remerciements

Merci à tous !!!

Nicolas Greliche [1,2,3,4]

J'essaie dans ce papier, de remercier toutes les personnes qui m'ont aidées ou soutenues. Pour ce faire j'ai effectué un recensement exhaustif de l'ensemble de mes faits et gestes sur les trois denières années, que j'ai relié récursivement à toutes les personnes y ayant participé directement ou indirectement, en précisant leur degré d'implication. J'ai ensuite rapproché ces données avec l'évolution jour par jour de l'avancement de ma thèse et de mon taux d'heureusité. Les résultats indiquent qu'un très grand nombre de personnes ont eu une influence positive sur ma thèse ou sur ma joie de vivre ($p=10^{-1983}$). Cela inclut très probablement la personne qui est en train de lire cet abstract et que je m'empresse donc de remercier.

Entre août 2009 et décembre 2012, j'ai passé un peu plus de trois années pleines d'aventures durant lesquelles j'ai énormément appris, tant d'un point de vue professionnel que d'un point de vue humain. Mon taux d'heureusité, a rarement été au dessous de 13 (échelle allant de 18 à 4, avec 11 comme taux maximal), et a très souvent été au dessus de 8, voir 14, allant même jusqu'à atteindre 11 à 3.07 reprises. De même, l'avancement de ma thèse, a beaucoup fluctué mais n'a finalement que rarement reculé, alors qu'il a par moment atteint des rythmes très élevés, notamment durant les derniers jours. J'en profite au passage pour tirer un coup de chapeau au créateur de l'espèce humaine (ou ce que vous voulez - notons juste que s'ils sont plusieurs, la prouesse est un peu moins impressionnante), qui a pensé aux doctorants en ajoutant une option "Allez, on ne dort plus pendant deux jours !" qui marche plutôt bien. De récentes études ont montré que les facteurs humains étaient parmi les principales raisons de la variabilité du bonheur et de l'évolution professionnelle[1,2]. C'est pourquoi, comme mes prédecesseurs[3,4,5], j'ai essayé d'en savoir plus et d'identifier les personnes qui ont été déterminantes durant ma thèse, afin de leur faire un gros bécot !

Pour ce faire, 34070 de mes faits et gestes ont été recensés à 87 endroits différents entre septembre 2009 et octobre 2012, ainsi que leurs dates, heures, durées et raisons lorsque cela était possible. Étrangement, alors qu'il y a très peu de données manquantes pour les trois premières caractéristiques, il ne fut pas rare que je ne trouve aucune logique à certains comportements. Pour relier l'ensemble à toutes les personnes y étant impliquées, j'ai commencé par identifier les individus ayant soit pris part aux actions, soit servis à les effectuer. J'ai ensuite identifié les personnes reliées aux actions de ces individus et celles reliées à leurs actions et ainsi de suite jusqu'à ne plus pouvoir trouver d'origine humaine. Afin d'éviter une trop grande perte de qualité d'information due à cette récursivité, j'ai cependant choisi de limiter cette recherche aux individus nés après 1859, date choisie au hasard mais qui s'avère remarquable car en y ajoutant 11*13 - 2 = 141, elle permet d'obtenir environ 2000. Au final, un peu plus de 3 milliards de personnes de tous origines ont été intégrées dans l'analyse. Le génotypage de ces individus, un temps envisagé a finalement été abandonné car je n'ai pas pu trouver de raison pour les justifier autre que pour avoir plus de données. Finalement, l'avancement de ma thèse et mon taux d'heureusité ont été mesurés chaque jour à l'aide de la puce Illuminetoi humanSun11 beadchip. La normalisation dnPic fut utilisée pour corriger les bruits de fond de "Hey Soul Sister" ou Gaetan Roussel. Ils furent remplacés par la reprise de "Sous le sunlight des tropiques" par Joyce Jonathan & Tony. Les modèles réduit et mixtes, ajustés pour la température et le nombre de centièmes de secondes d'ensoleillement sur Paris, furent utilisés pour expliquer les deux phénotypes étudiés et j'ai fait le 23 devant Top Gun, Le Karaté Kid, Rasta Rocket et Shining lorsque le cluster était en panne. Toutes ces analyses ont été effectuées avec les logiciels Notepad++, Inkscape, Sozi, Scribus, MobaXterm, R, OpenShot, Everything, AllMynotes, pdfXchangeViewer, pdfsam ainsi que les conseils du site du zéro et de Aaron Koblin. Tous les détails sont disponibles dans la méthode supplémentaire S1 que vous ne trouverez nulle part.

Les résultats montrent qu'une grande partie des personnes de l'étude ont eu un impact soit sur l'avancement de ma thèse, soit sur mon bien-être et parfois sur les deux et ce, même après correction pour tests multiple BH (cf. tableau 1). En particulier, l'analyse par modèle structural montre que David Trégouët a été un élément déterminant dans l'avancement de ma thèse. Curieusement, il s'avère également être mon directeur de thèse, ce qui suggèrerait un lien. J'invite les chercheurs à se pencher sur le sujet. En tout cas, je remercie David de m'avoir pris comme doctorant. Je pense que je n'ai pas dû être un doctorant facile à manager, ayant parfois des idées bien arrêtées sur ce que je veux faire. Je le remercie de m'avoir aidé à mener à bien cette thèse et d'avoir pensé à moi en voyant Mathieu Kassovitz. François Cambien apparaît également comme une personne essentielle dans la genèse de ma thèse en tant que directeur du laboratoire UMRS937. Je le remercie de m'avoir accueilli dans son laboratoire. Ensuite, il y a quatre

Individu	Chiffre aléatoire	P-value corrigée	Individu	Chiffre aléatoire	P-value corrigée	Individu	Chiffre aléatoire	P-value corrigée
David	0.840373	4.67E-306	Cindy	0.051740	5.17E-143	Guillaume	0.443134	4.43E-113
François	0.946667	9.47E-218	Aurélien	0.698038	6.08E-142	Marie	0.498460	4.98E-125
Gaelle	0.295842	2.96E-218	Vanessa	0.366509	3.67E-142	Ares	0.677163	6.77E-125
David	0.950471	9.50E-218	Antoine	0.960169	9.60E-142	Marine	0.747246	7.47E-125
Nadine	0.064770	6.48E-219	Colin	0.939088	9.39E-142	Sophie	0.351338	3.51E-125
Hervé	0.495049	4.95E-218	Jérome	0.184186	1.84E-142	Senaya	0.321830	3.22E-125
Maxime	0.560066	5.60E-186	Philippe	0.447017	4.47E-142	Santy	0.815120	8.15E-125
Raphaele	0.051473	5.15E-187	Thomas	0.143908	1.43E-142	Hervé	0.686368	6.86E-125
Guillemette	0.856749	8.57E-186	Christophe	0.079836	7.98E-143	Hervé	0.110098	1.10E-125
Dylan	0.103281	1.03E-186	Romain	0.281098	2.81E-142	Dominique	0.118103	1.18E-125
Viviane	0.671502	6.72E-186	Franck	0.941167	9.41E-142	Ewa	0.199207	1.99E-125
Ricardo	0.974779	9.75E-186	Ronan	0.759677	7.60E-142	Jean-Marc	0.609595	6.10E-125
Jessica	0.277466	2.77E-186	Bruno	0.841379	8.41E-142	Nathalie	0.986703	9.87E-125
Lim	0.723242	7.23E-186	Anne	0.384301	3.84E-142	Françoise	0.361700	3.62E-125
Isabelle	0.582817	5.83E-166	Lajon	0.258137	2.58E-142	Papa	0.543215	5.43E-116
Caroline	0.643974	6.44E-166	Françoise	0.531843	5.19E-142	Maman	0.620029	6.20E-116
Antoine	0.683412	6.83E-166	Cécile	0.131547	1.32E-142	Marie	0.118030	1.18E-116
Charlotte	0.521905	5.22E-166	Claire	0.609475	6.09E-138	Pascal	0.174185	1.74E-116
Ulrike	0.625354	6.25E-166	Olivier	0.578419	5.78E-138	Aurélien	0.872869	8.73E-116
Sylvia	0.077973	7.80E-054	Marc	0.347233	3.47E-138	Delphine	0.594240	5.94E-116
Viviane	0.784192	7.84E-153	Samuel	0.726200	7.26E-138	Valentin	0.644974	6.45E-116
MarieLise	0.818618	8.19E-153	JR	0.116883	1.17E-138	Théo	0.304112	3.04E-116
Vinh	0.852878	8.53E-153	Florian	0.644667	6.45E-138	Timéo	0.067874	6.79E-117
Nadjim	0.335030	3.35E-153	Luc	0.790268	7.90E-138	Lisa	0.302894	3.03E-116
Tiphaine	0.512839	5.13E-153	Valentina	0.499497	4.99E-138	Annie	0.868600	8.69E-116
Christine	0.162493	1.62E-153	Dobril	0.933619	9.34E-138	Mathieu	0.633326	6.33E-116
Monique	0.756330	7.56E-153	Paul	0.630970	6.31E-138	Grand-Père	0.431711	4.32E-116
Fanzin	0.490008	4.90E-153	Murielle	0.192323	1.92E-138	Guilhaume	0.908429	9.08E-116
Laurence	0.331050	3.31E-153	Pierre	0.817677	8.18E-128	Arnaud	0.486848	4.87E-100
Claire	0.209920	2.09E-153	Laureinne	0.895650	8.96E-128	Florian	0.834081	8.34E-100
Carole	0.798160	7.98E-153	Manon	0.679842	6.80E-128	Nicolas	0.560281	5.60E-100

Tableau 1 : 100 premières associations significatives

[1] INSERM UMR_S 937, Paris, France, [2] Université Pierre et Marie Curie (UPMC, Paris 6), Paris, France, [3] Universite Paris-Sud (Paris 11), Paris, France, [4] Université Paris-Diderot (Paris 7), Paris, France

Received 18 December 2012; accepted 18 January 2013; published outline 18 February 2013; doi:34.1000/ng.100

autres personnes qui ressortent de mes analyse et qui ont (ou plutôt vont avoir pour le moment) eu un rôle important dans l'aboutissement de ma thèse. Il s'agit des chercheurs qui ont accepté de faire partie de mon jury de thèse. Merci beaucoup à Gaëlle Lelandais d'abord, d'avoir accepté d'être rapporteur malgré un domaine un peu différent du mien... quoique l'homme a peut-être emprunté un peu d'ADN aux levures en mangeant des gâteaux ou en buvant de la bière, non ? Merci ensuite à David Cox, qui a accepté de venir de Lyon pour être rapporteur de ma thèse, surtout que ça fait classe d'avoir un tel nom sur sa page de couverture, qu'on fasse de la statistique ou non. Merci aussi à Nadine Andrieu et à Hervé Seitz qui vient lui de la ville du nouveau Champion de France de Ligue 1.

Outre David et François, il apparaît d'après les résultats des GEE que plusieurs autres personnes du laboratoire ont eu une influence positive sur ma thèse. Merci donc à Dominique, Hervé, Ewa plein d'tomates, Bio-wonderwoman Sophie, Electromagnet-Christine qui j'espère, réussira à maîtriser son pouvoir, Nathalie, Jean-Marc, Madame Marine Germain Lambert, Badrédine, Elmout-trouve pas d'appart-Ulrike, le gang des miss congélo : Laurence, Carole et Claire avec qui j'ai passé de très bons moments à discuter congélateur, Nadjim qui a une table, Ares qui ne sera bientôt plus espagnole mais... française (oui soyons sérieux, la catalogne ne peut quand même pas être un pays), Méthylman Dylan grand copain d'allergie qui prend le relais pour embêter David, Henri c'est toi le Hen - non moi c'est ri, maman tarte à la crème Guitoud et Vinh qui vient de manger un truc du frigo qui n'est pas a lui mais que je pardonne parce qu'il a bien relu ma thèse... et parce que ce qu'il vient de manger n'est sûrement pas à moi.

Si l'on regarde bien la figure 1 (ce n'est pas forcément évident à première vue), on se rend compte que de nombreux anciens du labo ont aussi contribuer à ce que ma thèse se passe plutôt bien. Vous pourrez en situer quelques uns sur la figure 2 représentant les gens qui m'ont accompagné dans le bureau. Je remercie donc Marie-Lise que j'ai remplacée pour embêter David, Viviane, Monique passe-partout, DrDr Guillemette qui m'a beaucoup aidé à barbujiller contrairement au lapin. Je la remercie particulièrement pour son soutien durant la fin de ma thèse et pour m'avoir nourri. Je remercie aussi Maria, Sylvia, Lynda, Rajaï, Soraya (en petit pour pas me faire taper), Tiphaine oudort plus beaucoup, Sonia Karabatikina, Sonia Lisandro Lopez, Marie Bretonne, Dr Raph à qui je prouverai bientôt une bonne fois pour toute le non-sens de son dessert préféré. Je la remercie notamment de m'avoir appris à discuter avec tout le monde au monoprix ce qui me permet de faire attendre ceux qui viennent avec moi. Je remercie aussi Dr Big-Boss-Master-Statman-Max, grand vainqueur du concours de longévité à mes côtés dans le bureau (voir figure 2) et qui fut toujours partant pour me suivre dans mes conneries (et inversement). Merci à Chili con Ricardo le nettoyeur, Farzin Benzebaby-foot, Brown-Cheese Linn et Jessica O'Comon Broccoli (merci pour tout Jess !!!!), sans oublier Cedric, Sana, Emilien et les super stagiaires Caroline, Isabelle, Antoine, Charlotte, Santy, Bathilde et Hélène. N'hésitez pas à passer me voir... mais n'oubliez pas les pains au chocolat cette fois.

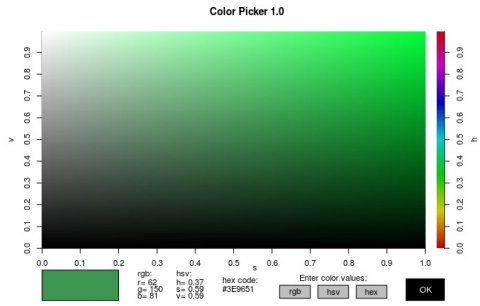


Figure 1 : Application R colorPicker 1.0

Les résultats les plus significatifs de l'analyse haplotypique, qui fut par ailleurs particulièrement difficile à réaliser en absence de données génotypique, furent un temps exclus pour cause d'écarts à l'équilibre d'Hardy-Weinberg. Une investigation plus poussée montra cependant que j'ai décidé de simplement ne pas me préoccuper de ce déséquilibre. Il en ressortit le résultat fort peu attendu d'un rôle majeur de mes parents dans mon bien-être personnel et professionnel. Aussi, je les remercie à fond de me laisser faire mes conneries et de continuer à me soutenir quoi qu'il arrive. Vous remarquerez que j'ai fait bien attention à ne pas dire la phrase bateau où je remercie mes parents sans qui je ne serai rien... mince. L'étude stratifiée des données de corrélations par le modèle de Cox-Simpsons (classe hein ! C'est parce qu'il y a Cox dedans.) m'incite aussi à faire un grand merci à l'ensemble de ma famille. Je remercie en particulier mon frère et ma soeur qui ont fait en sorte que l'on atteigne le nombre nécessaire de 11 joueurs pour faire une équipe de football (Valentin, Téo, Timéo et Axel se rajoutent en effet à Papou, Luc, Mathieu, Pascal, Pidane, Aurélien, Papitou et moi... oui, ça fait 12 mais il faut bien des remplaçants) Je leur propose pour la suite, de continuer à mettre l'accent sur les garçons pour permettre au plus vite l'organisation de matchs à 11 contre 11. Il ne me paraît en effet pas nécessaire de renforcer le club couture, déjà bien fourni (Mamido, Kiki, Annie, Marie, Nanou, Delphine, Miflo, Alice) et qui ne devrait pas avoir de mal à nous concocter de beaux petits maillots pour la prochaine saison. J'en profite aussi pour remercier les autres équipes et notamment celles de Lajon et Marielle, en

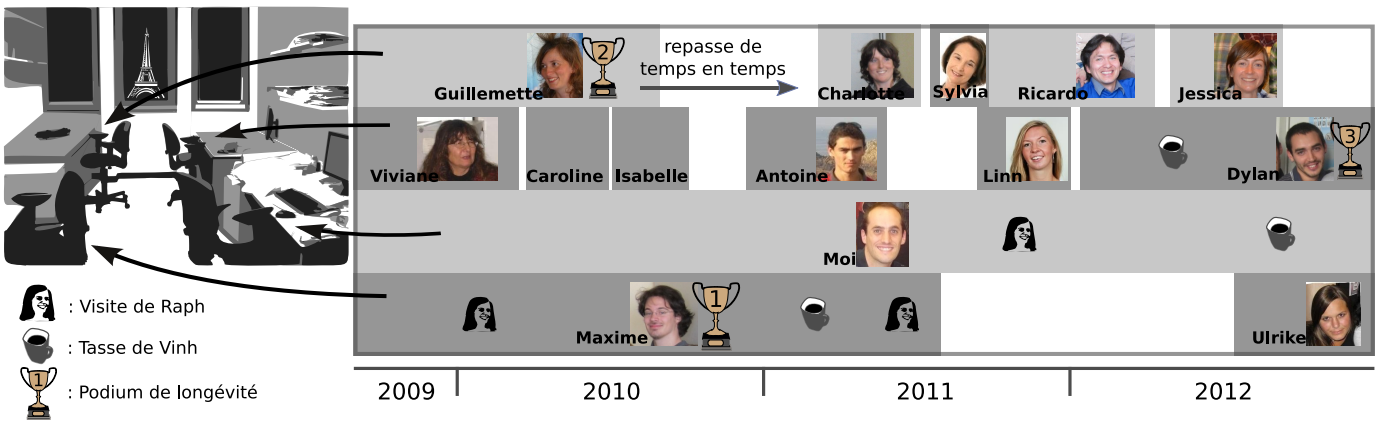


Figure 2 : Apperçu des différentes personnes qui ont passé plus d'une semaine dans le bureau pendant ma thèse

espérant les rencontrer bientôt. J'ai enfin une pensée pour mon grand-père qui vient de prendre sa retraite et qui peut désormais voir les match d'un peu plus haut. J'espère qu'il pensera à nous préparer des mini saucisses (demande à mami si tu ne sais pas faire) pour quand on viendra (Je lui conseille cependant des les mettre au congélateur pour l'instant parce qu'on ne va pas venir de suite).

En ce qui concerne l'analyse par Bootstrap-Jacknife, elle m'a permise de réaliser l'importance de mes potes de l'Ensai et leurs associé(e)s dans l'excellent déroulement de ces trois années. Un énorme merci donc à Arnaud et Guillaume, les colloc' forever, Brasil-Philippe, Florian la benz, Cindy ala..., Aurélien frappe de mule tant qu'il l'a pas mangée, Baby-Foot Vaness, Runner Coolin Colin, Antoine qui devrait bientôt faire sa crémaillère, Julien, Franck, Ronan, Thomas, Christophe, Romain, Jérome et Math-discussions de tarés dans le train pour Rennes-ieu.

Même si ils ont plus eu tendance à avoir un rôle en amont de ma thèse, je remercie aussi Olivier le trentenaire, Marc qui apprécierait le seuil de significativité qui suit, Sam G. qui doit passer le code, Flo qui soutient le même jour que moi !, JR qui va bientôt nous pondre le nouveau Another World et Sam A., qui s'est lui aussi lancé dans la galère thésardienne. Tous ressortent au seuil FDR de 10^{-69}.

Tous ces remerciements proviennent principalement de la recherche d'associations avec mon épanouissement personnel. Pour ce qui est de l'aspect plus professionnel, il a fallut que je m'adonne à l'utilisation de modèles mixte, aléatoires, multiniveaux, random, hiérarchiques, nested, à effets aléatoires, en split-plot. Il en a résulté ma gratitude envers deux clusters de personnes. Le premier, composé de Valentina Moskvina, Dobril Ivanov et Paul Buckland, que Bathilde a rejoint depuis et qui m'ont introduit au domaine de l'épidémiologie génétique. Le second est lui composé des professeurs et chargés de TD de biostatistique de Paris 7 et notamment Bruno Toupance et Anne Badel, qui m'ont fait découvrir les joies de l'enseignement.

Merci aussi à Pazu, Luffy, Onizuka, Shu, Sangoku, Vincent A. F., Ralph W., et bien d'autres qui m'ont parfois filé des bons coups de boost. Liste complète non disponible sur demande.

Pour finir, je tiens à remercier la personne qui est en train de lire ces lignes car elle a probablement eu un rôle même indirect dans ma thèse. Je tiens à ce qu'elle réalise l'énorme quantité de travail qui a été effectuée pour réaliser ces remerciements. Après avoir bien réfléchi à leur design pour limiter les éventuels problèmes d'interprétation, j'ai dû planifier et organiser la collecte et le stockage de trois années de données. J'ai dû nettoyer et filtrer ces données avant de pouvoir les analyser pour finalement en tirer les principaux résultats que je viens de vous exposer. De part la nature du travail, il m'a bien entendu été impossible de déléguer ces remerciements, si bien que j'espère que le lecteur m'excusera d'avoir quelque peu bâclé le reste de ce document faute de temps.

ACKNOWLEDGMENTS

Comme le but de ce papier n'est pas de pratiquer une inception, je vais éviter de remplir cette partie pour ne pas risquer de me perdre dans les limbes des remerciements. Je remercierai simplement Jean Bouyer et Audrey Bourgeois que je n'ai pas pu remercier auparavant.

COMPETING FINANCIAL INTERESTS

L'auteur déclare ne pas trop savoir comment il pourrait avoir des conflits d'intérêts avec cette publication. Il précise aussi qu'il ne le dirait de toute façon pas s'il en avait.

1. Nicolas Greliche, *Tous les gens sont gentils* (2008).
2. Nicolas Greliche, *Il y a peut-être des gens qui sont un peu moins gentils, mais ils ont leurs raisons* (2010)
3. Guillemette Antoni, *Identification de facteurs génétiques modulant deux phénotypes intermédiaires de la maladie thrombo-embolique veineuse : les taux de facteurs VIII et von Willebrand : Intérêt de l'utilisation de différentes approches de recherche pangénomique* (2012)
4. Raphaële Castagné, *Expression des gènes du chromosome X chez l'homme : approche intégrée par génomique et transcriptomique à haut-débit* (2011)
5. Maxime Rotival, *Approches intégrées du génome et du transcriptome dans les maladies complexes humaines* (2011)

... et plein d'autres trucs qui n'existent pas forcément et que vous ne lirez de toute façon jamais.

Merci et bonne chance à vous !

Sheldon L Cooper[1], Quinn R Mallory[2], Eleanor A Arroway[3], Emmett L Brown[4], Samuel Beckett[5]

Grâce à nos avancées sur la théorie des cordes, nous sommes parvenus à mettre au point une nouvelle machine multi-tâche, permettant de voyager dans le temps, dans l'espace ainsi que dans toutes les dimensions et ce, à près de 88 miles à l'heure. Malheureusement, elle n'a pour l'instant qu'une capacité de cinq places et vu que nous pouvons confirmer maintenant que les Mayas avaient raison (pour ceux qui auraient déjà oublié la phrase précédente, nous avons créé une machine à voyager un peu partout et notamment dans le temps), nous avons souhaité remercier les gens de cette planète de leur companie, avant qu'il ne périssent pour la nuit des temps... ainsi que leur souhaiter bonne chance, au cas où.

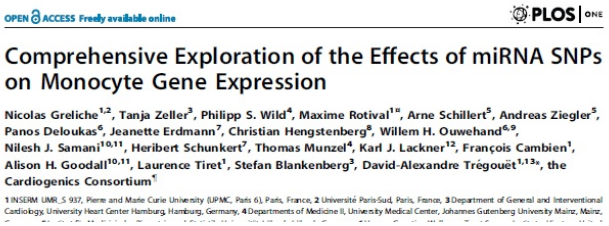
Les travaux de Minus et Cortex[1] sur les lasers cosmico-reducteurs avaient montré en 1998 la faisabilité du voyage gallaxio-temporel à bord de gruyères quantiques. Ce n'est cependant qu'en 2001, que Malcolm et Dewey[2], se basant sur les résultats des professeurs Shadoko[3] et Tournesol[4] (surtout du second en fait) ont pu adapter le concept au champ octo-dimensionnel, pour aboutir à la machine de Turing-Cox. Après avoir amélioré le matrice de passage intrafusionnelle[5], nous avons réussi en 2007 à naviguer hors du champ snikersien, vers Fantasia et Laputa[6,7] avant qu'en 2008, Will Hunting et al.[8] utilisent de la poudre d'azote liquide pour atteindre Santa Destroy. Finalement, nous venons de déchiffrer, grâce à la découverte du Boson de Higgs[9], la question ultime associée à la réponse 42 du sens de la vie. Cette question est la suivante :

[1]California Institute of Technology, Pasadena, California, USA. [2]California University, Palo Alto, California, USA. [3]Center for Search for Extraterrestrial Intelligence (SETI), Arecibo Observatory, Puerto Rico, USA. [4]Institute of Future Technology, Hill Valley, California, USA. [5]US Government, Project Quantum Leap, Stallion's Gate, New Mexico, USA.

Received 21 October 2015; accepted 26 October 1985; published online 18 February 2013; doi:05.1955/ng.11

Du même auteur

Greliche, N., Zeller, T., Wild, P. S., Rotival, M., Schillert, A., Ziegler, A., Deloukas, P., et al. (2012). *Comprehensive Exploration of the Effects of miRNA SNPs on Monocyte Gene Expression.* PloS one, 7(9)


OPEN ACCESS Freely available online PLOS ONE

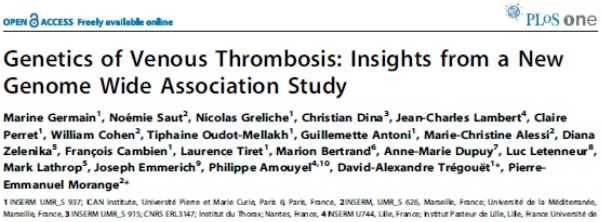
Comprehensive Exploration of the Effects of miRNA SNPs on Monocyte Gene Expression

Nicolas Greliche[1,2], Tanja Zeller[3], Philipp S. Wild[4], Maxime Rotival[1¤], Arne Schillert[5], Andreas Ziegler[5], Panos Deloukas[6], Jeanette Erdmann[7], Christian Hengstenberg[8], Willem H. Ouwehand[6,9], Nilesh J. Samani[10,11], Heribert Schunkert[7], Thomas Munzel[4], Karl J. Lackner[12], François Cambien[1], Alison H. Goodall[10,11], Laurence Tiret[1], Stefan Blankenberg[3], David-Alexandre Trégouët[1,13]*, the Cardiogenics Consortium[¶]

1 INSERM UMR_S 937, Pierre and Marie Curie University (UPMC, Paris 6), Paris, France, 2 Université Paris-Sud, Paris, France, 3 Department of General and Interventional Cardiology, University Heart Center Hamburg, Hamburg, Germany, 4 Departments of Medicine II, University Medical Center, Johannes Gutenberg University Mainz, Mainz,

Greliche, N. (2012). *Introduire des concepts statistiques en faisant appel à l'intuition.* Troisième colloque francophone international sur l'enseignement de la statistique (CFIES). Angers.

Germain, M., Saut, N., **Greliche, N.**, Dina, C., Lambert, J.-C., Perret, C., Cohen, W., et al. (2011). *Genetics of venous thrombosis: insights from a new genome wide association study.* PloS one, 6(9)


OPEN ACCESS Freely available online PLOS one

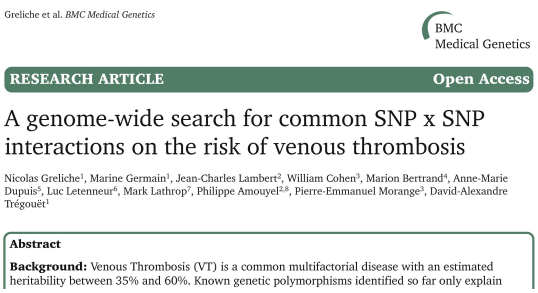
Genetics of Venous Thrombosis: Insights from a New Genome Wide Association Study

Marine Germain[1], Noémie Saut[2], Nicolas Greliche[1], Christian Dina[3], Jean-Charles Lambert[4], Claire Perret[1], William Cohen[2], Tiphaine Oudot-Mellakh[1], Guillemette Antoni[1], Marie-Christine Alessi[2], Diana Zelenika[5], François Cambien[1], Laurence Tiret[1], Marion Bertrand[6], Anne-Marie Dupuy[7], Luc Letenneur[8], Mark Lathrop[5], Joseph Emmerich[9], Philippe Amouyel[4,10], David-Alexandre Trégouët[1]*, Pierre-Emmanuel Morange[2]*

1 INSERM UMR_S 937; ICAN Institute, Université Pierre et Marie Curie, Paris 6, Paris, France, 2 INSERM, UMR_S 626, Marseille, France; Université de la Méditerranée, Marseille, France, 3 INSERM UMR_S 915; CNRS ERL3147; Institut du Thorax; Nantes, France, 4 INSERM U744, Lille, France; Institut Pasteur de Lille, Lille, France Université de

En cours de révision...

Greliche, N., Germain, M., Lambert, J.-C., Cohen, W., Bertrand, M., Dupuis, A.-M., Letenneur, L., et al. (soumis). *A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis.*


Greliche et al. *BMC Medical Genetics*

BMC Medical Genetics

RESEARCH ARTICLE Open Access

A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis

Nicolas Greliche[1], Marine Germain[1], Jean-Charles Lambert[2], William Cohen[3], Marion Bertrand[4], Anne-Marie Dupuis[5], Luc Letenneur[6], Mark Lathrop[7], Philippe Amouyel[2,8], Pierre-Emmanuel Morange[3], David-Alexandre Trégouët[1]

Abstract

Background: Venous Thrombosis (VT) is a common multifactorial disease with an estimated heritability between 35% and 60%. Known genetic polymorphisms identified so far only explain

A toi qui t'apprêtes à me lire... ou plus probablement à me feuilleter.

Avant-Propos

Durant mes trois années de thèse, j'ai eu le plaisir, en plus de mon travail de recherche, d'encadrer un projet statistique à l'Ensai, mon ancienne école, et d'effectuer une mission complémentaire d'enseignement. C'est ce que l'on appelait autrefois le « monitorat ». Cette mission permet à tout thésard d'arrondir un peu ses fins de mois en lui proposant une première expérience d'enseignement. Dans mon cas, cette mission a consisté à donner des TD et des TP de statistique à des étudiants de deuxième et troisième année de licence de biologie.

Dans le même temps, j'ai aussi beaucoup réfléchi. J'ai réfléchi au monde, à la science, à la statistique, à l'enseignement, aux gens... un peu à tout en y réfléchissant bien et ces réflexions m'ont emmené à considérer l'enseignement et la pédagogie bien haut dans la hiérarchie de mes priorités. À une époque où l'on parle beaucoup d'efficacité en recherche, je suis convaincu que pour faire avancer la science, mais aussi pour le simple bien de notre société, les chercheurs ont tout intérêt à « perdre » un peu de temps à effectuer de gros efforts de pédagogie envers leurs étudiants, la société ou même les autres chercheurs lors de la présentation de leurs résultats... mais je m'égare.

Cette mission d'enseignement et ces différentes réflexions ont été déterminantes dans l'orientation que j'ai pu donner à mon travail de recherche. Elles l'ont impacté par moment positivement, mais aussi parfois négativement en ce sens qu'elles ont usé de mon temps et de ma motivation au grand dam de mon directeur de thèse. Je pense en fait avec un peu de recul qu'elles font partie intégrante de mon travail de thèse et c'est la raison pour laquelle j'ai souhaité essayer de les intégrer dans ce document.

Ainsi, c'est délibérément que ce manuscrit de thèse, tout en essayant de ne pas dévier de son objectif principal, à savoir exposer le travail de trois années de recherches, est emprunt d'une tentative d'être abordable au novice, voire parfois ludique.

Ce document se décompose en les chapitres suivants :

1. Le fonctionnement du vivant
2. La variabilité génétique
3. L'épidémiologie génétique
4. Les tests statistiques
5. La gestion des tests multiples
6. Les données épidémiologiques utilisées
7. À la recherche de phénomènes d'interactions dans la maladie thromboembolique veineuse
8. Cap sur la recherche de polymorphismes liés aux microARNs
9. Discussions et perspectives

Dans les trois premiers chapitres, j'introduis les concepts biologiques (chapitre 1 et 2) et expose le contexte scientifique (chapitre 3) qui me semblent nécessaires à la compréhension du document. Les trois chapitres suivants introduisent les méthodes statistiques (chapitres 4 et 5) et les études (chapitre 6) que j'ai utilisées dans mes travaux de recherche. Les chapitres 7 et 8 présentent les résultats de ces travaux, enfin, dans le dernier chapitre, je discute ces résultats et propose quelques perspectives à mon travail de thèse.

J'ai essayé tout au long du document de garder une certaine homogénéité dans la construction de mes graphiques, notamment dans les couleurs et le fléchage. Vous pouvez commencer à vous y familiariser en jetant un oeil à la figure 1 qui vise en particulier à vous expliquer les significations des différentes flèches que vous rencontrerez bientôt.

En espérant ne pas choquer le lecteur averti, habitué aux thèses plus ardues, je vous souhaite mesdames, messieurs et autres, une excellente lecture !

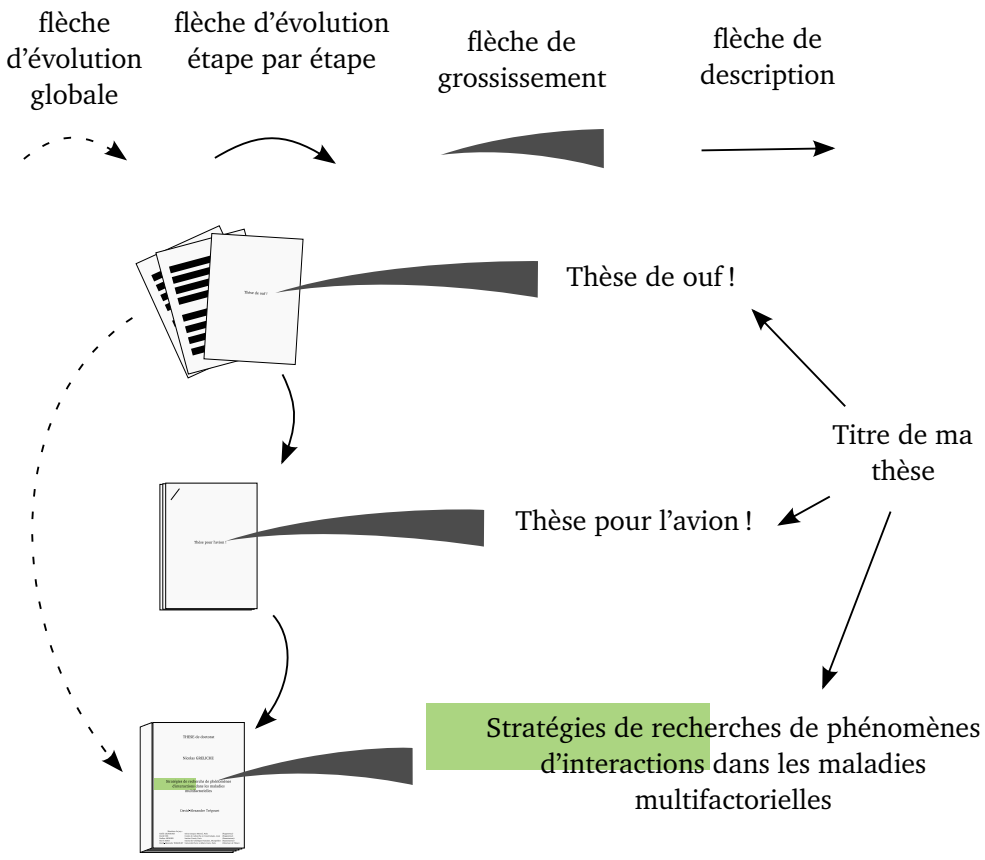


FIGURE 1 – Règle adoptée dans le fléchage des graphiques. Illustration par l'exemple de l'évolution du titre de ma thèse.

Table des matières

Chapitre 1

Le fonctionnement du vivant

Et voici la viiiie... la belle vie toute pressée d'éclooore.

Il était une fois... la vie (Générique)
http://youtu.be/m0pUKsMJYao

Le but de ce chapitre est d'introduire brièvement le fonctionnement de base du vivant dont l'élément essentiel est l'ADN, une grande molécule qui contient les instructions pour la production et la régulation de la production des protéines.

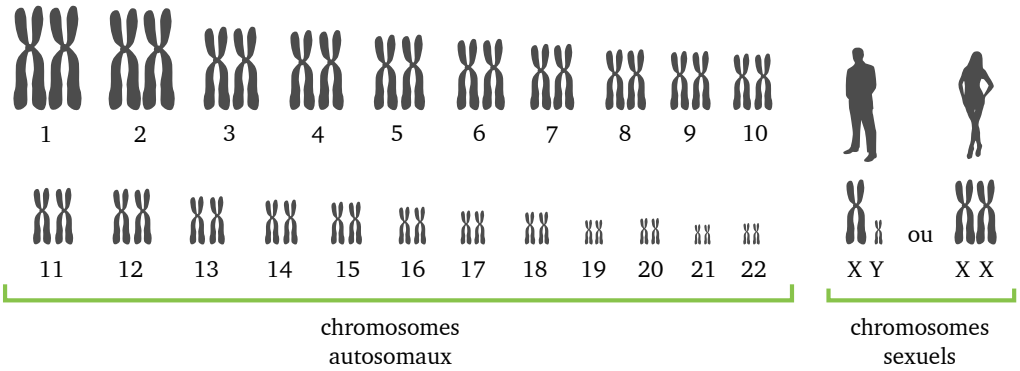
1.1 L'ADN est à la base de chaque cellule vivante

1.1.1 Structure de l'ADN

Tous les êtres vivants que nous connaissons sont constitués de cellules [1] et celles-ci ont toujours la même structure fondamentale leur permettant d'être la plus petite unité autonome et capable de se reproduire. En particulier, les hommes ont des dizaines de milliers de milliards de cellules [113], chacune renfermant un noyau [2], dans lequel réside 23 paires de chromosomes (22 paires de chromosomes autosomaux, et une paire de chromosomes sexuels, cf. figure 1.1). Selon la phase du cycle cellulaire à laquelle se trouve la cellule, ces chromosomes sont formés d'une unique ou de deux identiques immenses molécules d'acide désoxyribonucléique (ADN) enroulées à de multiples niveaux. La structure de chacune de ces molécules d'ADN est identique à savoir qu'elle consiste en une double hélice composée de

1. Certains organismes ne sont cependant constitués que d'une seule cellule
2. En réalité, il existe des cellules très spécialisées comme les globules rouges, qui ont perdu leur noyau

deux brins antiparallèles et complémentaires de nucléotides, où un nucléotide est lui-même constitué d'une base azotée, d'un sucre et d'un groupement phosphate. La



Source : Genome Reference Consortium, Assembly GRCh37.p10

FIGURE 1.1 – Les 23 paires de chromosomes de notre génome, représentées de manière à ce que la taille des chromosomes soit proportionnelle à la longueur de leur séquence

complémentarité des deux brins se fait au niveau des bases azotées (on parlera alors de paire de bases) alors que les groupements phosphates et les sucres permettent l'enchaînement des nucléotides de ces brins (cf. figure 1.2).

1.1.2 Ses bases azotées contiennent les instructions pour la fabrication des protéines

L'ensemble de nos chromosomes sont présents par paire, un provenant du père, l'autre de la mère. En tout, nous possédons deux copies d'environ trois milliards de paires de bases azotées. Chaque base azotée existe en quatre versions : l'adénine (A), la thymine (T), la guanine (G) et la cytosine (C), dont les bases complémentaires sont respectivement T, A, C et G. La séquence d'ADN, c'est-à-dire l'ordre dans lequel ces quatre différentes bases azotées s'enchaînent dans l'ADN forme ce que l'on appelle le génome. Il est identique pour toutes les cellules [1] et fournit les instructions de fabrication des protéines, les molécules qui assurent la plupart des fonctions nécessaires à la vie cellulaire.

1. A quelques variations près, car en réalité chaque molécule d'ADN peut se distinguer légèrement des autres si elle subit des erreurs de copie, des modifications de certaines bases ou de petits réarrangements entre certaines séquences

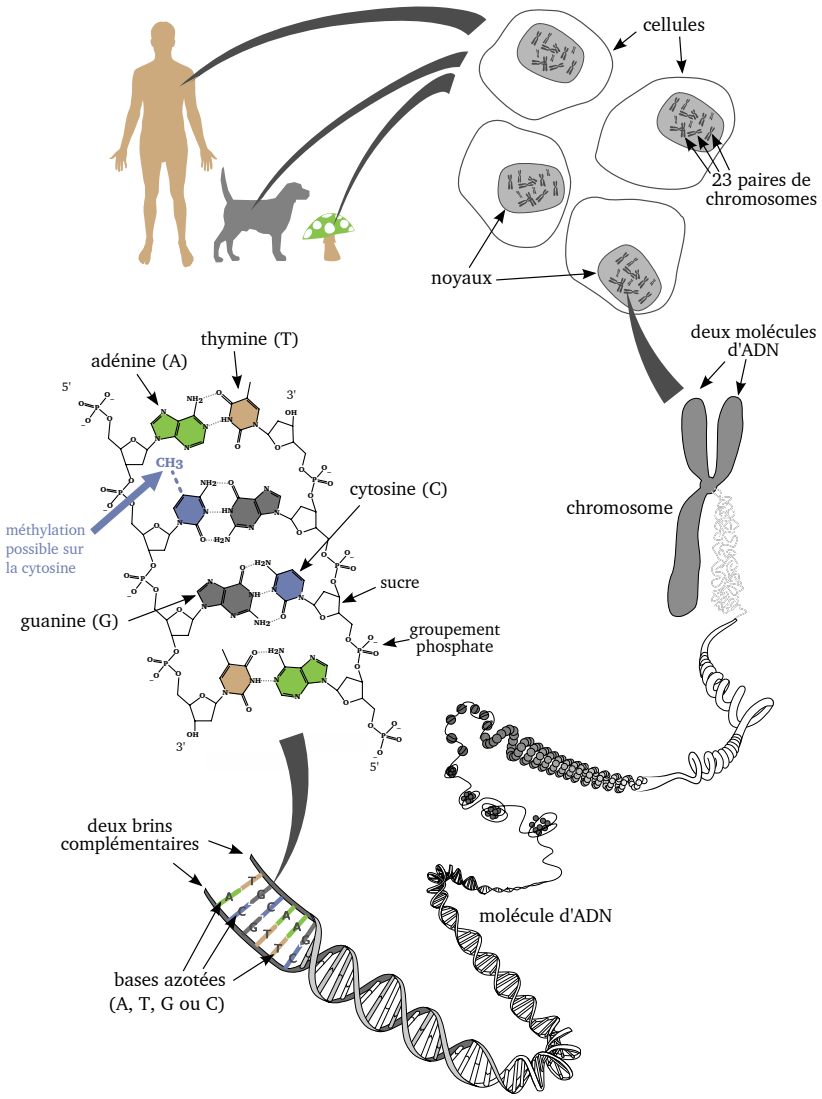


FIGURE 1.2 – L'ADN réside au sein de chacune de nos cellules

1.2 Des gènes aux protéines

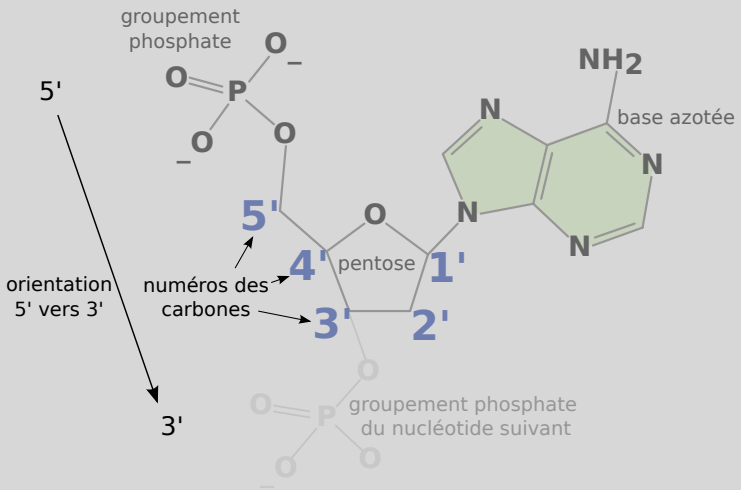
1.2.1 Les gènes

En fait, la séquence génétique permettant aux cellules de savoir comment fabriquer les protéines ne correspond pas à l'ensemble de la séquence d'ADN de notre génome, mais seulement à certaines portions appelées gènes. Pour procéder à la fabrication d'une protéine, le brin correspondant à la séquence d'un gène est copié par complémentarité de ses bases azotées en acide ribonucléique (ARN) à l'intérieur du noyau, lors de ce que l'on appelle la transcription. Ce processus est initié à l'extrémité du gène, dans la région appelée promotrice, sur laquelle peut se fixer l'ARN polymérase, un complexe composé de plusieurs protéines, qui se chargera

de copier l'ADN en ARN. Par ailleurs, la configuration de la molécule d'ADN n'étant pas symétrique (voir figure 1.2), les deux brins d'ADN sont orientés. Ils le sont en sens inverse l'un de l'autre et c'est cette orientation qui détermine entre autres, le sens de copie de l'ADN. Il est d'usage de décrire une séquence dans son orientation 5' vers 3' (voir encadré).

Extrémités 5' et 3'

Les extrémités 5' et 3' font référence aux carbones des sucres de l'ADN ou l'ARN. Chaque nucléotide est composé d'un sucre ayant cinq atomes de carbone (pentose). Par convention, ces atomes sont numérotés de 1 à 5 de sorte que la base azotée se lie au carbone 1 du pentose alors que le groupement phosphate est relié à l'atome 5. La séquence d'ADN ou d'ARN provient de la succession de nucléotides où chaque nucléotide est relié par son groupement phosphate, au carbone 3 du pentose du nucléotide qui le précède. Cette convention est importante, car les brins d'ADN et d'ARN sont orientés. Ils ne peuvent être synthétisés que dans le sens 5' vers 3' et il en est de même pour la traduction en protéine.



1.2.2 L'ARN messager

L'ARN est également une molécule constituée d'un enchaînement de nucléotides, mais contrairement à l'ADN, sa structure est simple brin et elle comporte une base azotée différente : la thymine de l'ADN est remplacée par l'uracile (U) dans l'ARN [1]. L'ARN transcrit à partir des gènes n'est pas conservé tel quel tout au long de sa vie,

1. L'ARN se différencie aussi de l'ADN par la substitution d'un atome d'hydrogène par un groupement hydroxyle en position 2' du sucre

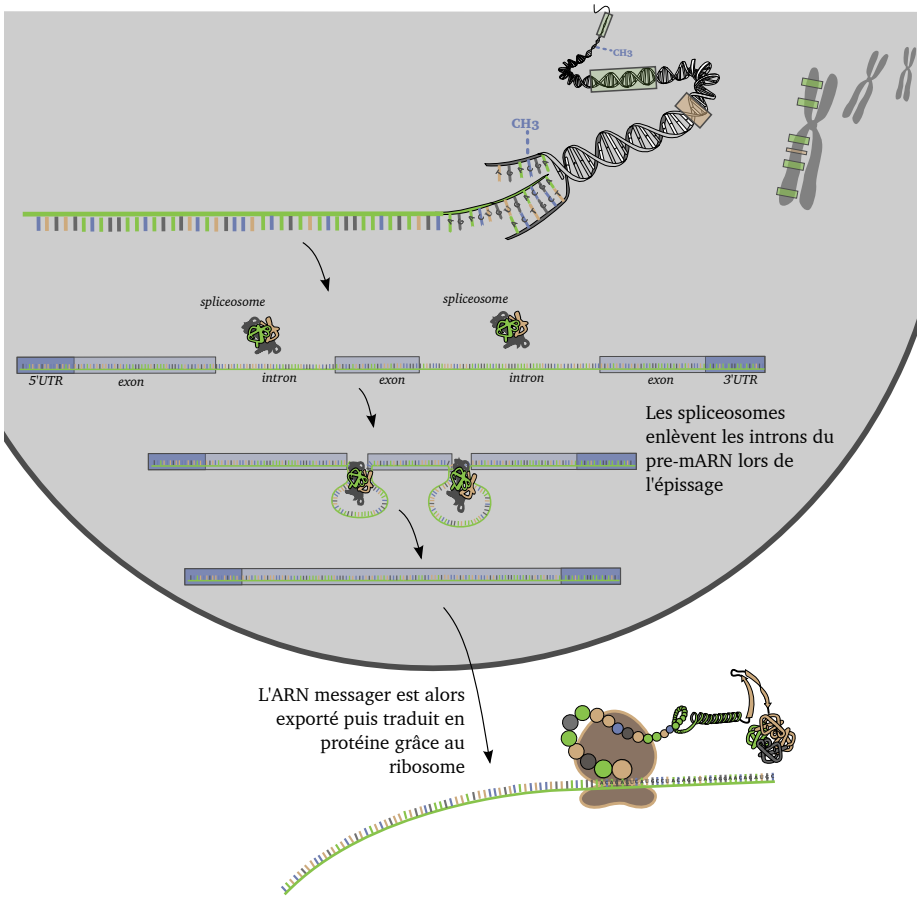


FIGURE 1.3 – Processus de maturation de l'ARN messager

mais subit des suppressions de certaines parties de ses séquences, les introns, lors de l'épissage (cf. figure 1.3). Ces suppressions peuvent varier d'un ARN à l'autre donnant lieu à des épissages dits alternatifs. L'ARN résultant de la transcription s'appelle ARN primaire (preARN) alors que celui issu de l'épissage s'appelle l'ARN mature. A la fin de l'épissage, l'ARN mature est composé de trois régions principales : La région non traduite située à l'extrémité 5' de l'ARN (5'UTR, pour 5' UnTranslated Region), la région codante, située au milieu [1] et la région non traduite située à l'extrémité 3' de l'ARN (3'UTR). Les deux régions 5'UTR et 3'UTR sont des éléments clés de la régulation de l'expression du gène.

1.2.3 Les protéines

Cet ARN mature est ensuite transporté à l'extérieur du noyau de la cellule où il fournira le mode d'emploi de fabrication de la protéine lors de ce que l'on appelle la traduction (cf. figure 1.4). L'ARN mature contient alors la séquence codante des

1. La région codante commence par le triplet de nucléotides ATG et se termine par un autre triplet appelé codon stop

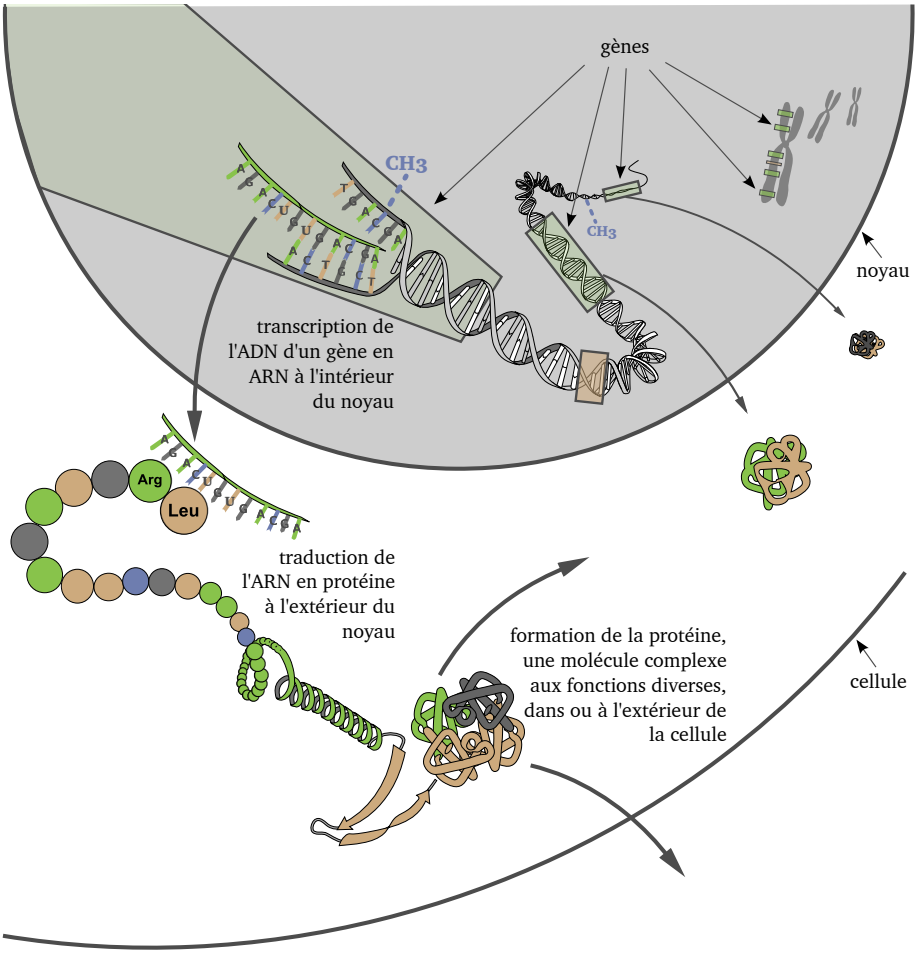


FIGURE 1.4 – L'ADN est transcrit en ARN qui est lui-même traduit en protéine

protéines. Ce processus s'effectue grâce au code génétique universel car identique chez la grande majorité des espèces vivantes et qui, à chaque triplet de nucléotides de l'ARN, associe un acide aminé. La succession des triplets présents dans l'ARN permet l'agrégation de centaines d'acides aminés pour former une protéine, une molécule très complexe qui peut avoir des fonctions très diverses dans la cellule ou en dehors. Par exemple, l'actine participe à la structure de notre corps, les histones permettent la compaction de l'ADN alors que les enzymes augmentent ou réduisent les vitesses des réactions chimiques de notre organisme.

1.3 Les microARNs : des régulateurs de la production de protéines

1.3.1 Rôle des microARNs

Il existe aussi dans notre génome des séquences d'ADN qui ne codent pas pour des protéines mais qui sont toutefois transcrites en ARN. Parmi ces ARNs, on trouve les ARN ribosomiques (ARNr), les ARN de transfert (ARNt), les petits ARN nucléolaires (snoARN), les petits ARN nucléaires (ARNsn) ou encore, ceux qui vont nous intéresser par la suite, les microARNs. Les microARNs sont une importante famille de petits ARNs, longs de 18 à 25 nucléotides, simple brin qui régulent l'expression des gènes après la transcription, en orientant la fixation d'un complexe protéique appelé RISC (RNA-induces silencing complex) vers une séquence d'ARN d'un gène codant qui est complètement ou partiellement complémentaire à la séquence du microARN. Le plus souvent, la séquence d'ARN ciblée par le microARN se situe dans la région 3'UTR de cet ARN cible, mais parfois elle peut aussi se trouver dans sa partie 5'UTR ou dans sa phase ouverte de lecture (ou ORF pour Open Reading Frame), c'est-à-dire dans des séquences potentiellement codantes pour des protéines. Si la complémentarité entre le microARN et l'ARN cible est parfaite, la fixation du complexe donne lieu à un clivage endonucléolytique ayant en général pour conséquence de dégrader fortement l'ARN ciblé. Si la complémentarité est partielle, le complexe RISC n'a tendance à dégrader que partiellement l'ARN, par une réaction exonucléolytique, mais permet en général d'empêcher la traduction de l'ARN en protéine. Dans les deux cas, la production de la protéine est réduite par l'action du microARN.

1.3.2 Importance des microARNs

Le premier microARN identifié, lin-4, a été découvert en 1993 chez le ver Caenorhabditis elegans [66]. Depuis, on a trouvé des microARNs chez la plupart des eucaryotes [8] et notamment chez l'homme où ils forment une des classes de petits ARNs inhibiteurs les plus importantes avec 1600 membres identifiés au moment de l'écriture de ce document, d'après miRBase, la base de registre des microARNs [43]. L'influence globale des microARNs sur notre organisme n'est pas encore bien connue, mais grâce à des algorithmes de prédictions (miRanda [56], TargetScan [70], Diana MicroT [76], PicTar [63]), on pense que la plupart des microARNs pourraient cibler des centaines de gènes et qu'inversement, chacun de ces gènes serait susceptible d'avoir des sites de fixation pour plusieurs microARNs. Au final, on estime que 50 % de nos gènes seraient sujets à une régulation par les microARNs [8, 38, 64], ce qui ferait de ces derniers les éléments régulateurs clés de la vie de la cellule.

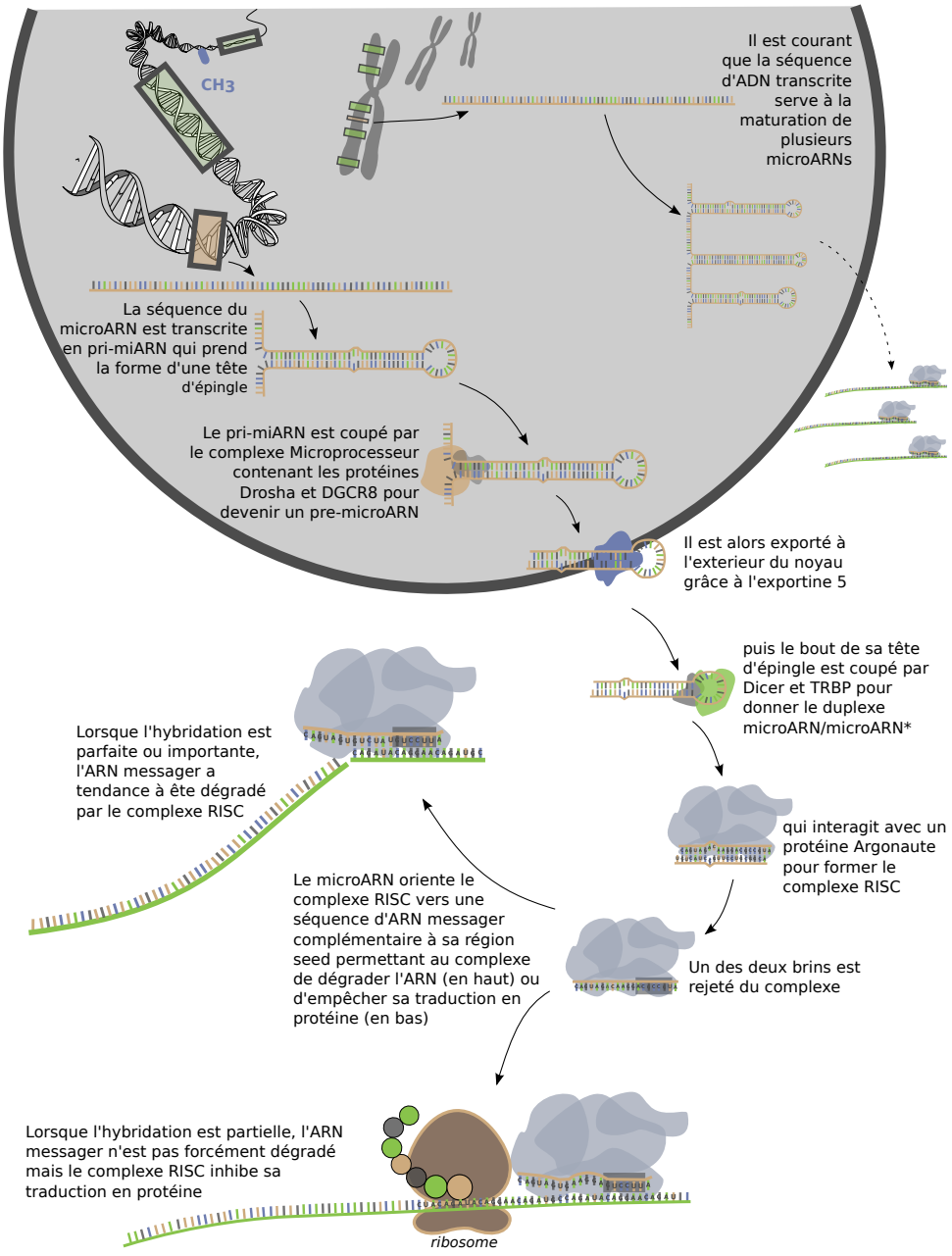


FIGURE 1.5 – Processus de maturation et mécanisme d'action du microARN

1.3.3 Processus de maturation du microARN

Depuis quelques années, notre compréhension de la biogenèse des microARNs a beaucoup progressé. On sait désormais que les microARNs proviennent de petites gènes non codants situées soit à l'extérieur des gènes codants, soit dans leur partie intronique. La séquence du microARN est d'abord transcrite en pri-microARN qui, après avoir pris la forme d'une tête d'épingle, est coupé par le complexe protéique Microprocessor, notamment composé de l'enzyme Drosha et de la protéine DGCR8,

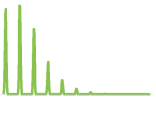
pour former une séquence de nucléotides double brins que l'on nomme pre-microARN. Notons qu'il n'est pas rare que la séquence du microARN serve en fait à la maturation de plusieurs microARNs. Dans un tel cas, la séquence est transcrite en un seul pri-microARN qui est par la suite divisé en plusieurs pre-microARNs par l'action des protéines Drosha et DGCR8. Le pre-microARN est alors exporté à l'extérieur du noyau dans le cytoplasme par les protéines Exportin-5 et RAN, puis coupé de nouveau, par l'enzyme Dicer et la protéine TRBP en le duplexe microARN/microARN* composé de deux séquences de nucléotides complémentaires d'environ 20 bases chacune. Un des deux brins du duplex interagit ensuite avec une protéine de la famille Argonaute, pour former le complexe RISC dans lequel le microARN désormais mature peut orienter la fixation du complexe vers une séquence d'ARN d'un gène codant qui lui est complémentaire (cf. figure 1.5). Il arrive que les deux brins du duplex microARN/microARN* puissent cibler des séquences d'ARNs, aussi on les distingue généralement en étoilant le nom de la version la moins couramment rencontrée [1]. Le microARN, lorsqu'il est intégré dans le complexe RISC, va cibler des séquences d'ARN qui sont complémentaires avec les nucléotides 2 à 7 de son extrémité 5', appelée région « seed ». Après hybridation du microARN sur la séquence cible, le complexe RISC va alors participer au processus de régulation de la production de protéines dans la cellule, en dégradant la séquence d'ARN messager, en particulier lors d'une complémentarité parfaite, ou en empêchant sa traduction en protéine.

1.4 Ce que renferme notre ADN

1.4.1 Notre génome en chiffres

Notre génome est constitué d'environ 23 000 gènes longs de quelques centaines à plusieurs centaines de milliers de paires de bases [53] codant pour un nombre probablement bien plus important de protéines grâce aux épissages alternatifs. Mais au final, la partie codante de tous ces gènes ne représente qu'environ 1.5 % du génome, le reste de notre séquence étant composé d'introns (ce qui est supprimé par l'épissage), de séquences qui codent pour des ARNs non traduits comme les microARNs, de séquences participants au recrutement des différents acteurs du processus de fabrication et de régulation des protéines et enfin d'ADN, constitué majoritairement de séquences répétitives, qui étaient il y a peu appelées « poubelle » mais qui semblent finalement avoir bien des fonctions [118].

1. La définition des versions étoilées et non étoilées peut parfois être complexe du fait de rapports d'abondance entre miARN et miARN* variables entres les tissus [96]



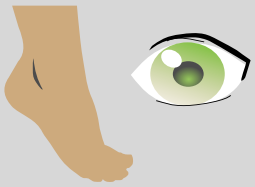
1.4.2 Les dernières nouvelles du génome

Très récemment, en septembre 2012, les chercheurs du projet Encode (pour Encyclopedia of DNA Elements), dont l'objectif est de trouver et déterminer la fonction de tous les éléments fonctionnels du génome humain, ont publié une série d'articles donnant un meilleur aperçu global de notre génome. L'une de leurs découvertes est qu'au-delà des séquences codantes pour des protéines, près de 80 % du génome humain serait finalement fonctionnel, notamment en participant à la régulation différentielle des quantités de protéines produites selon le type cellulaire [13]. D'après les résultats de leurs recherches, des séquences régulatrices pour un gène dans un type cellulaire, pourraient chevaucher des séquences régulatrices pour un autre gène dans un autre type cellulaire, ce qui les amène à proposer une redéfinition du concept de gène [30].

1.4.3 La régulation de la production de protéines dépend du type cellulaire

Les résultats du projet Encode ne sont pas si surprenants que ça. En effet, l'ADN est le même dans chaque cellule d'un individu mais, à certains endroits du corps, nous avons des doigts, à d'autres le c$\frac{1}{2}$ur ou le cerveau et ceux-ci ne se ressemblent pas du tout. Ceci est dû au fait qu'avec le temps et au contact d'environnements cellulaires et extérieurs différents, nos cellules se sont différenciées. Aussi, certains types cellulaires régulent l'expression de certains gènes permettant de fabriquer un grand nombre de certaines protéines, alors que d'autres types au contraire inhibent leur expression permettant de mieux répondre aux besoins de la cellule et de son environnement (cf. figure 1.6). Cette régulation différentielle selon les types cellulaires n'est pas encore bien comprise mais il est probable qu'au-delà des gènes et des microARNs, une grande partie du génome y soit sollicitée, de même que certains phénomènes épigénétiques, c'est-à-dire des événements qui ne sont pas codés par la séquence d'ADN mais qui peuvent cependant se transmettre. Le principal exemple est celui la méthylation consistant en des modifications de conformation de la molécule d'ADN lorsque des groupements méthyles se fixent sur certaines bases azotées de type cystéine.

Différentes protéines pour différents types cellulaires



Toutes les cellules du pied ont exactement la même séquence d'ADN que les cellules de l'oeil[a]. Ce qui change, ce sont les quantités de protéines produites à partir de cette même séquence d'ADN. C'est cela qui permet au pied d'avoir une fonction différente de celle de l'oeil.

a. En réalité, comme dit précédemment, il peut y avoir de petites variations

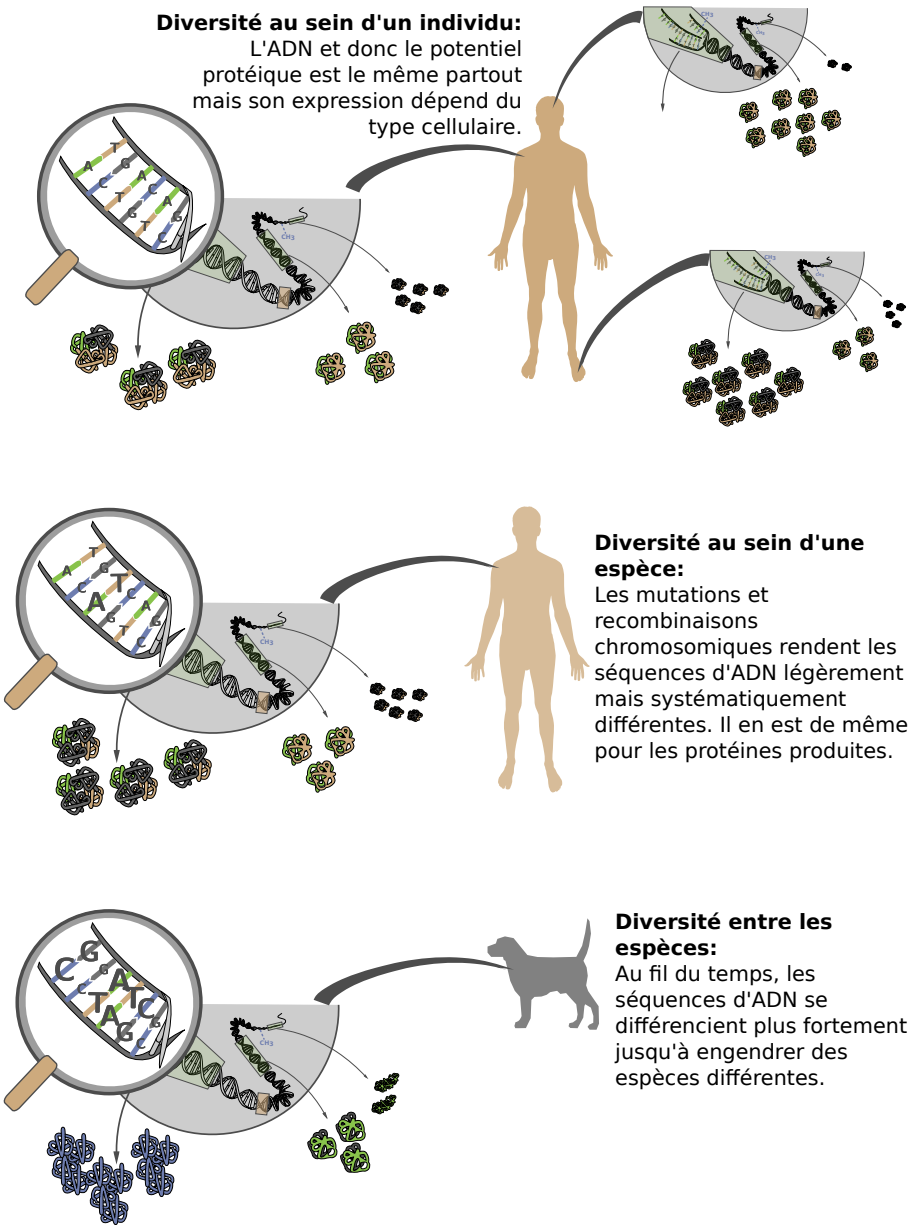


FIGURE 1.6 – La diversité du vivant

Chapitre 2

La variabilité génétique

Tout avantage a ses inconvénients et réciproquement.

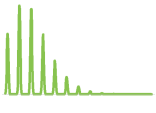
Devise Shadok

Dans le chapitre précédent, nous avons vu que l'ADN est l'élément clé de la vie, car il contient toutes les instructions dont les cellules vivantes ont besoin pour fabriquer les bonnes quantités de chaque protéine. Dans ce chapitre, nous allons voir que les variations au sein de cet ADN sont à l'origine de la diversité du vivant, du mécanisme d'évolution, mais aussi de certaines maladies.

2.1 Les sources de variabilité génétique

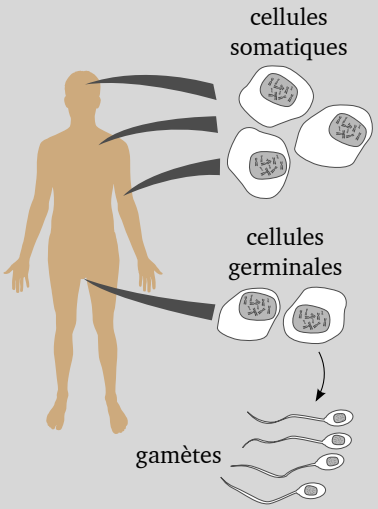
2.1.1 Les mutations

Lors de la division cellulaire, mécanisme permettant le développement de l'individu ou le renouvellement de ses cellules, il arrive parfois que l'ADN ne soit pas copié parfaitement et que certaines bases soient transformées, insérées ou omises. C'est ce que l'on appelle des mutations. Lorsqu'une mutation survient au sein d'un chromosome d'une cellule germinale (voir encadré) d'un individu, celle-ci peut alors se transmettre à sa descendance par la transmission de ce chromosome. Comme chaque individu transmet la moitié de son matériel génétique à sa descendance (l'autre moitié venant de son partenaire), l'enfant qui hérite d'une mutation a une chance sur deux de la transmettre à son tour et c'est ainsi que les mutations peuvent se propager de génération en génération.



Cellules germinales

Les cellules germinales sont les cellules qui permettent la formation des gamètes, telles que les spermatozoïdes et les ovocytes. Ce sont les cellules susceptibles d'être transmises à l'embryon lors de la reproduction. Inversement, le matériel génétique des cellules dites somatiques ne sera jamais transmis à la descendance d'un individu. L'immense majorité des cellules d'un individu sont somatiques.



La fréquence des mutations est relativement faible. On estime que le taux de mutations par paire de bases et par génération est d'environ 2.5×10^{-8} [85]. Ce taux varie cependant suivant l'endroit du génome et l'exposition à certains événements environnementaux. L'activité cérébrale pourrait par ailleurs modifier les motifs de méthylation des individus [44], alors que ceux-ci semblent corrélés à certaines instabilités au sein du génome [72, 79]. Ces éléments suggèrent qu'il ne serait pas impossible qu'un individu puisse agir sur la fréquence de mutation de son ADN, menant pourquoi pas à l'idée que les individus puissent influer et accélérer l'évolution de leur espèce [40].

2.1.2 Les recombinaisons chromosomiques

Lors de la méiose avant la formation des gamètes, les deux versions de chaque paire de chromosomes peuvent se mélanger et s'échanger du matériel génétique pour former de nouveaux chromosomes parfaitement uniques. C'est ce que l'on appelle les recombinaisons chromosomiques. Les mutations et recombinaisons sont les deux sources de la variabilité de notre génome (cf. figure 2.1).

2.2 Les conséquences de cette variabilité génétique

2.2.1 Des individus uniques

Cette variabilité génétique permet à chaque individu d'avoir une séquence d'ADN qui lui est propre et ainsi des protéines et traits physiques uniques, notamment lorsque ces différences apparaissent au sein des gènes. Chez l'homme, deux individus

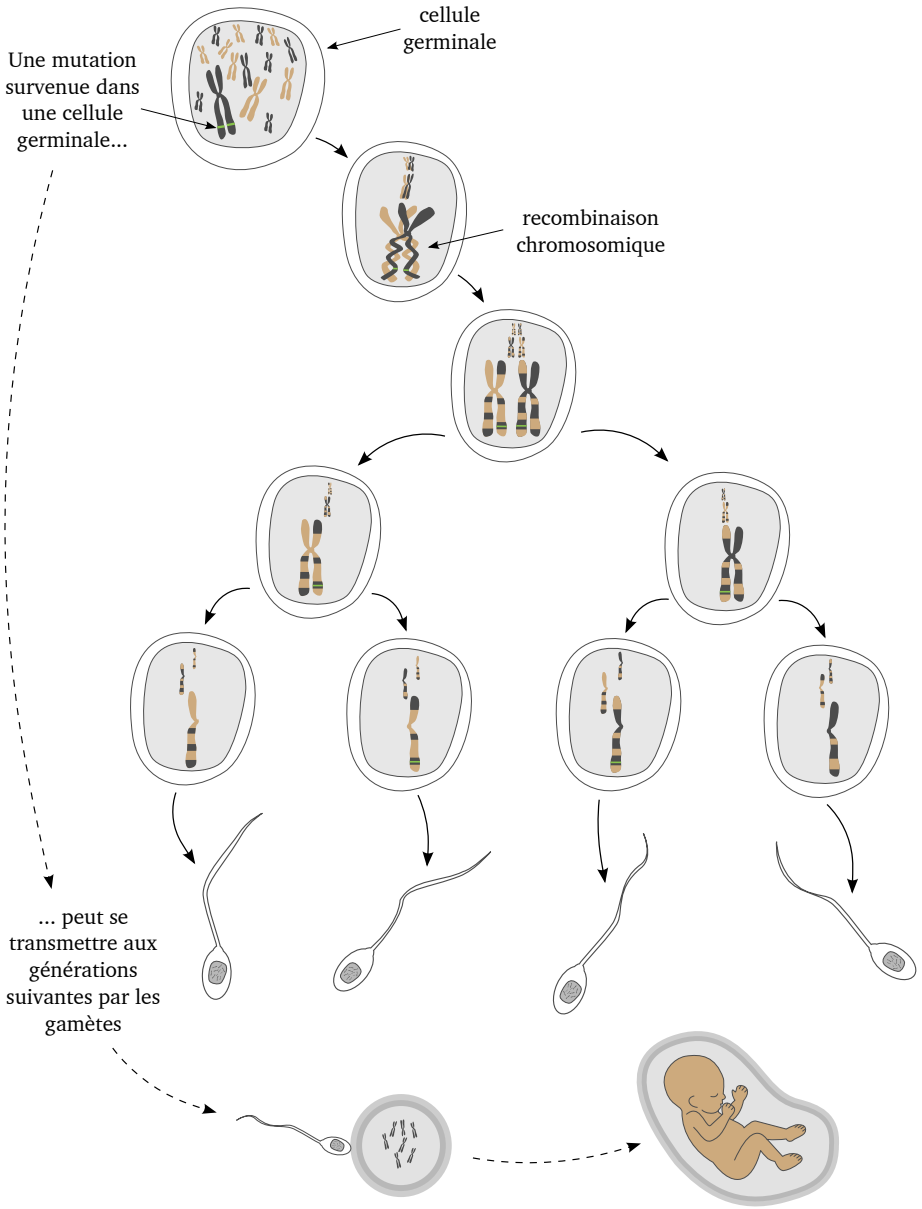
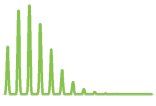


FIGURE 2.1 – Aperçu des différentes étapes de la méiose. Grâce aux mutations qui surviennent dans les cellules germinales et aux recombinaisons chromosomiques qui s'opèrent lors de la méiose, le matériel génétique de chaque gamète et donc de chaque individu devient unique.

ont environ 99.9 % de leurs séquences d'ADN en commun [124]. Ceci représente un pourcentage de similarité important mais toutes ces similitudes laissent tout de même des différences sur plusieurs millions de paires de bases qui participent à la diversité que l'on peut observer au sein de notre espèce telle que les différences de couleurs, de silhouettes ou d'aptitudes. Cette variabilité génétique n'est pas la seule responsable de nos différences. Elle agit de concert avec l'environnement extérieur dont l'influence lui est parfois bien supérieure.



2.2.2 Le mécanisme de l'évolution

Lorsque plusieurs populations éloignées sont soumises à des environnements différents, les mutations et recombinaisons qui donnent à certaines populations un avantage en termes de survie ont tendance à se conserver plus que dans une autre population où elles peuvent s'avérer néfastes. L'accumulation au cours de milliers d'années de ces modifications aboutit à une différentiation importante du génome entre les populations allant jusqu'à l'incompatibilité sexuelle et la création de nouvelles espèces. C'est le principe de l'évolution introduit par Charles Darwin dans *On the Origin of Species*[27]. Notre classification des espèces vivantes passe d'ailleurs depuis la seconde moitié du XXeme siècle par une analyse dite « phylogénétique » de reconstruction de la différenciation des gènes ou des expressions [1] des gènes [68] au cours du temps.

2.2.3 Les maladies génétiques

L'héritabilité

Grâce à des études sur des familles ou des jumeaux, en observant des phénotypes [2] plus semblables pour les individus génétiquement plus proches, on a pu découvrir qu'un certain nombre de maladies avaient une part d'origine génétique, provenant de certaines mutations ou recombinaisons.

Ces études permettent notamment d'estimer la part de la variabilité de la maladie qui est due à la génétique par rapport à celle qui est due à l'environnement extérieur comme l'alimentation, la pollution, ou les virus. Cette part des facteurs génétiques dans la variabilité d'un trait phénotypique s'appelle l'héritabilité. La figure 2.2 donne les estimations de l'héritabilité de quelques maladies ou traits communs.

Les causes biologiques

D'après le processus de synthèse des protéines que nous avons vu dans le premier chapitre, il n'est pas surprenant que les mutations et recombinaisons puissent aussi entraîner la survenue de certaines maladies. En effet, si par exemple, une mutation survient dans la séquence codante d'un gène et que cette mutation entraîne la modification d'un ou de plusieurs acides aminés lors de la traduction, il est probable que la formation de la protéine soit affectée. Ceci peut engendrer la survenue d'une maladie si la protéine est non fonctionnelle alors qu'elle est nécessaire à l'organisme (cf. figure 2.3).

1. L'expression d'un gène est l'ensemble de ce qui est produit par une cellule à partir de la séquence de ce gène. Par la suite, j'utiliserai quelque peu abusivement ce terme pour désigner la quantité d'ARN produite par un gène dans un type cellulaire donné.
2. Un phénotype est l'état d'un individu en ce qui concerne un caractère observable

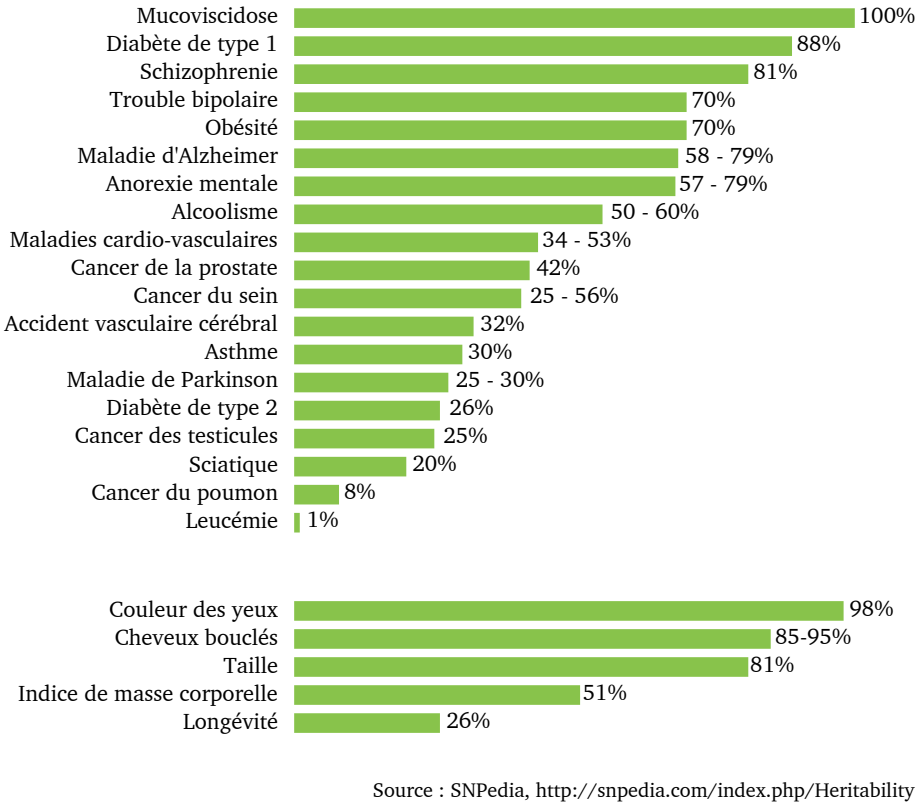


FIGURE 2.2 – Estimations de l'héritabilité de quelques traits communs ou pathologiques

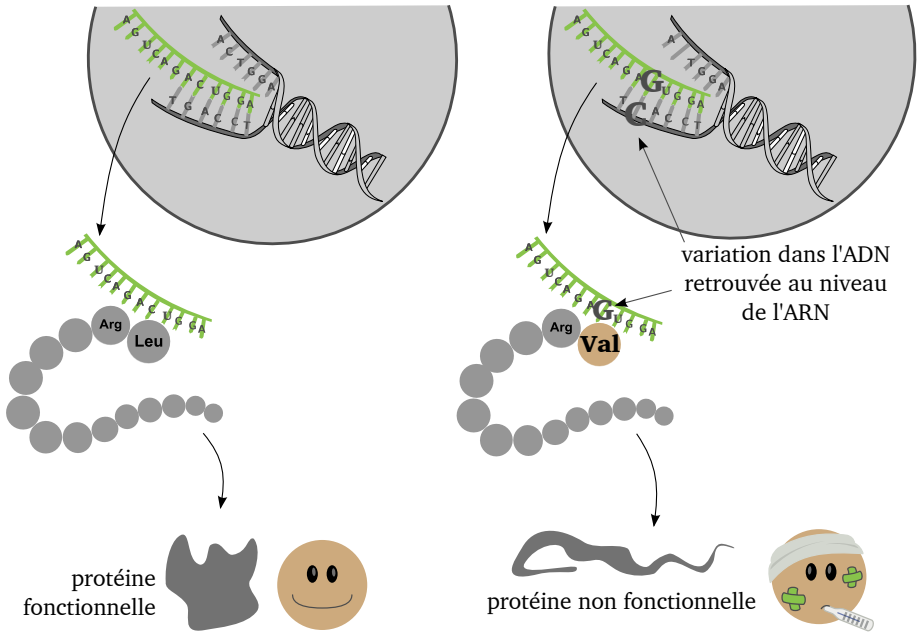


FIGURE 2.3 – Une mutation dans la séquence codante peut engendrer la formation d'une protéine non fonctionnelle causant l'apparition d'une maladie

2.3 Définitions et caractéristiques liées à la variabilité génétique

2.3.1 Quelques définitions

Lorsque la séquence d'ADN à un endroit du génome (que l'on appelle un locus) peut prendre plusieurs formes au sein d'une population, on appelle la diversité en à cet endroit un polymorphisme génétique. Les différentes formes qu'il peut prendre sont appelées des allèles. La forme la plus couramment rencontrée est l'allèle majeur et celle la moins fréquente, l'allèle mineur. Chaque individu ayant deux copies de chaque chromosome autosomal, il possède aussi deux versions de chaque polymorphisme. Si ces deux versions (ou allèles) sont identiques on dira qu'il est homozygote pour ce polymorphisme, si elles sont différentes, on dira qu'il est hétérozygote. L'ensemble des deux allèles d'un individu pour un polymorphisme donné représente son génotype. Enfin, si l'on considère plusieurs polymorphismes, l'ensemble des allèles situés sur un même chromosome d'un individu est l'un de ses deux haplotypes pour ces polymorphismes (cf. figure 2.4).

2.3.2 Substitution d'une base par une autre

Le plus souvent, une variation génétique consiste en la simple substitution d'un nucléotide par un autre. Lorsqu'une variation de ce type est présent au sein d'une population, les individus de cette population se retrouvent à avoir plusieurs formes possibles (parmi A, C, G ou T) pour le nucléotide situé au locus de la variation. Comme les mutations sont un phénomène très peu fréquent [1], il est extrêmement rare que deux mutations surviennent exactement à la même position. Les polymorphismes les plus couramment rencontrés sont donc des variations d'une seule paire de bases ne prenant que deux formes et appelées SNP (pour Single Nucleotide Polymorphism).

2.3.3 Insertions, délétions et répétitions de bases nucléotidiques.

Parfois, une variation génétique peut consister en la suppression ou l'addition d'un ou de plusieurs nucléotides. On parlera alors d'insertion et de délétion. Lors de la recombinaison chromosomique, l'échange du matériel génétique entre les deux chromosomes d'une même paire s'effectue au niveau de séquences similaires. Aussi, il n'est pas rare qu'en des endroits du génome constitués de séquences répétées, les recombinaisons ne s'effectuent pas exactement aux mêmes locus [2]

1. C'est le grand nombre de paires de bases de notre génome qui fait qu'un taux de mutations même faible permet au final d'observer un relativement grand nombre de différences entre les individus.

2. Le pluriel de locus est loci en latin comme me le faisait remarquer mon directeur de thèse. Cependant, j'ai pris parti ici de suivre les suggestion d'Albert Jacquard[54] estimant qu'étant adopté

sur les deux chromosomes résultant en des insertions et délétions des séquences répétées. Il en résulte des variations du nombre de copies de ces séquences répétées au sein de la population. On appelle CNV (pour Copy Number Variation) ce type de polymorphisme.

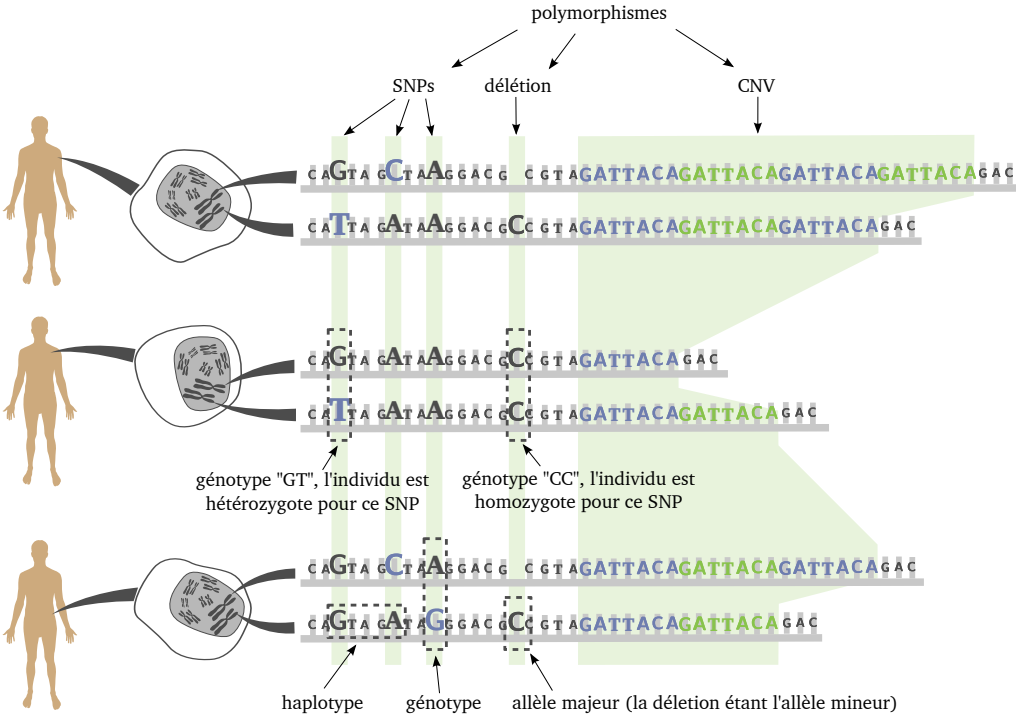


FIGURE 2.4 – Visualisation de quelques termes liés à la variabilité génétique

2.3.4 Le déséquilibre de liaison

Mutations et recombinaisons créent le déséquilibre de liaison

Si les mutations étaient le seul phénomène expliquant la variabilité de notre génome, lorsqu'une mutation apparaîtrait au sein du génome d'un individu, le chromosome contenant la mutation serait transmis de génération en génération inchangé (excepté pour les rares mutations nouvellement apparues) et la mutation serait alors intimement liée à l'ensemble des autres mutations de ce chromosome. Autrement dit, tous les individus qui auraient un certain allèle pour un polymorphisme d'un de leurs chromosomes auraient de grandes chances d'avoir également les mêmes allèles pour les autres polymorphismes de ce chromosome. On appelle cette liaison entre les polymorphismes, le déséquilibre de liaison. Cependant, chez l'homme, il n'y a que sur le petit génome mitochondrial (voir encadré) que les

par la langue française, on peut tout à fait appliquer au mot locus, les règles de cette langue, d'où mon emploi de « locus » au pluriel.

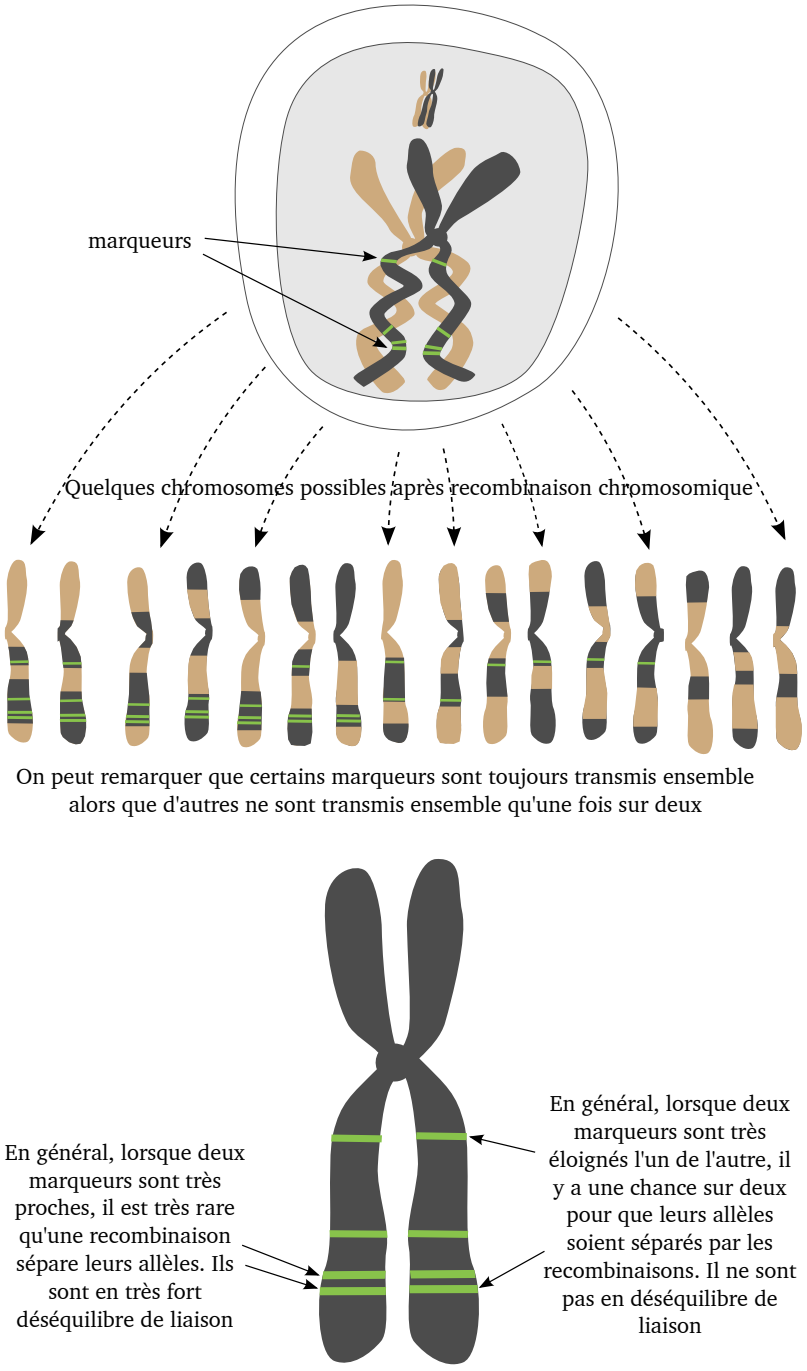


FIGURE 2.5 – Le déséquilibre de liaison

mutations sont l'unique mécanisme de variabilité. Sur le génome nucléaire (celui dont on parle dans tout ce document), les mutations s'accompagnent de recombinaisons chromosomiques. Comme les recombinaisons s'effectuent aléatoirement à certains endroit des chromosomes, si deux polymorphismes sont éloignés l'un de l'autre, il y a plus de chances qu'une ou plusieurs recombinaisons surviennent entre eux,

ce qui aura pour conséquence sur la population générale, de réduire fortement le déséquilibre de liaison entre ces deux polymorphismes. À l'inverse, si deux polymorphismes sont très proches l'un de l'autre sur un chromosome, alors il n'y a que peu de recombinaisons possibles qui permettent le réarrangement de leurs allèles entre les deux chromosomes de la paire considérée et le déséquilibre de liaison entre ces deux polymorphismes restera très fort pendant longtemps dans la population. Ce phénomène est résumé sur la figure 2.5.

Génome mitochondrial

Le génome mitochondrial humain est un petit génome transmis uniquement par la mère et constitué de 16 569 paires de bases. Il ne contient que quelques dizaines de gènes, mais est très utilisé pour son caractère peu variable du fait de l'absence de mécanisme de recombinaison génétique. Il est par exemple très utilisé dans les recherches des ancêtres communs entre les espèces ou au sein de l'espèce humaine ainsi que dans l'identification de suspects dans les enquêtes policières.

Mesure du déséquilibre de liaison

Il y a plusieurs façons de mesurer le déséquilibre de liaison entre deux polymorphismes [29]. Afin de les introduire, commençons par noter f(x) la fréquence de x dans la population (où x est un allèle ou un haplotype). Considérons ensuite deux locus bi-alléliques A et B :

- le locus A peut avoir l'allèle A_1 avec une fréquence $f(A_1)$ ou l'allèle A_2 avec une fréquence $f(A_2) = 1 - f(A_1)$.
- le locus B peut avoir l'allèle B_1 avec une fréquence $f(B_1)$ ou l'allèle B_2 avec une fréquence $f(B_2) = 1 - f(B_1)$.

Lorsque les deux locus sont indépendants, les fréquences de chaque haplotype sont simplement les produits des fréquences de chaque allèle :

$$f(A_1B_1) = f(A_1)f(B_1)$$

$$f(A_1B_2) = f(A_1)f(B_2) = f(A_1)(1 - f(B_1))$$

$$f(A_2B_1) = f(A_2)f(B_1) = (1 - f(A_1))f(B_1)$$

$$f(A_2B_2) = f(A_2)f(B_2) = (1 - f(A_1))(1 - f(B_1))$$

Alors, il est directement visible d'après le tableau 2.6, qu'une déviation D de la fréquence d'un de ces haplotypes détermine complètement les déviations des fréquences des autres haplotypes :

$$\begin{aligned} D &= f(A_1B_1) - f(A_1)f(B_1) \\ &= -f(A_1B_2) + f(A_1)f(B_2) \\ &= -f(A_2B_1) + f(A_2)f(B_1) \\ &= f(A_2B_2) - f(A_2)f(B_2) \end{aligned}$$

	A_1	A_2	Total
B_1	$f(A_1)f(B_1)$ + D	$f(A_2)f(B_1)$ - D	$f(B_1)$
B_2	$f(A_1)f(B_2)$ - D	$f(A_2)f(B_2)$ + D	$f(B_2)$
Total	$f(A_1)$	$f(A_2)$	1

FIGURE 2.6 – Fréquences alléliques et haplotypiques

Pour visualiser le lien entre cette mesure D et la recombinaison chromosomique, considérons que le polymorphisme au locus B est plus récent qu'au locus A. Au moment de sa naissance, il existe deux allèles A_1 et A_2 au locus A mais au locus B, il n'y a que l'allèle B_1 qui est présent dans la population, avant que l'allèle B_2 soit crée par une mutation survenue (puis transmise) chez un individu. Supposons que la mutation soit survenue sur un chromosome sur lequel est présent l'allèle A_2 du locus A. Alors, lors de sa création, l'allèle B_2 est toujours lié à l'allèle A_2 et le locus B est ainsi en déséquilibre de liaison complet avec le locus A. On a $f(A_1B_2) = 0$ et d'après le tableau, D est maximal et vaut $D_0 = f(A_1)f(B_2)$. Imaginons maintenant que le taux de recombinaisons entre les deux locus A et B soit égal à θ, avec $\theta \in [0; 0,5]$. Alors, d'une génération k à la suivante k+1, la fréquence de l'haplotype A_1B_2 dans la population passe de $f_k(A_1B_2)$ à :

$$f_{k+1}(A_1B_2) = (1-\theta)f_k(A_1B_2) + \theta f(A_1)f(B_2)$$

ce qui peut se réécrire

$$f_{k+1}(A_1B_2) - f(A_1)f(A_2) = (1-\theta)(f_k(A_1B_2) - f(A_1)f(A_2))$$

soit

$$D_{k+1} = (1-\theta)D_k$$

et ainsi, à la génération n, on a

$$D_n = (1-\theta)^n D_0$$

Le déséquilibre de liaison D diminue donc de génération en génération et d'autant plus rapidement que le taux de recombinaison est fort.

Une critique de la mesure D est que celle-ci n'est pas standardisée. Ainsi, un déséquilibre de liaison important entre deux polymorphismes peut prendre aussi bien des valeurs proches de 1, que des valeurs très faibles si les fréquences des allèles en jeu sont faibles. C'est pourquoi Lewontin proposa d'utiliser la mesure D' [71] :

$$D' = \frac{D}{D_{\max}}$$

où $D_{\max}$ est la valeur que prendrait D si le déséquilibre de liaison était complet :

$$D_{\max} = \begin{cases} \max\left(-f(A_1)f(B_1), -f(A_2)f(B_2)\right) & \text{si D<0} \\ \min\left(f(A_1)f(B_2), f(A_2)f(B_1)\right) & \text{si D>0} \end{cases}$$

Une autre mesure normalisée est le coefficient de corrélation au carré dont le lien avec D est donné par :

$$r^2 = \frac{D^2}{f(A_1)f(A_2)f(B_1)f(B_2)}$$

Dans ce manuscrit, c'est cette dernière mesure que j'utiliserai pour décrire le déséquilibre de liaisons entre deux SNPs.

Chapitre 3

L'épidémiologie génétique

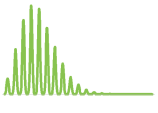
Mais où est donc passé Snippy ?

Dans le chapitre précédent, nous avons vu que les mutations et recombinaisons chromosomiques permettent l'évolution et la diversité du vivant mais qu'ils peuvent aussi être en partie responsables de certaines maladies. Dans ce chapitre, nous allons voir comment les évolutions scientifiques, technologiques ou informatiques permirent de découvrir certains des polymorphismes impliqués dans ces maladies. J'y présenterai également la stratégie adoptée durant ma thèse pour essayer de détecter une partie des nombreux variants qui restent à identifier.

3.1 Rappel historique

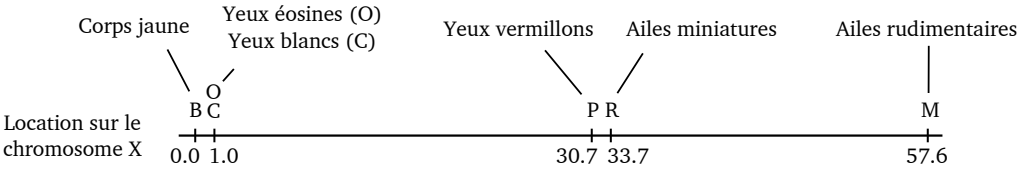
3.1.1 Principe général

Le but de l'épidémiologie génétique est d'identifier les déterminants génétiques des caractères héréditaires observables des individus. Pour ce faire, les épidémiologistes commencent par identifier certains caractères que l'on sait déterminés (au moins en partie) par la génétique, sur un certain nombre de personnes hétérogènes en ce qui concerne le trait à étudier. On appelle ces caractéristiques des marqueurs génétiques. Lors de la reproduction, les locus qui sont proches auront tendance à moins subir de recombinaisons que ceux qui sont éloignés. De ce fait, si l'on observe que les individus similaires pour certains marqueurs partagent souvent le même phénotype, cela indique une certaine proximité de ces marqueurs aux locus impliqués dans la variation du phénotype. C'est par ce biais-là que l'on a pu localiser des variations génétiques impliquées dans le caractère étudié.



3.1.2 Découverte des premiers marqueurs génétiques

Avant que Oswald Avery, Colin MacLeod et Maclyn McCarty ne démontrent en 1944 que l'ADN est le support de l'information génétique [5], les scientifiques savaient déjà que l'hérédité était transmise par les chromosomes. Ceci avait été démontré par Théodor Boveri au milieu des années 1880 et soutenu par Walter Sutton [117] pour donner la « Boveri-Sutton Chromosome Theory ». Peu après, William Bateson et Reginald Punnett avaient également pu montrer que certains caractères héréditaires étaient liés [9] ce qui contredisait ainsi la loi d'indépendance de Gregor Mendel [80], le fondateur de la génétique. Aussi, à partir de la description du phénomène d'enjambement chromosomique (« crossing-over ») par Frans Alfons Janssens en 1909 [55], Thomas Hunt Morgan put développer une correspondance entre la fréquence de « crossing-over » entre deux caractères et leur distance sur un chromosome [82]. Ceci rendit alors possible la création de la première carte génétique composée de six marqueurs, par Alfred Sturtevant en 1913 [116]. Cette carte génétique est présentée en figure 3.1.



Source : adapté de la carte originale de Sturtevant

FIGURE 3.1 – Carte génétique du chromosome X de la mouche drosophile, réalisée par Sturtevant. C'est la première carte génétique réalisée. Sturtevant y positionna six gènes qu'il nomma B, C, O, P, R et M. Le gène O semblant complètement lié au gène C, les gènes C et O sont situés au même endroit.

3.1.3 Les techniques d'ingénierie génétique

La localisation des régions chromosomiques susceptibles d'influencer certains caractères héréditaires devint ainsi possible grâce aux travaux de Morgan. Cependant, pour être efficace, la technique de Morgan nécessite que l'on dispose de marqueurs relativement proches des gènes impliqués dans le trait étudié. On peut s'assurer d'avoir ce genre de marqueurs en augmentant la densité de marqueurs sur le génome mais ceci nécessite l'identification de nombreux marqueurs génétique, ce qui s'avéra difficile jusqu'à la découverte de la structure en double hélice de l'ADN par James Watson et Francis Crick en 1953 [26]. Cette découverte permit l'essor de l'ingénierie génétique.

L'hybridation

Watson et Crick ont découvert que l'ADN est composé de deux brins antiparallèles associés par complémentarité de leurs bases azotées (A avec T, C avec G), par des liaisons hydrogènes. De par cette structure, deux fragments de brins d'ADN libres et complémentaires auront tendance à s'associer pour former un fragment double brins. C'est ce que l'on appelle l'hybridation et c'est le mécanisme qui est à la base des techniques d'identification de polymorphismes que sont le « southern blot » et la puce à ADN ou ARN. En effet, un polymorphisme génétique présent au sein d'une population engendre des différences de séquences qui peuvent être testées par hybridation. Si deux individus n'ont pas les mêmes allèles, leurs séquences ne s'hybrideront pas ou en tout cas moins bien que si les séquences étaient identiques.

La fragmentation de l'ADN

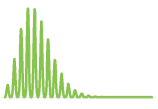
Une découverte importante fut celle de Hamilton O. Smith lorsqu'il isola en 1970 la première enzyme de restriction [110], une protéine capable de couper une petite séquence d'ADN bien déterminée. Smith ainsi que Daniel Nathans et Werner Arber, en découvriront de nombreuses autres et fournirent ainsi aux chercheurs une méthode puissante et rapide de fragmentation de l'ADN.

L'amplification en chaîne par polymérase

Quelques années auparavant, vers la fin des années 1950, Arthur Kornberg avait découvert que lors de la division cellulaire, l'ADN se dédouble grâce à la copie de ses deux brins d'ADN par l'ADN polymérase [67]. En 1983, Kary Mullis eut l'idée d'utiliser cette enzyme pour augmenter artificiellement et rapidement le nombre de copies d'un fragment d'ADN. Ce procédé fut appelé « amplification en chaîne par polymérase », ou plus simplement « PCR » (pour polymerase chain reaction) [84]. Avant les PCRs, dans les années 1970, Stanley N. Cohen et Herbert W. Boyer avaient déjà rendu possible la copie de l'ADN par la technique d'ADN recombinant [23]. Cette technique consiste à introduire un fragment d'ADN dans la séquence d'ADN d'une cellule étrangère afin d'engendrer sa réplication de façon naturelle dans ce corps étranger. La PCR en est une alternative puissante qui est souvent utilisée par les chercheurs.

HTT, le premier gène de prédisposition localisé

Toutes ces avancées permirent alors de fragmenter, amplifier puis hybrider des séquences d'ADN, afin de les comparer et ainsi faciliter l'identification de multiples polymorphismes. C'est ainsi que les cartes génétiques purent se densifier en marqueurs et le premier gène de prédisposition à une maladie, la maladie de



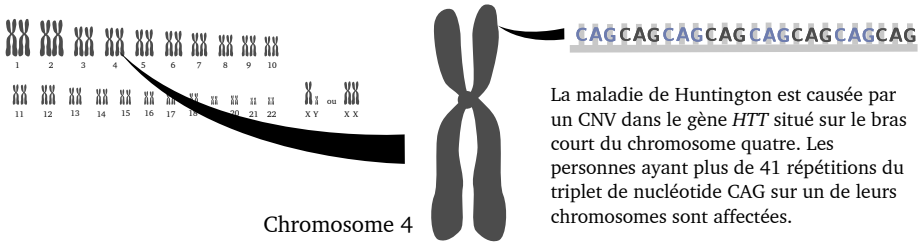


FIGURE 3.2 – Locus de prédisposition à la maladie de Huntington

Huntington, put être localisé sur le génome, approximativement d'abord en 1983, puis précisément dix ans plus tard, en 1993 (cf. figure 3.2), [127].

3.1.4 Puces à ADN/ARN

La densification du nombre de marqueurs sur le génome facilita donc la localisation des variations responsables de certains caractères, mais elle impliqua également le besoin de génotyper (identification des allèles) les individus pour tous ces marqueurs, ce qui était très fastidieux. Ce sont les technologies des puces miniatures à ADN et ARN (aussi appelées biopuces), apparues au milieu des années 1990 [102] qui apportèrent la solution. Le principe des puces à ADN est relativement simple. On commence par produire un grand nombre de fragments d'ADN simples brins, à partir d'une ou plusieurs séquences d'ADN de référence, par amplification. Ensuite, on attache ces fragments sur des puces rigides. On les appelle alors des sondes. Comme les fragments sont simples brins, ils peuvent s'hybrider avec d'autres fragments simples brins, si ceux-ci leur sont complémentaires. Finalement, on fragmente et amplifie l'ADN des individus et on les dispose sur les puces (voir figure 3.3). La mesure de l'hybridation des fragments des individus sur les puces permet de connaître les génotypes des individus pour les marqueurs présents sur les puces (ceux présents dans les séquences de référence) : si pour un individu, on observe une hybridation bien plus importante sur les sondes contenant un premier allèle, que sur les sondes contenant un autre allèle, cela signifie qu'il y a de grandes chances pour que l'individu soit homozygote pour le premier allèle. Si au contraire, on n'observe pas de différences notables dans l'hybridation, c'est certainement que l'individu est hétérozygote.

Le principe des puces à ARN est sensiblement le même à cela près que les fragments de référence sont construits à partir d'ARN (ce sont des fragments d'ADN complémentaires aux fragments d'ARN et appelés ADNc pour ADN complémentaires) et les fragments des individus sont construits à partir d'ARN également. Une séquence d'ARN de référence représentant un gène, une forte hybridation indiquera une forte expression du gène chez l'individu et au contraire, une faible hybridation indiquera une expression faible ou nulle. Les expressions des gènes peuvent également

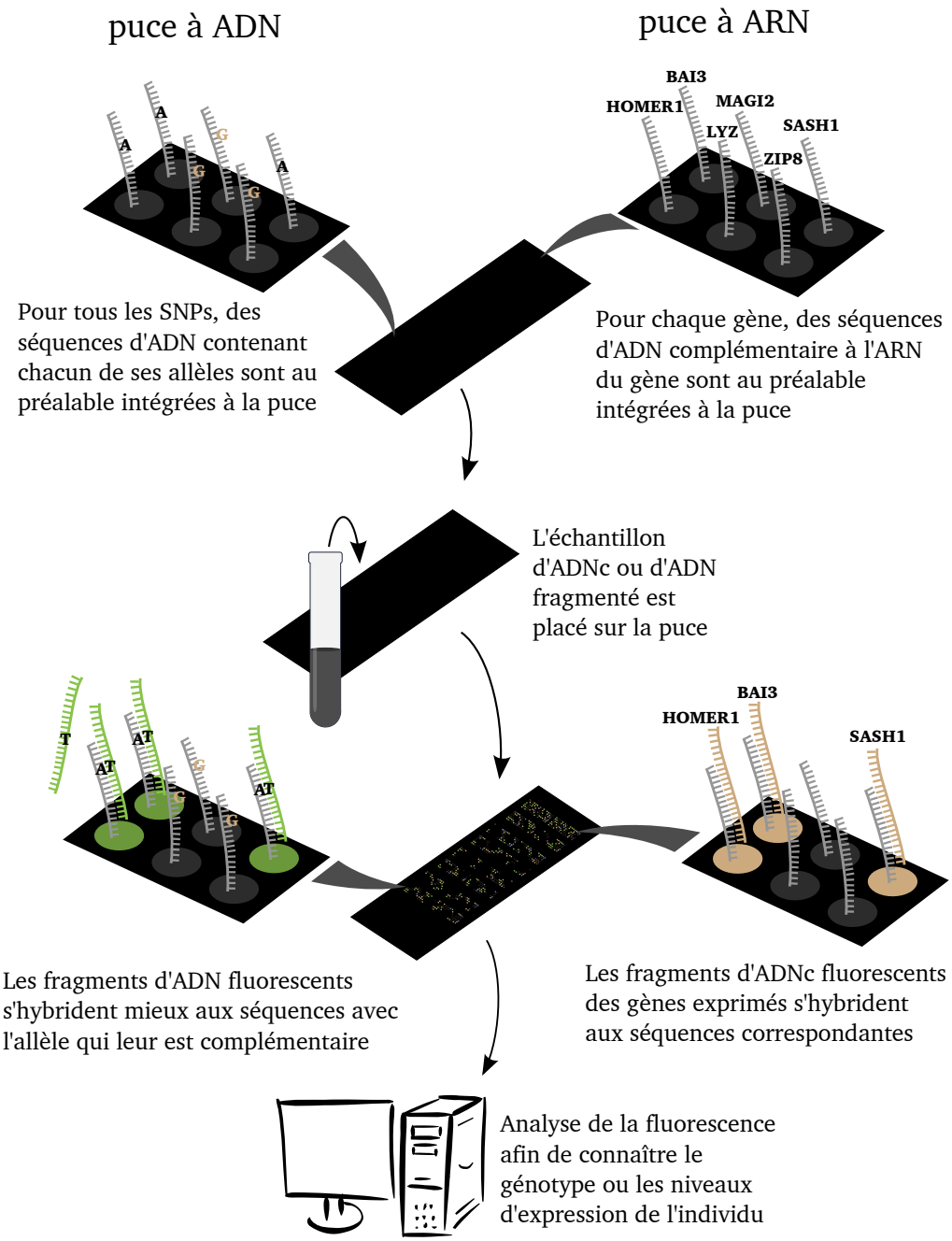


FIGURE 3.3 – La technologie des puces à ADN et ARN. A gauche, l'individu semble être de génotype homozygote TT pour l'un des SNPs de la puce. A droite, il semble que les gènes *HOMER1*, *BAI3* et *SASH1* soient exprimés, au contraire des gènes *MAGI2*, *LYZ* et *ZIP8* [1].

permettre de détecter les gènes impliqués dans certains phénotypes. En effet, si l'on observe que les individus qui ont un gène très exprimé, ont des phénotypes

1. Pour illustrer la technologie des puces à ARN, j'ai choisi ici d'utiliser des noms de gènes faisant références aux travaux de certains collègues que j'ai croisés ou au désormais plus long sitcom, en termes d'épisodes, de l'histoire de la télévision [150].

différents des autres individus, cela suggère donc une implication du gène dans le phénotype. Il peut aussi être intéressant d'identifier les polymorphismes qui affectent l'expression des gènes, car ceux-ci ont alors de bonnes chances d'avoir des effets sur certains phénotypes. On verra un peu plus tard que c'est ce second objectif que j'ai visé lorsque j'ai utilisé des données d'expression au cours de cette thèse.

3.1.5 Le séquençage

Les techniques d'ingénierie génétique évoquées précédemment permirent aussi l'apparition dans les années 1970, des premières techniques de séquençage, développées par Frederick Sanger [99], Allan Maxam et Walter Gilbert [78]. Au contraire du génotypage, qui vise « seulement » à connaître les allèles des individus pour certains marqueurs polymorphiques, le séquençage a pour but de déterminer complètement les séquences génétiques des individus. Rapidement, la technique mise au point par Sanger se popularisa. Elle permit en particulier le lancement en 1990, du projet génome humain avec pour mission de séquencer entièrement notre génome. La partie gauche de la figure 3.4 résume succinctement le principe de cette méthode de séquençage. Après avoir fragmenté l'ADN d'un individu, chaque fragment est amplifié puis mis en contact avec une enzyme d'ADN polymérase, des amorces pour entamer la copie des brins, des nucléotides de chaque type, ainsi qu'un seul des quatre didésoxyribonucléotides (ddNTPs) A, C, G ou T. Les ddNTPs sont des nucléotides qui ne possèdent pas de groupe hydroxyle, à leur extrémité 3', ce qui les empêche de se lier avec un nucléotide supplémentaire. Dans le milieu ainsi formé, chaque brin complémentaire au fragment initial se met à croître grâce à l'ajout de nucléotides par l'ADN polymérase, jusqu'à l'ajout d'un ddNTP de type particulier. On se retrouve ainsi avec des fragments de tailles variables, mais qui correspondent aux morceaux de séquences du fragment initial qui se terminent par la base complémentaire au type de ddNTP intégré. Cette opération est effectuée quatre fois, avec chaque type de ddNTP, puis on compare les poids des différents fragments obtenus en les faisant migrer dans un gel par l'application d'un champ électrique. Comme les fragments de petites tailles migrent plus rapidement que les fragments de grandes tailles, après avoir arrêté le champ électrique, on peut déterminer le type de ddNTP correspondant au fragment qui a migré le plus loin. Ce type de ddNTP est aussi le type de la première base. On fait de même pour le second fragment qui a migré le plus loin et ainsi de suite, de manière à déterminer la séquence complète du fragment initial (en l'occurrence de son complémentaire). Finalement, on répète cette expérience pour chaque fragment de la séquence d'ADN de départ et par similarité des extrémités des fragments séquencés, on peut retrouver la séquence d'ADN de départ. Il aura fallu un peu plus de 10 ans pour que le projet génome humain aboutisse avec la publication officielle de la première séquence d'ADN complète de

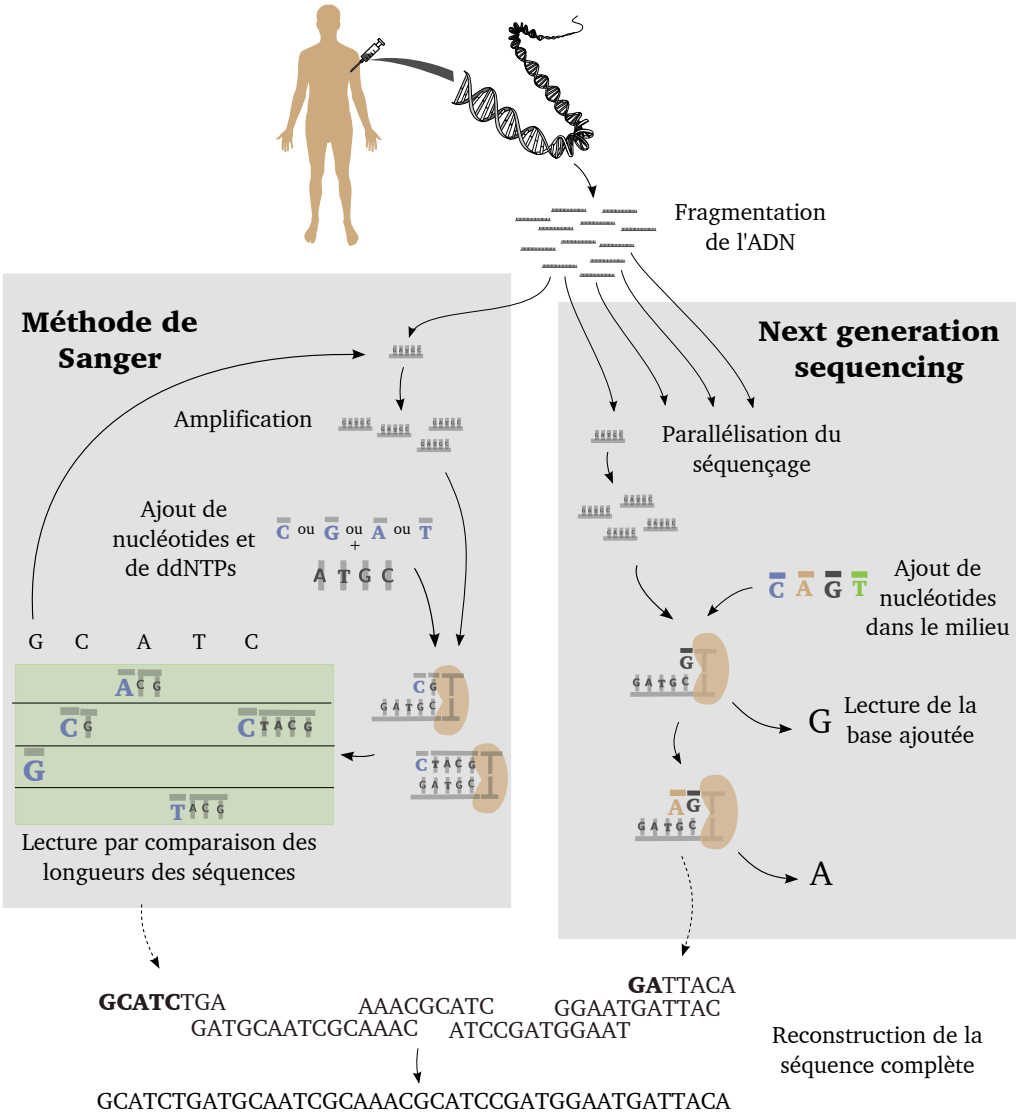


FIGURE 3.4 – Les techniques de séquençage : à gauche, la méthode de Sanger, à droite, le principe des séquenceurs de nouvelle génération

l'homme en 2004 (une première séquence ayant été pré-publiée en 2001)[53]. Dès lors, les scientifiques eurent à disposition une séquence de référence ce qui facilita grandement la découverte de nouveaux polymorphismes. Ils bénéficièrent également de l'avènement des technologies internet et de l'arrivée d'ordinateurs de plus en plus puissants et le tout permit de faciliter grandement la communication, l'efficacité de la communauté scientifique et l'enrichissement rapide des bases de données de biologie moléculaire, comme en témoigne la figure 3.5. De nos jours la technique de séquençage créée par Sanger est supplantée par le séquençage de nouvelle génération (communément appelé « next generation sequencing ») qui accélère grandement le processus par le séquençage de nombreux fragments d'ADN en parallèle (voir la partie droite de la figure 3.4). Cette « next generation sequencing », sans oublier

l'arrivée prochaine de la « third generation sequencing » [101], permet désormais de détecter facilement et rapidement n'importe quel type de polymorphisme qu'il soit connu ou inconnu et ainsi d'avoir à disposition un très grand nombre de marqueurs. Ces technologies, en assurant en théorie le séquençage des variants causaux tendent aussi à rendre inutile le raisonnement en terme de marqueurs génétiques.

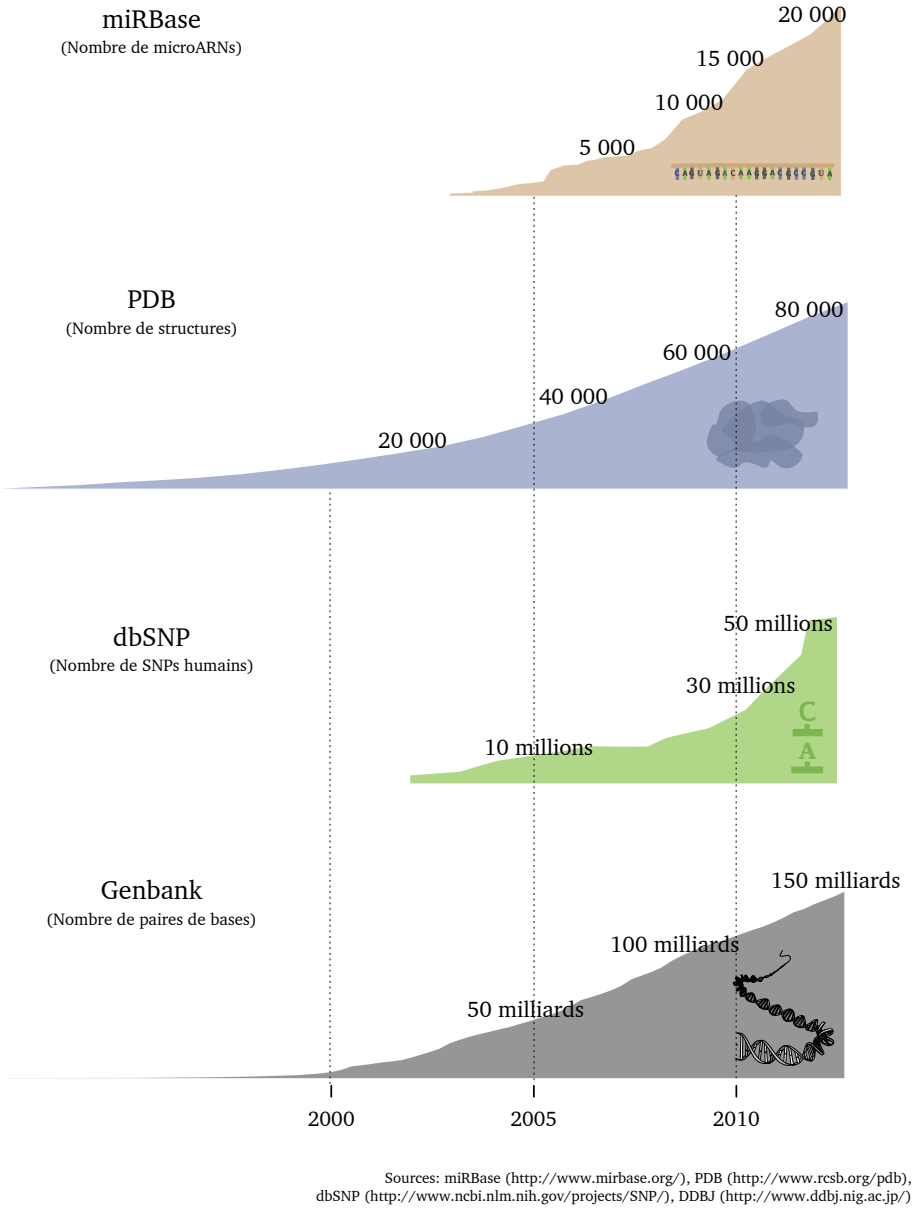


FIGURE 3.5 – Évolution des quantités de données intégrées dans quatre des plus importantes bases de données de biologie moléculaire : miRBase [12, 43] (base de données répertoriant l'ensemble des microARNs identifiés), PDB [12] (base de données répertoriant l'ensemble des structures 3D de macromolécules biologiques publiquement disponibles), dbSNP [105](base de données répertoriant l'ensemble des polymorphismes identifiés) et GenBank [11](base de données répertoriant l'ensemble des séquences de nucléotides publiquement disponibles)

3.1.6 Les stratégies d'analyse en épidémiologie génétique

La recherche en épidémiologie génétique nécessite l'utilisation de ce que l'on appelle des études. Ces études consistent au recrutement d'individus présentant déjà une variabilité phénotypique (comme pour les études cas-témoins [2]) ou qui présenteront une variabilité phénotypique (études de cohorte [3]) et pour lesquels on identifie le génotype pour certains marqueurs génétiques (tels que des SNPs) et on peut récupérer certaines autres caractéristiques (tels que l'âge, le sexe ou des mesures biologiques).

Les analyses de liaisons

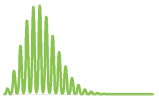
Lorsque l'on a peu de marqueurs génétiques à notre disposition, une manière de pouvoir tout de même identifier les régions du génome impliquées dans le trait étudié, est de recruter et génotyper des familles pour ces marqueurs. En effet, comme l'ADN des individus d'une même famille provient d'ancêtres communs très récents, cet ADN n'a pas pu subir de nombreuses recombinaisons chromosomiques. Aussi, les individus qui partagent le même phénotype auront tendance à recevoir les mêmes allèles pour tous les marqueurs qui ne sont pas trop éloignés du variant responsable du phénotype. On appelle ce type d'approche l'analyse de liaisons. Les analyses de liaisons furent très utilisées jusqu'au début des années 2000 afin de localiser les régions du génome susceptibles d'abriter les variants impliqués dans les traits étudiés. Après avoir identifié ces régions, on pouvait alors y rechercher de nouveau marqueurs afin de réaliser des analyses de liaisons plus fines sur ces régions. Il était aussi courant d'utiliser une autre approche, l'analyse d'associations.

Les analyses d'associations

Les analyses d'associations se basent sur le déséquilibre de liaison, plutôt que sur la liaison génétique familiale, ce qui fait qu'elles ne nécessitent pas l'utilisation de données familiales. Leur intérêt provient du fait que, dans la population générale, le déséquilibre de liaison entre polymorphismes se réduit rapidement à mesure que les polymorphismes s'éloignent. Ainsi, si l'on découvre que les individus qui partagent un même phénotype partagent souvent le même allèle pour un certain marqueur, cela indique que ce marqueur est très proche du polymorphisme impliqué dans le

2. Dans les études cas-témoins, deux groupes d'individus sont recrutés puis comparés, un composé de patients porteurs de la maladie (les cas) et l'autre de sujets seins (les témoins) mais similaires par ailleurs eux individus cas.

3. Dans les études de cohorte, on observe les évolutions au cours du temps du phénotype étudié et des autres caractéristiques mesurées sur un ensemble d'individus recruté aléatoirement. Les études de cohortes sont très utilisées pour déterminer les causes génétiques de certaines maladies fréquentes comme par exemple le cancer du sain [25].



phénotype. On arrive donc avec les analyses d'associations à une localisation plus fine des variants causaux. La contrepartie est que ces études nécessitent une forte densité de marqueurs, ce qui limitait, jusqu'au milieu des années 2000, leur utilisation à de petites régions du génome.

Les études d'associations et d'expressions en génome entier

Cependant, l'augmentation rapide des capacités en marqueurs des puces à ADN, passant de quelques centaines, à plusieurs centaines de milliers de polymorphismes a permis, à partir de 2004, la réalisation des premières études d'associations en génome entier (communément appelées GWAS, pour Genome-Wide Association Study) [61]. Celles-ci n'ont alors cessé de se multiplier comme on peut le voir sur la figure 3.6. L'une des plus remarquables est peut-être la GWAS publiée par le Wellcome Trust Case Control Consortium (WTCCC) en 2007 [128], qui révéla un bon nombre de nouveaux gènes de susceptibilité pour pas moins de sept maladies. Les analyses de liaisons restent cependant encore utilisées aujourd'hui, car les données familiales ont certains avantages comme celui de fournir des populations très homogènes. Les années 2000 ont aussi vu l'arrivée des premières études d'expression en génome entier (GWES, pour Genome-Wide Expression Study), dont le principe est de mesurer les expressions de tous les gènes du génome, par des puces à ARN. Il fut ainsi possible de combiner des données de génotypage avec des données d'expression pour détecter les polymorphismes susceptibles d'être impliqués dans les traits étudiés.

3.2 La recherche d'interactions pour tenter d'expliquer l'héritabilité manquante

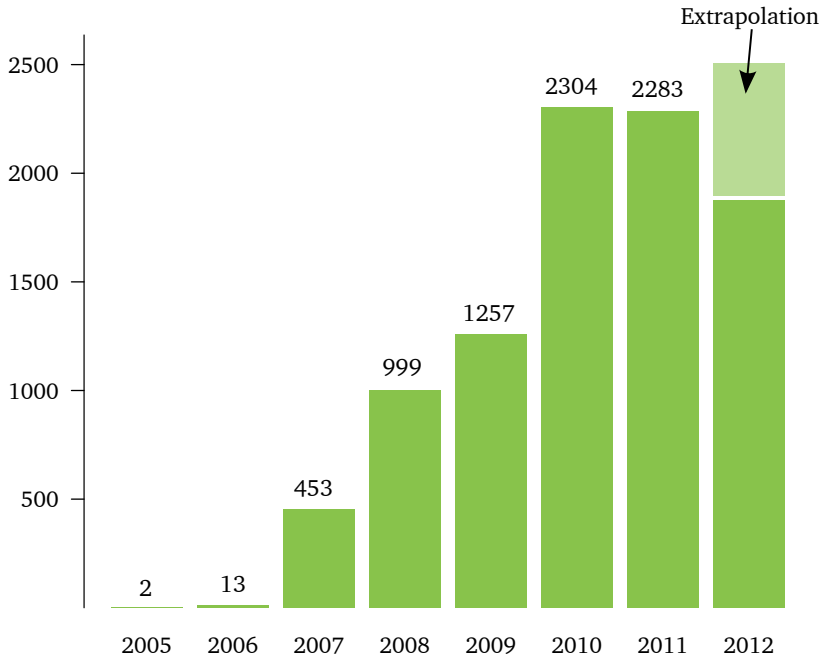
3.2.1 L'héritabilité manquante dans les maladies complexes

Les maladie mendéliennes

Lorsqu'une maladie est la conséquence de variation(s) génétique(s) au sein d'un faible nombre de gènes, on dit que cette maladie est « mendélienne ». Grâce aux études de liaisons et d'associations, nous sommes parvenus à identifier un nombre relativement important de gènes responsables de ces maladies. Ceux du chromosome 7 sont indiqués dans la figure 3.7. On y voit par exemple le gène *CFTR* en vert, dont l'une de ses formes est connue pour entraîner l'apparition de la mucoviscidose [60].

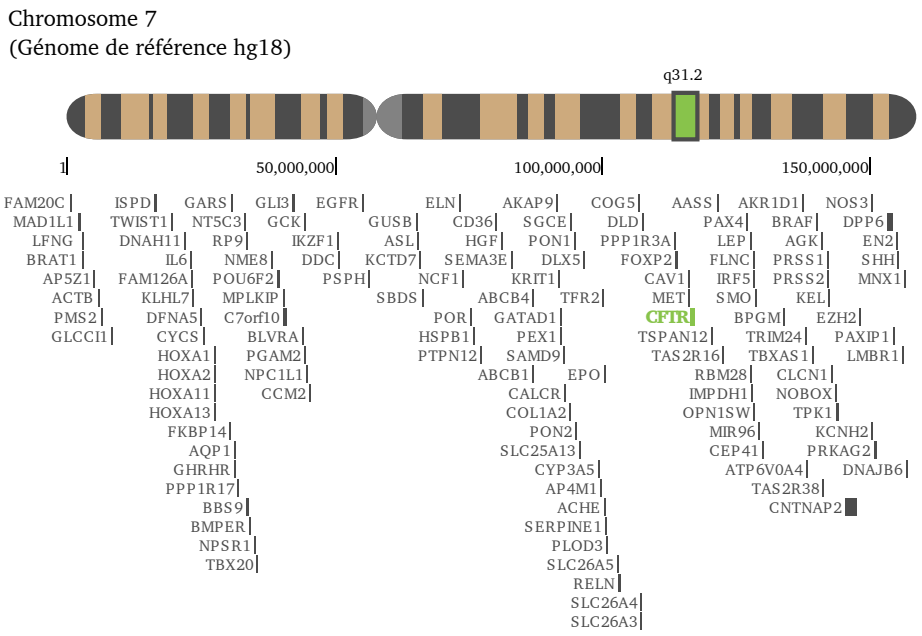
Les maladies complexes

Les maladies complexes en revanche sont des maladies dont les causes sont à la fois génétiques et environnementales avec des possibles interactions entre ces facteurs



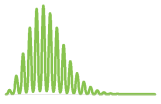
Source : Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwastudies/)

FIGURE 3.6 – Nombre de GWAS réalisées chaque année. Ayant réalisé ce graphique avant la fin de l'année 2012, le nombre de GWAS réalisées en 2012 est une extrapolation du nombre de GWAS réalisées au moment de la création du graphique.



Source : adapté de http://genome.ucsc.edu/

FIGURE 3.7 – Gènes du chromosome 7 pour lesquels on connaît une ou plusieurs mutations responsables de maladies mendéliennes d'après la base de données OMIM (Online Mendelian Inheritance in Man) [153]



de risque. L'obésité, le diabète, les maladies cardio-vasculaires, la schizophrénie ou encore la maladie d'Alzheimer sont quelques exemples de maladies complexes. Bien qu'elles soient responsables d'une part de plus en plus importante des décès dans le monde, on ne connaît encore que très peu leurs déterminants, notamment génétiques, et le grand nombre de GWAS réalisées n'a pour l'instant permis d'expliquer qu'une faible part, généralement, inférieure à 10 % de leur héritabilité (voir la figure 2.2 pour les estimations d'héritabilité pour certains traits complexes). Cela pousse à se demander pourquoi nous n'arrivons pas à expliquer totalement cette héritabilité [75].

3.2.2 Les possibles causes de cette héritabilité manquante

Des polymorphismes plus difficiles à trouver que Charlie

Où est Charlie ? « Où est Charlie ? » est une série de jeux/bandes dessinées créée par Martin Handford en 1987 et diffusée en France à partir de 1989 [45], dans laquelle le lecteur doit retrouver le personnage de Charlie, un jeune homme portant des lunettes et habillé d'un bonnet et d'un pull à rayures horizontales rouges et blanches. La difficulté du jeu réside dans le fait que sur chaque page où l'on doit trouver Charlie, celui-ci se retrouve entouré de centaines d'autres personnages et objets. On peut voir un exemple d'imitation de la série dans la figure 3.8. Si vous n'y trouvez pas Charlie, vous pouvez trouver la solution à la fin de cette thèse [155]. À première vue, la recherche de polymorphismes impliqués dans un phénotype parmi l'ensemble des polymorphismes d'une étude génome entier peut sembler un peu similaire à la recherche de Charlie parmi tous les personnages présents sur une même image. Cependant, comme nous allons le voir, la quantité de données et la complexité d'une recherche de polymorphismes dans une GWAS est bien supérieure à ce qui peut se trouver dans un tel jeu et c'est peut-être l'une des raisons pour lesquelles nous sommes encore très loin d'avoir trouvé tous les facteurs génétiques des maladies complexes.

Où est Snippy ? Si notre recherche consistait en un jeu de type « Où est Charlie ? », les règles en seraient cependant forcément un peu différentes :

- Le but ne consisterait pas en la recherche de « Charlie » mais d'un nombre non communiqué de Charlies... avec la possibilité qu'il n'y en ait aucun (si par exemple les causes de la maladie sont de nature épigénétique).
- Le nombre de personnages sur une page ne serait pas de quelques centaines, mais de plusieurs centaines de milliers (les puces à ADN couramment utilisées permettent en général le génotypage de près d'un million de polymorphismes)

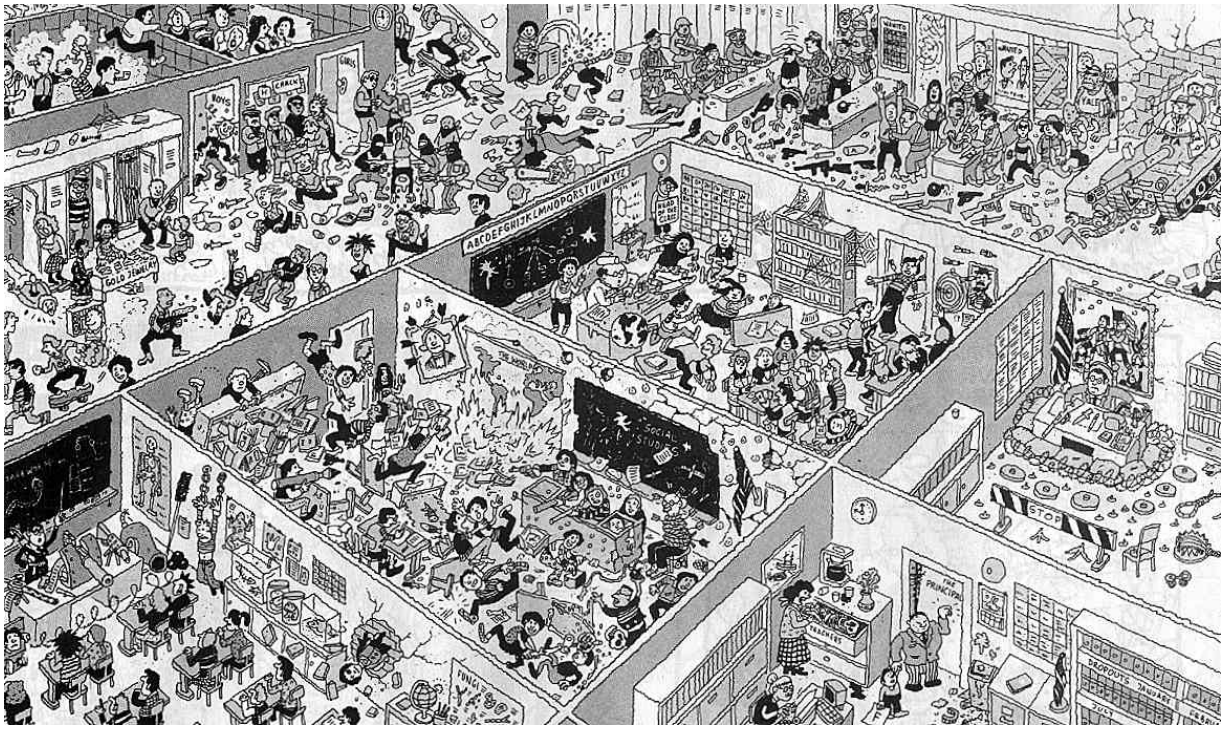


FIGURE 3.8 – Où est Charlie à l'école : une des imitations du célèbre jeu « Où est Charlie ? »

- Les habits des différents personnages auraient tendance à déteindre sur les habits de leurs voisins (à cause du déséquilibre de liaison)
- Les vrais Charlies ne seraient pas forcément eux-même visibles sur le dessin, mais cachés par d'autres personnages et il faudrait les deviner en devinant les personnages sur lesquels leurs habits auraient déteints (les variants causaux ne sont pas forcément génotypés).
- Il se pourrait tout à fait que les Charlies se partagent leurs vêtements et qu'il faille chercher le pantalon à un endroit et le pull à un autre endroit (il est possible qu'il y aient des phénomènes d'interactions entre polymorphismes comme nous allons d'ailleurs le supposer par la suite).
- Et, pour compliquer l'affaire, les Charlies ne seraient pas tous habillés parfaitement comme le Charlie de la figure 3.9 et inversement certains autres personnages auraient parfois des habits qui pourraient laisser croire qu'ils sont des Charlies (Problème des tests multiples en statistique qui seront abordés dans le chapitre 5 : sur un million de SNPs, il y a forcément des SNPs qui sembleraient, pris séparément, être associés à la maladie)

... enfin, pour que cette analogie soit complète, il faudrait rajouter une dernière règle, non des moindres :

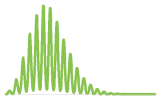


FIGURE 3.9 – Charlie. Afin de le distinguer du Charlie du jeu traditionnel, il a quitté ses rayures rouges et blanches pour enfiler des rayures vertes et blanches et il tient dans ses mains une cellule remplie de chromosomes

- Chaque personnage ne serait en fait pas visible sur une seule page, mais ses habits et objets seraient répartis sur plusieurs centaines, milliers voire centaines de milliers de pages (l'ensemble des génotypes d'un individu d'une GWAS n'est qu'une petite part de l'information nécessaire à la découverte de variants. C'est la combinaison astucieuse des génotypes de tous les individus, plusieurs milliers voire centaines de milliers[34, 65, 111], qui permet d'y arriver).
- Pour nous aider, on pourrait demander des pages supplémentaires, mais il faudrait les payer (Avoir plus d'individus dans une GWAS permet d'avoir plus de chances de détecter certaines variations, mais il en résulte un coût également supérieur).

En dehors des différences de règle du jeu, on observerait également des différence d'état de la bande dessinée :

- La BD (pavé) ne serait pas forcément livrée en excellent état (données manquantes, erreurs de génotypage) et il faudrait enlever certaines pages abîmées (filtrage sur les individus) ou même reboucher certains trous traversant le livre (filtrage sur les variants) avant de pouvoir le feuilleter correctement.
- Parfois l'éditeur pourra avoir malencontreusement associé les pages du livre avec celles d'un autre livre (problème de population non homogène). Comme

le dessinateur ne s'embête pas trop et remet toujours les mêmes personnages dans ses livres (on a souvent les mêmes polymorphismes sur la plupart des puces à ADN), en ne changeant que les vêtements et encore, parfois très peu (les fréquences des polymorphismes ne sont pas toujours différentes entre les populations), il serait cependant envisageable d'utiliser le livre entier en plaçant des calques différents sur les pages provenant de livres différents (méthodes d'ajustement pour corriger ce problème de stratification).

Les hypothèses biologiques avancées pour expliquer cette héritabilité manquante ?

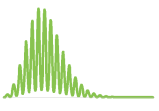
A cette complexité statistique, s'ajoute le fait que l'on n'a probablement pas cherché l'héritabilité de certains traits complexes là où elle se trouvait [58, 75]. Parmi les hypothèses les plus souvent évoquées, l'héritabilité inexpliquée proviendrait :

- de variants rares aux effets forts. La plupart des marqueurs génétiques présents sur les puces à ADN étant relativement fréquents (fréquences de l'allèle mineur supérieure à 1 %), l'utilisation de ces puces ne permet pas de détecter l'effet de variants rares. Les méthodes statistiques classiquement utilisées sont aussi souvent peu adaptées à ce genre d'analyse.
- de très nombreux polymorphismes aux effets faibles, non détectés car le grand nombre tests effectués dans les analyses de GWAS implique des corrections pour tests multiples sévères ne permettant pas de détecter des effets faibles. L'augmentation des tailles des études devrait a priori permettre de détecter ce genre de polymorphismes.
- de phénomènes épigénétiques qui ne sont pas détectables par les puces à ADN classiques. Il existe par exemple maintenant des puces spécifiquement adaptées à la détection de la méthylation.
- des phénomènes d'interactions entre gènes ou avec l'environnement, pas systématiquement testés et qui sont difficiles à détecter du fait de l'augmentation importante des nombre de tests qu'ils engendrent.

L'ensemble de ces hypothèses est résumé sur la figure 3.10.

3.2.3 La stratégie adoptée dans ce travail de thèse

Partant du constat de cette grande part d'héritabilité génétique encore inexpliquée par les approches classiques, nous avons décidé dans ce travail de thèse d'investiguer l'hypothèse d'une héritabilité manquante se situant dans des phénomènes d'interactions entre gènes.



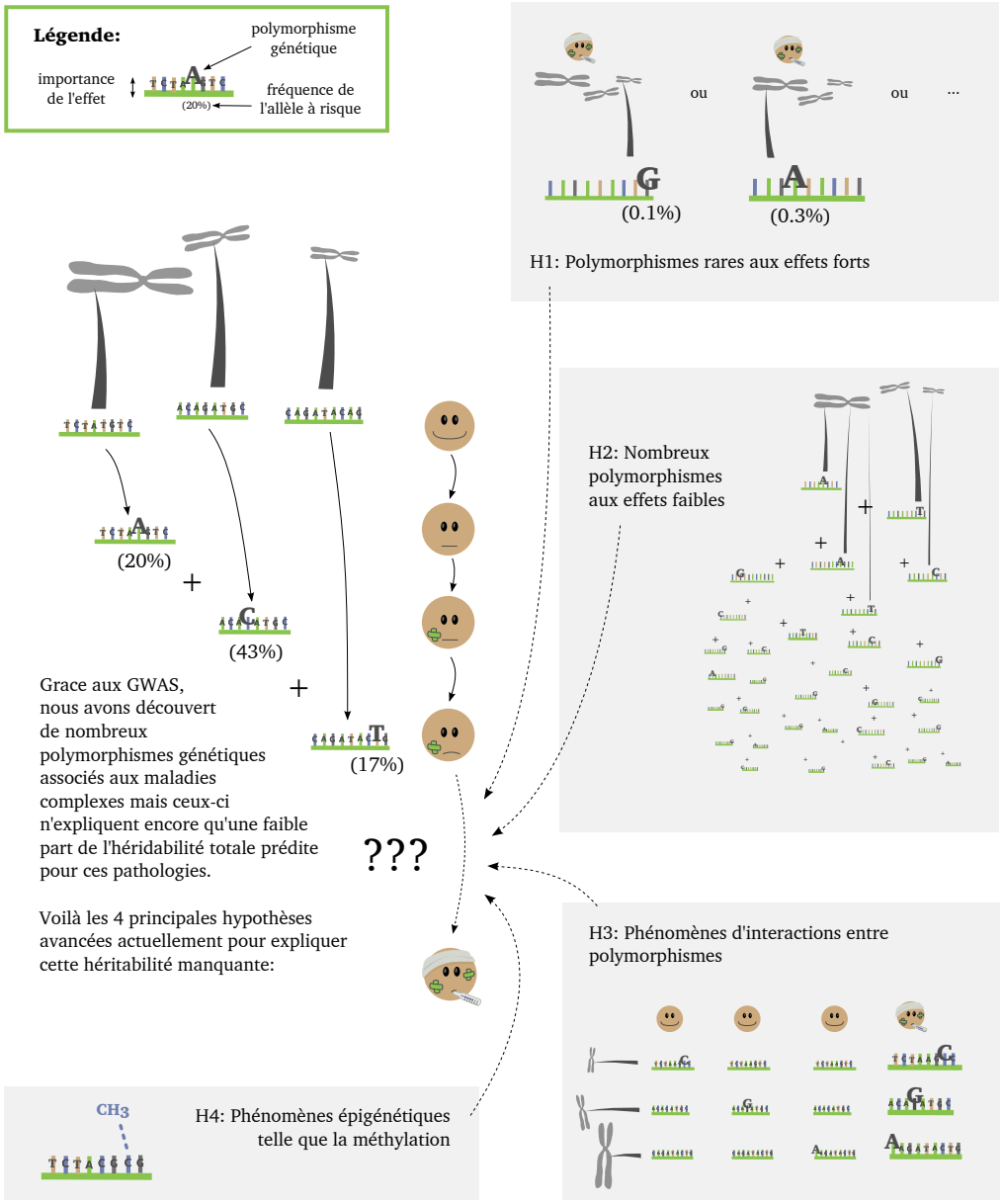
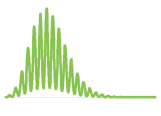


FIGURE 3.10 – Hypothèses les plus couramment avancées pour expliquer où se trouve l'héritabilité manquante.

- Nous avons commencé par rechercher si des phénomènes d'interactions entre polymorphismes ne pouvaient pas être impliqués dans la thrombose veineuse. Ceci nous a amené à tenter d'améliorer la détection de phénomènes d'interactions par des considérations statistiques (chapitre 7).
- Puis, nous avons essayé de nous limiter à certains éléments biologiques nous paraissant plus à même d'être impliqués dans ce genre de phénomène, en recherchant les polymorphismes liés aux microARNs qui pourraient affecter l'expression de nos gènes (chapitre 8).

Les trois prochains chapitres visent à introduire les méthodes statistiques (chapitres 4 et 5) et les études (chapitre 6) utilisées lors de ces travaux de recherches.



Chapitre 4

Les tests statistiques

C'est un très bon test pour nous.

Olivier Giroud (avant-match Montpellier-PSG 2011)

Les trois chapitres précédents ont permis de décrire la base de l'épidémiologie génétique à savoir que pour localiser des gènes ou polymorphismes impliqués dans un phénotype, on recherche des marqueurs pour lesquels les individus similaires pour le phénotype sont aussi similaires pour ces marqueurs. Plus spécifiquement, afin de savoir si cette similarité conjointe peut-être attribuable au hasard ou est le reflet d'un réel rôle dans le trait étudié, on teste nos hypothèses par des tests statistiques. Le but de ce chapitre est d'introduire le principe du test statistique.

4.1 Introduction

4.1.1 Notre raisonnement au Pile ou Face

Supposons que l'on nous demande d'établir si il y a tricherie ou non dans un jeu du pile ou face.

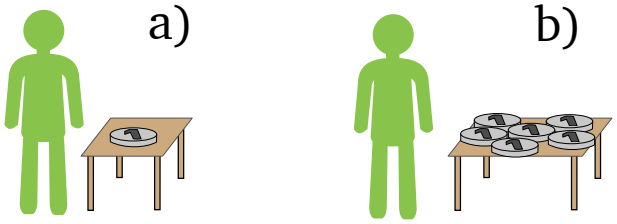


FIGURE 4.1

Une personne lance une pièce de monnaie. Elle obtient pile (figure 4.1– a). Nous ne sommes absolument pas surpris, tout comme si elle avait obtenue face d'ailleurs. Par contre, si elle lance 6 fois une pièce et obtient 6 fois pile (figure 4.1– b), nous aurions tendance à penser qu'elle a triché. Chaque pile indépendamment nous paraît possible mais c'est la combinaison qui nous surprend car sur un tel nombre d'observations, on s'attend à voir des piles, mais aussi des faces.

Lorsque la situation se complique, on a recours à un test statistique dont le raisonnement est similaire. Dans un test statistique, on a deux hypothèses (non tricherie et tricherie dans l'exemple précédant) et l'on cherche a :

- **combiner les observations** de sorte à pouvoir pencher vers l'une ou l'autre de nos hypothèses, ce qui n'aurait en général pas été possible en analysant les observations indépendamment.
- **évaluer si la combinaison semble possible sous nos hypothèses**, ce qui nous donne des indications sur l'hypothèse la plus probable.

Après avoir brièvement développé chacun des points en gras dans un contexte global, je les aborderai plus spécifiquement dans le contexte de mon sujet de thèse.

4.1.2 Quelques termes utilisés dans la suite de ce chapitre

Hypothèses Dans un test statistique, on a toujours deux hypothèses, l'hypothèse appelée H0 qui est notre hypothèse par défaut et l'hypothèse appelée H1 qui est l'hypothèse alternative vers laquelle on penchera si H0 ne nous semble pas correcte.

Statistique Le résultat numérique d'une combinaison des observations s'appelle une statistique.

Distribution Déterminer les valeurs que la statistique peut prendre et avec quelles fréquences revient à en connaître sa distribution. Lorsque la distribution est bien déterminée, on peut aussi parler de loi de distribution.

Modèle statistique Un modèle statistique consiste en une supposition de forme de lien entre différentes variables où la force du lien est intégrée dans des paramètres. Ce sont ces paramètres qui différencient donc les différentes hypothèses intégrées dans le modèle. Lorsque le modèle consiste à expliquer une variable en particulier, comme la survenue d'une maladie ou le niveau d'expression d'un gène, on utilise plus précisément le terme « modèle de régression ». Tous les modèles décrits dans ce document sont de ce type.

Espérance L'espérance est la valeur moyenne parmi toutes les valeurs que peut prendre une variable.

Variance La variance est une mesure de la variabilité d'une variable. Plus précisément, si x est une variable et $E(x)$ est son espérance, alors, la variance de x est l'espérance du carré des écarts entre x et $E(x)$.

4.2 Les différentes approches

4.2.1 Comment combiner des observations ?

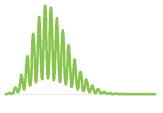
Dans un test statistique, on souhaite donc trouver une combinaison des observations qui discrimine bien les hypothèses. Il y a deux façons de procéder :

directement : On peut essayer de trouver directement une combinaison qui permet de bien différencier les hypothèses. Par exemple, dans l'exemple du pile ou face, le maximum du nombre de pile et du nombre de face semble être une bonne statistique. Une personne qui triche aura tendance à obtenir une statistique élevée contrairement à une personne qui ne triche pas. Cette façon de combiner est à la base de la plupart des tests développés. Parmi les plus connus, on peut ainsi citer le t-test [115], le test du χ^2 d'indépendance [89] ou encore les tests de Lévene [69] et d'Hardy-Weinberg [46] qui sont décrits dans la section 4.5.

en utilisant un modèle : On peut aussi essayer de proposer un lien entre les observations, dont les paramètres varient en fonction de nos hypothèses. C'est ce que l'on appelle un modèle. Par exemple, on pourrait dire que la probabilité d'obtenir pile est égal à 0.5 plus un paramètre a, qui vaut zéro si l'on est sous l'hypothèse de la non tricherie et est différent de zéro sinon.

$$P(\text{pile}) = 0.5 + a$$

Établir un modèle nécessite d'être plus explicite sur nos hypothèses mais permet plus de transparence et en passant par des modèles complexes, de répondre à des questions qui le sont tout autant. L'autre avantage est qu'en passant par un modèle, on a des statistiques évidentes que sont les estimations des paramètres du modèle (a pour l'exemple ici). Il y a plusieurs méthodes pour les calculer. La plus courante est le maximum de vraisemblance visant à trouver les valeurs des paramètres qui permettent aux observations d'être les moins surprenantes possibles. C'est surtout cette deuxième méthode que j'ai utilisée pour combiner les observations dans cette



thèse. Les modèles utilisés ainsi que les estimations des paramètres sont décrits en section 4.3.

4.2.2 Évaluer si la combinaison obtenue est possible

distribution exacte : Si l'on connaît la distribution des observations sous H0, alors, il est possible dans certains cas, d'en déduire la distribution de la combinaison effectuée. C'est la méthode sur laquelle se base le test de Levene (voir section 4.5) et que je décris un peu plus dans le prochain chapitre.

distribution asymptotique : Si l'on ne connaît pas la distribution des observations, il n'est alors pas possible de connaître la distribution exacte de la combinaison sous H0. Cependant, la combinaison la plus intuitive et la plus pertinente consiste souvent en une somme effectuée sur les observations. Dans une telle situation, la variabilité de chaque observation est en partie compensée par celle des autres observations et à mesure que le nombre d'observations augmente, la somme effectuée tend à avoir une distribution normale (voir figure 4.2). C'est le théorème central limite. Vous pouvez voir une petite illustration de ce phénomène en feuilletant rapidement le coin en bas à droite de ce document. Pour toutes les pages numérotées k, à partir de la table des matières, le dessin du coin bas droit représente la distribution de la somme de k variables distribuées selon la distribution présentée à la page numérotée 1. A mesure que l'on s'approche de la fin du document, la distribution se rapproche clairement d'une distribution normale. De très nombreux tests se basent sur cette approximation. C'est en particulier sur ce théorème qu'est basé le test de Wald [126] qui est utilisé dans la plupart des analyses effectuées dans cette thèse et que je décris en section 4.4.

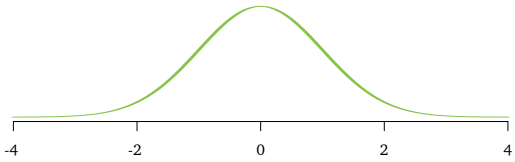


FIGURE 4.2 – Distribution normale de moyenne 0 et de variance 1.

distribution estimée empiriquement : Enfin, si l'on est capable de simuler des observations sous H0, on peut alors simuler la statistique sous H0 et ainsi estimer sa distribution. En épidémiologie génétique, il est souvent facile de simuler des observations sous notre hypothèse H0 en réassignant aléatoirement le phénotype étudié aux individus. Ainsi, les associations réelles sont « cassées » et toutes les statistiques que l'on pourra calculer sur ces données simulées permettront d'estimer la distribution de la statistique sous H0. C'est la méthode d'estimation de la distribution

par « permutations ». Elle permet d'estimer n'importe quelle distribution mais en contre partie, pour que l'estimation soit précise, elle requiert de très nombreuses simulations ce qui peut parfois prendre beaucoup de temps et nécessiter des capacités de calculs importantes.

Conclusion du test : la valeur de probabilité

Généralement, la conclusion d'un test statistique consiste au calcul de la valeur de probabilité couramment appelée p-value (c'est comme cela que je l'appellerai par la suite) grâce à la distribution de la statistique que l'on a déterminée auparavant. Par définition, la p-value est « la probabilité, si H0 est vraie, d'observer pour un test une statistique plus extrême que celle véritablement observée ». En fonction de la valeur de cette p-value (plus elle est faible plus on considérera que notre hypothèse n'est pas plausible) et des objectifs du test, on décide alors de rejeter ou non notre hypothèse H0. Lorsque le test effectué nous permet de rejeter l'hypothèse H0, on dit que le test est significatif.

4.3 Les modèles utilisés et l'estimation de leurs paramètres

Après cet aperçu des différentes façons de tester une hypothèse, je vais maintenant présenter les modèles et les tests utilisés dans mes travaux de recherches.

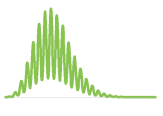
Modèle linéaire

Le modèle linéaire est un modèle liant linéairement un caractère quantitatif, tel que le niveau d'expression d'un gène, à des variables explicatives telles que le nombre de copies d'un allèle d'un SNP (cf. figure 4.3). On l'explicite par une équation mathématique du type

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdot + a_p x_p + \epsilon$$

où y est la variable à expliquer, x_1, x_2, ... x_p sont les variables dites explicatives, qui influencent de manière linéaire y, ϵ représente une variable aléatoire que l'on suppose normale et de moyenne nulle. Elle englobe le caractère aléatoire de toute mesure qui n'est pas due aux autres variables. Enfin, a_0, a_1, a_2, ... ,a_p sont les paramètres du modèle, représentant la magnitude moyenne globale de y et les magnitudes des liens entre y et x_1, x_2, ..., x_p, respectivement.

Le modèle linéaire est le modèle le plus utilisé lorsque l'on souhaite modéliser un phénotype quantitatif. C'est aussi celui utilisé dans ce document pour tenter d'expliquer les niveaux d'expression des gènes et quelques autres caractéristiques biologiques.



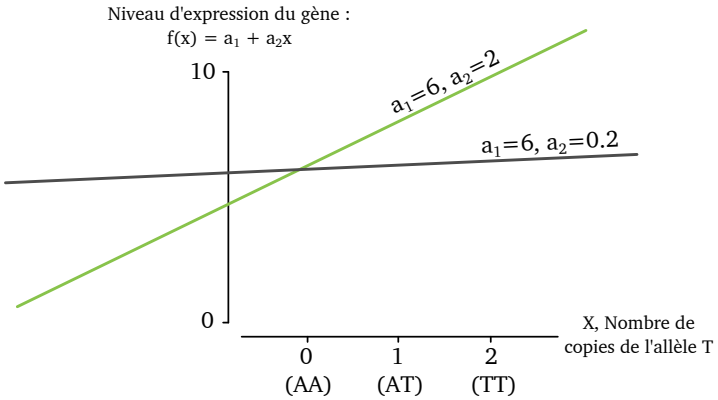


FIGURE 4.3 – Relation linéaire entre le nombre de copies de l'allèle mineur d'un SNP et le niveau d'expression d'un gène. En gris, le SNP n'a pas (ou peu) d'effet sur l'expression du gène. À l'inverse, en vert, le SNP a un effet additif relativement fort sur son niveau d'expression.

Modèle logistique

Le modèle logistique est un modèle liant une variable binaire tel que le caractère « malade/non malade » à des variables explicatives telles que les génotypes des individus pour un SNP (cf. figure 4.4). Il se définit mathématiquement par l'équation

$$P(y=1) = \frac{\exp a_0 + a_1 x_1 + a_2 x_2 + \cdot + a_p x_p}{1 + \exp a_0 + a_1 x_1 + a_2 x_2 + \cdot + a_p x_p}$$

où y est la variable binaire recodée en 0/1 alors que x_1, x_2, ... x_p et a_0, a_1, a_2, ..., a_p représentent comme pour le modèle linéaire, les variables explicatives et leurs paramètres associés. Notons que l'incertitude qui était comprise dans la variable ϵ dans le modèle linéaire, est directement intégrée à la variable que l'on tente d'expliquer dans le modèle logistique puisque l'on ne modélise pas y, mais sa probabilité de valoir 1. Notons qu'il est courant de transformer les paramètres du modèle logistique en odds-ratios (voir encadré) pour faciliter leur interprétation.

Les odds-ratios

L'odds-ratio (OR) est une mesure de l'effet d'une variable explicative sur une variable binaire que l'on souhaite comprendre (par exemple le phénotype « malade/non malade »). Pour un SNP ayant les allèles A et T, on peut définir l'odds-ratio de l'association entre le SNP et la maladie de la manière suivante :

Si la probabilité d'être malade est p lorsque l'on possède l'allèle A et q lorsque l'on possède l'allèle T, alors, l'odds-ratio associé à l'allèle A est :

$$OR = \frac{\frac{p}{1-p}}{\frac{q}{1-q}}$$

Si l'allèle A est à risque, alors, $\frac{p}{1-p}$ sera plus grand que 1 au contraire de $\frac{q}{1-q}$. L'OR sera donc supérieur à 1. Si par contre, l'allèle A n'est pas à risque, alors $\frac{p}{1-p}$ sera sensiblement égal à $\frac{q}{1-q}$ et l'OR sera proche de 1.

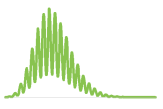
Ce qui est intéressant avec cette mesure, c'est que dans un modèle logistique, le logarithme de l'odds-ratio associé à une variable correspond au paramètre estimé associé à la variable.
Remarque : le paramètre estimé associé au terme d'interaction dont nous parlerons un peu plus tard n'est en fait pas le logarithme d'un odds-ratio mais celui d'un ratio d'odds-ratio. L'interprétation est cependant similaire et par souci de simplicité, dans la suite du document, j'emploierai également le terme d'odds-ratio pour décrire les mesures des associations impliquant des interactions.

Le modèle logistique est le modèle le plus utilisé lorsque l'on souhaite modéliser le risque de survenue d'une maladie. C'est aussi celui utilisé dans ce document lorsque le phénotype à expliquer est de ce type.

4.3.1 Les variables du modèle

Les génotypes

Chaque individu ayant deux copies de chaque chromosome autosomal, le rôle joué par un gène, un microARN, un SNP ou tout autre élément variable du génome, résulte en réalité des actions combinées des deux versions de ces éléments (les deux allèles de l'individu). Parfois, comme l'avait mis en évidence Grégor Mendel [80], cette combinaison est complètement dominée par un seul des allèles, auquel cas



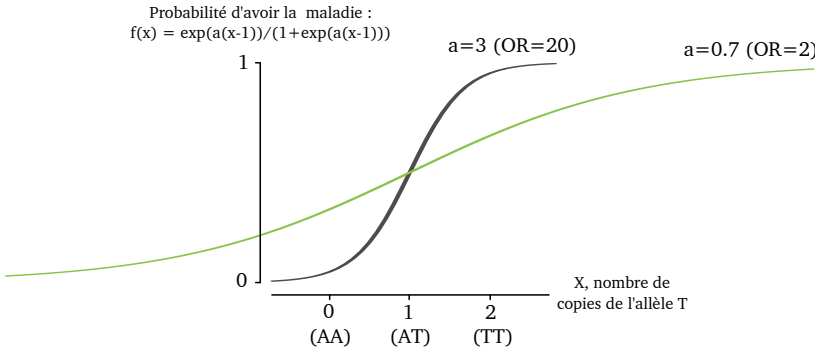


FIGURE 4.4 – Relation logistique entre le génotype d'un individu pour un SNP et sa probabilité d'être atteint par la maladie. En gris, un effet additif à pénétrance complète, observable dans certaines maladies mendéliennes. En vert, un effet additif fort, observable dans certaines maladies complexes.

on parlera d'allèle dominant, les autres allèles étant récessifs. D'autres fois, on a des allèles codominants, dont les effets se combinent lorsqu'ils sont ensemble. À la position d'un SNP qui ne peut avoir que deux allèles (par exemple C ou T), un individu a trois génotypes possibles : CC, CT ou TT. Suivant l'hypothèse faite sur l'effet du génotype sur le caractère étudié, on choisira un codage du génotype plutôt qu'un autre. En général, on utilise l'un des trois codages suivants :

- **0/1/1** pour CC/CT/TT lorsque l'on fait l'hypothèse d'un effet dominé par l'allèle T (l'allèle C sera alors récessif). À ce moment là, soit l'individu a une (ou plusieurs) copies de l'allèle T et l'on pense que l'on devrait observer l'effet de l'allèle T, soit il n'en a aucune et l'effet ne devrait pas être perçu.
- **0/0/1** lorsque l'on fait au contraire l'hypothèse d'un effet dominé par l'allèle C.
- **0/1/2** lorsque l'on émet l'hypothèse d'une codominance avec un effet intermédiaire lorsque le génotype est CT. On parlera aussi d'effet additif, car on peut l'interpréter comme une accentuation de l'effet à mesure que le nombre de copies de l'allèle T (ou C) augmente.

Codage des génotypes en 0,1,2

Dans tout mon travail de thèse, j'ai choisi d'utiliser un codage additif. Les raisons en sont d'une part biologiques, car une bonne partie de mon travail de recherche a porté sur l'influence des polymorphismes génétiques liés aux microARN sur le transcriptome [1] et l'on peut imaginer qu'étant donné le caractère quantitatif de l'effet d'un microARN (il régule la production de protéine d'un gène), les allèles d'un SNP lié à un microARN auraient plutôt tendance à avoir des effets qui s'ajoutent donc

1. Le transcriptome est l'ensemble des ARN messagers qui sont exprimés dans un type cellulaire.

additifs. D'autre part, contrairement à un codage en 0/1/1, un codage additif fait bien la différence entre avoir aucune ou une version d'un allèle (codage 0 ou 1) et avoir deux versions d'un allèle (codage 2). Inversement, le codage additif différencie aussi l'absence (codage 0) et la présence d'un allèle (codage 1 ou 2), au contraire d'un codage en 0/0/1. Ainsi, le codage additif permet également dans une moindre mesure, de détecter des effets récessif et dominants. (cf. figure 4.5).

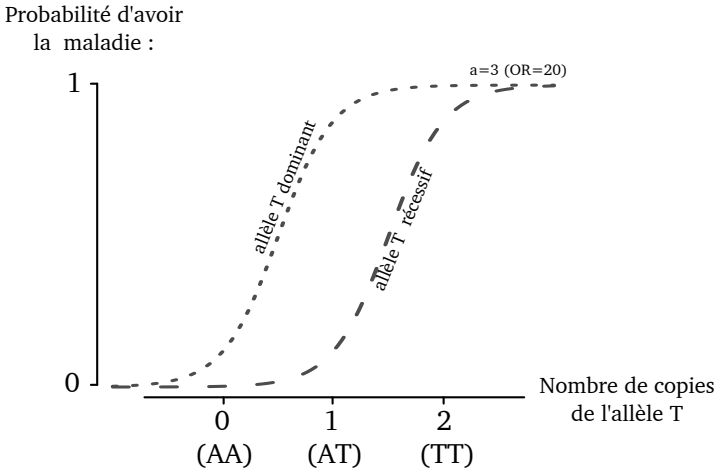


FIGURE 4.5 – Un codage en 0,1,2 peut détecter les différences qu'il peut y avoir les individus AA et les individus AT ou TT (effet dominant). Il peut également détecter les différences entre individus génotypés AA ou AT et ceux génotypés TT (effet récessif).

Les ajustements

Lorsqu'un modèle est éloigné de la réalité, les valeurs des paramètres qui sont les plus en accord avec les observations tendront à être celles qui annulent les liens décrits par le modèle. Aussi, si l'on connaît déjà certains facteurs de risque d'une maladie, même si ce ne sont pas ceux qui nous intéressent, il est important de les inclure dans les modèles d'explication de la maladie afin que les modèles et donc les hypothèses soient les plus susceptibles d'être conformes à la réalité. On dira alors que l'on ajuste les modèles pour ces facteurs de risque. Par exemple, l'âge et le sexe sont souvent des facteurs de risque importants pour les maladies complexes et on les utilise ainsi souvent pour ajuster les différents modèles que l'on teste.

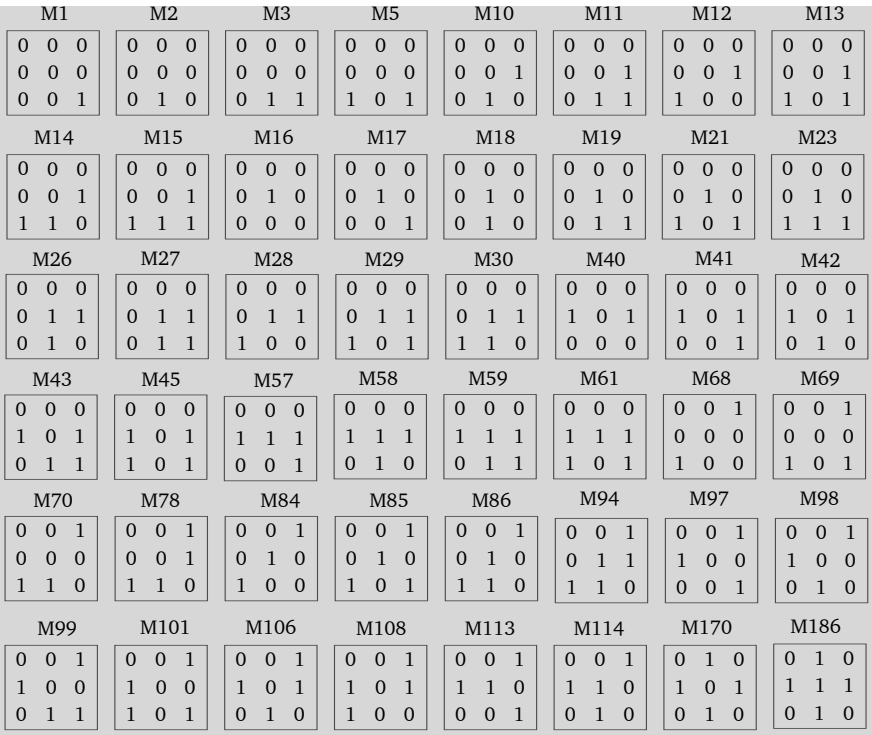
Terme d'interaction

Détecter des interactions entre gènes

Définition Une interaction représente une action réciproque entre plusieurs éléments. D'un point de vue biologique, dès que plusieurs molécules entrent en contact, inévitablement, elles interagissent. D'un point de vue statistique

cependant, pour qu'il y ait interaction, il faut que les effets de certaines variables sur le caractère étudié dépendent des valeurs d'autres variables. Par exemple, pour qu'il y ait interaction entre des polymorphismes génétiques, il faut que les effets de certains allèles de ces polymorphismes soient modifiés selon la présence ou non de certains autres allèles. Dans cette thèse, c'est la définition statistique qui va nous intéresser, car elle décrit les interactions que nous pouvons détecter et qui ont un réel impact sur le caractère que l'on étudie.

Types d'interactions entre polymorphismes Il existe de nombreuses façons selon lesquelles ne serait-ce que deux polymorphismes peuvent interagir entre eux, y compris des SNPs qui n'ont que deux allèles possibles. Par exemple, considérons une interaction entre deux SNPs. Le premier, au locus 1, peut prendre les allèles A et T, le second au locus 2, les allèles C et G. Au locus 1, un individu peut donc avoir les génotypes AA, AT ou TT et au locus 2, il peut avoir les génotypes CC, CG ou GG. Il se peut alors que le génotype TT au locus 1 entraîne la maladie, sauf lorsque le génotype GG au locus 2 est présent. Mais il se peut aussi qu'il entraîne la maladie uniquement lorsque ce génotype GG est présent ou encore lorsque c'est le génotype CG qui est présent. En tout il y a 48 types de modèles uniques à pénétrance complète[a] impliquant une interaction [35]. Or, plus probablement, dans les maladies complexes, la présence d'un allèle ou de plusieurs allèles n'est pas responsable de la maladie mais simplement augmente le risque d'en être atteint, ce qui augmente considérablement le nombre de modèles d'interactions possibles.



Adapté de : Evans et al. (2006) *Two-stage two-locus models in genome-wide association*, PLoS Genetique

Les 48 types de modèles d'interaction à pénétrance complète. Chaque ligne correspond au génotype pour le premier locus et chaque colonne, à celui pour le second locus, avec les 1 représentant la présence (ou absence) de la maladie. En tout il y a $2^9 = 512$ possibilités, mais du fait de symétries et de modèles sans interaction, 48 sont des modèles d'interaction uniques.

Il y a un peu plus de trois ans, Heather Cordell publiait une revue de la littérature sur les méthodes de détection d'interactions entre gènes impliqués dans les maladies humaines [24]. Elle y disait notamment que le sujet était très vaste et qu'il était nécessaire de passer par plusieurs revues de littérature pour avoir une vision d'ensemble des méthodes existantes. Depuis, il y a eu une explosion de nouvelles méthodes et il n'est clairement pas possible de ne serait-ce que de donner une vue de l'ensemble des méthodes de détection d'interactions gène-gène [112]. On peut cependant lister quelques-unes des méthodes des plus populaires :

La méthode « classique » La méthode que je qualifierai de « classique », est celle que nous avons utilisée. Elle consiste à construire un modèle de régression linéaire ou logistique (suivant si le phénotype est quantitatif ou binaire), dans lequel on inclut un terme d'interaction, le plus souvent entre deux polymorphismes. On estime alors le paramètre associé à l'interaction avant de

déterminer si ce paramètre peut-être considéré comme étant différent de zéro : On se demande si l'estimation obtenue aurait pu arriver si l'interaction n'avait aucun effet sur le phénotype étudié.

Les méthodes « random forests » Les méthodes du type random forest consistent à chercher des arbres de décisions. Un premier polymorphisme est sélectionné aléatoirement et sépare les individus en deux groupes suivant leur génotype. Pour chaque groupe, un second polymorphisme est sélectionné qui va séparer chaque groupe en deux et ainsi de suite, comme un arbre, jusqu'à avoir un certain nombre de branches. Cette opération est réalisée un grand nombre de fois de manière à tester un grand nombre d'arbres, dans le but de trouver un arbre qui révèle des groupes d'individus aux phénotypes bien différents [74].

La méthode « hypercube » Les méthodes du type « hypercube » consistent à disposer les individus dans un espace ayant autant de dimensions que de polymorphismes, puis à créer des hypercubes en fixant aléatoirement des contraintes sur certaines des dimensions de l'espace (en fixant par exemple le génotype d'un polymorphisme). Le but de la méthode est de trouver des hypercubes qui contiennent des individus ayant des phénotypes différents de ceux (les individus) qui sont en dehors de l'hypercube.

De nombreuses méthodes (comme les deux dernières) évaluent ensuite la pertinence des interactions identifiées en faisant ce que l'on appelle de la validation croisée. C'est-à-dire que les interactions sont recherchées sur une partie des données, puis leur validité est testée sur le reste des données. Pour ce qui est de comparer des méthodes, on peut simuler des données sous différentes hypothèses et tester les différentes méthodes sur ces données [3].

a. On dit qu'une maladie est à pénétrance complète lorsque les individus porteurs de la variation « causale » sont tous malades et qu'aucun individu non porteur n'est malade.

Codage de l'interaction Rappelons que l'on a opté pour un codage en 0,1,2 des génotypes des SNPs (qui permet de détecter des effets additifs, mais également des effets dominants et récessifs relativement forts). Un modèle d'interaction qui apparaît mathématiquement intuitif est le modèle où les effets des deux SNPs sont multipliés. C'est aussi un modèle relativement intuitif d'un point de vue génétique et qui se trouve être là encore un compromis entre les modèles multiplicatifs d'effets dominants (figure 4.6 b)) et récessifs (figure 4.6 c)), deux autres modèles intuitifs d'un point de vue génétique.

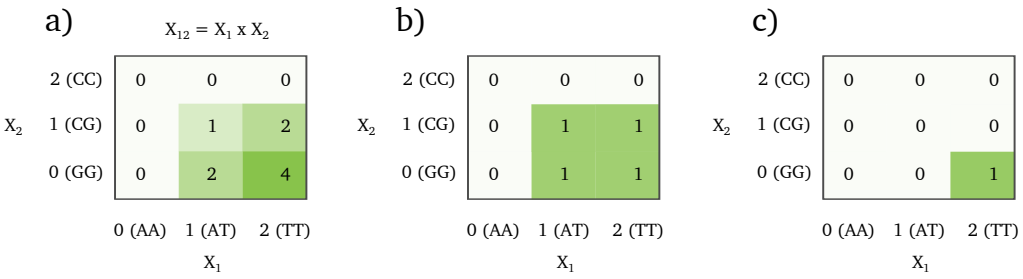


FIGURE 4.6 – a) le modèle d'interaction multiplicatif lorsque les effets marginaux sont additifs b) modèle multiplicatif lorsque les effets marginaux sont dominants c) modèle multiplicatif lorsque les effets marginaux sont récessifs

4.3.2 Modèles utilisés

Finalement, lorsque nous avons tenté de détecter des phénomènes d'interaction SNP–SNP, nous avons utilisé les modèles multiplicatifs (tels que décrits précédemment) linéaires (pour les expressions des gènes) ou logistiques (pour le phénotype « malade/non malade »), avec un codage additif des génotypes, ajustés sur les génotypes marginaux en plus des ajustements classiques, tels que l'âge ou le sexe. Pour les recherches d'associations simples entre les génotypes et la variable à expliquer, nous avons utilisé les mêmes types de modèles en prenant soin d'exclure le terme d'interaction.

Modèle linéaire

$$\text{phénotype} = a_0 + a_1 SNP_1 + a_2 SNP_2 + a_3 SNP_1 \times SNP_2 + \text{ajustements} + \epsilon$$

Modèle logistique

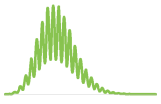
$$P(\text{malade}) = \frac{(\exp a_0 + a_1 SNP_1 + a_2 SNP_2 + a_3 SNP_1 \times SNP_2 + \text{ajustements})}{1 + \exp(a_0 + a_1 SNP_1 + a_2 SNP_2 + a_3 SNP_1 \times SNP_2 + \text{ajustements})}$$

où SNP_1 et SNP_2 représentent les génotypes (codés 0,1,2) aux locus 1 et 2 respectivement. Notons qu'il est courant de transformer les paramètres du modèle logistique en odds-ratios (voir encadré) pour faciliter leur interprétation.

4.3.3 Estimation des paramètres

Maximum de vraisemblance

La technique du maximum de vraisemblance [36] est donc probablement la méthode statistique la plus connue et la plus utilisée pour estimer des paramètres d'un modèle. Comme expliqué précédemment, elle consiste à rechercher les valeurs des paramètres qui rendent les observations les plus probables possibles d'après



le modèle supposé. Classiquement, pour ce faire, on établit la vraisemblance de nos observations d'après le modèle utilisé, c'est-à-dire la probabilité d'apparition de nos observations en fonction des paramètres du modèle. Puis, on recherche les paramètres la maximisant en égalant la dérivée de cette vraisemblance, ou plus souvent, son logarithme à zéro.

Estimation des paramètres d'un modèle linéaire

Supposons un modèle de régression linéaire tel que défini auparavant (voir section 4.3) entre une variable Y (qui sera classiquement l'expression d'un gène dans ce document) et des variables X_1, X_2 (typiquement, des génotypes pour un SNP, codés additivement) et d'ajustement (l'âge, le sexe), où ϵ représente une variable aléatoire supposée normale et de moyenne nulle. On a alors pour chaque individu :

$$P(Y = y_i) = P(\epsilon = y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i} + a_3 x_{1i} x_{2i} + \text{ajustements}))$$

$$= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i} + a_3 x_{1i} x_{2i} + \text{ajustements}))^2}{2\sigma_i^2} \right)$$

où σ_i est la variance de ϵ_i. Dans la suite, on simplifiera l'écriture en appelant x_i le vecteur $(1, x_{1i}, x_{2i}, x_{1i}x_{2i}, \text{ajustements})$ et a le vecteur $(a_0, a_1, a_2, a_3, \text{ajustements})$ permettant d'avoir $ax_i^t = a_0 + a_1 x_{1i} + a_2 x_{2i} + a_3 x_{1i} x_{2i} + \text{ajustements}$.

Si l'on considère que les observations sont indépendantes (les individus ne sont pas apparentés), on peut faire le produit des probabilités de chacune des observations pour calculer la probabilité parmi tous les échantillons possibles, d'obtenir notre échantillon d'observations. C'est ce que l'on appelle la vraisemblance des observations :

$$V(Y/X) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - ax_i^t)^2}{2\sigma_i^2} \right)$$

en passant au logarithme, on obtient la log-vraisemblance de l'échantillon qui consiste alors en une somme de termes, plus facile à manipuler :

$$\log V(Y/X) = \sum_i^n \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \right) - \frac{(y_i - ax_i^t)^2}{2\sigma_i^2}$$

La fonction logarithme étant strictement croissante, le maximum de la vraisemblance est donc aussi le maximum de la log-vraisemblance et l'on peut estimer les valeurs des paramètres qui permettent d'atteindre ce maximum en annulant la dérivée de la

log-vraisemblance. La dérivée par rapport à a_j (j valant 0, 1, ..., k) est :

$$\frac{\partial \log V(Y/X)}{\partial a_j} = \sum_i^n \frac{x_j^t(y_i - ax_i^t)}{\sigma_i^2}$$

et si pour tous les individus, $\sigma_i = \sigma$, alors, celle-ci s'annule lorsque

$$\sum_i^n x_j^t(y_i - ax_i^t) = 0$$

On obtient ainsi un système de k+1 équations linéaires résoluble analytiquement et nous donnant les estimations des paramètres du modèle.

Estimation des paramètres d'un modèle logistique

L'estimation des paramètres d'un modèle logistique est similaire. Si l'on appelle Y la variable binaire (malade/non malade par exemple), avec $y_i = 1$ si l'individu est malade et $y_i = 0$ sinon, si de plus on appelle X1 et X2 les variables génotypiques, codées additivement, pour les SNPs 1 et 2 et que l'on suppose un lien logistique entre le risque de survenue de la maladie et les génotypes pour ces deux SNPs ainsi que leur interaction, alors, la probabilité qu'un individu i soit malade peut s'écrire :

$$P(Y=1) = \frac{\exp(a_0 + a_1x_{1i} + a_2x_{2i} + a_3x_{1i}x_{2i} + \text{ajustements})}{1 + \exp(a_0 + a_1x_{1i} + a_2x_{2i} + a_3x_{1i}x_{2i} + \text{ajustements})}$$

et la probabilité que l'individu i ne soit pas malade est :

$$P(Y=0) = \frac{1}{1 + \exp(a_0 + a_1x_{1i} + a_2x_{2i} + a_3x_{1i}x_{2i} + \text{ajustements})}$$

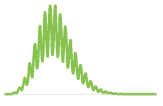
où a_0, a_1, a_2 et a_3 sont les paramètres liés à l'effet global, le génotype du SNP 1, celui du SNP 2 et à l'interaction entre ces deux génotypes respectivement.

Comme précédemment, on simplifiera l'écriture en appelant x_i le vecteur $(1, x_{1i}, x_{2i}, x_{1i}x_{2i}, \text{ajustements})$ et a le vecteur $(a_0, a_1, a_2, a_3, \text{ajustements})$ permettant d'avoir $ax_i^t = a_0 + a_1x_{1i} + a_2x_{2i} + a_3x_{1i}x_{2i} + +\text{ajustements}$

Si les observations sont indépendantes on obtient la vraisemblance en faisant le produit des probabilités de chacune des observations :

$$V(Y/X) = \prod_i^n \left(\frac{\exp(ax_i^t)}{1+\exp(ax_i^t)}\right)^{y_i} \left(\frac{1}{1+\exp(ax_i^t)}\right)^{1-y_i}$$

$$= \prod_i^n \frac{\exp(ax_i^t)^{y_i}}{1+\exp(ax_i^t)}$$



la log-vraisemblance est alors :

$$\log V(Y/X) = \sum_{i}^{n} y_i a x_i^t - \log\left(1 + \exp(a x_i^t)\right)$$

et sa dérivée par rapport à a_j :

$$\frac{\partial \log V(Y/X)}{\partial a_j} = \sum_{i}^{n} y_i x_{ij} - x_{ij} \frac{\exp(a x_i^t)}{1 + \exp(a x_i^t)}$$

En cherchant analytiquement si cela est possible ou numériquement sinon, les valeurs des paramètres qui annulent cette dérivée, on obtient les estimations du maximum de vraisemblance des paramètres de ce modèle logistique.

4.4 Distribution de la statistique

Principe général

Lorsque l'on a combiné les observations de sorte à avoir une statistique qui discrimine bien les hypothèses, il faut ensuite savoir si la valeur observée de la statistique est cohérente avec l'hypothèse H0. Pour y arriver on a besoin de connaître le genre de valeurs que peut prendre la statistique sous H0, c'est-à-dire qu'il faut connaître sa distribution. Si la valeur observée se trouve dans les (disons 5 % de) valeurs les plus extrêmes de la distribution de la statistique sous H0, alors, cela nous poussera plutôt à rejeter cette hypothèse H0.

Test de Wald

On a vu que l'estimation par maximum de vraisemblance consistait en la résolution d'un système d'équations impliquant les sommes des variables du modèle. Comme en général, on ne connaît pas la distribution exacte de ces variables, il n'est pas possible d'en déduire la distribution exacte des estimations. En revanche si le nombre d'observations est suffisamment important, on peut se servir du fait que les estimations sont calculées à partir d'une somme de variables. Par le théorème central limite, on peut les approcher par une distribution normale. Il y a trois tests généraux principaux qui sont couramment utilisés et qui utilisent ces estimations des paramètres comme statistiques : le test de Wald, le test du rapport de vraisemblance [131] et le test du score. Ces trois tests utilisent ce type d'approximation. Le test classique effectué dans la majorité des logiciels statistiques lorsque l'on effectue une régression linéaire ou logistique (c'est-à-dire lorsque l'on estime les paramètres d'un modèle linéaire ou logistique) est le test de Wald. C'est aussi celui-ci que j'utilise dans ce document. Si l'on considère le cas d'un paramètre unique pour simplifier, il

consiste à utiliser $\hat{a}$ [1], l'estimation du paramètre a dans la statistique

$$\frac{\hat{a} - a_{H0}}{var(\hat{a})^{1/2}}$$

où a_{H0} correspond à la valeur du paramètre sous l'hypothèse H0. Pour nous, l'hypothèse H0 sera toujours l'absence de lien et on prendra donc $a_{H0} = 0$ alors que $var(\hat{a})^{1/2}$ représente la racine carrée de la variance du paramètre estimé. Abraham Wald montra que cette statistique avait asymptotiquement (c'est-à-dire lorsque le nombre d'observations tend vers l'infini) une distribution normale de moyenne 0 et de variance 1 [126]. Lorsque l'on a k paramètres à estimer, c'est le carré de cette statistique qui est en général calculé et qui est distribué selon une loi du χ^2 à k degrés de liberté [2]. Il suffit alors de calculer les probabilités qu'une variable avec cette distribution obtienne des valeurs plus extrêmes que celles que l'on observe, pour savoir si les variables associées aux paramètres estimés peuvent être considérées ou non comme étant associées au caractère que l'on essaye d'expliquer.

Calcul de la statistique de Wald

Nous avons déjà vu comment l'on pouvait estimer les valeurs des paramètres de nos modèles. Il nous reste donc à voir comment on peut calculer leur variance, afin de déterminer la valeur de la statistique de Wald. Une méthode classique pour y arriver est de passer par le calcul de l'information de Fisher I, associée au paramètre a, qui peut être défini sous certaines conditions assez souples par :

$$I(a) = -\mathrm{E}\left[\frac{\partial^2}{\partial a^2}\log V(Y/X)\middle| a\right]$$

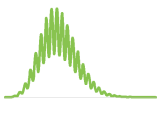
où E désigne l'espérance. Son inverse est la borne de Cramér-Rao [94]. C'est la plus petite variance que l'on peut atteindre pour un paramètre estimé. Lorsque le nombre d'observations est important, l'estimation par maximum de vraisemblance tend vers cette borne. C'est donc la variance que l'on va essayer de calculer ici :

En reprenant les notations et les calculs introduits dans la section 4.3.3 lors de l'estimation des paramètres du modèle logistique avec interaction, on a donc la dérivée par rapport à a_j de la log-vraisemblance des observations :

$$\frac{\partial \log V(Y/X)}{\partial a_j} = \sum_i^n y_i x_{ij} - x_{ij}\frac{\exp(a x_i^t)}{1+\exp(a x_i^t)}$$

1. Il est courant de mettre un chapeau sur le nom d'un paramètre pour représenter une estimation de ce paramètre.

2. Lorsqu'une variable suite une loi normale de centre 0 et de variance 1, le carré de cette variable suit une loi du χ^2 a un degré de liberté.



On peut alors de nouveau calculer sa dérivée par rapport à a_k :

$$\frac{\partial^2 \log V(Y/X)}{\partial a_k \partial a_j} = -\sum_i^n x_{ij} x_{ik} \frac{\exp(a x_i^t)}{(1+\exp(a x_i^t))^2}$$

pour laquelle, l'espérance n'est autre que l'opposé de l'information de Fisher.

Il est alors possible de calculer la borne inférieure (la borne asymptotique) de la variance des estimations des paramètres du modèle en prenant l'inverse de l'information de Fisher et l'on peut alors construire un des éléments de la statistique de Wald asymptotique :

$$\hat{a} \sqrt{\mathrm{E}\left(\frac{\partial^2 \log V(Y/X)}{\partial a_k \partial a_j}\right)}$$

Qui suit donc une loi normale de moyenne nulle et de variance 1, sous H0.

4.5 Quelques tests qui ne sont pas basés sur des modèles

4.5.1 Le test de Levene

À quoi sert-il ?

Le test de Levene est un test permettant de détecter des différences de variances entre plusieurs groupes. Guillaume Paré a suggéré que des différences de variances entre génotypes pour un phénotype quantitatif, pouvaient être un indicateur de la présence d'une interaction entre ces génotypes et le phénotype [88]. Nous avons utilisé ce test pour pondérer (voir chapitre 5) nos résultats lors de notre recherche de phénomènes d'interactions entre polymorphismes liés aux microARNs (voir chapitre 8).

La statistique du test

La statistique L, du test de Levene se base sur les valeurs

$$z_{ij} = \left| y_{ij} - \frac{\sum_j y_{ij}}{n_i} \right|$$

où y_{ij} représente la valeur de la j-ieme observation du groupe i pour le phénotype étudié et n_i le nombre d'observation dans ce même groupe i. $\frac{\sum_j y_{ij}}{n_i}$ représente donc la moyenne du phénotype pour le groupe i et z_{ij}, l'écart absolu de la j-ieme observation du phénotype, à la moyenne du groupe.

L'idée du test est que si les variances sont différentes entre les différents groupes (hypothèse H1), ces écarts z_i devraient varier plus fortement entre les groupes qu'au

sein des groupes, d'où la statistique du test de Levene :

$$L = \frac{\frac{1}{k-1}\sum_i n_i(\sum_j z_{ij} - \sum_i \sum_j z_{ij})^2}{\frac{1}{\sum_i (n_i-1)}\sum_i \sum_j (z_{ij} - \sum_j z_{ij})^2}$$

avec k, le nombre de groupes. Comme nous travaillons sur des génotypes pour des SNPs, pour nous, k sera égal à 3.

La distribution de L sous H0

Levene montra que si le phénotype y suit une distribution normale dans chaque groupe, alors, lorsqu'il n'y a pas de différences de variances entre les groupes (lorsque l'on est sous H0), la statistique L suit une distribution de Fisher Snedecor à $k-1$ et $\sum_i (n_i - 1)$ degrés de liberté [69](nous sommes donc dans le cas d'un test où l'on déduit la distribution exacte de la statistique à partir de la distribution des observations). Le graphique 4.7 montre la distribution d'une telle statistique lorsque $k = 3$ et le nombre total d'observations est de 1467 (nombre d'individus utilisé dans le chapitre 8). Notons que bien que ce test se base sur une hypothèse de normalité du phénotype étudié, Howard Levene le décrivit cependant comme relativement robuste au non respect de cette hypothèse [69].

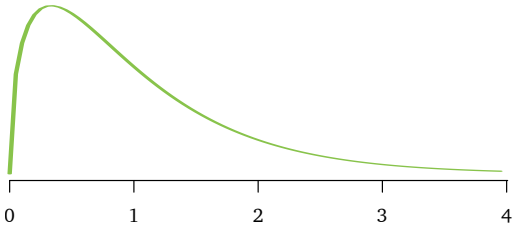
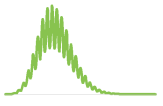


FIGURE 4.7 – Loi de distribution de Fisher-Snedecor à 3 et 1467 degrés de liberté.

4.5.2 Le test d'Hardy-Weiberg

À quoi sert-il ?

Dans une population, si un polymorphisme a deux allèles (un SNP par exemple) A et a avec des fréquences respectives f et $(1-f)$, alors en supposant que les couples de parents se forment aléatoirement, on s'attendrait pour un individu donné, à ce que ses allèles lui aient été transmis de manière indépendantes et donc que sa probabilité d'avoir le génotype AA soit f^2, Aa, $2f(1-f)$ et aa, $(1-f)^2$. C'est ce que l'on appelle l'équilibre d'Hardy-Weinberg [46]. Il peut arriver parfois que certains polymorphismes ne semblent pas respecter cet équilibre. Une telle situation pourrait s'expliquer si le polymorphisme en question est sujet à sélection, par exemple, si il a un effet récessif fort sur une maladie mortelle provoquant une sous-représentation



des individus homozygotes pour l'allèle à risque. Cependant il est souvent bien plus probable que ce déséquilibre provienne d'une erreur lors du génotypage. Aussi, souvent, les épidémiologistes tendent à ne pas garder les polymorphismes qui ne vérifient pas l'équilibre d'Hardy-Weinberg dans les analyses. Le test d'Hardy-Weinberg est un test statistique visant à détecter si un polymorphisme s'écarte de cet équilibre.

La statistique du test

Il existe plusieurs statistiques pour tester l'écart à l'équilibre d'Hardy-Weinberg. La statistique H la plus utilisée est assez intuitive puisqu'elle se base sur les écarts relatifs entre les effectifs observés et attendus :

$$H = n\left(\frac{(f(AA)-f^2)^2}{f^2} + \frac{(f(Aa)-2f(1-f))^2}{2f(1-f)} + \frac{(f(aa)-(1-f)^2)^2}{(1-f)^2)^2}\right)$$

où n est le nombre d'individus et $f(AA)$, $f(Aa)$, $f(aa)$ sont les fréquences observées de chacun des phénotypes. Les valeurs de la statistique pour les SNPs qui ne sont pas à l'équilibre d'Hardy-Weinberg devraient être plus élevées que pour ceux qui sont à l'équilibre.

La distribution de la statistique sous H0

Karl Pearson montra que sous l'hypothèse H0, cette statistique suit asymptotiquement (on n'est donc pas dans le cas précédent d'une distribution exacte) une distribution du χ^2 à 1 degré de liberté (on enlève 2 degrés de liberté à cette somme de 3 termes du fait de l'utilisation dans la distribution théorique de la fréquence observée f de l'allèle A, en plus du nombre n d'individus) [89]. Ce résultat se base sur des approximations similaire au théorème central limite. Le graphique 4.8 montre la distribution d'une telle statistique. Lorsque le nombre d'observations est important, ce qui est notre cas avec des effectifs de plusieurs centaines d'individus, la distribution du χ^2 à 1 degré de liberté est une très bonne approximation de la distribution de H.

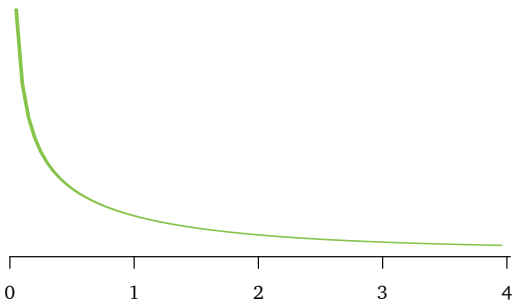
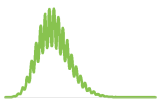


FIGURE 4.8 – Loi de distribution du χ^2 à 1 degré de liberté.

Logiciels de travail

Pour la grande majorité des calculs et analyses statistiques effectués dans mes travaux de recherches, j'ai utilisé le logiciel R [92]. Il m'est cependant également arrivé d'utiliser les logiciels PLINK [91] pour certaines recherches d'interactions gourmandes en temps de calcul et Thesias [119] lorsqu'il j'ai eu à manipuler des haplotypes.



Chapitre 5

La gestion des tests multiples

Hoagie : *Doc, can't you just send Bernard ?*

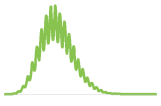
Dr. Fred : *No, you must all go to increase the odds that one of you will make it there alive.*

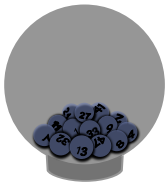
Day of the Tentacle
`http://lucasartsoldgames.free.fr/dott/`

Le chapitre précédent a permis d'introduire le principe du test statistique. En épidémiologie génétique cependant, on ne cherche généralement pas à tester nos hypothèses concernant l'effet d'un seul polymorphisme dans la variation d'un phénotype, mais on formule au contraire une multitude d'hypothèses d'effets. Ceci nous amène à réaliser une multitude de tests, dont les résultats ne peuvent pas être interprétés comme si ils étaient uniques. Ce chapitre vise à expliquer comment on peut gérer ces tests multiples.

L'analogie du loto

Supposons qu'une personne joue au loto les numéros 16, 27, 42, 47, 49. Il y a très peu de chances pour que sans tricherie (sous H0) ces numéros sortent au tirage (bien moins de 5 % de chances). Aussi la personne sera extrêmement surprise (probablement agréablement) si ils sortent : de son point de vue, la p-value associée au tirage est alors extrêmement faible et il est possible qu'elle associe un tel tirage à une action divine ou quelque-chose du genre. Si elle est statisticienne, elle aura simplement tendance à penser que l'on est sous H1.





Résultat du tirage du SUPER LOTO du vendredi 13 Juillet 2012

42 27 47 16 49 8

	Nombre de grilles gagnantes	Gain par grille gagnante
5 bons numéros + N° CHANCE	0	Pas de gagnant
5 bons numéros	5	117 403,20 €
4 bons numéros	989	1 277,40 €
3 bons numéros	46571	11,70 €
2 bons numéros	687845	5,60 €
N° CHANCE gagnant	1154692	grille à 2 € remboursée

Pourtant, il y a fort à parier que si cette personne joue d'autres numéros, elle ne sera pas vraiment surprise d'apprendre dans le journal que quelqu'un a trouvé les bons numéros : il y a tellement de gens qui jouent qu'il y en a forcément certains qui ont de la chance et même beaucoup de chance... et si elle est statisticienne, elle aura tendance à penser qu'elle n'a pas assez d'éléments pour rejeter H0.

- **Effectuer plusieurs tests augmente les probabilités de voir les combinaisons rares**, si bien qu'il arrive souvent lorsque l'on effectue un très grand nombre de tests, que l'on ne parvienne plus à bien discriminer les hypothèses. Cela nous emmène à essayer de
- **combiner nos résultats** entre ou au sein de nos études.
- **sélectionner ou pondérer** les tests effectués.

5.1 Les corrections pour tests multiples

La p-value est la probabilité sur **UN** test et sous H0, d'observer une valeur de statistique plus extrême que celle réellement observée mais ce n'est pas la probabilité sous H0 d'observer sur n tests, une valeur de statistique plus extrême qu'une de celles véritablement observées. Or c'est cette seconde définition dont on aimerait avoir une mesure et que l'on appellera par la suite FWER pour Family-Wise Error Rate. Classiquement, si sur n tests, on n'a que 5 % de chances sous H0, d'observer une valeur de statistique plus extrême que celle que l'on a observée sur le test k (FWER<5 %), alors on pourra être relativement confiant que l'on n'est pas sous H0 sur le k-ieme test. Dans le cas contraire, il est difficile de se prononcer car cela veut dire que la statistique que l'on observe peut probablement être arrivée juste par chance (sous H0). L'idée des corrections pour tests multiples consiste à définir de nouveaux seuils pour la p-value, qui nous assurent d'avoir une valeur de FWER assez

faible pour avoir une faible probabilité de se tromper lorsque l'on déclare un test significatif.

5.1.1 La correction de Bonferroni

La correction la plus simple et l'une des plus couramment utilisée est la correction de Bonferroni [14]. Elle repose sur le fait que si on choisit un seuil $\alpha = 0.05/n$ où n est le nombre total de tests réalisés, alors on s'assure que sous H0,

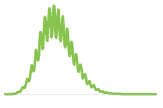
$$
\begin{aligned}
FWER &= P(\text{une } p\text{-value} < \alpha) \\
&= P(p\text{-value}_1 < \alpha \text{ ou } \ldots \text{ ou } p\text{-value}_n < \alpha) \quad (1) \\
&\leq \sum_{i=1}^{n} P(p\text{-value}_i < \alpha) \quad (2) \\
&= n\alpha \\
&= 0.05
\end{aligned}
$$

Ainsi, la valeur de FWER restera inférieure à 5 %. Cette démonstration repose sur l'inégalité du passage de (1) à (2) qui peut parfois s'avérer relativement forte, notamment lorsque les tests sont corrélés positivement. Quelques méthodes alternatives ont été proposées pour réduire cette correction tels que les corrections de Sidak [107] ou de Bonferroni-Holm [49].

5.1.2 Le taux de faux positifs (FDR) comme alternative au FWER

Dans certaines situations, il peut arriver que l'on soit sous l'hypothèse alternative H1 dans un grand nombre des tests effectués. Par exemple, certains gènes appelé facteurs de transcription sont impliqués dans le processus de transcription de l'ADN en ARN, si bien que la variation de leur expression impacte les expressions d'un très grand nombre d'autres gènes. Si l'on testait l'association entre l'expression d'un de ces facteurs de transcription et les expressions de n autres gènes, on serait effectivement sous l'hypothèse H1 d'une association, sur une bonne partie des n tests.

Dans une telle situation, s'assurer que la probabilité de se tromper soit faible en déclarant un des tests significatifs lorsque tous sont sous H0 paraît peu pertinente car au contraire nous avons beaucoup de tests sous H1. On préférerait simplement s'assurer de ne pas avoir trop de faux positifs, c'est à dire de ne pas trop nous tromper en déclarant un des tests significatifs et ce, quelque soit l'hypothèse sous laquelle on se trouve. Benjamini et Hochberg [10] proposèrent pour ce faire d'estimer ce taux



de faux positifs (couramment appelé FDR pour False Discovery Rate), en calculant le ratio entre le nombre attendu de tests significatifs par erreur et le nombre de tests k significatifs :

$$\widehat{FDR} = \frac{\text{Nombre attendu de tests significatifs par erreur}}{\text{Nombre de tests déclarés significatifs}} = \frac{n\alpha}{k}$$

où n est le nombre total de tests, α est le risque d'erreur sous H0 et k est le nombre de tests effectivement déclarés significatifs. On peut alors

- choisir comme seuil de significativité pour chaque test la plus grande valeur de α en dessous de laquelle le taux de faux positifs est inférieur au seuil α_{FDR} que l'on s'est fixé (typiquement 5 %).
- pour chaque test, estimer la q-value : le taux de faux positifs parmi tous les tests ayant des p-values plus petites que celle du test. La q-value peut alors être interprétée comme la p-value, à savoir que tous les tests ayant une q-value plus petite que α_{FDR} sont déclarés significatifs.

Lorsque le nombre de tests sous H1 est faible, le seuil de significativité obtenu par le FDR aura tendance à se rapprocher du seuil de Bonferroni. Lorsqu'il y a beaucoup de tests sous H1 en revanche, ce seuil aura tendance à être bien moins stringent.

En général, quelque soit la correction, il n'est pas rare que l'on ne soit plus capable de détecter les observations qui ne sont pas sous H0, par peur de se tromper si on les affirme sous H1. C'est le problème du manque de puissance.

5.1.3 La puissance

Définition

La puissance d'un test statistique est la probabilité de rejeter l'hypothèse H0 (ie : d'avoir une statistique plus extrême que ce que l'on attendrait) lorsque l'on n'est effectivement pas sous H0.

Pour l'analogie du loto précédente, la puissance du test serait la probabilité de détecter une quelconque tricherie au tirage du loto (ie : d'être très surpris du tirage) lorsqu'il y a réellement eu tricherie. Lorsque l'on effectue de nombreux tests ou qu'il y a de nombreuses personnes qui jouent au loto, on sait que certaines statistiques seront très extrêmes (certaines personnes auront beaucoup de chances) et il faudra que la statistique soit très extrême (la tricherie soit très flagrante) pour que l'on arrive à rejeter H0 (suspecter une tricherie). Ainsi, notre puissance de détection diminue lorsque l'on augmente le nombre de tests, car du fait des corrections pour tests multiples cela diminue le seuil.

Calcul de la puissance d'un test

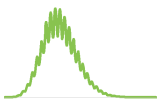
Lorsque l'on connaît la distribution d'une statistique de test sous H0, on peut trouver les valeurs seuil au delà desquelles on déclarera le test significatif. Si en plus on connaît la distribution sous H1, on peut alors calculer la probabilité que cette statistique soit significative lorsque l'on est sous H1, c'est-à-dire, la puissance du test. Abraham Wald montra qu'indépendamment des hypothèses, la statistique de Wald suit asymptotiquement une distribution normale. On peut donc calculer la puissance que l'on a de détecter une interaction dans nos modèles linéaires et logistiques. Cela m'a servi à déterminer notre puissance de détection de certains phénomènes d'interaction (voir chapitre 7).

La statistique de Wald W est significative lorsque sa valeur se trouve être plus extrême que ce que l'on attendrait et ainsi la p-value plus faible que le risque α que l'on s'est fixé. Comme la statistique de Wald est distribuée normalement, on s'attendrait à ce que la valeur observée w se situe vers le centre de la distribution et l'on déclarera le test significatif si elle se trouve dans les queues de la distribution, plus précisément, si :

$$|w| > \phi^{-1}(1-\alpha/2)$$

où ϕ est le fonction de répartition de la loi normale de moyenne 0 et de variance 1. Ainsi, si l'on appelle $\hat{\gamma}$ l'estimation d'un paramètre γ non nul sous H1, la puissance du test de Wald pour détecter la non-nullité de ce paramètre γ est :

$$\begin{aligned}
\text{Puissance} &= P(\text{statistique de Wald significative}|H1) \\
&= P(|w| > \phi^{-1}(1-\alpha/2)|H1) \\
&= P\left(\left|\frac{\hat{\gamma}}{var(\hat{\gamma})^{1/2}}\right| > \phi^{-1}(1-\alpha/2)\middle|H1\right) \\
&= P\left(\frac{\hat{\gamma}}{var(\hat{\gamma})^{1/2}} > \phi^{-1}(1-\alpha/2)\middle|H1\right) + P\left(\frac{\hat{\gamma}}{var(\hat{\gamma})^{1/2}} < -\phi^{-1}(1-\alpha/2)\middle|H1\right) \\
&\approx P\left(\frac{\hat{\gamma}}{var(\hat{\gamma})^{1/2}} > \phi^{-1}(1-\alpha/2)\middle|H1\right) \text{ , si l'on suppose } \gamma > 0 \\
&= P\left(\frac{\hat{\gamma}-\gamma}{var(\hat{\gamma})^{1/2}} > \phi^{-1}(1-\alpha/2) - \frac{\gamma}{var(\hat{\gamma})^{1/2}}\middle|H1\right) \\
&= P\left(\epsilon > \phi^{-1}(1-\alpha/2) - \frac{\gamma}{var(\hat{\gamma})^{1/2}}\middle|H1\right) \text{ où } \epsilon \sim N(0,1)
\end{aligned}$$



$$= 1 - \phi\left(\phi^{-1}(1-\alpha/2) - \frac{\gamma}{var(\hat{\gamma})^{1/2}}\right)$$

ce qui dépend du paramètre γ, du risque α, et du nombre d'observations par $var(\hat{\gamma})$

5.2 Comment augmenter la puissance de détection d'un test ?

- en améliorant le modèle ou la statistique du test pour être plus en phase avec la réalité
- en augmentant le nombre d'observations : plus on accumule des observations, plus la statistique s'éloignera de la distribution de l'hypothèse sous laquelle on ne se trouve pas.
- en acceptant un risque α plus important lorsque l'on est sous H0.
- en recherchant des effets forts plutôt que des effets faibles

Les deuxième et troisième points sont ceux sur lesquels on peut influer lorsque l'on effectue plusieurs tests, en combinant certains ou en effectuant des sélections ou des pondérations. C'est ce que l'on va voir maintenant.

5.2.1 Combiner des tests

Même principe que combiner des observations

Le résultat d'un test (par exemple la p-value) peut aussi être vu comme une observation et comme pour une seule observation, il est possible qu'un unique test ne permette pas de différencier nos hypothèses, mais que la combinaison de plusieurs tests le puisse. C'est ce que l'on tente de faire lorsque l'on combine des tests.

Combiner quoi et pourquoi ?

Le but de combiner des tests est d'augmenter la puissance de ces tests en augmentant le nombre d'observations. Lorsque la combinaison s'effectue entre tests d'une même étude, elle permet aussi de réduire le nombre de tests et par la même, de limiter la correction pour tests multiples à effectuer sur ces tests.

Combiner des tests identiques provenant d'études différentes Lorsque l'on n'a pas assez de puissance pour détecter un effet, il est tentant d'augmenter le nombre de sujets en utilisant les individus d'une autre étude. Cependant, il n'est souvent

pas souhaitable de former une seule grande étude à partir de plusieurs études indépendantes. D'une part, il y a souvent des différences de variables (utilisation de puces à ADN différentes par exemple) et d'autre part, les études entreprises séparément ne sont en général jamais construites exactement de la même manière. Il en résulte des populations parfois très différentes qui, étudiées ensemble, peuvent faire apparaître de fausses associations. Dans de tels cas, il est préférable d'effectuer les analyses sur chaque étude séparément, puis de combiner les tests. On appelle ce type de combinaison, des méta-analyses.

Combiner des tests différents au sein d'une même étude Parfois, il se peut que lorsqu'un test est significatif, on suspecte certains autres tests d'avoir de bonnes chances d'être aussi significatifs. Par exemple, si un SNP d'un gène est associé à un phénotype, il nous paraît relativement probable que d'autres SNPs du gène puissent l'être également. Dans une telle situation, il peut être une bonne idée de combiner les tests en question.

Comment combiner des tests ?

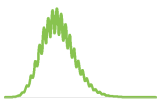
On peut combiner des tests de la même façon que l'on peut combiner des observations. Cependant, en règle général, on ne voudra pas combiner un grand nombre de tests ensemble et l'on ne pourra ainsi pas utiliser les théorèmes asymptotiques tels que le théorème central limite. Aussi on aura le choix entre effectuer des permutations pour estimer la distribution de la statistique construite ou trouver une statistique dont nous connaissons la distribution sous H0. L'un des avantages que l'on a lorsque l'on combine des tests par rapport à combiner des observations est que nous connaissons la distribution de la p-value du test sous H0. Cela nous permet de connaître la distribution exacte de certaines combinaisons de tests.

Distribution de la p-value

Rappelons la définition de la p-value : c'est la probabilité d'observer sous H0, une statistique plus extrême que celle que l'on a calculé sur nos données.

Supposons que l'on ait bien choisi notre statistique de test et que l'on soit sous H1. Alors, une statistique calculée sous H0 devrait avoir peu de chances d'être plus extrême que la statistique que l'on a calculée (qui est sous H1), autrement dit, la p-value de notre test statistique aura une plus grande probabilité d'être faible que d'être forte (voir figure 5.1).

Maintenant, si au lieu d'être sous H1, on est sous H0, une statistique calculée sous H0 devrait avoir autant de chances d'être plus extrême que celle que l'on a



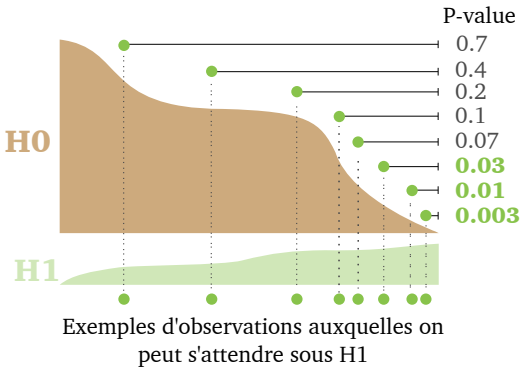


FIGURE 5.1 – Distribution de la p-value sous H1.

calculée. C'est à dire que la p-value de notre test devrait avoir les même probabilités d'être faible que forte. En fait, la p-value de notre test a une distribution uniforme (voir figure 5.2).

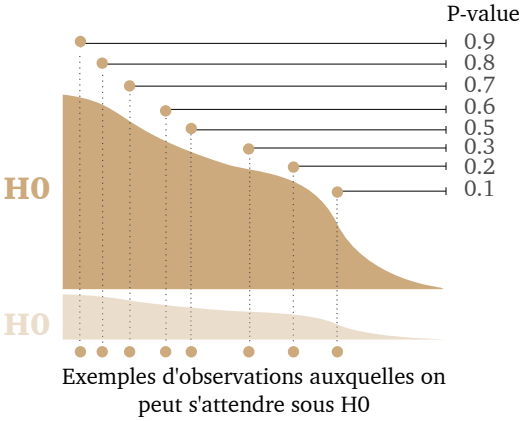


FIGURE 5.2 – Distribution de la p-value sous H0.

Transformation de la p-value

Plus généralement, ce raisonnement s'applique à n'importe quelle variable après transformation par sa fonction de répartition :

Si une variable X suit une distribution **D** et que sa fonction de répartition est F, alors, F(X) aura une distribution uniforme sur [0,1]

Ainsi, on peut obtenir n'importe quel type de distribution à partir d'une distribution uniforme sur [0,1], simplement en lui appliquant la fonction de répartition inverse de la distribution souhaitée. En particulier, comme sous H0, la p-value suit une loi uniforme sur [0,1], si on la transforme par une fonction de répartition inverse, on obtient une variable qui suit la distribution correspondante à cette fonction.

Par exemple, la fonction F définie par

$$F(x) = -2\ln(x)$$

est la fonction inverse de répartition de la distribution du χ^2 à 2 degrés de liberté Du coup, sous H0, $-2\ln(p\text{-value})$ suit une loi du χ^2 à 2 degré de liberté.

Combinaisons possibles

Comme on est capable de transformer une p-value en une variable suivant n'importe quelle distribution, il suffit désormais de connaître la distribution de certaines combinaisons de variables pour pouvoir appliquer cette combinaison à nos tests. Comme cela a pu être dit précédemment, les combinaisons les plus intuitives sont les sommes de variables. Il se trouve qu'il existe des distributions pour lesquelles nous connaissons la distribution de leur somme. Par exemple,

- la distribution normale : la somme de n variables indépendantes distribuées normalement, de moyennes μ_i et de variance σ_i^2 suit une loi normale de moyenne $\sum \mu_i$ et de variance $\sum \sigma_i^2$
- la distribution du χ^2 : la somme de n variables indépendantes distribuées selon une loi du χ^2 à k degrés de liberté, suit une loi du χ^2 à k*n degrés de liberté.
- la distribution gamma qui est une généralisation de la distribution du χ^2

Ainsi, par exemple, si l'on a n tests indépendants, alors, sous H0, la statistique

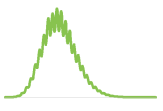
$$-2\sum_{i=1}^{n} ln(p\text{-value}_i)$$

a une distribution de χ^2 à 2n degrés de liberté. C'est la méthode de combinaison des p-values proposée par Fisher [37].

De même, si ϕ est la fonction de répartition de la loi normale de centre 0 et de variance 1, sous H0, la statistique

$$\frac{\sum_{i=1}^{n} w_i \phi^{-1}(p\text{-value}_i)}{\sqrt{\sum_{i=1}^{n} w_i^2}}$$

où les w_i sont les poids accordés à chaque test, a une distribution normale de centre 0 et de variance 1. Cette méthode introduite par Stouffer [114] est par exemple implémentée dans le logiciel METAL [132]. C'est aussi la méthode que j'ai utilisée pour effectuer la méta-analyse des études dans le chapitre 7.



En déterminant la distribution par permutations Enfin, l'on peut toujours estimer la distribution sous H0 en faisant des permutations et donc sans avoir besoin de connaître théoriquement la distribution : si F est une fonction de répartition, on peut estimer la distribution sous H0 de $\sum F^{-1}(p\text{-value})$, par des permutations (voir chapitre précédent).

5.2.2 Sélectionner et pondérer des tests

Une autre piste pour augmenter la puissance de nos tests consiste à effectuer une sélection sur ces tests. Cela permet de réduire le nombre de tests et les corrections pour tests multiples.

Sélection selon la *p*-value

Étant donné le grand nombre de tests qu'impliquent les recherches d'interaction, les chercheurs se limitent souvent à des recherches d'interaction entre sous-ensemble de SNPs, notamment les SNPs qui ressortent les plus significatifs en analyse simple, sans interaction. En fait, dans le chapitre 7, je montre que cette méthode de sélection n'est pas forcément optimale d'un point de vue statistique. En revanche, il est vrai que d'un point de vue biologique, s'il y a une interaction entre deux éléments, qui impacte une maladie, on peut alors s'attendre à ce que ces éléments pris séparément aient aussi une influence sur la pathologie. Par contre, s'il existe des phénomènes de pures interactions, sans apparents effets marginaux, ceux-ci ne pourront être détectés.

5.2.3 Pondération

Une autre méthode qui peut permettre de réduire les corrections pour tests multiples, ou en tout cas, faire ressortir certains tests qui ne seraient pas ressortis à cause d'une correction trop stringente, est la pondération. La pondération consiste à donner une certaine priorité à certains tests par rapport à d'autres. Lorsque l'on dispose des p-values de n test et que l'on attribue à chacun des tests i un poids w_i, alors, les p-values pondérées deviennent :

$$p\text{-value}'_i = \frac{p\text{-value}_i \times \sum w_j}{n w_i}$$

où n est le nombre de tests effectués. L'interprétation est alors la même que pour les p-values originales car les nouvelles p-values ont le même seuil que les anciennes. Ce qui change, c'est l'ordre d'importance des tests.

Le choix des pondérations se fait en fonction de critères du même ordre que ceux utilisés pour effectuer les sélections des tests. D'ailleurs, la méthode de sélection est

un cas particulier de la méthode de pondération, où la pondération est la même pour tous les tests sélectionnés et est nulle pour les autres. On peut imaginer de très nombreuses façons de pondérer les tests de manière à faire ressortir certaines hypothèses dans lesquelles on a plus confiance. Il faut cependant faire attention à utiliser des pondérations qui reposent sur de l'information indépendante des données afin de pouvoir garder les mêmes risques d'erreurs qu'avant pondération. Dans ce travail de thèse, j'ai utilisé divers types de pondérations tels qu'une pondération par la p-value du test marginal de Levene (voir chapitre précédent) ou par les fréquences alléliques.

5.2.4 La corrélation

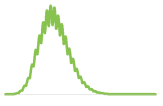
En statistique, la corrélation est une mesure de la liaison qu'il peut y avoir entre deux variables. Si à chaque fois que l'on fait évoluer une variable dans une direction, une autre variable a tendance à évoluer aussi dans le même sens, alors, ces deux variables sont corrélées. Le déséquilibre de liaison par exemple est une forme de corrélation. Il est d'ailleurs très courant d'utiliser le r^2 pour quantifier la corrélation entre deux variables.

La corrélation entre les SNPs

Une des spécificités des données génotypiques par rapport à d'autres types de données est que leurs variables (notamment les SNPs) sont potentiellement très nombreuses et parfois fortement corrélées entre elles. Cette corrélation vient du déséquilibres de liaisons existant entre les polymorphismes et de l'augmentation des capacités des puces à ADN qui a pour conséquence l'inclusion dans les études de SNPs parfois très proches les uns des autres. Cette corrélation est un avantage car elle nous permet de ne pas avoir à génotyper l'ensemble des polymorphismes de notre génome. Si un SNP n'est pas sur notre puce à ADN, il est probable qu'un SNP qui lui est proche y soit et puisse bien le représenter. Dans la suite du document lorsque j'utiliserai un SNP d'une puce pour représenter un SNP qui n'est pas disponible, j'appellerai ce SNP un proxySNP. Les proxySNPs sont identifiés grâce aux projets de reconstruction d'haplotypes, notamment les projets HapMap [52] et 1000 génomes [125], qui fournissent les informations de déséquilibre de liaison entre les polymorphismes.

La corrélation entre les tests

Les méthodes de corrections pour tests multiples décrites au début de ce chapitre, ainsi que les différentes techniques permettant d'augmenter la puissance globale de détection des effets recherchés, sont performantes lorsque les tests ne sont pas corrélés entre eux sous H0. Elles le sont cependant moins lorsqu'il y a beaucoup



de corrélations. Prenons l'exemple extrême de deux SNPs totalement corrélés, c'est-à-dire que lorsqu'un individu a un certain allèle pour l'un, il a toujours le même allèle pour l'autre et inversement. Si l'on teste séparément les effets de ces deux SNPs sur le risque d'apparition d'une maladie, les résultats de ces deux tests vont être exactement identiques. On aura effectué deux tests alors qu'un seul suffisait et si l'on ne fait pas d'ajustement, la correction pour tests multiples devient bien trop stringente. Certains chercheurs suggèrent d'estimer le nombre théorique de tests effectués (appelé nombre effectifs de tests) et d'utiliser ce nombre pour effectuer les corrections pour tests multiples [83]. Une autre solution serait d'effectuer une sélection sur les tests, pour ne garder que ceux qui ne sont pas trop corrélés entre eux. C'est l'approche que j'ai utilisée dans le chapitre 7.

Chapitre 6

Les données épidémiologiques utilisées

Dilbert : *Studies have shown that accurate numbers aren't any more useful than the ones you make up.*

Boss : *How many studies showed that ?*

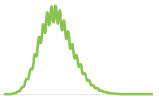
Dilbert : *Eighty-seven.*

Dilbert
http://www.dilbert.com/

En abordant les tests statistiques et la gestion des tests multiples, j'ai pu, dans les deux derniers chapitres, présenter les méthodes statistiques que j'ai utilisées pour effectuer mes recherches de phénomènes d'interactions. Dans ce chapitre, je présente les données épidémiologiques sur lesquelles j'ai appliqué ces méthodes, à savoir les données de quatre études différentes : l'Early-Onset Venous Thrombosis (EOVT) et l'étude MARTHA d'une part, et la Gutenberg Health Study (GHS) et l'étude Cardiogenics d'autre part.

6.1 Les études EOVT et MARTHA

Les études EOVT et MARTHA ont pour objectif de découvrir de nouveaux facteurs de risque de thrombose veineuse. Je les ai utilisées dans le cadre de mes recherches de phénomènes d'interactions entre polymorphismes pouvant affecter l'apparition de la maladie thrombo-embolique veineuse (voir chapitre 7).



6.1.1 L'Early-Onset Venous Thrombosis (EOVT)

L'étude EOVT est une étude d'association génome-entier composée de deux échantillons de cas et de témoins d'origine européenne et résidant en France. L'échantillon de cas contient 453 patients recrutés dans quatre centres médicaux à Grenoble, Marseille, Montpellier et Paris entre 1999 et 2006 avec pour critères d'inclusion, une apparition de la maladie avant 50 ans et une absence de facteurs de risque majeur de la thrombose veineuse : pas de déficit en AntiThrombine (AT), Protéine C (PC), ou Protéine S (PS) et non homozygosité pour les mutations du facteur V Leiden ou du facteur II [39, 120]. L'échantillon de témoins consiste pour sa part en 1 327 sujets sains choisis aléatoirement parmi les 13 017 sujets volontaires qui participèrent à l'étude Suvimax, une étude qui consistait à tester si la prise de vitamines et minéraux antioxydants avait un effet sur l'incidence des maladies cardiovasculaires et des cancers dans la population générale [48]. Les deux échantillons furent génotypés pour plus de 300 000 SNPs avec la puce à ADN Illumina Sentrix HumanHap300.

Les données utilisées dans ce travail de thèse ont par ailleurs été filtrées pour exclure les individus qui semblaient apparentés ou dont l'origine européenne nous paraissait discutable. Au final, en s'assurant d'un taux de génotypage réussi par individu de plus de 95 %, les analyses que j'ai effectuées sur cette étude reposent sur 411 cas et 1 228 témoins (599 hommes et 1 040 femmes).

Critères de qualité des SNPs

Tous les SNPs qui n'avaient pas une p-value pour le test d'Hardy-Weiberg supérieure à 10^{-5}, une fréquence de l'allèle mineur supérieure à 1% chez les cas et 1% chez les témoins ainsi qu'un taux de succès lors du génotypage d'au moins 99% ont été exclu lors des analyses effectuées à partir de cette étude. Le nombre de SNPs restant est de 268 356.

C'est la première étude que j'ai utilisée pour rechercher des phénomènes d'interactions liés à la thrombose veineuse. Je l'ai ensuite étudié en méta-analyse avec l'étude MARTHA

6.1.2 L'étude MARTHA

L'étude MARTHA (pour MARseille THrombosis Association) provient du projet du même nom, mis en place par Pierre Emmanuel Morange en 1994 et financé par le Programme Hospitalier de Recherche Clinique (PHRC). Son objectif est de découvrir de nouveaux facteurs de risque de la maladie thrombo-embolique veineuse en réalisant, notamment, des études d'association génome-entier. L'étude est composée de deux échantillons indépendants de patients d'origine européenne recrutés au

centre de thrombophilie de l'hôpital de la timone à Marseille parmi les malades ne présentant aucun des facteurs de risque principaux décrit précédemment. Chacun des individus de l'étude a été génotypé pour plus de 600 000 SNPs. Le premier échantillon appelé MARTHA08 est composé de 1 006 patients recrutés entre 1994 et 2008 et génotypés avec la puce à ADN Illumina Human 610-Quad alors que le second, MARTHA10, consiste en 586 patients recrutés entre 2008 et 2010 et génotypés avec la puce à ADN Illumina Human 660W-Quad [39, 87].

Les patients de l'étude MARTHA ont été comparés à un groupe de témoins provenant de l'étude prospective des 3 cités (3C). L'étude des 3C avait pour objectif d'investiguer les éventuels liens entre la démence et les facteurs de risque vasculaires. Elle est composée de sujets sains de plus de 65 ans recrutés aléatoirement entre janvier 1999 et mars 2001 à partir des listes électorales de trois villes françaises : Bordeaux, Montpellier et Dijon [1]. L'échantillon des témoins utilisé dans ce projet de thèse est composé de 1 140 individus tirés aléatoirement parmi les 8 707 sujets de l'étude des 3C ne présentant aucune maladie chronique apparente et pour lesquels un prélèvement sanguin avait été réalisé.

Afin d'éviter la présence d'individus apparentés ou d'origine non européenne, un filtrage des donnée a été effectué par clustering et positionnement multidimensionnel (MDS) si bien qu'après avoir gardé uniquement les individus dont le taux de génotypage réussi était supérieur à 95 %, les analyses qui suivent portent pour cette étude sur 1 542 cas et 1 110 témoins (870 hommes et 1 782 femmes).

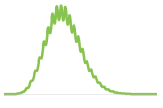
Critères de qualité des SNPs

Le filtrage effectué sur les SNPs a consisté pour cette étude à ne garder que ceux dont la p-value associée au test d'Hardy-Weinberg était supérieure à 10^{-5}, dont la fréquence de l'allèle mineur dépassait les 1 % et dont le taux de réussite lors du génotypage était supérieur à 99 % pour chaque échantillon.

Biomarqueurs

Un biomarqueur est une caractéristique mesurable liée à un état biologique. Par exemple, les individus ayant un haut niveau plasmatique de facteur VIII ont des plus grands risques de thrombose veineuse. La mesure du taux de facteur VIII est donc un biomarqueur de cette maladie.

En plus des données génotypiques, l'étude MARTHA inclut pour certains individus cas, les mesures de certains biomarqueurs de la maladie [145], dont un bon nombre sont liés à des protéines participant à la cascade de coagulation du sang illustrée par la figure 6.1 :



- le dosage de la protéine C (PC) : La protéine C est une protéine jouant un rôle important dans la régulation de la coagulation du sang.
- le dosage de la protéine S (PS) : La protéine S est une protéine agissant sur la protéine C.
- l'Agkistrodon contortrix venum test normalisé (ACVN) : C'est une mesure du ratio entre le temps de coagulation en présence d'un produit (le venin de l'Agkistrodon contortrix, une espèce de serpent) activateur de la protéine C et en absence de ce produit.
- le dosage du facteur VIII (VIII) : Le facteur VIII est une protéine participant à la formation du caillot sanguin lorsque le processus en cascade de coagulation du sang est entamé.
- le dosage du facteur de von Willebrand par antigène (VWF) : le facteur de von Willebrand est une protéine essentielle à l'hémostase primaire, le mécanisme permettant l'adhésion des plaquettes à la veine lésée, avant la coagulation proprement-dite. Il permet entre autres le transport du facteur VIII.
- le dosage de l'antithrombine (AT) : l'antithrombine est la principale protéine inhibitrice de la thrombine, déclencheur de la première phase de la cascade de coagulation, afin d'éviter l'apparition de thromboses veineuses ou artérielles. Elle inhibe également les facteurs Xa, IX et XIa, également présents dans la cascade de coagulation.
- le dosage du fibrinogène (FIB). La protéine fibrinogène, aussi appelée facteur I, est une protéine impliquée dans le processus en cascade de coagulation du sang. Elle se transforme en fibrine, principal constituant du caillot sanguin, sous l'action de la thrombine.
- le temps de thrombine (PT). C'est une mesure du temps d'apparition du caillot de fibrine après ajout d'une faible quantité de thrombine.
- le temps de céphaline activée (TCA) : C'est une mesure du temps de coagulation d'un plasma en présence notamment de céphaline.

L'étude MARTHA a été utilisée en combinaison avec l'étude EOVT pour rechercher des phénomènes d'interaction liés à la thrombose veineuse. Nous l'avons également utilisée pour rechercher des associations entre certaines interactions et les biomarqueurs que je viens de décrire.

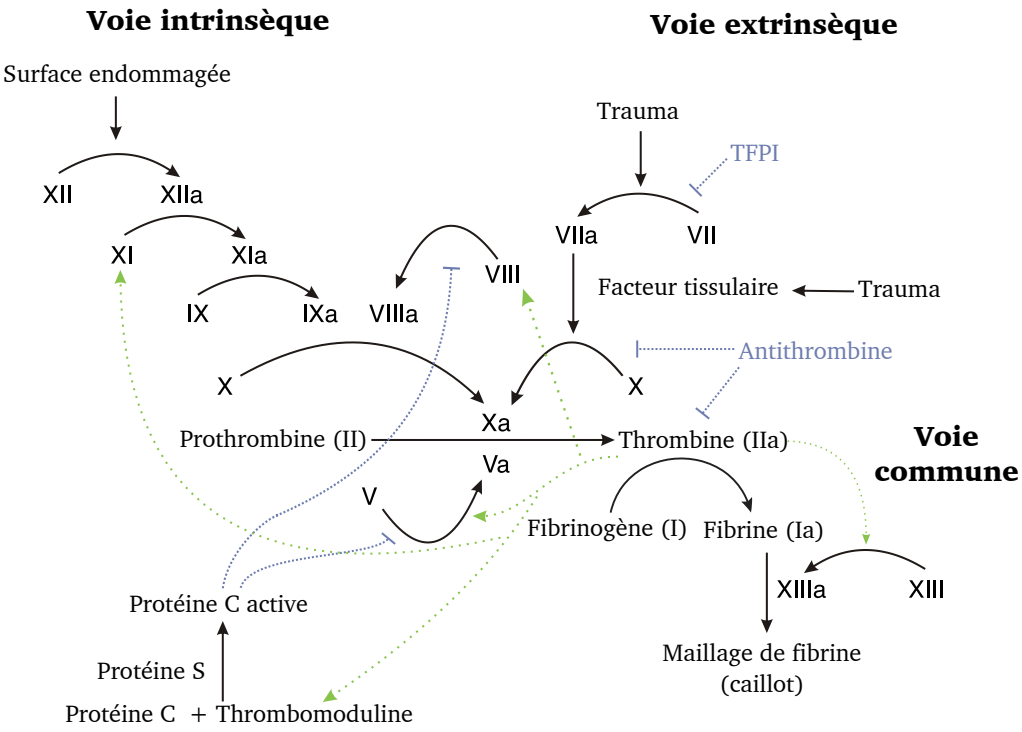


FIGURE 6.1 – Cascade de coagulation du sang.

6.2 Les études GHS et Cardiogenics

Les objectifs des études GHS et Cardiogenics sont de découvrir de nouveaux facteurs de risque des maladies cardiovasculaires. Je les ai utilisées pour rechercher des phénomènes d'interactions entre des polymorphismes liés aux microARNs, qui pourraient affecter l'expression des gènes (voir chapitre 8).

6.2.1 La Gutenberg Health Study (GHS)

L'étude GHS est une grande étude prospective, initiée en 2006 par le docteur Stefan Blankenberg. Elle vise plus spécifiquement à connaître l'état de santé général des habitants de la région de Mayence, en Allemagne, ainsi qu'à identifier de nouveaux facteurs de risque pour diverses pathologies, avec une attention particulière pour les maladies cardiovasculaires [151]. Le design de l'étude consiste à recruter entre 2006 et 2012 près de 17 000 hommes et femmes sains, âgés de 35 à 74 ans et de les soumettre à deux examens médicaux approfondis au centre médical universitaire de Mayence, le premier lors de leur recrutement et le second 5 ans plus tard. Dans le même temps, un certain nombre de prélèvements biologiques, notamment sanguins, sont effectués permettant le génotypage des individus et pour certains sujets, la mesure de leur expression génique dans le monocyte [130].

Les données de cette étude que j'ai utilisées pour mon travail de thèse sont issues

des prélèvements et mesures effectués sur les 3 300 premiers sujets recrutés. Leur génotypage pour environ 900 000 SNPs a été effectué à l'aide de la puce à ADN Affymetrix 6.0 tandis que l'expression de plus de 35 000 gènes provenant de cellules monocytaires [1] a été mesuré pour environ la moitié de ces individus grâce quelques 48 000 sondes contenues dans la puce à ARN Illumina HT-12 v3. Seuls les individus d'origine européenne et pour lesquels les données de génotypage et d'expression étaient disponibles ont été utilisées dans ce travail de thèse si bien qu'au final, mes analyses ont porté sur 750 hommes et 717 femmes.

Critères de qualité des sondes et SNPs

Seules les sondes étant annotées comme ne contenant pas de SNPs et ayant un score de qualité dit « perfect » d'après ReMOAT [6, 161] (Reannotation and Mapping of Oligonucleotide Arrays Technologies) ont été conservées pour l'analyse. En ce qui concerne le filtrage au niveau des SNPs, celui-ci a consisté à ne conserver que les SNPs au taux de génotypage réussi supérieur à 98 %, situés sur les chromosomes autosomaux, dont la fréquence de l'allèle mineur était supérieure à 1 % et la p-value associée au test d'Hardy-Weinberg était supérieure à 10^{-4}.

C'est l'étude principale sur laquelle je me suis appuyé pour rechercher des polymorphismes liés aux microARNs qui pourraient, seuls ou en interaction avec d'autres polymorphismes, agir sur les expressions de nos gènes.

6.2.2 L'étude Cardiogenics

L'étude Cardiogenics est issue du projet européen du même nom, financé par le 6ème programme cadre pour la recherche et le développement technologique (FP6). Ce projet résulte de la collaboration de 15 partenaires européens et a pour objectif de découvrir de nouveaux variants génétiques associés aux cardiopathies coronariennes, afin de mieux comprendre les mécanismes impliqués dans cette maladies et ainsi aider au développement de nouveaux traitements [142]. Au contraire de l'étude GHS, l'étude Cardiogenics est une étude cas-témoins et résulte ainsi du regroupement de deux échantillons. Le premier est composé de 370 sujets âgés de 26 à 87 ans et atteints du syndrome coronaire aigu tel que défini par la Société Européenne de Cardiologie. Leur recrutement s'est effectué parmi les patients coronariens des hôpitaux de Leicester, Lübeck, Paris et Regensburg. Le second échantillon est composé de 403 sujets sains recrutés à Cambridge parmi les donneurs de sang volontaires en s'assurant d'une distribution d'âges similaire à l'échantillon des cas [47, 103]. Plusieurs dizaines d'informations et de mesures biologiques ont été collectées pour

1. Les monocytes sont des globules blancs qui évoluent en macrophages pour participer à la destruction des débris cellulaires et des agents infectieux.

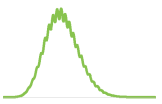
chaque individu. Chaque sujet a été génotypé pour environ 600 000 SNPs à l'aide d'une des deux puces à ADN suivantes : la puce Illumina Sentrix Human Custom 1.2M et la puce Human 610 quad. Enfin, la mesure d'expression d'environ 18 000 gènes sur les cellules du monocyte et du macrophage ont été réalisées en utilisant les 24 516 sondes incluses dans la puce à ARN Illumina Ref8 v3.

Lors de mon travail de thèse, j'ai utilisé uniquement les données des individus d'origine européenne et pour lesquels les données de génotypage et d'expression dans le monocyte étaient disponibles, à savoir 363 sujets coronariens et 395 sujets sains.

Critères de qualité des sondes et SNPs

Le filtrage des sondes fut identique à celui de l'étude GHS. Pour ce qui est des SNPs, afin d'éviter au maximum des problèmes ultérieurs d'interprétations, seuls les SNPs autosomaux avec une fréquence allélique mineure supérieure à 1 %, un taux de succès lors du génotypage dépassant les 95 % et pour lesquels la p-value associée au test d'équilibre d'Hardy-Weinberg était supérieure à 10^{-5} furent conservés.

Cette étude m'a servi pour la réplication des résultats issus de l'étude GHS.



Chapitre 7

À la recherche de phénomènes d'interactions dans la maladie thromboembolique veineuse

C'est pas faux.

Perceval (Kaamelott)
http://www.kaamelott.com/

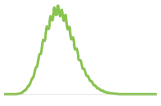
Les chapitres précédents m'ont permis d'introduire les données et méthodes statistiques que j'ai utilisées dans mes recherches de phénomènes d'interactions. Les résultats de ces travaux sont l'objet des deux prochains chapitres. En particulier, dans ce chapitre, après avoir brièvement introduit la maladie thromboembolique veineuse, je donne les résultats de mes recherches d'interactions entre polymorphismes qui pourraient être impliquées dans cette pathologie. Une partie des résultats présentés ici a fait l'objet d'un article en cours de révision et pour lequel je suis premier auteur [41].

7.1 Motivations et stratégie de recherche

7.1.1 Description de la maladie

La thrombose veineuse est une maladie complexe touchant 1 à 2 personnes sur 1 000 chaque année. Elle consiste, comme on peut le voir sur la figure 7.1, en la formation de caillots sanguins dans les veines [1]. On distingue la thrombose veineuse

1. Les veines sont les vaisseaux transportant le sang des organes vers le coeur, au contraire des artères qui amènent le sang du coeur vers les organes.



profonde de l'embolie pulmonaire sa principale complication qui survient lorsque les caillots de sang migrent vers les poumons. L'embolie pulmonaire est caractérisée par un taux de mortalité à un an d'environ 10 % lorsque l'on exclut les individus présentant des symptômes pour d'autres pathologies [129].

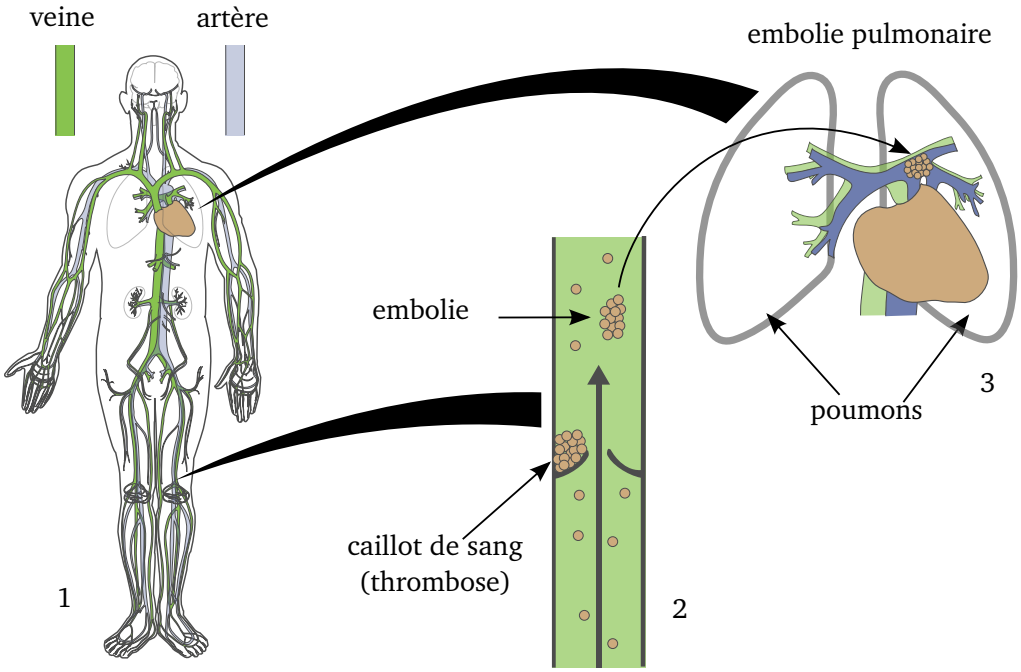


FIGURE 7.1 – 1 : les veines transportent le sang des organes vers le coeur ; les artères, du coeur vers les organes. 2 : la thrombose veineuse consiste en l'apparition d'un caillot sanguin qui, s'il se détache de la paroi, crée ce que l'on appelle une embolie. 3 : l'embolie pulmonaire survient lorsqu'un caillot de sang atteint et obstrue l'artère pulmonaire

7.1.2 Facteurs de risque

L'âge, les longues immobilisations ainsi que la présence de diverses autres anomalies biologiques ou pathologiques sont les principaux facteurs de risque avérés de la maladie, et comme pour la plupart des maladies complexes, les facteurs génétiques identifiés à ce jour (le groupe ABO, *FII*, *FV*, *FGG*, *GP6*, *HIVEP1*, *KNG1*, *STAB2*, *STXBP5* ou encore *VWF*) n'expliquent encore qu'une faible part de l'héritabilité estimée de la maladie [39, 81, 120].

7.1.3 L'hypothèse de nombreuses interactions

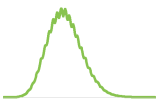
Cette héritabilité manquante pourrait se trouver au niveau d'interactions entre les facteurs de risque génétiques et environnementaux. En effet, il est désormais acquis que les déficits en protéines importantes dans la cascade de la coagulation sanguine ne sont pas suffisants pour expliquer l'apparition de thrombose veineuse.

Les déficits en protéine C [62], protéine S [140] ou antithrombine [121] semblent agir en interaction avec un ou plusieurs autres facteurs de risque, notamment génétiques, pour augmenter le risque de thrombose ce qui suggère que de multiples facteurs génétiques et environnementaux interagissent pour contribuer au risque de la maladie [17, 97]. Par exemple, il a été montré que le risque d'apparition de la maladie était accru lorsque les femmes porteurs de la mutation G20210A du facteur FII (codant pour la protéine prothrombine) ou de la mutation du facteur V Leiden, utilisaient des moyens contraceptifs oraux [77, 123]. Pour ce qui est des interactions entre polymorphismes, il a par exemple été montré que la mutation du facteur II combinée avec celle du facteur V Leiden accroissait le risque de thrombose récurrente chez les personnes ayant déjà été affectées par la maladie [28]. Une étude plus récente a aussi rapporté plusieurs interactions potentielles entre 86 polymorphismes sur une étude de cohorte de 439 individus parmi lesquels 43 développèrent la maladie [137]. Le tout suggérerait qu'il y a potentiellement de nombreuses interactions entre polymorphismes qui peuvent agir sur le risque de la maladie thromboembolique veineuse. Pourtant, à ce jour et à notre connaissance, aucune recherche d'interaction en génome entier n'a été réalisée sur le risque de thrombose veineuse. C'est ce que nous avons cherché à faire ici avec les données des études EOVT et MARTHA.

7.1.4 Stratégie de recherche

Comme on peut le voir sur la figure 7.2, la stratégie de recherche adoptée peut être résumée en plusieurs étapes :

- Dans un premier temps, j'ai identifié et sélectionné dans l'étude EOVT, les SNPs non redondants qui n'étaient en fort déséquilibre de liaison ($r^2 < 0.9$) avec aucun autre SNP conservé, ceci de manière à réduire le nombre de SNPs utilisé à 243 189 (contre 268 356 auparavant) et ainsi réduire la correction pour tests multiples effectuée.
- Nous avons ensuite testé l'ensemble des $243\,189 \times 243\,188/2 \approx 2.96 \times 10^{10}$ interactions entre SNPs sur le statut malade/non malade dans EOVT.
- Les 2 126 084 interactions ayant une p-value inférieure à 10^{-4} furent ensuite testées dans MARTHA en prenant soin d'utiliser des proxySNPs lorsque les SNPs de EOVT n'étaient pas disponibles dans MARTHA.
- Nous avons alors recherché des associations entre les interactions ressortant le plus de notre analyse et les biomarqueurs mesurés dans l'étude MARTHA.
- Enfin, nous avons essayé diverses méthodes de pondérations pour tenter d'augmenter notre puissance de détection d'interactions.



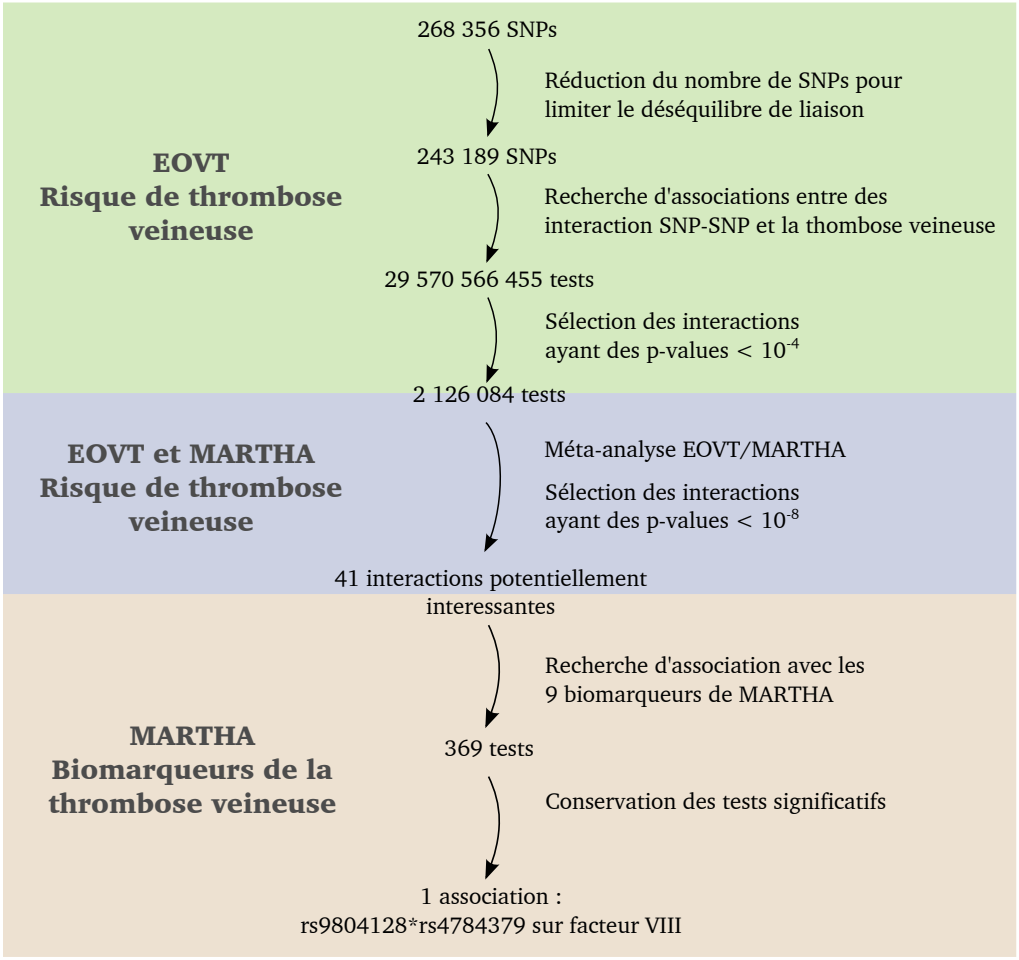


FIGURE 7.2 – Stratégie de recherche d'interactions SNP-SNP associées avec la thrombose veineuse.

7.2 Une puissance trop faible dans EOVT

Après correction de Bonferroni, aucun des ~30 milliards de tests effectués n'est resté significatif, ce qui nous a poussé à nous demander le genre de magnitude d'effet qu'il était possible ou n'était pas possible de détecter par ce type d'approche, sur les données de l'étude EOVT.

7.2.1 Validation du calcul de puissance

Pour ce faire, j'ai commencé par m'assurer par simulation de la validité des calculs de puissance décrits dans le chapitre 5. J'ai d'abord simulé des individus en générant aléatoirement des génotypes pour deux SNPs non corrélés. Je leur ai ensuite attribué le statut malade ou non malade avec une probabilité calculée à partir

des paramètres des effets marginaux [1] et d'interaction dans les modèles sans et avec le terme d'interaction respectivement, et ce, jusqu'à obtenir 411 cas et 1228 témoins (comme dans l'étude EOVT). J'ai effectué 10 000 simulations, pour divers modèles (codage additif mais aussi récessif ou dominant), et différentes fréquences de SNPs, odds-ratios marginaux, effets d'interaction et seuils de significativité. Pour chaque modèle, la proportion de simulations pour lesquelles le test de Wald (décrit dans le chapitre 4) est significatif fournit une estimation de la puissance qui est ensuite comparée à mon calcul théorique. Quelques résultats de ces simulations sont donnés dans la figure 7.3.

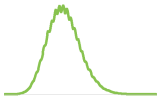
Fréquence allélique		Odds-ratio marginal		logarithme du paramètre d'intéraction	erreur de type I	Puissance	
SNP_1	SNP_2	SNP_1	SNP_2			théorique	observée
0.4	0.4	1.1	1.1	5	1.E-12	1.00	1.00
0.4	0.2	1.1	1.1	5	1.E-12	0.96	0.99
0.2	0.2	1.1	1.1	5	1.E-12	0.83	0.88
0.4	0.4	2.0	0.8	5	1.E-12	1.00	1.00
0.4	0.2	2.0	0.8	5	1.E-12	0.64	0.71
0.2	0.2	2.0	0.8	5	1.E-12	0.64	0.68
0.4	0.4	1.1	1.1	3	1.E-12	0.86	0.90
0.4	0.2	1.1	1.1	3	1.E-12	0.35	0.33
0.2	0.2	1.1	1.1	3	1.E-12	0.11	0.07
0.4	0.4	2.0	0.8	3	1.E-12	0.71	0.76
0.4	0.2	2.0	0.8	3	1.E-12	0.10	0.04
0.2	0.2	2.0	0.8	3	1.E-12	0.06	0.03
0.4	0.4	1.1	1.1	4	1.E-10	1.00	1.00
0.4	0.2	1.1	1.1	4	1.E-10	0.94	0.98
0.2	0.2	1.1	1.1	4	1.E-10	0.75	0.79
0.4	0.4	2.0	0.8	4	1.E-10	1.00	1.00
0.4	0.2	2.0	0.8	4	1.E-10	0.65	0.71
0.2	0.2	2.0	0.8	4	1.E-10	0.58	0.61
0.4	0.4	1.1	1.1	2	1.E-10	0.18	0.16
0.4	0.2	1.1	1.1	2	1.E-10	0.03	0.02
0.2	0.2	1.1	1.1	2	1.E-10	0.00	0.00
0.4	0.4	2.0	0.8	2	1.E-10	0.11	0.08
0.4	0.2	2.0	0.8	2	1.E-10	0.01	0.00
0.2	0.2	2.0	0.8	2	1.E-10	0.00	0.00

FIGURE 7.3 – Résultats des simulations

7.2.2 Effets détectables et non détectables dans EOVT

Les puissances simulées et calculées étant très proches quelque soit les fréquences, odds-ratios ou seuils de significativité choisis, ces simulations m'ont permis de

1. Ici, j'appelle effet marginal l'effet estimé d'un SNP seul, sans autre effet de SNPs ni terme d'interaction



m'assurer de la pertinence de mes calculs. Étant donné le lien étroit existant entre la puissance d'un test d'interaction et la magnitude de l'effet d'interaction testé, il est par ailleurs aisé de déterminer les magnitudes des effets d'interaction détectables avec une puissance fixée plutôt que l'inverse. J'ai donc calculé les magnitudes minimums des effets d'interactions SNP-SNP qui avaient 80 % de chances d'être détectées (puissance de 80 %) par les tests de Wald effectués dans l'étude EOVT, c'est-à-dire en faisant l'hypothèse d'un modèle additif et en choisissant un seuil de significativité de $0.05/(2.96\times10^{10}) = 1.7\times10^{-12}$ (seuil de bonferroni pour les ~30 milliards de tests). Afin de raccourcir les temps de calculs, nous n'avons en fait pas opté pour effectuer les calculs des effets détectables sur l'ensemble des couples de SNPs de l'étude, mais sur des classes de couples de fréquences alléliques et odds-ratios marginaux similaires. Nous avons choisi des classes de fréquences alléliques de largeur 0.025 et allant de 0.1 à 0.5. Nous avons opté en ce qui concerne les odds-ratios marginaux, pour des classes de largeurs 0.01, allant de 1 à 2.4 [1]. Ce faisant, nous sommes arrivés à un ensemble de 11 726 classes différentes et donc à $11726(11726+1)/2 = 6.9\times10^{7}$ calculs.

Odds-ratios détectables

La courbe noire de la figure 7.4 (en partie confondue avec la courbe beige), représente la densité des magnitudes minimales détectables pour les effets d'interactions entre SNPs avec une puissance de 80 % : On y voit une densité qui est nulle entre 1 et 2.5, puis qui augmente pour atteindre son maximum en 3 et diminuer ensuite. Cela signifie qu'aucune interaction entre les SNPs de l'étude EOVT ne pouvait être détectée avec plus de 80 % de probabilité si leur effet sur la maladie n'était pas supérieur à 2.5 en terme d'odds-ratio et même 3 pour la plupart des couples de SNPs. On y voit aussi qu'un certain nombre de d'interactions SNP-SNP nécessite des odds-ratios bien plus élevés, du type de ceux observés dans les maladies mendéliennes mais qu'on ne s'attendrait sans doute pas à trouver dans les maladies complexes.

Détection plus facile pour les allèles fréquents

Les autres courbes de la figure sont tracées après un filtrage des SNPs selon plusieurs critères. Ils permettent de repérer les critères influant le plus sur la probabilité de détection d'une interaction SNP-SNP. On y voit, sur la courbe beige, que les interactions impliquant les 5 % de SNPs aux plus petites p-values marginales (p-value associée au test des effets de chaque SNP seul, dans un modèle sans autre SNP

1. Nous ne considérons ici que les odds-ratios marginaux supérieurs à 1 car ceux étant inférieurs à 1 leur sont identiques en terme d'effet, par symétrie.

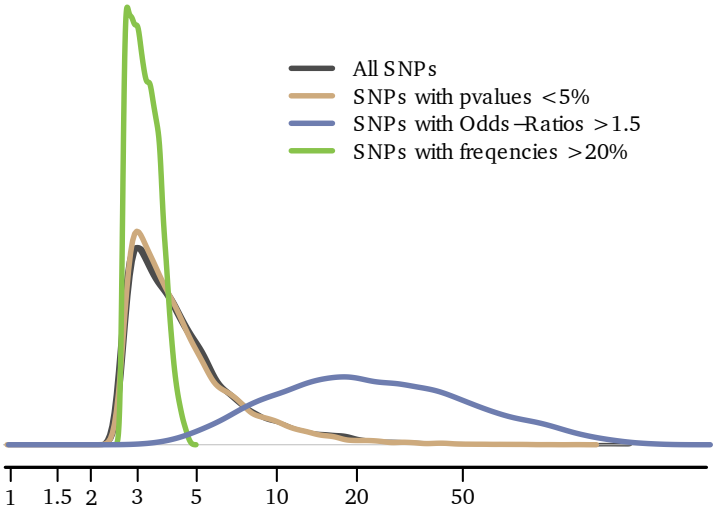


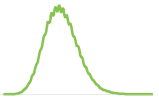
FIGURE 7.4 – Densité des effets d'interaction minimum détectables en suivant le critère de sélection des SNPs, en ratio d'odds-ratio (échelle logarithmique).

et sans terme d'interaction) ont le même profil de magnitude d'effet détectable que l'ensemble des interactions. Ce n'est en revanche pas le cas pour les SNPs fréquents (courbe verte) pour lesquels on peut être confiant que si ils affectent fortement la maladie (odds-ratios supérieurs à 5 sur la figure), en interaction avec d'autres SNPs fréquents, nous seront en mesure de détecter de tels effets. Ces résultats suggèrent que d'un point de vue statistique, il est préférable de sélectionner des SNPs fréquents pour tester des interactions SNP-SNP, que de sélectionner des SNPs qui seuls, semblent associés à la maladie.

7.3 Associations dans l'étude MARTHA - méta-analyse

Nous avons ensuite décidé de tester dans l'étude MARTHA toutes les associations ayant eu une p-value inférieure à 10^{-4} dans l'étude EOVT. La plus petite p-value (6.73×10^{-7}) n'apparaît pas significative après correction pour tests multiples (seuil de Bonferroni à $0.05/2126084 \approx 2.35 \times 10^{-8}$). Nous avons donc décidé d'augmenter la puissance de nos tests, en effectuant une méta-analyse par la méthode de Stouffer décrite dans le chapitre 5. Celle-ci ne permit pas non plus de trouver des interactions significatives après correction de Bonferroni mais nous a cependant mené à considérer 41 interactions potentiellement intéressantes : les interactions aux p-values inférieures à 10^{-8} (voir figure 7.5). La plus petite p-value (p-value=6.00×10^{-11}) est observée pour deux SNPs (rs493014[1] et rs886090) au voisinage du gène *SURF6*, lui-même proche du gène *ABO* (à environ 40 000

1. Par convention, les noms de la majorité des SNPs de notre génome consistent en un numéro, précédé des deux lettre « rs »



bases) qui comme cela a été indiqué au début de ce chapitre est un des principaux facteurs de risque de thrombose veineuse. En ajustant cette interaction sur la variable « groupe sanguin ABO », l'association disparaît (p-value=0.37) ce qui suggère que cette interaction caractérise en fait, grâce au déséquilibre de liaison, l'effet de cette variable ABO. On peut noter que deux SNPs (rs8176746 et rs505922) rapportés comme représentant bien le groupe ABO [4] sont présents dans l'étude EOVT. Ils sont en déséquilibre de liaison avec le SNP rs493014 ($D' = 0.70$ avec rs8176746 et $D' = 0.69$ avec rs505922), moins avec le SNP rs886090 ($D' = 0.27$ et $r^2 = 0.04$).

						EOVT				MARTHA					
						MAF[2]		Interaction		MAF		Interaction		Combinés	
rsID	Allèles[1]	CHR	rsID	Allèles	CHR	SNP1	SNP2	OR[3]	P	SNP1	SNP2	OR	P	OR[4]	P
rs493014	T/G	9	rs886090	G/A	9	0,31	0,33	1,72	1.85 x 10^{-5}	0,30	0,31	1,60	6.73 x 10^{-7}	1,64	6.00 x 10^{-11}
rs1336472	G/A	1	rs4715555	A/G	6	0,41	0,39	1,64	4.10 x 10^{-5}	0,40	0,38	1,49	2.00 x 10^{-6}	1,54	4.24 x 10^{-10}
rs380904	G/A	8	rs8086028	G/A	18	0,30	0,27	1,96	3.76 x 10^{-6}	0,29	0,31	1,55	1.12 x 10^{-5}	1,67	4.51 x 10^{-10}
rs6815916	A/G	4	rs6092326	C/T	20	0,09	0,48	2,37	4.32 x 10^{-5}	0,09	0,47	1,98	2.95 x 10^{-6}	2,10	6.84 x 10^{-10}
rs2282015	T/G	10	rs13050454	G/A	21	0,41	0,43	1,81	3.52 x 10^{-7}	0,41	0,42	1,37	7.68 x 10^{-5}	1,50	8.36 x 10^{-10}
rs7648704	T/G	3	rs4868644	C/T	5	0,33	0,49	1,64	7.36 x 10^{-5}	0,33	0,49	1,52	2.88 x 10^{-6}	1,56	9.89 x 10^{-10}
rs1985317	T/C	9	rs827637	G/A	10	0,39	0,46	0,55	7.13 x 10^{-7}	0,41	0,46	0,72	7.73 x 10^{-5}	0,66	1.32 x 10^{-9}
rs2321744	A/G	13	rs6497540	T/G	16	0,09	0,41	0,43	8.61 x 10^{-5}	0,10	0,42	0,52	2.98 x 10^{-6}	0,49	1.38 x 10^{-9}
rs315122	T/G	12	rs884483	T/C	15	0,29	0,11	2,61	1.92 x 10^{-5}	0,31	0,12	1,87	7.90 x 10^{-6}	2,05	1.42 x 10^{-9}
rs1423386	A/G	5	rs6491679	T/G	13	0,20	0,29	1,92	7.24 x 10^{-5}	0,20	0,29	1,66	4.17 x 10^{-6}	1,73	1.63 x 10^{-9}
rs6491679	T/G	13	rs1423386	A/G	5	0,29	0,20	1,92	7.24 x 10^{-5}	0,29	0,20	1,66	4.17 x 10^{-6}	1,73	1.63 x 10^{-9}
rs7714670	T/C	5	rs12880735	G/A	14	0,44	0,34	1,75	4.59 x 10^{-6}	0,44	0,36	1,42	3.32 x 10^{-5}	1,52	1.75 x 10^{-9}
rs12880735	G/A	14	rs7714670	T/C	5	0,34	0,44	1,75	4.59 x 10^{-6}	0,36	0,44	1,42	3.32 x 10^{-5}	1,52	1.75 x 10^{-9}
rs9392653	C/T	6	rs7780976	A/C	7	0,27	0,18	2,14	2.28 x 10^{-6}	0,29	0,19	1,57	5.49 x 10^{-5}	1,74	1.83 x 10^{-9}
rs9804128	A/G	1	rs4784379	G/A	16	0,27	0,25	1,97	2.73 x 10^{-5}	0,26	0,24	1,60	9.45 x 10^{-6}	1,71	1.90 x 10^{-9}
rs1364505	G/A	7	rs1204660	G/A	20	0,30	0,16	2,14	2.32 x 10^{-5}	0,33	0,16	1,67	1.11 x 10^{-5}	1,80	2.10 x 10^{-9}
rs2288073	A/G	2	rs10771022	G/T	12	0,30	0,34	1,71	7.94 x 10^{-5}	0,29	0,34	1,55	5.51 x 10^{-6}	1,60	2.11 x 10^{-9}
rs1367228	C/A	2	rs3905075	C/T	13	0,43	0,41	1,61	9.44 x 10^{-5}	0,45	0,40	1,44	4.22 x 10^{-6}	1,49	2.20 x 10^{-9}
rs536477	G/A	1	rs1937920	A/G	10	0,43	0,26	0,57	3.27 x 10^{-5}	0,43	0,27	0,67	1.40 x 10^{-5}	0,63	2.93 x 10^{-9}
rs2710201	A/G	7	rs3780293	G/A	9	0,06	0,34	0,35	6.84 x 10^{-5}	0,06	0,36	0,43	9.92 x 10^{-6}	0,40	3.30 x 10^{-9}
rs12541254	G/A	8	rs305009	G/A	15	0,35	0,23	1,99	3.15 x 10^{-6}	0,34	0,23	1,50	7.63 x 10^{-5}	1,65	3.33 x 10^{-9}
rs4507975	A/G	1	rs9914518	G/A	17	0,29	0,46	0,61	9.59 x 10^{-5}	0,29	0,47	0,67	7.95 x 10^{-6}	0,65	3.58 x 10^{-9}
rs2771051	T/G	9	rs827637	G/A	10	0,37	0,46	0,52	9.27 x 10^{-8}	0,37	0,46	0,75	4.59 x 10-4	0,67	3.82 x 10^{-9}
rs10516089	T/C	5	rs11072930	T/C	15	0,32	0,28	0,51	2.66 x 10^{-6}	0,31	0,30	0,69	7.19 x 10^{-5}	0,63	3.86 x 10^{-9}
rs10504130	G/A	8	rs2847351	A/G	18	0,15	0,30	2,46	1.04 x 10^{-5}	0,14	0,32	1,69	3.07 x 10^{-5}	1,88	4.46 x 10^{-9}
rs318497	G/A	6	rs7019259	A/G	9	0,49	0,07	0,29	2.56 x 10^{-6}	0,49	0,07	0,51	8.40 x 10^{-5}	0,43	4.54 x 10^{-9}
rs6695223	T/C	1	rs1763510	C/T	6	0,12	0,39	2,49	6.00 x 10^{-6}	0,13	0,39	1,66	4.31 x 10^{-5}	1,86	4.70 x 10^{-9}
rs1336708	A/G	13	rs1423386	A/G	5	0,26	0,20	0,51	6.77 x 10^{-5}	0,25	0,20	0,61	1.20 x 10^{-5}	0,58	4.85 x 10^{-9}
rs1423386	A/G	5	rs1336708	A/G	13	0,20	0,26	0,51	6.77 x 10^{-5}	0,20	0,25	0,61	1.19 x 10^{-5}	0,58	4.85 x 10^{-9}
rs6771316	G/A	3	rs10986432	T/C	9	0,14	0,18	2,41	4.64 x 10^{-5}	0,13	0,17	1,99	2.20 x 10^{-5}	2,13	5.26 x 10^{-9}
rs664910	A/G	3	rs877228	G/A	15	0,30	0,47	1,63	6.05 x 10^{-5}	0,30	0,44	1,44	1.92 x 10^{-5}	1,50	6.63 x 10^{-9}
rs9945428	C/A	18	rs4823535	G/A	22	0,30	0,28	0,58	7.47 x 10^{-5}	0,30	0,26	0,65	1.85 x 10^{-5}	0,62	6.88 x 10^{-9}
rs1910358	T/C	5	rs9981595[5]	T/G	21	0,23	0,12	2,21	9.60 x 10^{-5}	0,23	0,11	1,93	1.63 x 10^{-5}	2,03	7.14 x 10^{-9}
rs6771725	G/T	3	rs10507246	G/T	12	0,26	0,08	2,60	4.02 x 10^{-5}	0,28	0,09	2,04	3.77 x 10^{-5}	2,22	8.60 x 10^{-9}
rs16865717	C/T	2	rs2009579	C/T	20	0,27	0,36	1,90	5.22 x 10^{-6}	0,29	0,36	1,43	9.59 x 10^{-5}	1,56	8.82 x 10^{-9}
rs2028385	A/G	12	rs2038227	A/C	16	0,16	0,39	2,19	3.36 x 10^{-7}	0,16	0,37	1,47	7.11 x 10^{-4}	1,69	8.82 x 10^{-9}
rs10476160	A/G	5	rs1707420	C/T	8	0,21	0,48	0,56	6.35 x 10^{-5}	0,20	0,48	0,65	2.48 x 10^{-5}	0,62	9.09 x 10^{-9}
rs971572	C/A	1	rs10828151	A/C	10	0,32	0,07	0,35	3.43 x 10^{-5}	0,32	0,07	0,47	4.38 x 10^{-5}	0,42	9.30 x 10^{-9}
rs6858430	C/T	4	rs4800250	A/G	18	0,20	0,40	1,86	2.44 x 10^{-5}	0,21	0,40	1,52	5.16 x 10^{-5}	1,62	9.67 x 10^{-9}
rs467650	T/C	5	rs7153749	T/C	14	0,36	0,44	0,59	1.69 x 10^{-5}	0,37	0,44	0,71	6.00 x 10^{-5}	0,67	9.91 x 10^{-9}
rs7153749	T/C	14	rs467650	T/C	5	0,44	0,36	0,59	1.69 x 10^{-5}	0,44	0,37	0,71	6.00 x 10^{-5}	0,67	9.91 x 10^{-9}

FIGURE 7.5 – Les 41 interactions ayant une p-value inférieure à 10^{-8} dans la méta-analyse des études EOVT et MARTHA. L'échantillon global est ainsi composé de 1953 cas et de 2338 témoins. [1] Allèle majeur/mineur. [2] Fréquence de l'allèle mineur. [3] Odds ratio de l'interaction pour le risque de thrombose veineuse dans un modèle logistique avec des effets alléliques additifs. [4] Odds ratio combiné en pondérant par l'inverse de la variance. [5] rs2836978 est un proxySNP pour rs9981595 ($r^2 = 1$).

7.4 Associations avec certains biomarqueurs de la maladie

Bien que l'on n'ait pas pu détecter d'interaction significative par la méta-analyse des études EOVT et MARTHA, il reste cependant possible que certaines interactions soient intéressantes d'un point de vue biologique mais que leur effet sur la maladie ne soit pas suffisant pour être détecté par l'approche adoptée. Aussi, nous avons essayé de voir si ces 41 interactions n'étaient pas liées aux neuf biomarqueurs de la thrombose veineuse mesurés dans MARTHA, en prenant soin d'ajuster les modèles pour l'âge, le sexe, le groupe sanguin ABO et la présence des mutations du facteur II et du facteur V Leiden. Ceci nous a amené à effectuer $41 \times 9 = 369$ tests, pour un seuil de Bonferroni de $0.05/369 = 1.35 \times 10^{-4}$. Une interaction en est ressortie significative (p-value$= 4.82 \times 10^{-5}$). Il s'agit de l'interaction entre le SNP rs9804128, situé dans la région promotrice du gène *IGSF21* et le SNP rs4784379 qui se trouve à 130 000 bases en amont du gène *IRXS*. Cette interaction ressort associée à la mesure du niveau de facteur VIII, les patients porteurs des allèles G et A (haplotype G/A) pour les SNPs rs9804128 et rs4784379 respectivement ayant des niveaux plus élevés que les autres individus. L'haplotype G/A semble par ailleurs protecteur du risque de thrombose veineuse comme l'illustre une fréquence environ double chez les témoins (0.083) par rapport à chez les cas (0.046) (voir figure 7.6).

		EOVT Fréquence		MARTHA Fréquence		Combinés Fréquence		Patients de MARTHA [1] Moyenne haplotypique attendue pour FVIII	
		Témoins	Cas	Témoins	Cas	Témoins	Cas		
rs9804128	rs4784379	N =1228	N =419	N =1110	N =1542	N =2338	N =1961	Fréquence	[95%CI]
A	G	0,56	0,53	0,58	0,55	0,57	0,55	0,55	68.77 [66.27 - 71.26]
A	A	0,17	0,2	0,17	0,19	0,17	0,19	0,18	62.34 [58.03 - 66.64]
G	G	0,19	0,24	0,17	0,21	0,18	0,22	0,22	62.09 [56.35 - 67.83]
G	A	0,08	0,04	0,09	0,05	0,08	0,05	0,05	91.95 [92.98 - 100.9]
		$p^{(2)} = 2.73\ 10^{-5}$		$p^{(2)} = 9.45\ 10^{-6}$		$p^{(3)} = 1.90\ 10^{-9}$			$p^{(4)} = 6.89\ 10^{-5}$

FIGURE 7.6 – Effet des allèles des SNPs rs9804128 et rs4784379, en interaction, sur le risque de thrombose veineuse et le niveau plasmatique de facteur VIII. [1] Dans MARTHA, le niveau de facteur VIII a été mesuré pour 699 patients. [2] P-value du terme d'interaction entre les deux SNPs sous un modèle logistique de risque de thrombose veineuse, avec des effets alléliques additifs. [3] P-value obtenue par la méta-analyse des deux études. [4] P-value du terme d'interaction entre les deux SNPs dans le modèle d'association linéaire avec le taux de facteur VIII, ajusté sur l'âge, le sexe, le groupe sanguin ABO ainsi que la présence des mutations des gènes F2 et F5.

En regardant ces résultats au niveau génotypique (figure 7.7), on peut observer que les taux de facteur VIII les plus élevés concernent les individus de génotypes GA/AA, GG/AA ou GG/AG pour les SNPs rs9804128/rs4784379. Ces combinaisons sont celles pour lesquelles, les individus sont assurément porteurs de l'haplotype G/A.

Enfin, en termes d'odds ratios, l'association entre le SNP rs4784379 (allèles G ou

	rs4784379		
rs9804128	AA	AG	GG
AA	115.91 (32.80) N =34	132.70 (49.75) N =231	136.16 (51.35) N =321
GA	155.93 (77.17) N =16	141.42 (56.03) N =144	131.76 (47.11) N =266
GG	156.00 (68.98) N =4	150.17 (42.90) N =23	122.90 (60.11) N =52

FIGURE 7.7 – Moyennes et écart-types (entre parenthèses), des niveaux plasmatiques de facteur VIII, par combinaison génotypique des SNPs rs9804128 et rs4784379. L'effectif est précisé en dessous.

A) et la maladie passe de 1,18 chez les individus porteurs de l'allèle A pour le SNP rs9804128, à 0,46 chez ceux qui sont porteurs de l'allèle G (voir figure 7.8)

rs9804128	Fréquence	Odds-Ratio rs4784379 (G/A) OR [95%CI]	*p*-value
A	0.74	1.18 [1.04 - 1.35]	0.01
G	0.26	0.46 [0.35 - 0.59]	$<10^{-6}$

FIGURE 7.8 – Odds-ratios (et *p*-value associée) de l'association entre le SNP rs4784379 et la thrombose veineuse pour chaque allèle du SNP rs9804128.

7.5 Pondérations et combinaisons

7.5.1 Les interactions du chromosome 20

Nous avons ensuite essayé de tester quelques méthodes de pondérations ou de combinaison de tests pour tenter d'augmenter la puissance de détection de phénomènes d'interactions. Une première étape consista à effectuer une sélection plus drastique des SNPs, d'une part, afin de réduire les temps de calculs et de faciliter la manipulation des données nécessaire à ce genre d'analyse et d'autre part, afin de limiter la correction pour tests multiples à effectuer sur les résultats des tests. Aussi, nous avons opté pour une recherche d'interactions entre les SNPs du chromosome 20 uniquement, car bien qu'il soit relativement petit, ce chromosome semblerait pouvoir contribuer à près de 7 % de l'héritabilité génétique de la maladie thromboembolique veineuse [39]. Nous avons ainsi testé les interactions entre 6 092 SNPs sur la maladie, menant à $6\,092 * 6\,091/2 = 18\,553\,186$ tests et un seuil de Bonferroni à 2.70×10^{-9}. Ces tests ont été effectués séparément dans les études EOVT et MARTHA. Les 15 premiers résultats de ces sont donnés dans la figure 7.9.

On y voit que les plus petites *p*-values ne passent pas le seuil de Bonferroni (2.09×10^{-7} dans EOVT et 6.06×10^{-8} dans MARTHA).

EOVT							MARTHA						
SNP1	freq	$P^{(1)}$	SNP2	freq	$P^{(1)}$	P-int$^{(2)}$	SNP1	freq	$P^{(1)}$	SNP2	freq	$P^{(1)}$	P-int$^{(2)}$
rs6043659	0.34	0.601	rs3746337	0.46	0.414	2.09E-07	rs7264608	0.16	0.039	rs6128273	0.09	0.454	6.06E-08
rs487377	0.21	0.228	rs6075458	0.21	0.409	2.09E-07	rs4811206	0.45	0.796	rs1293144	0.46	0.959	9.17E-08
rs761901	0.40	0.938	rs975137	0.22	0.264	2.12E-07	rs4811206	0.45	0.796	rs1293143	0.42	0.876	1.45E-07
rs2326660	0.17	0.900	rs6123082	0.29	0.621	3.53E-07	rs6054992	0.24	0.215	rs1983702	0.2	0.251	3.57E-07
rs979242	0.37	0.872	rs6021083	0.30	0.576	6.22E-07	rs6038151	0.28	0.582	rs3092379	0.42	0.035	3.78E-07
rs6064733	0.25	0.136	rs2284803	0.17	0.594	6.43E-07	rs1777361	0.33	0.064	rs6110458	0.21	0.587	5.12E-07
rs1984279	0.40	0.832	rs1291211	0.09	0.772	7.23E-07	rs214833	0.27	0.260	rs6126251	0.21	0.480	5.49E-07
rs6132784	0.19	0.379	rs6125111	0.41	0.959	8.74E-07	rs8120756	0.36	0.711	rs2567608	0.48	0.276	6.40E-07
rs6088177	0.42	0.476	rs6062014	0.15	0.620	9.44E-07	rs2327449	0.24	0.299	rs4809607	0.22	0.899	7.61E-07
rs6078239	0.04	0.439	rs6041821	0.18	0.418	9.57E-07	rs6033471	0.31	0.090	rs6021293	0.28	0.361	7.85E-07
rs910901	0.33	0.040	rs2268879	0.43	0.063	9.73E-07	rs742754	0.41	0.826	rs1293144	0.46	0.959	8.79E-07
rs4814489	0.38	0.865	rs975137	0.22	0.264	1.07E-06	rs6107581	0.12	0.204	rs7260918	0.35	0.764	9.37E-07
rs11086869	0.21	0.092	rs2224272	0.22	0.098	1.14E-06	rs421630	0.38	0.069	rs2766641	0.46	0.598	9.52E-07
rs6020391	0.31	0.104	rs6513544	0.06	0.638	1.42E-06	rs1998105	0.27	0.040	rs761382	0.48	0.686	1.06E-06
rs3212198	0.43	0.299	rs3787537	0.24	0.913	1.60E-06	rs742754	0.41	0.826	rs1293143	0.42	0.876	1.07E-06

FIGURE 7.9 – Les 15 interactions entre les SNPs du chromosome 20 qui ressortent les plus associées à la thrombose veineuse dans EOVT (à gauche) et dans MARTHA (à droite). $^{(1)}p$-value marginale associée à chaque SNP. $^{(2)}p$-value liée au terme d'interaction.

7.5.2 Pondérations sur chaque étude

Par les fréquences alléliques

Les résultats de nos calculs de puissance sur l'étude EOVT ont montré que d'un point de vue purement statistique, il était plus facile de détecter des interactions entre SNPs aux allèles fréquents qu'entre SNPs aux allèles rares. Il apparaît ainsi pertinent d'essayer de pondérer les p-values des tests effectués sur le chromosome 20 par les fréquences alléliques des SNPs impliqués dans ces tests. Afin de prendre en compte les fréquences alléliques de chaque SNP, j'ai opté pour une pondération par le produit des fréquences des allèles mineurs :

$$w = \text{freq}_1 \times \text{freq}_2$$

où freq_1 et freq_2 sont les fréquences des allèles mineurs des SNPs 1 et 2 respectivement. La figure 7.10 montre les 15 tests qui ressortent le plus après cette pondération dans EOVT et dans MARTHA.

On peut y voir que les fréquences alléliques sont plus fortes dans la figure 7.10 que dans la figure 7.9. Ceci confirme bien qu'une pondération par le produit des fréquences alléliques favorise les hypothèses impliquant des interactions entre SNPs fréquents. Par ailleurs, on peut noter que dans les deux études cette pondération améliore légèrement la significativité des premiers résultats.

Par les p-values marginales

Il paraît intéressant aussi de considérer une pondération pas les p-values marginales associées à chaque SNP des tests. En effet, si l'on a vu que d'un point de vue statistique, il n'y avait pas d'intérêt à privilégier les interactions impliquant

EOVT								MARTHA							
SNP1	freq	P$^{(1)}$	SNP2	freq	P$^{(1)}$	P-int$^{(2)}$	P-pond$^{(3)}$	SNP1	freq	P$^{(1)}$	SNP2	freq	P$^{(1)}$	P-int$^{(2)}$	P-pond$^{(3)}$
rs487377	0.21	0.228	rs6075458	0.21	0.409	2.09E-07	2.04E-07	rs7264608	0.16	0.039	rs6128273	0.09	0.454	6.06E-08	4.19E-08
rs6043659	0.34	0.601	rs3746337	0.46	0.414	2.09E-07	3.47E-07	rs6038151	0.28	0.582	rs3092379	0.42	0.035	3.78E-07	2.71E-07
rs761901	0.40	0.938	rs975137	0.22	0.264	2.12E-07	3.50E-07	rs6054992	0.24	0.215	rs1983702	0.2	0.251	3.57E-07	3.42E-07
rs910901	0.33	0.040	rs2268879	0.43	0.063	9.73E-07	3.76E-07	rs1777361	0.33	0.064	rs6110458	0.21	0.587	5.12E-07	4.35E-07
rs11086869	0.21	0.092	rs2224272	0.22	0.098	1.14E-06	5.61E-07	rs6033471	0.31	0.090	rs6021293	0.28	0.361	7.85E-07	6.40E-07
rs6064733	0.25	0.136	rs2284803	0.17	0.594	6.43E-07	5.90E-07	rs214833	0.27	0.260	rs6126251	0.21	0.480	5.49E-07	7.36E-07
rs6105852	0.46	0.027	rs2268879	0.43	0.063	2.53E-06	9.14E-07	rs1998105	0.27	0.040	rs761382	0.48	0.686	1.06E-06	8.17E-07
rs2423011	0.35	0.214	rs6096260	0.38	0.006	3.15E-06	1.09E-06	rs878198	0.33	0.000	rs6068770	0.06	0.894	3.93E-06	8.20E-07
rs6034465	0.17	0.337	rs12624715	0.13	0.039	2.27E-06	1.21E-06	rs421630	0.38	0.069	rs2766641	0.46	0.598	9.52E-07	8.32E-07
rs6020391	0.31	0.104	rs6513544	0.06	0.638	1.42E-06	1.21E-06	rs4811206	0.45	0.796	rs1293144	0.46	0.959	9.17E-08	9.49E-07
rs3810510	0.14	0.306	rs10485442	0.21	0.019	2.84E-06	1.28E-06	rs4814789	0.17	0.018	rs6128273	0.09	0.454	1.81E-06	1.05E-06
rs6078239	0.04	0.439	rs6041821	0.18	0.418	9.57E-07	1.30E-06	rs673261	0.29	0.108	rs6127376	0.14	0.001	3.41E-06	1.06E-06
rs2326660	0.17	0.900	rs6123082	0.29	0.621	3.53E-07	1.40E-06	rs8120756	0.36	0.711	rs2567608	0.48	0.276	6.40E-07	1.10E-06
rs6041386	0.20	0.290	rs6067931	0.20	0.122	2.11E-06	1.46E-06	rs2745756	0.19	0.046	rs6127015	0.43	0.212	1.84E-06	1.11E-06
rs2249353	0 34	0 314	rs10485569	0 12	0 241	1 68E-06	1 51E-06	rs4811206	0 45	0 796	rs1293143	0 42	0 876	1 45E-07	1 12E-06

FIGURE 7.10 – Les 15 interactions entre les SNPs du chromosome 20 qui ressortent les plus associées à la thrombose veineuse dans EOVT (à gauche) et dans MARTHA (à droite) après pondération par les fréquences alléliques. $^{(1)}p$-value marginale associée à chaque SNP. $^{(2)}p$-value liée au terme d'interaction. $^{(3)}p$-value du terme d'interaction, pondérée par les fréquences alléliques.

des SNPs qui semblent déjà associés à la maladie, d'un point de vue biologique, il paraîtrait assez logique que les SNPs impliqués en interaction dans un phénotype, le soit également séparément. Comme ce sont les p-values faibles que nous souhaitons privilégier, nous avons opté pour une pondération par l'opposé du logarithme du produit des p-values marginales :

$$w = -\log(p\text{-value}_1 \times p\text{-value}_2)$$

où p-value$_1$ et p-value$_2$ sont les p-values marginales associées aux modèles marginaux incluant uniquement les SNPs 1 et 2 respectivement.

EOVT								MARTHA							
SNP1	freq	P$^{(1)}$	SNP2	freq	P$^{(1)}$	P-int$^{(2)}$	P-pond$^{(3)}$	SNP1	freq	P$^{(1)}$	SNP2	freq	P$^{(1)}$	P-int$^{(2)}$	P-pond$^{(3)}$
rs6043659	0.34	0.601	rs3746337	0.46	0.414	2.09E-07	8.70E-08	rs4811206	0.45	0.796	rs1293144	0.46	0.959	9.17E-08	2.85E-08
rs761901	0.40	0.938	rs975137	0.22	0.264	2.12E-07	1.56E-07	rs4811206	0.45	0.796	rs1293143	0.42	0.876	1.45E-07	4.89E-08
rs487377	0.21	0.228	rs6075458	0.21	0.409	2.09E-07	3.12E-07	rs6038151	0.28	0.582	rs3092379	0.42	0.035	3.78E-07	2.09E-07
rs979242	0.37	0.872	rs6021083	0.30	0.576	6.22E-07	3.63E-07	rs8120756	0.36	0.711	rs2567608	0.48	0.276	6.40E-07	2.41E-07
rs910901	0.33	0.040	rs2268879	0.43	0.063	9.73E-07	4.37E-07	rs7264608	0.16	0.039	rs6128273	0.09	0.454	6.06E-08	2.84E-07
rs2326660	0.17	0.900	rs6123082	0.29	0.621	3.53E-07	4.69E-07	rs742754	0.41	0.826	rs1293144	0.46	0.959	8.79E-07	3.03E-07
rs6132784	0.19	0.379	rs6125111	0.41	0.959	8.74E-07	7.03E-07	rs421630	0.38	0.069	rs2766641	0.46	0.598	9.52E-07	3.54E-07
rs6105852	0.46	0.027	rs2268879	0.43	0.063	2.53E-06	8.13E-07	rs742754	0.41	0.826	rs1293143	0.42	0.876	1.07E-06	4.01E-07
rs4814489	0.38	0.865	rs975137	0.22	0.264	1.07E-06	8.29E-07	rs1777361	0.33	0.064	rs6110458	0.21	0.587	5.12E-07	4.76E-07
rs6131222	0.43	0.586	rs734532	0.28	0.570	1.77E-06	9.43E-07	rs6054992	0.24	0.215	rs1983702	0.2	0.251	3.57E-07	4.93E-07
rs283273	0.43	0.296	rs1739591	0.41	0.590	2.62E-06	9.49E-07	rs1998105	0.27	0.040	rs761382	0.48	0.686	1.06E-06	5.25E-07
rs6064733	0.25	0.136	rs2284803	0.17	0.594	6.43E-07	9.61E-07	rs6126343	0.39	0.287	rs6061928	0.47	0.383	1.64E-06	5.85E-07
rs6088177	0.42	0.476	rs6062014	0.15	0.620	9.44E-07	9.71E-07	rs6033471	0.31	0.090	rs6021293	0.28	0.361	7.85E-07	5.94E-07
rs3212198	0.43	0.299	rs3787537	0.24	0.913	1.60E-06	1.01E-06	rs214833	0.27	0.260	rs6126251	0.21	0.480	5.49E-07	6.18E-07
rs6085054	0.40	0.608	rs6108790	0.29	0.542	1.91E-06	1.04E-06	rs1475670	0.51	0.682	rs1293144	0.46	0.959	2.27E-06	6.20E-07

FIGURE 7.11 – Les 15 interactions entre les SNPs du chromosome 20 qui ressortent les plus associées à la thrombose veineuse dans EOVT (à gauche) et dans MARTHA (à droite) après pondération par les p-values marginales. $^{(1)}p$-value marginale associée à chaque SNP. $^{(2)}p$-value liée au terme d'interaction. $^{(3)}p$-value du terme d'interaction, pondérée par les p-values marginales.

On peut voir sur la figure 7.11 que comme attendu, ce sont cette fois les hypothèses pour lesquelles les p-values marginales sont faibles qui sont favorisées.

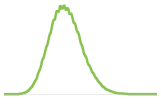
On remarque aussi que cette pondération ne semble pas apporter d'amélioration par rapport aux résultats non pondérés.

7.5.3 Combinaison des études

Pour ce qui est de la combinaison des tests entre les deux études, j'ai utilisé la méthode classique de Fisher (voir chapitre 5). On peut voir les résultats de cette combinaison sur la figure 7.12. L'amélioration par rapport aux résultats non combinés et non pondérés (figure 7.9) n'est pas flagrante (partie gauche du tableau), ce qui suggérerait une absence de phénomènes d'interactions réels parmi les hypothèses testées. La partie droite du tableau montre cependant qu'en favorisant les hypothèses impliquant des SNPs fréquents, il y a là encore une amélioration. Aussi, bien que notre puissance de détection est trop faible pour les détecter, il se peut qu'il y aient des phénomènes d'interactions entre les SNPs du chromosome 20 qui soient faiblement impliqués dans la maladie thromboembolique veineuse.

Sans pondération préalable					*En utilisant la pondération par les fréquences au préalable*				
		P-value					P-value		
SNP1	SNP2	EOVT	MARTHA	Combiné	SNP1	SNP2	EOVT	MARTHA	Combiné
rs1033807	rs6070829	8.63E-04	1.10E-05	1.64E-07	rs1033807	rs6070829	8.63E-04	1.10E-05	6.83E-08
rs1033807	rs6070933	2.14E-03	3.91E-06	1.85E-07	rs1033807	rs6070933	2.14E-03	3.91E-06	8.19E-08
rs10485756	rs1418927	2.52E-05	3.05E-03	2.44E-07	rs172470	rs2426778	2.58E-04	4.94E-05	1.03E-07
rs16995641	rs6018718	2.67E-03	2.30E-05	3.76E-07	rs1777361	rs6110458	3.93E-02	5.12E-07	1.07E-07
rs172470	rs2426778	2.58E-04	4.94E-05	5.96E-07	rs4811206	rs1293143	3.25E-01	1.45E-07	1.43E-07
rs1777361	rs6110458	3.93E-02	5.12E-07	6.23E-07	rs4811206	rs1293144	4.64E-01	9.17E-08	1.79E-07
rs214833	rs6126251	6.39E-02	5.49E-07	6.37E-07	rs6013469	rs9760	6.94E-06	1.11E-02	3.27E-07
rs4811206	rs1293143	3.25E-01	1.45E-07	6.39E-07	rs6043659	rs3746337	2.09E-07	2.09E-01	3.39E-07
rs4811206	rs1293144	4.64E-01	9.17E-08	7.55E-07	rs6054545	rs1327231	5.51E-05	2.15E-03	3.43E-07
rs6013469	rs9760	6.94E-06	1.11E-02	7.65E-07	rs6074012	rs4810671	3.20E-04	2.46E-04	4.43E-07
rs6034465	rs12624715	2.27E-06	1.55E-02	7.84E-07	rs6083931	rs2795025	1.38E-05	2.48E-03	4.59E-07
rs6043659	rs3746337	2.09E-07	2.09E-01	8.41E-07	rs6083931	rs803880	4.88E-05	3.34E-03	4.88E-07
rs6083931	rs2795025	1.38E-05	2.48E-03	1.08E-06	rs6098930	rs348793	2.33E-04	1.17E-03	5.47E-07
rs6135844	rs16998508	6.71E-06	6.26E-03	1.33E-06	rs6115830	rs1971447	5.90E-03	1.95E-05	5.58E-07
rs7264608	rs6128273	5.39E-01	6.06E-08	1.34E-06	rs8120756	rs2567608	2.61E-01	6.40E-07	5.78E-07

FIGURE 7.12 – *P*-values combinées par la méthode de Fisher sans (à gauche) et avec (à droite) pondération au préalable.



Chapitre 8

Cap sur la recherche de polymorphismes liés aux microARNs

C'est pas la taille qui compte.

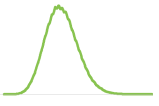
Mini ciabattas tomates & origan (blague Monoprix)

Le chapitre précédent a exposé les résultats des mes recherches d'interactions SNP-SNP impliquées dans la thrombose veineuse. Dans ce chapitre, je m'attaque à ce qui était au début de ma thèse mon principal projet : la recherche d'associations entre les polymorphismes liés aux microARNs et l'expression des gènes du monocyte. Les résultats présentés ici ont fait l'objet d'une publication pour laquelle je suis premier auteur [42].

8.1 Motivations et stratégie de recherche

8.1.1 Implication des microARNs dans de nombreuses maladies

Compte tenu de ce rôle régulateur important, il n'est pas surprenant que de nombreux microARNs soient rapportés comme étant associés à de nombreuses maladies. La base de données des maladies liées aux microARNs humains en répertorie près de 400 [73] et recense en particulier des associations récurrentes avec de nombreux cancers [98, 108]. Une partie de l'attention semble s'être récemment tournée vers le lien entre les microARNs et les maladies cardiovasculaires. Une simple recherche des mots-clés « MicroRNAs » et « Cardiovascular Diseases » dans la base



de données PubMed [160], par le moteur de recherche GoPubMed [1] [31, 149] m'a donné, au moment où j'écrivais ce document, 1 260 résultats dont 1 018 provenant d'articles publiés lors des trois dernières années. Il semble que de nombreux microARNs soient exprimés et jouent un rôle dans le bon fonctionnement des tissus du système cardiovasculaire [22]. Plusieurs articles rapportent leur implication dans les maladies cardiovasculaires [109] comme l'hypertrophie ventriculaire (miR-1, miR-133a) [50], l'infarctus du myocarde (miR-1, miR-133a, miR-133b, miR-208 [15], mir-199a [93], miR-320 [95]), la fibrose cardiaque (miR-21 [21], miR-29 [122]) ou le trouble du rythme cardiaque (miR-1 [136]).

8.1.2 La faute aux SNPs ?

Il semble désormais acquis qu'un SNP lié à un microARN peut affecter un phénotype. Cela fut montré pour la première fois en 2005, où un SNP situé dans un site de fixation pour le microARN hsa-miR-189, dans le gène *SLITRK1* fut trouvé associé au syndrome de Tourette [2]. Depuis de nombreux autres SNPs liés à des microARNs ont été rapportés comme associés à des maladies. En particuliers, des polymorphismes liés à certains microARNs (miR-196-a2, miR-146a, miR-27a) ont été identifiés à plusieurs reprises, avec de hauts niveau de significativité [108] comme associés à certains cancers.

8.1.3 Mécanisme d'action

Ces SNPs peuvent se situer dans un site de fixation de microARN (le plus souvent une région 3'UTR d'un ARN messager), mais aussi dans la séquence d'un microARN mature, d'un pré-microARN [106] ou d'un pri-microARN [133], en affectant la stabilité, l'efficacité ou la maturation du microARN [32]. Lorsqu'un SNP se situe dans la séquence d'un microARN mature ou dans une de ces cibles potentielles, il a en général pour effet d'altérer la fixation du microARN sur la cible, mais il peut aussi arriver qu'il ait pour effet la création d'un nouveau site de fixation [18].

8.1.4 Des exemples dans les maladies cardiovasculaires

Des SNPs liés à des microARNs ont aussi été rapportés comme associés à des maladies cardiovasculaires. Un SNP rare situé dans la séquence du microARN hsa-mir-499 semble par exemple altérer le fonctionnement de certains organes cardiaques [33]. D'autres polymorphismes, rs11614913 et rs3746444 situés dans les microARNs hsa-mir-196a2 et hsa-mir-499 respectivement, ont été trouvés associés

1. GoPubMed questionne la base de données de publications biologiques et médicales PubMed et propose notamment à l'utilisateur des termes de nomenclatures en fonction des mots-clés qu'il a renseignés, ceci afin d'améliorer la pertinence des résultats

aux cardiopathie congénitale [135] et coronarienne [139]. Le SNP rs4846049 dans le gène *MTHFR* fut également trouvé associé au risque de cardiopathie coronarienne, possiblement par le biais d'une modification d'un site de fixation de microARN, et en particulier de hsa-mir-149 [134]. Enfin, plusieurs polymorphismes situés dans des régions de fixations pour microARNs sur des gènes du système rénine-angiotensine-aldostérone (RAAS) semblent associés à des maladies cardiovasculaires. En particulier certains SNPs communs semblent influencer la pression artérielle et le risque d'infarctus [86] alors que le SNP rs5186 situé dans un site de fixation du gène *AGTR1* serait associé avec l'ataxie de Friedreich [59].

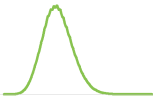
Tous ces éléments suggèrent que les variations situées dans les différentes séquences associées aux microARNs ou dans celles de leurs cibles pourraient, seules ou en interaction, avoir un rôle important dans les variations d'expression des gènes des tissus cardiovasculaires. À notre connaissance, très peu d'études ont consisté à la recherche sur l'ensemble du génome, de tels polymorphismes et c'est ce que nous avons tenté de faire dans ce chapitre, avec les données des études GHS et Cardiogenics.

8.1.5 Stratégie de recherche

- Dans un premier temps, j'ai identifié sur la puce à ADN utilisée dans GHS, les SNPs (ou leurs proxys) situés aux alentours des séquences des pri-microARNs ainsi que ceux situés dans les régions 3'UTR des gènes présents sur la puce à ARN.
- J'ai ensuite commencé par tester l'association des SNPs situés autours des microARNs, avec l'expression des gènes du monocyte.
- Puis, j'ai répliqué les résultats obtenus dans GHS, dans l'étude Cardiogenics.
- Enfin, j'ai testé l'ensemble de ces SNPs en interaction avec ceux situés dans les régions 3'UTR de gènes, sur l'expression de ces gènes.
- Avant de répliquer ces résultats dans l'étude Cardiogenics.

8.2 Identification des polymorphismes

Pour la localisation des SNPs, des microARNs, des gènes et des régions 3'UTR j'ai utilisé le génome de référence GRCH37 [53] (voir encadré). J'ai utilisé la base de données RefSeq [90] de NCBI (pour National Center for Biotechnology Information) pour identifier les gènes et leurs régions 3'UTR dans le génome de référence alors que la 17eme version de la base de données miRBase [43] m'a permis d'identifier les pre-microARNs. Comme il n'y avait pas à ma connaissance de base de données

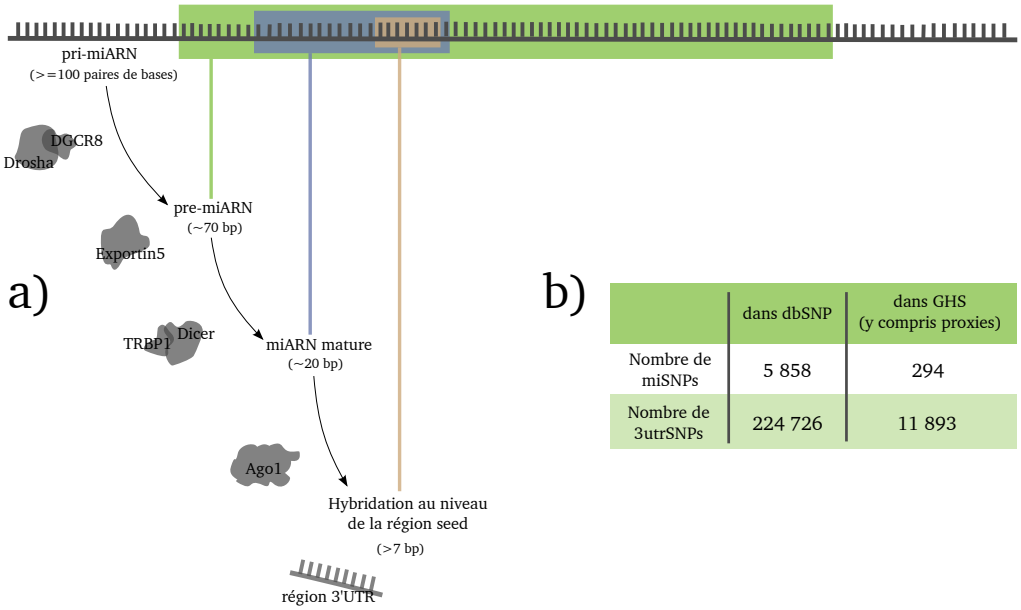


de pri-microARNs, j'ai simplement considéré comme faisant partie du pri-microARN, toutes les bases situées à moins de 200 bases du pre-microARN. Ce choix, quelque peu arbitraire, permet de s'assurer la capture de la majeure partie des pri-microARNs, tout en évitant d'ajouter dans l'analyse, un trop grand nombre de SNPs n'y étant pas réellement. Enfin, j'ai utilisé la version 131 de la base de données dbSNP [105] pour identifier l'ensemble des SNPs localisés dans les différentes régions concernées.

Le génome de référence GRCH37

Un génome de référence consiste en la séquence complète d'acides nucléiques d'un génome. C'est sur cette séquence complète que les scientifiques se basent ensuite pour déterminer les positions d'autres séquences particulières comme les gènes. GRCH37 (pour Genome Reference Consortium Human Genome - build 37) est le génome humain de référence produit par le GRC (Genome Reference Consortium) en Mai 2010, à partir du séquençage de 13 individus anonymes. C'est actuellement probablement le génome de référence humain le plus couramment utilisé.

Les nombres totaux de SNPs identifiés sont renseignés dans la figure 8.1 b). Les SNPs situés dans ou autour des microARNs sont par la suite appelés miSNPs, ceux situés dans les régions 3'UTR sont appelés 3utrSNPs.



b)

	dans dbSNP	dans GHS (y compris proxies)
Nombre de miSNPs	5 858	294
Nombre de 3utrSNPs	224 726	11 893

FIGURE 8.1 – a) Récapitulatif visuel des différents acteurs du processus de maturation et d'action des microARNs. b) Nombre de miSNPs et 3utrSNPs identifiés dans dbSNP et dans GHS.

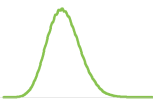
proxySNPs Comme cela a été expliqué dans le chapitre 5, lorsque l'on souhaite tester l'association entre un SNP en particulier et un certain phénotype, il n'est pas nécessaire que ce SNP soit inclus sur la puce à ADN utilisée. Il suffit que celle-ci contienne un SNP qui lui est fortement corrélé. C'est ce qui a été fait ici : Pour chaque SNP considéré comme lié aux microARNs, nous avons cherché, lorsqu'il n'était pas présent sur la puce à ADN de GHS Affymetrix 6.0, un SNP inclus sur la puce qui lui était corrélé ($r^2 > 0.9$). Aussi, alors qu'il n'y a que très peu de SNPs présents sur la puce utilisée parmi les SNPs identifiés précédemment, le nombre de ces SNPs qui sont corrélés à des SNPs présents permet finalement d'étudier l'association d'un certain nombre de SNPs liés aux microARNs. Les données de corrélations proviennent de l'application en ligne SNAP (SNp Annotation and Proxy search) [57]. Dans la suite du chapitre, j'utiliserai les termes proxy-utrSNP et proxy-miSNP pour désigner les SNPs de la puces représentant un utrSNP ou un miSNP (respectivement) non disponibles sur la puce.

8.3 L'association de ces SNPs sur l'expression des gènes

La première étape de l'analyse a consisté à tester l'ensemble des associations entre les miSNPs (c'est à dire les SNPs ou les marqueurs des SNPs situés dans ou à moins de 200 bases d'un pre-microARN) et les expressions des sondes de la puce à ARN avec comme hypothèse un lien linéaire et additif entre le miSNP et l'expression de la sonde (voir chapitre 4). Il en a résulté 294 miSNPs x 22 004 sondes = 6 469 176 tests qui ne sont en fait qu'un sous ensemble des résultats d'association génome entier déjà publiés par ailleurs [138]. Étant donné le nombre important de tests, même en absence totale d'association entre les miSNPs et l'expression des gènes, on s'attendrait à trouver par chance des p-values relativement faibles (voir chapitre 5). C'est pourquoi nous avons appliqué une correction pour tests multiples, en l'occurrence, la correction de Bonferroni. Nous avons déclaré significatifs les tests ayant des p-values inférieures à $0.05/6469176 = 7.73 \times 10^{-9}$.

8.3.1 De nombreuses associations significatives après correction de Bonferroni

Il résulte de cette analyse 57 associations significatives au seuil de Bonferroni (voir figure 8.2). Cependant, les interprétations de 48 d'entre elles paraissent relativement délicates car les proxy-miSNPs impliqués dans ces associations sont localisés dans des régions proches des gènes avec lesquels ils semblent associés. On peut alors facilement imaginer une association dite en « cis » (voir encadré), où le SNP responsable de l'association se trouve dans une région régulatrice du gène et affecte ainsi son expression sans passer par un microARN. Afin d'investiguer un



miSNPs				Expression de gène associée						GHS						Cardiogenics			
miSNP	miRNA	CHR	position	probe	gène	CHR	Début	Fin	best cis eSNP[1]	miProxy	beta[2]	SE	p[3]	p[4]	r²[5]	miProxy	beta[2]	SE	p[3]
rs1889470	hsa-mir-197	1	110141421	ILMN_1696003	GNAI3	1	110091185	110138451	rs6663547	rs1889470	-0.24	0.011	$2.03\ 10^{-87}$	0.024	0.82	rs2304355	-0.28	0.019	$8.20\ 10^{-42}$
rs7522956	hsa-mir-4742	1	224585958	ILMN_1714759	CNIH4	1	224544594	224563690	rs12731630	rs12731630	-0.09	0.011	$6.30\ 10^{-17}$	1	"-[6]	rs12731630	-0.06	0.023	0.015
rs6709245	hsa-mir-4784	2	132248876	ILMN_1766902	FAM128A	2	132241532	132250063	rs12463867	rs12463867	0.03	0.004	$1.37\ 10^{-14}$	1	"-[6]	rs6709245	NA	NA	NA
rs1829515	hsa-mir-4448	3	183604723	ILMN_1688464	MAP6D1	3	183533663	183543392	rs262998	rs6775950	0.07	0.007	$5.06\ 10^{-23}$	0.899	0.78	rs1554397	0.10	0.013	$2.42\ 10^{-13}$
rs1829515	hsa-mir-4448	3	183604723	ILMN_1651964	ABCC5	3	183637721	183735726	rs1402003	rs6775950	0.09	0.009	$2.80\ 10^{-22}$	0.893	0.44	rs1554397	0.10	0.017	$2.39\ 10^{-9}$
rs1829515	hsa-mir-4448	3	183604723	ILMN_1706531	ABCC5	3	183637721	183735726	rs17817736	rs6775950	0.08	0.010	$1.41\ 10^{-16}$	0.277	0.23	rs1554397	0.09	0.019	$1.29\ 10^{-6}$
rs10934682	hsa-mir-544b	3	124451312	ILMN_1757437	UMPS	3	124449212	124464039	rs614664	rs9878092	0.03	0.004	$3.12\ 10^{-9}$	0.026	0.33	rs13146	0.05	0.008	$2.33\ 10^{-10}$
rs778594	hsa-mir-3655	5	140027304	ILMN_1678957	WDR55	5	140044383	140050552	rs2563335	rs753280	0.03	0.008	$2.30\ 10^{-22}$	0.911	0.89	rs1583005	0.02	0.010	$7.75\ 10^{-16}$
rs778594	hsa-mir-3655	5	140027304	ILMN_2180624	TMCO6	5	140019011	140024988	rs17208187	rs753280	0.08	0.005	$2.30\ 10^{-11}$	$1.32\ 10^{-3}$	0.27	rs1583005	0.08	0.008	$6.03\ 10^{-3}$
rs778594	hsa-mir-3655	5	140027304	ILMN_1715113	HARS2	5	140071017	140078889	rs2245643	rs753280	0.05	0.007	$8.07\ 10^{-11}$	0.965	0.97	rs1583005	0.02	0.013	0.086
rs17165405	hsa-mir-4461	5	134263701	ILMN_1714599	CAMLG	5	134074205	134087846	rs11950943	rs3733897	0.07	0.013	$2.83\ 10^{-9}$	0.617	0.58	rs9285933	NA	NA	NA
rs403569	hsa-mir-1236	6	31924880	ILMN_1765532	RDBP	6	31919863	31926863	rs522162	rs550513	-0.07	0.009	$2.32\ 10^{-15}$	1	1.00	rs438999	-0.09	0.023	$2.03\ 10^{-4}$
rs403569	hsa-mir-1236	6	31924880	ILMN_1666512	SKIV2L	6	31926580	31937531	rs387608	rs550513	0.11	0.014	$3.78\ 10^{-14}$	0.817	0.60	rs438999	0.11	0.022	$2.78\ 10^{-6}$
rs107822	**hsa-mir-219-1**	**6**	**33175575**	**ILMN_1749070**	**HLA-DPB1**	**6**	**33043759**	**33054975**	**rs3128963**	**rs213208**	**0.18**	**0.019**	$\mathbf{1.34\ 10^{-19}}$	$\mathbf{6.56\ 10^{-17}}$	**0.36**	**rs439205**	**0.10**	**0.027**	$\mathbf{2.44\ 10^{-4}}$
rs4351242	hsa-mir-3135b	6	32717722	ILMN_1661266	HLA-DQB1	6	32627656	32634465	rs9272346	rs5014418	0.50	0.084	$4.13\ 10^{-9}$	0.533	0.03	rs7774954	0.51	0.109	$3.71\ 10^{-6}$
rs1049622	hsa-mir-4640	6	30858857	ILMN_1812262	DDR1	6	30851860	30867932	rs4711249	rs2074512	0.05	0.005	$3.02\ 10^{-27}$	0.773	0.82	rs3873334	0.04	0.019	0.055
rs221299	hsa-mir-153-2	7	157366855	ILMN_1728844	PTPRN2	7	157331749	158380481	rs221296	rs221296	0.21	0.012	$3.84\ 10^{-59}$	1	"-[6]	rs221277	0.29	0.020	$2.53\ 10^{-42}$
rs1649215	hsa-mir-548n	7	34980643	ILMN_1791296	DPY19L1	7	34968492	35077652	rs328928	rs1637670	-0.06	0.009	$5.62\ 10^{-12}$	$1.88\ 10^{-5}$	0.53	rs1186717	NA	NA	NA
rs10243209	**hsa-mir-548o**	**7**	**102046328**	**ILMN_1657317**	**POLR2J**	**7**	**102113546**	**102119380**	**rs11764413**	**rs6958795**	**-0.07**	**0.016**	$\mathbf{1.66\ 10^{-79}}$	$\mathbf{2.84\ 10^{-76}}$	**0.02**	**NA**	**NA**	**NA**	**NA**
rs10243209	hsa-mir-548o	7	102046328	ILMN_1804490	PRKRIP1	7	102036803	102067128	rs6958795	rs6958795	0.09	0.010	$1.47\ 10^{-17}$	1	"-[6]	NA	0.09	NA	NA
rs10243209	hsa-mir-548o	7	102046328	ILMN_1699049	PMS2L3	7	75137068	75157393	NA	rs6958795	0.32	0.008	$1.99\ 10^{-17}$	NA	NA	NA	0.43	NA	NA
rs2898254	hsa-mir-598	8	10892530	ILMN_1725338	CLDN23	8	8559665	8561616	rs940031	rs2898254	-0.07	0.011	$8.79\ 10^{-11}$	0.885	0.13	rs10100449	-0.04	0.018	0.028
rs2104533	hsa-mir-4670	9	95290456	ILMN_1748241	CENPP	9	95087740	95377445	rs10992312	rs2296666	0.02	0.003	$4.18\ 10^{-13}$	0.423	0.36	rs2895219	0.04	0.005	$8.45\ 10^{-17}$
rs7911488	hsa-mir-1307	10	105154089	ILMN_1773313	USMG5	10	105148813	105156222	rs2271751	rs2271751	0.07	0.009	$<10^{-400}$	1	"-[6]	rs1163073	NA	0.013	$4.12\ 10^{-314}$
rs7911488	hsa-mir-1307	10	105154089	ILMN_1772706	AS3MT	10	104629209	104661655	rs4919690	rs2271751	0.89	0.011	$1.99\ 10^{-11}$	0.053	0.14	rs1163073	0.88	NA	NA
rs3008373	hsa-mir-202	10	135060872	ILMN_1775074	TUBGCP2	10	135093137	135122643	rs3008334	rs3008309	-0.15	0.018	$1.65\ 10^{-10}$	0.345	0.51	rs3008373	-0.19	0.019	$8.90\ 10^{-3}$
rs3008373	hsa-mir-202	10	135060872	ILMN_1782352	VENTX	10	135051407	135055432	rs6537595	rs3008309	-0.12	0.025	$4.46\ 10^{-9}$	0.063	0.29	rs3008373	-0.05	0.040	$2.08\ 10^{-6}$
rs4919510	hsa-mir-608	10	102734778	ILMN_1678974	MRPL43	10	102737578	102747271	rs3824783	rs4919510	0.37	0.006	$<10^{-400}$	0.873	0.96	rs3740487	NA	NA	NA
rs4919510	hsa-mir-608	10	102734778	ILMN_2258774	MRPL43	10	102737578	102747271	rs3740488	rs4919510	-0.27	0.010	$2.38\ 10^{-139}$	0.071	0.96	rs3740487	-0.26	0.016	$3.74\ 10^{-50}$
rs11032942	**hsa-mir-1343**	**11**	**34963459**	**ILMN_2128293**	**APIP**	**11**	**34903841**	**34937938**	**rs2915178**	**rs11032937**	**-0.11**	**0.010**	$\mathbf{1.31\ 10^{-27}}$	$\mathbf{2.06\ 10^{-19}}$	**0.04**	**rs12282098**	**-0.13**	**0.018**	$\mathbf{7.22\ 10^{-13}}$
rs11032942	hsa-mir-1343	11	34963459	ILMN_1793598	APIP	11	34903841	34937938	rs2915214	rs11032937	-0.10	0.014	$1.38\ 10^{-13}$	$2.49\ 10^{-6}$	0.04	rs12282098	NA	NA	NA
rs1463335	**hsa-mir-1279**	**12**	**69667075**	**ILMN_1748730**	**CTRC**	**1**	**15764937**	**15773152**	**rs6693417**	**rs317657**	**-0.03**	**0.004**	$\mathbf{1.39\ 10^{-13}}$	$\mathbf{4.76\ 10^{-15}}$	**0.00**	**rs998022**	**-0.06**	**0.007**	$\mathbf{1.54\ 10^{-15}}$
rs1463335	hsa-mir-1279	12	69667075	ILMN_2252021	LRRFIP1	2	238536223	238690289	NA	rs317657	-0.05	0.004	$1.50\ 10^{-35}$	NA	NA	rs998022	-0.12	0.010	$6.65\ 10^{-32}$
rs1463335	hsa-mir-1279	12	69667075	ILMN_1699317	CNTN6	3	1134628	1445277	NA	rs317657	-0.02	0.003	$1.16\ 10^{-12}$	NA	NA	rs998022	-0.04	0.006	$7.56\ 10^{-12}$
rs1463335	hsa-mir-1279	12	69667075	ILMN_1740494	PCDHA6	5	140207649	140391928	NA	rs317657	-0.04	0.003	$9.44\ 10^{-33}$	NA	NA	rs998022	-0.10	0.008	$2.67\ 10^{-31}$
rs1463335	hsa-mir-1279	12	69667075	ILMN_1663381	TRAF3IP2	6	111880142	111927320	NA	rs317657	-0.03	0.003	$2.74\ 10^{-17}$	NA	NA	rs998022	-0.06	0.007	$5.23\ 10^{-17}$
rs1463335	hsa-mir-1279	12	69667075	ILMN_2114422	NOD1	7	30464142	30518392	NA	rs317657	0.05	0.008	$7.25\ 10^{-9}$	NA	NA	rs998022	0.12	0.013	$7.83\ 10^{-19}$
rs1463335	hsa-mir-1279	12	69667075	ILMN_1731063	ST5	11	8714898	8932497	NA	rs317657	-0.06	0.007	$2.05\ 10^{-18}$	NA	NA	rs998022	-0.22	0.019	$2.51\ 10^{-30}$
rs1463335	**hsa-mir-1279**	**12**	**69667075**	**ILMN_1815205**	**LYZ**	**12**	**69742133**	**69748012**	**rs11177644**	**rs317657**	**0.20**	**0.010**	$\mathbf{1.36\ 10^{-76}}$	$\mathbf{6.17\ 10^{-11}}$	**0.34**	**rs998022**	**NA**	**NA**	**NA**
rs1463335	hsa-mir-1279	12	69667075	ILMN_1801387	YEATS4	12	69753531	69784575	rs11177644	rs317657	0.15	0.010	$1.32\ 10^{-46}$	0.734	0.34	rs998022	0.19	0.020	$3.27\ 10^{-21}$
rs1463335	hsa-mir-1279	12	69667075	ILMN_1792568	KRT9	17	39722092	39728309	NA	rs317657	-0.04	0.006	$1.15\ 10^{-15}$	NA	NA	rs998022	-0.11	0.016	$1.11\ 10^{-11}$
rs1463335	**hsa-mir-1279**	**12**	**69667075**	**ILMN_1667361**	**COPZ2**	**17**	**46103532**	**46115151**	**rs17617053**	**rs317657**	**-0.03**	**0.005**	$\mathbf{2.33\ 10^{-11}}$	$\mathbf{7.68\ 10^{-12}}$	**0.00**	**rs998022**	**-0.10**	**0.011**	$\mathbf{2.06\ 10^{-18}}$
rs4566029	hsa-mir-4305	13	40238387	ILMN_1779530	COG6	13	40229763	40365801	rs1028671	rs4598779	-0.07	0.007	$1.69\ 10^{-23}$	0.015	0.60	rs4569133	-0.11	0.010	$6.51\ 10^{-26}$
rs2168519	hsa-mir-4513	15	75080872	ILMN_1679495	ULK3	15	75128458	75135551	rs936230	rs7180484	0.10	0.006	$1.83\ 10^{-63}$	0.105	0.79	rs2168519	0.10	0.016	$9.23\ 10^{-11}$
rs3785626	hsa-mir-1288	17	16185248	ILMN_1760332	PIGL	17	16120508	16229572	rs1476810	rs3785626	0.04	0.004	$2.31\ 10^{-37}$	0.109	0.79	rs2109262	0.07	0.008	$4.47\ 10^{-11}$
rs3785626	hsa-mir-1288	17	16185248	ILMN_1703946	ADORA2B	17	15848230	15879209	rs2324142	rs3785626	0.06	0.006	$7.39\ 10^{-11}$	0.839	0.61	rs2109262	0.05	0.012	$9.69\ 10^{-9}$
rs2271891	hsa-mir-3185	17	46802008	ILMN_1810274	HOXB2	17	46620016	46622392	rs12939811	rs4793601	-0.08	0.013	$1.63\ 10^{-10}$	0.727	0.07	rs1054072	-0.08	0.022	$3.18\ 10^{-4}$
rs6505162	hsa-mir-423	17	28444183	ILMN_1672834	SSH2	17	27952964	28257017	rs4474741	rs6505162	0.10	0.008	$5.63\ 10^{-36}$	0.678	0.66	NA	NA	NA	NA
rs8078913	hsa-mir-4520a	17	6558768	ILMN_1794987	ALOX15P	17	6560519	6564747	rs9889941	rs2040847	-0.02	0.003	$2.12\ 10^{-13}$	$1.75\ 10^{-3}$	0.43	NA	NA	NA	NA
rs2916076	hsa-mir-640	19	19545990	ILMN_1685631	KIAA0892	19	19431629	19469562	rs968525	rs2916074	-0.11	0.007	$5.89\ 10^{-52}$	0.257	0.35	NA	NA	NA	NA
rs2916076	hsa-mir-640	19	19545990	ILMN_2134224	ATP13A1	19	19756006	19774502	rs16996185	rs2916074	0.05	0.013	$6.63\ 10^{-36}$	$1.86\ 10^{-3}$	0.17	NA	NA	NA	NA
rs2916076	hsa-mir-640	19	19545990	ILMN_1662741	LPAR2	19	19734463	19739038	rs880090	rs2916074	-0.09	0.010	$5.64\ 10^{-22}$	0.323	0.34	NA	NA	NA	NA
rs2916076	hsa-mir-640	19	19545990	ILMN_1673682	GATAD2A	19	19496641	19619740	rs4808967	rs2916074	0.17	0.007	$2.77\ 10^{-11}$	0.16	0.98	NA	NA	NA	NA
rs2284385	hsa-mir-4755	20	32637001	ILMN_1787410	EIF6	20	33866724	33872593	rs6060244	rs2284390	0.15	0.024	$7.86\ 10^{-10}$	0.062	0.20	rs2038123	0.13	0.030	$1.46\ 10^{-5}$
rs5750504	**hsa-mir-659**	**22**	**38243675**	**ILMN_1757467**	**H1F0**	**22**	**38201113**	**38203442**	**rs763137**	**rs4821723**	**-0.06**	**0.018**	$\mathbf{3.58\ 10^{-20}}$	$\mathbf{8.89\ 10^{-14}}$	**0.02**	**rs9610864**	**-0.08**	**0.027**	$\mathbf{7.66\ 10^{-14}}$
rs5750504	hsa-mir-659	22	38243675	ILMN_1724437	GCAT	22	38203954	38212903	rs2285177	rs4821723	0.16	0.003	$1.88\ 10^{-13}$	0.376	0.51	rs9610864	0.20	0.005	$2.63\ 10^{-7}$
rs5750504	hsa-mir-659	22	38243675	ILMN_1735788	TRIOBP	22	38092994	38172562	rs4821708	rs4821723	0.03	0.009	$7.12\ 10^{-11}$	0.922	0.44	rs9610864	0.03	0.013	$2.21\ 10^{-9}$

FIGURE 8.2 – Associations significatives entre les miSNPs et les expressions monocytaires. (1) SNPs situé dans une région d'un million de bases autour du miSNP et affichant la plus forte association (avec une p-value $< 5 \times 10^{-5}$) avec l'expression de la sonde associée. (2) Estimation du paramètre associé au miSNP après ajustement sur l'âge et le sexe. (3) p-value de l'association entre le miSNP et l'expression du gène. (4) p-value après ajustement sur le best cis eSNP. (5) Corrélation entre le miSNP et le best cis eSNP dans GHS. (6) Le best cis eSNP coïncide avec le miSNP.

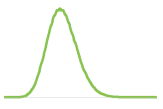
peu ce problème d'interprétation, nous avons recherché les SNPs proches (à moins d'un million de bases) et les plus associés aux expressions des gènes concernés, dans GHS. Ils seront appelés par la suite best cis-SNPs. Six des proxy-miSNPs impliqués dans les 48 cis-associations sont également des best cis-SNPs. Pour les 42 autres associations, on a essayé de savoir si celles-ci étaient indépendantes de l'association avec le best cis-SNP en ajustant les modèles avec ce best cis-SNP. Ceci a eu pour conséquences de faire disparaître 35 associations. L'hypothèse d'une action d'un microARN pour les 7 autres associations (en gras dans la tableau 8.2) doit cependant être considérée avec prudence, car l'analyse effectuée ne nous permet pas d'exclure la possibilité d'associations dues à un déséquilibre de liaison entre le proxy-miSNP et le vrai cis-SNP. Une implication du microARN reste néanmoins plausible pour celles dont le microARN concerné se situe dans l'intron du gène avec lequel le miSNP semble associé.

Association en « cis » / en « trans »

On dit qu'une association est en « cis » (signifiant « du même côté » en latin) lorsque l'association se fait entre un gène et un élément variable (ici un SNP) qui lui est proche. Ce type d'association est à opposer à l'association en « trans » (« de l'autre côté » en latin) où l'association implique deux éléments éloignés sur le génome. En général, on considère que les associations entre les miSNPs et les expressions des gènes sont en « trans », car leurs séquences respectivess ne sont a priori pas adjacentes.

8.3.2 Un cluster d'associations intriguant

En ce qui concerne les neuf associations en « trans » significatives après correction de Bonferroni (encadrés dans la figure 8.2), il est intéressant de remarquer que toutes impliquent le miSNP rs1463335 situé dans le pri-microARN hsa-mir-1279 et marqué par le proxy-miSNP rs317657, présent sur la puce de GHS (avec une corrélation parfaite entre le miSNP et son proxy-miSNP : $r^2 = 1.0$). Ce miSNP semble donc associé en « trans » avec les gènes *CNTN6* (p-value=1.16×10^{-12}), *CTRC* (1.39×10^{-13}), *COPZ2* (2.33×10^{-11}), *KRT9* (1.15×10^{-15}), *LRRFIP1* (1.5×10^{-35}), *NOD1* (7.25×10^{-9}), *PCDHA6* (9.44×10^{-33}), *ST5* (2.05×10^{-18}) et *TRAF3IP2* (2.74×10^{-17}) alors qu'il est aussi associé en « cis » avec *LYZ* (1.39×10^{-76}) et *YEATS4* (1.32×10^{-46}) (voir figure 8.3). Ces associations sont relativement fortes, en témoignent les carrés du coefficient de corrélation (R^2) entre le proxy-miSNP rs317657 et les expressions des gènes associés allant d'environ 2 % pour l'association avec *NOD1* à 10 % pour *LRRFIP1* et même 20 % pour celle avec *LYZ*. Les expressions



de *LYZ*, *YEATS4* et *NOD1* sont augmenté avec la présence de l'allèle C de ce SNP tandis qu'elle fait décroître les expressions des autres gènes cités.

					GHS			Cardiogenics		
Probe	Gène	CHR	Début	Fin	beta[2]	SE	P[3]	beta[2]	SE	P[3]
ILMN_1748730	*CTRC*	1	15764937	15773152	-0,03	0,004	$1.39\ 10^{-13}$	-0,06	0,01	$1.54\ 10^{-15}$
ILMN_2252021	*LRRFIP1*	2	238536223	238690289	-0,05	0,004	$1.50\ 10^{-35}$	-0,12	0,01	$6.65\ 10^{-32}$
ILMN_1699317	*CNTN6*	3	1134628	1445277	-0,02	0,003	$1.16\ 10^{-12}$	-0,04	0,01	$7.56\ 10^{-12}$
ILMN_1740494	*PCDHA6*	5	140207649	140391928	-0,04	0,003	$9.44\ 10^{-33}$	-0,10	0,01	$2.67\ 10^{-31}$
ILMN_1663381	*TRAF3IP2*	6	111880142	111927320	-0,03	0,003	$2.74\ 10^{-17}$	-0,06	0,01	$5.23\ 10^{-17}$
ILMN_2114422	*NOD1*	7	30464142	30518392	0,05	0,008	$7.25\ 10^{-9}$	0,12	0,01	$7.83\ 10^{-19}$
ILMN_1731063	*ST5*	11	8714898	8932497	-0,06	0,007	$2.05\ 10^{-18}$	-0,22	0,02	$2.51\ 10^{-30}$
ILMN_1815205	*LYZ*[1]	12	69742133	69748012	0,20	0,010	$1.36\ 10^{-76}$	NA	NA	NA
ILMN_1801387	*YEATS4*[1]	12	69753531	69784575	0,15	0,010	$1.32\ 10^{-46}$	0,19	0,02	$3.27\ 10^{-21}$
ILMN_1792568	*KRT9*	17	39722092	39728309	-0,04	0,006	$1.15\ 10^{-15}$	-0,11	0,02	$1.11\ 10^{-11}$
ILMN_1667361	*COPZ2*	17	46103532	46115151	-0,03	0,005	$2.33\ 10^{-11}$	-0,10	0,01	$2.06\ 10^{-18}$

FIGURE 8.3 – Associations entre le miSNP rs1463335 et les gènes *CTRC*, *LRRFIP1*, *CNTN6*, *PCDHA6*, *TRAF3IP2*, *NOD1*, *ST5*, *LYZ*, *YEATS4*, *KRT9* et *COPZ2*, dans GHS et Cardiogenics.

8.3.3 Des associations significatives, même après ajustements

L'association avec *LYZ* étant la plus forte, j'ai recherché son meilleur cis-SNP. Après avoir ajusté l'expression de *LYZ* pour son best cis-SNP, son association en cis avec le proxy-miSNP reste significative ($p = 6.17 \times 10^{-11}$) tandis que celle de *YEATS4* disparaît ($p = 0.734$). D'après TargetScan, un programme en ligne de prédiction de cibles de microARNs, les positions 648 à 654 de la région 3'UTR de *LYZ* sont complémentaires sur 8 bases au microARN hsa-mir-1279. Ce type de complémentarité appelé 8mer est assez habituel dans les séquences réellement ciblés par les microARN ce qui renforce l'hypothèse d'une régulation du microARN sur *LYZ*. Il est important d'avoir conscience cependant que ce genre de configuration est relativement courant et n'assure en aucune manière que le microARN qui est complémentaire à la séquence du gène sur 8mer le régule ce qui empêche une interprétation évidente d'action du miSNP sur *LYZ*. Après ajustement des associations en « trans » sur *LYZ*, la plupart de ces associations restent significatives ($p = 3.88 \times 10^{-11}, 1.15 \times 10^{-7}, 2.52 \times 10^{-6}, 1.65 \times 10^{-10}, 7.16 \times 10^{-29}, 2.44 \times 10^{-5}, 8.23 \times 10^{-28}, 1.81 \times 10^{-13} et 5.66 \times 10^{-10}$ respectivement pour *CNTN6*, *CTRC*, *COPZ2*, *KRT9*, *LRRFIP1*, *NOD1*, *PCDHA6*, *ST5* et *TRAF3IP2*). L'ajustement sur *YEATS4* donnant des p-values respectives égales à $1.86 \times 10^{-9}, 1.72 \times 10^{-11}, 6.45 \times 10^{-9}, 9.48 \times 10^{-12}, 6.10 \times 10^{-28}, 3.76 \times 10^{-13}, 1.59 \times 10^{-28}, 2.33 \times 10^{-13} et 5.10 \times 10^{-8}$. Les ajustement sur *LYZ* et *YEATS4* ensemble ne changeant pas fondamentalement les associations 2.98×10^{-6} pour *COPZ2* à 6.55×10^{-27} pour *PCDHA6*. La figure 8.4 donne une représentation de la région autour du miSNP rs1463335. On peut aussi voir dans la table 8.5 que ces neuf gènes ne sont pas très corrélés entre eux, comme ils ne le sont pas fortement non plus avec *LYZ*, le gène dans lequel se trouve le proxy-miSNP rs317657.

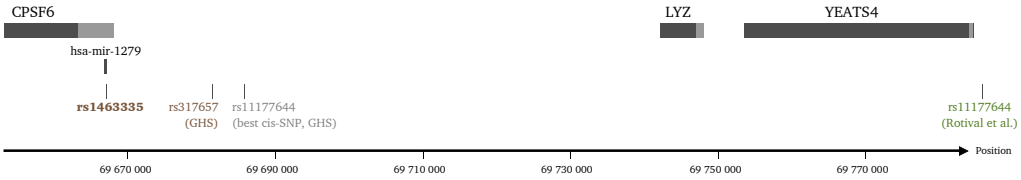


FIGURE 8.4 – Région de l'association entre le miSNP rs1463335 et les gènes *LIZ* et *YEATS4*, sur le chromosome 12. La légende de la figure est la même que celle de la figure 8.12.

	CTRC	LRRFIP1	CNTN6	PCDHA6	TRAF3IP2	NOD1	ST5	LYZ	YEATS4	KRT9
LRRFIP1	0,20	1,00								
CNTN6	0,14	0,24	1,00							
PCDHA6	0,20	0,45	0,20	1,00						
TRAF3IP2	0,13	0,27	0,20	0,27	1,00					
NOD1	0,23	-0,13	0,05	-0,06	0,03	1,00				
ST5	0,21	0,52	0,19	0,41	0,27	-0,18	1,00			
LYZ	-0,16	-0,14	-0,07	-0,13	-0,17	0,11	-0,13	1,00		
YEATS4	-0,08	-0,16	-0,11	-0,11	-0,25	-0,07	-0,14	0,56	1,00	
KRT9	0,22	0,49	0,17	0,40	0,30	-0,17	0,74	-0,13	-0,12	1,00
COPZ2	0,19	0,40	0,13	0,34	0,24	-0,14	0,59	-0,14	-0,09	0,59

FIGURE 8.5 – Corrélations entre les gènes du cluster.

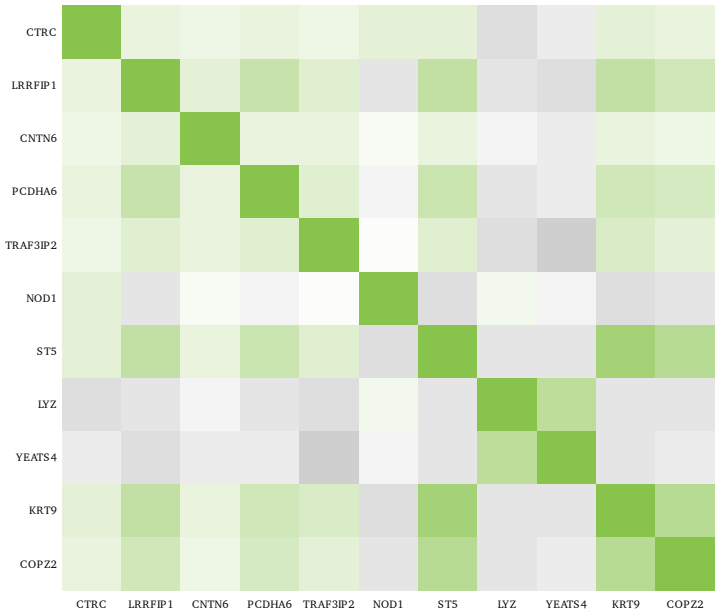


FIGURE 8.6 – Représentation de la corrélations entre les gènes du cluster de *LYZ*.

8.3.4 Réplication dans l'étude Cardiogenics

Nous avons ensuite essayé de répliquer ces résultats dans l'étude Cardiogenics. Le miSNP rs1463335 n'étant pas inclus dans la puce utilisée pour cette étude, nous avons trouvé un marqueur qui lui était corrélé ($r^2 = 0.9$), le proxy-miSNP rs998022. Sa corrélation avec le proxy-miSNP de GHS était par ailleurs de 0.84. La sonde

mesurant l'expression de *LYZ* utilisée dans GHS n'était pas non plus présente dans Cardiogenics mais toutes les autres associations ont pu être répliquées et semblent confirmer les associations trouvées dans GHS : La présence de l'allèle G du proxy-miSNP rs998022 (marqueur pour l'allèle C du proxy-miSNP rs317657 de GHS) est associée à une augmentation de l'expression des gènes *YEATS4* ($p = 3.21 \times 10^{-21}$) et *NOD4* ($p = 7.83 \times 10^{-19}$), et à une diminution des gènes *CNTN6* ($p = 7.56 \times 10^{-12}$), *CTRC* ($p = 1.54 \times 10^{-15}$), *COPZ2* ($p = 2.06 \times 10^{-18}$), *KRT9* ($p = 1.11 \times 10^{-11}$), *LRRFIP1* ($p = 6.65 \times 10^{-32}$), *PCDHA6* ($p = 2.67 \times 10^{-31}$), *ST5* ($p = 2.51 \times 10^{-30}$) et *TRAF3IP2* ($p = 5.23 \times 10^{-17}$) (voir figure 8.3). Ces associations sont aussi bien présentes chez les cas que chez les témoins de l'étude Cardiogenics (voir figure 8.7).

Expression de gène associée					Cas			Témoins		
Probe	Gène	CHR	Début	Fin	beta[2]	SE	P[3]	beta[2]	SE	P[3]
ILMN_1748730	CTRC	1	15764937	15773152	-0.05	0.01	$1.5\ 10^{-6}$	-0.07	0.01	$9.0\ 10^{-11}$
ILMN_2252021	LRRFIP1	2	238536223	238690289	-0.11	0.01	$9.6\ 10^{-14}$	-0.14	0.01	$1.0\ 10^{-20}$
ILMN_1699317	CNTN6	3	1134628	1445277	-0.03	0.01	$8.2\ 10^{-6}$	-0.04	0.01	$3.3\ 10^{-7}$
ILMN_1740494	PCDHA6	5	140207649	140391928	-0.08	0.01	$4.8\ 10^{-11}$	-0.12	0.01	$3.1\ 10^{-23}$
ILMN_1663381	TRAF3IP2	6	111880142	111927320	-0.06	0.01	$4.9\ 10^{-8}$	-0.07	0.01	$5.7\ 10^{-11}$
ILMN_2114422	NOD1	7	30464142	30518392	0.11	0.02	$5.1\ 10^{-10}$	0.12	0.02	$2.2\ 10^{-11}$
ILMN_1731063	ST5	11	8714898	8932497	-0.20	0.03	$2.2\ 10^{-12}$	-0.25	0.03	$6.5\ 10^{-21}$
ILMN_1815205	LYZ[1]	12	69742133	69748012	NA	NA	NA	NA	NA	NA
ILMN_1801387	YEATS4[1]	12	69753531	69784575	0.20	0.03	$2.4\ 10^{-10}$	0.18	0.02	$9.7\ 10^{-13}$
ILMN_1792568	KRT9	17	39722092	39728309	-0.12	0.02	$2.3\ 10^{-7}$	-0.10	0.02	$3.6\ 10^{-6}$
ILMN_1667361	COPZ2	17	46103532	46115151	-0.09	0.02	$8.1\ 10^{-8}$	-0.12	0.02	$5.3\ 10^{-13}$

FIGURE 8.7 – Associations entre le miSNP rs1463335 et les gènes *CTRC*, *LRRFIP1*, *CNTN6*, *PCDHA6*, *TRAF3IP2*, *NOD1*, *ST5*, *LYZ*, *YEATS4*, *KRT9* et *COPZ2* chez les cas ainsi que chez les témoins d'après l'étude Cardiogenics.

8.4 Recherche d'interactions SNP-SNP impliquées dans la variabilité de l'expression des gènes

Chacun des 3utrSNPs fut ensuite testé en interaction avec tous les miSNPs, sur les expressions des sondes du gène dans lequel il se trouve. Comme pour la recherche d'association directe, le modèle utilisé est un modèle linéaire intégrant les deux SNPs ainsi que leur terme d'interaction, le tout ajusté sur l'âge et le sexe. Ces modèles sont décrits plus en détail dans le chapitre 4. Le nombre total d'interactions testées fut de 4 890 102.

8.4.1 Correction pour tests multiples avec pondération

Au lieu d'appliquer la correction pour Bonferroni standard pour gérer le grand nombre de tests effectué, nous avons suivi la suggestion de Pare et al. [88] en adoptant une correction pour Bonferroni pondérée par la *p*-value du test de Levene.

La méthode de pondération a été décrite dans le chapitre 5 alors que le test de Levene a été présenté au chapitre 4. Pour rappel, le test de Levene permet de détecter des différences de variances entre plusieurs groupes. Ici les groupes sont définis par les trois différents génotypes de chaque 3utrSNP. S'il y a des différence de variances pour les expressions des gènes associés (auquel cas la p-value du test de Levene sera faible), entre les génotypes, cela suggère peut-être que ce génotype fait apparaître ou inhibe l'effet d'une autre variable et ce 3utrSNP interagit donc avec cette variable. Pondérer par la p-value du test de Levene consiste ensuite à donner plus de poids aux tests pour lesquels la variance des expression change en suivant le génotype. Sous l'hypothèse qu'un résultat significatif du test de Levene représente une telle interaction, cette procédure devrait permettre de détecter plus facilement les interactions entre 3utrSNPs et utrSNPs.

8.4.2 Résultats de l'analyse dans GHS

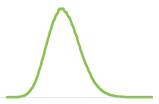
Après avoir appliqué cette correction de Bonferroni pondérée (seuil de significativité à 1.02×10^{-8}), 51 interaction miSNP-3utrSNPs se révèlent être significatifs (voir tableau 8.8). En utilisant la correction de Bonferroni standard, seules 31 interactions passent ce seuil de significativité. L'utilisation de la pondération par la p-value du test de Levene modifie les rangs des différents tests et permet ici d'augmenter sensiblement notre puissance de détection de phénomènes d'interactions.

Parmi les 51 interactions significatives, 17 impliquent le 3utrSNP rs13053624 du gène *RFPL1* pour moduler l'expression de la sonde ILMN_1797383 [1]. Ce 3utrSNP interagirait notamment avec le microARN hsa-mir-3674 et d'après la base de données microSNiPer [7], le gène *RFPL1* aurait un SNP (le SNP rs13053817) dans un site potentiel de fixation pour ce microARN. D'après la base de données SNAP, ce SNP est en fort déséquilibre de liaison avec notre 3utrSNP rs13053624 ($r^2 = 0.90$). Nous n'avons pas pu trouver d'information parmi les bases de données de prédictions de sites de fixation pour microARNs allant dans le sens de nos résultats pour les 30 autres interactions.

8.4.3 Réplication des résultats dans Cardiogenics

Nous avons essayé de répliquer les 51 interactions significatives dans Cardiogenics mais du fait de puces différentes, seules huit de ces interactions ont pu effectivement être testées. Parmi ces interactions, aucune n'impliquait le 3utrSNP rs13053624 du gène *RFPL1* (représenté par la sonde ILMN_1797383). En utilisant le même

1. Les noms des sondes provenant des puces à ARN de la société Illumina consistent en un numéro, précédé des lettres ILMN et d'un tiret bas.



							GHS				
Gène	CHR	Probe	3utrSNP	miARN	CHR	miSNP	miProxy	3utrProxy	P [1]	P-value de Levene	P-value Pondérée [2]
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-592	7	rs11563750	rs11563505	rs13053817	$1.04\ 10^{-35}$	$3.22\ 10^{-5}$	$1.50\ 10^{-36}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-3920	11	rs12275715	rs12283329	rs13053817	$1.21\ 10^{-26}$	$3.22\ 10^{-5}$	$1.74\ 10^{-27}$
TXNDC5	6	ILMN_1769082	rs8643	hsa-mir-125b-2	21	rs2823897	rs2211981	rs8643	$8.95\ 10^{-18}$	$3.39\ 10^{-1}$	$1.23\ 10^{-17}$
TXNDC5	6	ILMN_1769082	rs1043784	hsa-mir-125b-2	21	rs2823897	rs2211981	rs3734589	$1.26\ 10^{-17}$	$3.18\ 10^{-1}$	$1.64\ 10^{-17}$
LYZ	12	ILMN_1815205	rs710794	hsa-mir-1279	12	rs1463335	rs317657	rs710794	$4.13\ 10^{-15}$	$4.51\ 10^{-23}$	$1.20\ 10^{-16}$
ASB1	2	ILMN_1683096	rs1044561	hsa-mir-125b-2	21	rs2823897	rs2211981	rs2334004	$1.45\ 10^{-16}$	$8.91\ 10^{-1}$	$1.87\ 10^{-15}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4656	7	rs3750013	rs17135110	rs13053817	$2.28\ 10^{-14}$	$3.22\ 10^{-5}$	$3.29\ 10^{-15}$
ASB1	2	ILMN_1683096	rs2278768	hsa-mir-3119-1	1	rs17349873	rs1330387	rs2278768	$3.71\ 10^{-14}$	$1.34\ 10^{-6}$	$4.10\ 10^{-15}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-30c-1	1	rs16827546	rs16827546	rs13053817	$2.89\ 10^{-14}$	$3.22\ 10^{-5}$	$4.16\ 10^{-15}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-1307	10	rs7911488	rs2271751	rs9287035	$2.98\ 10^{-13}$	$9.07\ 10^{-46}$	$4.29\ 10^{-15}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-125b-1	11	rs2081443	rs2081443	rs13053817	$2.40\ 10^{-13}$	$3.22\ 10^{-5}$	$3.47\ 10^{-14}$
PKD1L2	16	ILMN_1742788	rs1901818	hsa-mir-4272	3	rs9868022	rs9868022	rs7198127	$8.92\ 10^{-14}$	$8.80\ 10^{-2}$	$5.47\ 10^{-14}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-4670	9	rs2104533	rs2296666	rs9287035	$5.16\ 10^{-12}$	$9.07\ 10^{-46}$	$7.42\ 10^{-14}$
ASB1	2	ILMN_1683096	rs2278768	hsa-mir-125b-2	21	rs2823897	rs2211981	rs2278768	$5.30\ 10^{-12}$	$1.34\ 10^{-6}$	$5.85\ 10^{-12}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4300	11	rs11603185	rs7944477	rs13053817	$2.02\ 10^{-11}$	$3.22\ 10^{-5}$	$2.92\ 10^{-12}$
SPRY1	4	ILMN_2329914	rs300574	hsa-mir-4666	1	rs16841344	rs4653963	rs300555	$1.52\ 10^{-11}$	$1.16\ 10^{-2}$	$5.10\ 10^{-12}$
HLA-DPB1	6	ILMN_1749070	rs1042448	hsa-mir-219-1	6	rs107822	rs213208	rs3128923	$1.26\ 10^{-10}$	$4.11\ 10^{-8}$	$1.11\ 10^{-11}$
ASB1	2	ILMN_1683096	rs2278768	hsa-mir-4636	5	rs257095	rs6555591	rs2278768	$1.09\ 10^{-10}$	$1.34\ 10^{-6}$	$1.20\ 10^{-11}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4292	9	rs2811749	rs2811749	rs13053817	$1.98\ 10^{-10}$	$3.22\ 10^{-5}$	$2.86\ 10^{-11}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-624	14	rs11156654	rs11156654	rs13053817	$2.20\ 10^{-10}$	$3.22\ 10^{-5}$	$3.18\ 10^{-11}$
GPRC5C	17	ILMN_1724211	rs2706527	hsa-mir-3667	22	rs135771	rs135775	rs2706526	$5.46\ 10^{-9}$	$5.08\ 10^{-79}$	$4.52\ 10^{-11}$
H1F0	22	ILMN_1757467	rs1894644	hsa-mir-659	22	rs5750504	rs2899293	rs763137	$2.98\ 10^{-10}$	$1.30\ 10^{-1}$	$2.18\ 10^{-10}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-548n	7	rs1649215	rs1637670	rs9287035	$1.64\ 10^{-8}$	$9.07\ 10^{-46}$	$2.37\ 10^{-10}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-521-1	19	rs4803178	rs4803178	rs13053817	$2.88\ 10^{-9}$	$3.22\ 10^{-5}$	$4.16\ 10^{-10}$
GPRC5C	17	ILMN_2352090	rs2706527	hsa-mir-3667	22	rs135771	rs135775	rs2706526	$1.06\ 10^{-7}$	$6.63\ 10^{-102}$	$6.80\ 10^{-10}$
GPRC5C	17	ILMN_2352090	rs2706527	hsa-mir-107	10	rs17481096	rs17481096	rs2706526	$1.20\ 10^{-7}$	$6.63\ 10^{-102}$	$7.69\ 10^{-10}$
HLA-DPB1	6	ILMN_1749070	rs1042448	hsa-mir-219-1	6	rs213210	rs213210	rs3128923	$8.98\ 10^{-9}$	$4.11\ 10^{-8}$	$7.88\ 10^{-10}$
MXRA7	17	ILMN_1743836	rs10473	hsa-mir-490	7	rs6963819	rs2350780	rs7221855	$2.66\ 10^{-7}$	$6.10\ 10^{-167}$	$1.04\ 10^{-9}$
SPRY1	4	ILMN_1651610	rs300574	hsa-mir-4666	1	rs16841344	rs4653963	rs300555	$3.82\ 10^{-9}$	$6.28\ 10^{-3}$	$1.12\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-1236	6	rs403569	rs550513	rs13053817	$7.89\ 10^{-9}$	$3.22\ 10^{-5}$	$1.14\ 10^{-9}$
GPRC5C	17	ILMN_2352090	rs2706527	hsa-mir-941-1	20	rs2427555	rs2427554	rs2706526	$2.03\ 10^{-7}$	$6.63\ 10^{-102}$	$1.30\ 10^{-9}$
POGZ	1	ILMN_2329309	rs3811409	hsa-mir-4666	1	rs16841344	rs4653963	rs3811409	$2.24\ 10^{-9}$	$1.12\ 10^{-1}$	$1.53\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4643	6	rs16884450	rs16884450	rs13053817	$1.28\ 10^{-8}$	$3.22\ 10^{-5}$	$1.85\ 10^{-9}$
ASB1	2	ILMN_1683096	rs1044561	hsa-mir-3973	11	rs262404	rs16928224	rs2334004	$1.60\ 10^{-10}$	$8.91\ 10^{-1}$	$2.06\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-3646	20	rs11574730	rs11574730	rs13053817	$1.70\ 10^{-8}$	$3.22\ 10^{-5}$	$2.45\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-4460	5	rs13171514	rs13171514	rs9287035	$2.47\ 10^{-7}$	$9.07\ 10^{-46}$	$3.55\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-3674	8	rs7003112	rs6558541	rs13053817	$2.55\ 10^{-8}$	$3.22\ 10^{-5}$	$3.67\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-1205	8	rs9649959	rs9649959	rs13053817	$2.78\ 10^{-8}$	$3.22\ 10^{-5}$	$4.02\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4656	7	rs17829969	rs17829969	rs13053817	$2.82\ 10^{-8}$	$3.22\ 10^{-5}$	$4.07\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-4784	2	rs6709245	rs12463867	rs9287035	$3.22\ 10^{-7}$	$9.07\ 10^{-46}$	$4.63\ 10^{-9}$
AAK1	2	ILMN_1880387	rs13427243	hsa-mir-3667	22	rs135771	rs135775	rs13427243	$7.28\ 10^{-9}$	$1.04\ 10^{-1}$	$4.80\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-604	10	rs2368392	rs3758371	rs13053817	$3.69\ 10^{-8}$	$3.22\ 10^{-5}$	$5.32\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-215	1	rs3820455	rs34406824	rs9287035	$3.88\ 10^{-7}$	$9.07\ 10^{-46}$	$5.58\ 10^{-9}$
RBM12	20	ILMN_1670841	rs6060539	hsa-mir-4755	20	rs2284385	rs2284390	rs2425125	$4.06\ 10^{-7}$	$1.65\ 10^{-47}$	$5.62\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-2113	6	rs9375085	rs9375085	rs9287035	$4.02\ 10^{-7}$	$9.07\ 10^{-46}$	$5.79\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-1269b	17	rs7210937	rs2240567	rs13053817	$4.93\ 10^{-8}$	$3.22\ 10^{-5}$	$7.10\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-4705	13	rs7337292	rs7337292	rs9287035	$5.10\ 10^{-7}$	$9.07\ 10^{-46}$	$7.33\ 10^{-9}$
PKD1L2	16	ILMN_1742788	rs1901818	hsa-mir-4473	9	rs16938058	rs16938057	rs7198127	$1.24\ 10^{-8}$	$8.80\ 10^{-2}$	$7.60\ 10^{-9}$
MRPL43	10	ILMN_1678974	rs2295716	hsa-mir-608	10	rs4919510	rs4919510	rs3824783	$3.06\ 10^{-7}$	$9.68\ 10^{-22}$	$9.44\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-520d	19	rs2217653	rs9304754	rs9287035	$6.62\ 10^{-7}$	$9.07\ 10^{-46}$	$9.52\ 10^{-9}$
ASB1	2	ILMN_1683096	rs1044561	hsa-mir-4636	5	rs257095	rs6555591	rs2334004	$7.57\ 10^{-10}$	$8.91\ 10^{-1}$	$9.74\ 10^{-9}$

FIGURE 8.8 – Les 51 interactions significatives d'après la *p*-value pondérée par le test de Levene et après correction de Bonferroni.

modèle que pour l'étude GHS, mais en ajustant en plus sur le statut « malade/non malade », deux des huit interactions furent répliquées au seuil de Bonferroni (p-value $= 6.25 \times 10^{-3}$) dans Cardiogenics (voir figure 8.9).

8.4.4 Interaction *HLA-DPB1*/hsa-mir-219-1

Localisation

La première interaction répliquée implique le 3utrSNP rs1042448 du gène *HLA-DPB1* et le miSNP rs107822 du microARN hsa-mir-219-1. Les miSNPs et 3utrSNPs ne

miSNP x 3utrSNP	miRNA (CHR)	Gène (CHR)	Probe	Cas			Témoins		
				Proxies	beta[1]	P-value [2]	Proxies	beta[1]	P-value [2]
rs17349873 rs2278768	hsa-mir-3119-1 (1)	ASB1 (2)	ILMN_1683096	rs1330387 rs2278768	0.04	0.83	rs6703198 rs10084192	0.30	0.23
rs107822 rs1042448	hsa-mir-219-1 (6)	HLA-DPB1 (6)	ILMN_1749070	rs213208 rs3128923	-0.25	**6.6 10^{-6}**	rs439205 rs3117222	-0.29	**8.9 10^{-9}**
rs257095 rs2278768	hsa-mir-4636 (5)	ASB1 (2)	ILMN_1683096	rs6555591 rs2278768	0.02	0.89	rs257095 rs10084192	0.07	0.44
rs5750504 rs1894644	hsa-mir-659 (22)	H1F0 (22)	ILMN_1757467	rs2899293 rs763137	-0.26	1.4 10^{-4}	rs6000905 rs1894644	-0.25	**1.0 10^{-4}**
rs6963819 rs10473	Hsa-mir-490 (7)	MXRA7 (7)	ILMN_1743836	rs2350780 rs7221855	0.00	0.98	rs2350780 rs9910052	0.02	0.53
rs262404 rs1044561	hsa-mir-3973 (11)	ASB1 (2)	ILMN_1683096	rs16928224 rs2334004	-0.14	0.37	rs262407 rs10084192	0.11	0.50
rs2284385 rs6060539	hsa-mir-4755 (20)	RBM12 (20)	ILMN_1670841	rs2284390 rs2425125	0.09	0.21	rs2038123 rs6121015	0.10	0.24
rs257095 rs1044561	hsa-mir-4636 (5)	ASB1 (2)	ILMN_1683096	rs6555591 rs2334004	0.02	0.89	rs257095 rs10084192	0.07	0.44

FIGURE 8.9 – Réplication dans l'étude Cardiogenics des interactions miSNP x 3utrSNPs détectées dans l'étude GHS.

sont ni sur la puce de GHS ni sur celle de Cardiogenics. Cependant, le 3utrSNP est bien représenté par les proxySNPs rs3128923 dans GHS et rs213208 dans Cardiogenics alors que le miSNP est lui, en fort déséquilibre de liaison avec les proxySNPs rs3117222 dans GHS et rs439205 dans Cardiogenics (voir figure 8.10). Le miSNP et le 3utrSNP sont assez proches (environ distants de 100 000 bases) et leurs proxySNPs respectifs sont en léger déséquilibre de liaison ($r^2 = 0.58$ dans GHS et $r^2 = 0.56$ dans Cardiogenics).

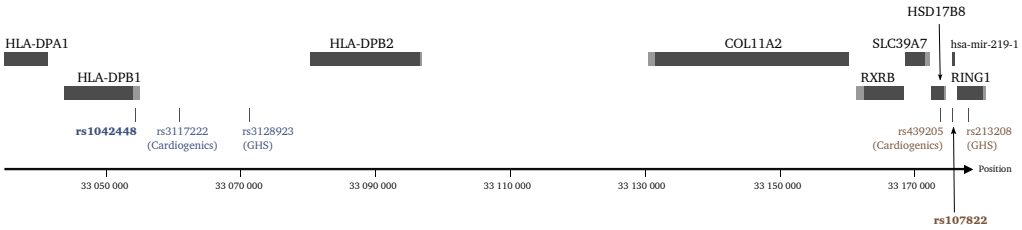


FIGURE 8.10 – Localisation des gènes, 3utrSNP, miSNP et proxySNPs situés dans la région de l'interaction *HLA-DPB1*/hsa-mir-219-1, sur le chromosome 6. La légende de la figure est la même que celle de la figure 8.12.

Analyse haplotypique

Dans GHS, l'analyse haplotypique des proxySNPs révèle que l'allèle A du 3utrSNP rs1042448 est associé à une forte augmentation de l'expression du gène *HLA-DPB1* ($\beta = 0.61$, p-value$= 1.64 \times 10^{-105}$) lorsqu'il est associée à l'allèle C du miSNP rs107822 (voir figure 8.11). Inversement, lorsqu'il est associé avec l'allèle T du miProxy rs107822, l'effet de l'allèle A du 3utrSNP rs1042448 du gène *HLA-DPB1* est significativement réduit (p-value$= 1.88 \times 10^{-20}$) et passe à $\beta = 0.18$ (p-value=

3.49×10^{-8}), ce qui illustre bien le phénomène d'interaction identifié par la régression. Cette interaction reste significative (p-value$= 2.81 \times 10^{-12}$) si l'on ajuste l'analyse haplotypique par le best cis-SNP affectant l'expression de HLA-DPB1, rs3128963 (p-value$= 2.30 \times 10^{-151}$, voir la base de données GHS_Express [138]). Les mêmes motifs se retrouvent dans Cardiogenics (voir figure 8.11) : l'augmentation du niveau d'expression du gène *HLA-DPB1* est importante lorsque l'allèle A du 3utrSNP est porté sur le même haplotype que l'allèle C du miSNP ($\beta = 0.63$, p-value$= 5.24 \times 10^{-62}$). A l'inverse, s'il se trouve avec l'allèle A du miSNP, cette augmentation est fortement réduite (p-value$= 2.68 \times 10^{-20}$) et ne passe plus le seuil de significativité ($\beta = 0.05$, p-value$= 0.23$).

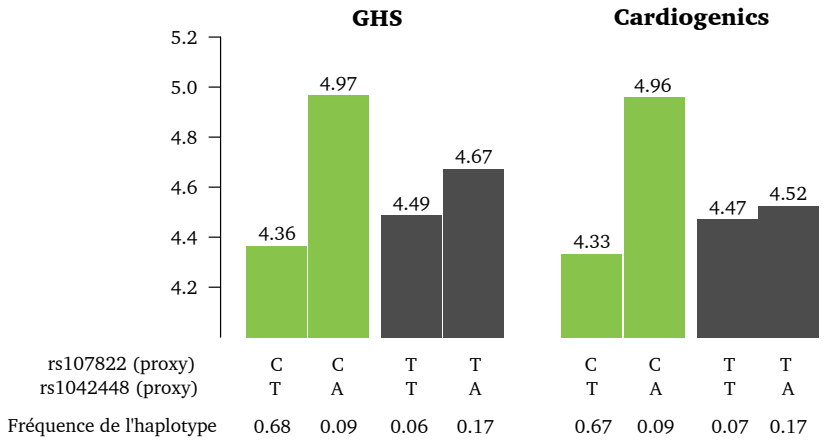


FIGURE 8.11 – Niveaux d'expressions du gène *HLA-DPB1* dans le monocyte, selon les haplotypes dérivés du 3utrSNP rs1042448 du gène *HLA-DPB1* et du miSNP rs107822 du microARN hsa-mir-219-1. La paire de SNPs rs1042448/rs107822 est représentée par la paire rs3128923/rs213208 dans GHS et par la paire rs3117222/rs439205 dans Cardiogenics.

8.4.5 Interaction *H1F0*/hsa-mir-659

Localisation

La seconde interaction répliquée dans Cardiogenics implique le 3utrSNP rs1894644 du gène *H1F0* et le miSNP rs5750504 du microARN hsa-mir-659. Ces deux SNPs ne sont pas sur la puce de GHS et sont représentés par les proxySNPs rs763137 et rs2899293 pour le 3utrSNP et le miSNP respectivement. Le 3utrSNP rs1894644 est en revanche présent sur la puce de Cardiogenics où le proxySNP rs6000905 fut utilisé comme marqueur du miSNP rs5750504 (voir figure 8.12). Les locus du 3utrSNP et du miSNPs sont distants d'environ 40 000 bases et leurs proxy respectifs sont en faible déséquilibre de liaison ($r^2 = 0.15$ dans GHS, $r^2 = 0.14$ en Cardiogenics).

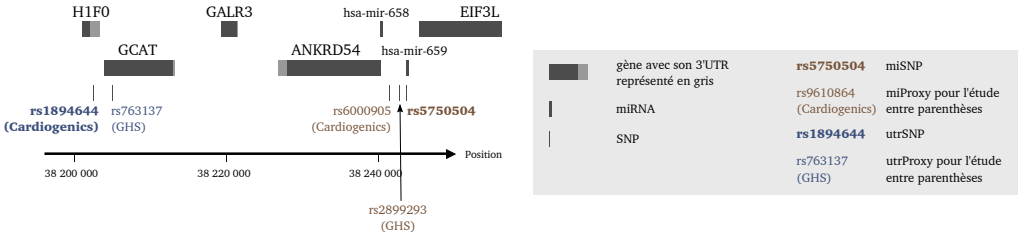


FIGURE 8.12 – Localisation des gènes, 3utrSNP, miSNP et proxySNPs situés dans la région de l'interaction *H1F0*/hsa-mir-659, sur le chromosome 22.

Analyse haplotypique

Dans les deux études, l'allèle T du 3utrSNP est associé à une forte augmentation de l'expression du gène *H1F0* ($\beta = +0.65$, p-value$= 1.71 \times 10^{-53}$ dans GHS et $\beta = +0.79$, p-value$= 1.36 \times 10^{-40}$ dans Cardiogenics) lorsqu'il est porté avec l'allèle T du miSNP rs5750504 (voir figure 8.13). Inversement, lorsque l'allèle T de ce 3utrSNP se trouve sur le même haplotype que l'allèle A du miSNP, l'augmentation d'expression de H1F0 est plus faible ($\beta = +0.23$, p-value$= 9.74 \times 10^{-13}$ dans GHS et $\beta = +0.26$, p-value$= 7.25 \times 10^{-8}$ dans Cardiogenics). On peut aussi noter que dans GHS, le proxySNP rs763137 qui représente le 3utrSNP est aussi le best cis-SNP pour le gène *H1F0* (p-value$=1.1010^{-62}$). Ces réplications dans Cardiogenics sont homogènes entre les cas et les témoins (voir 8.14).

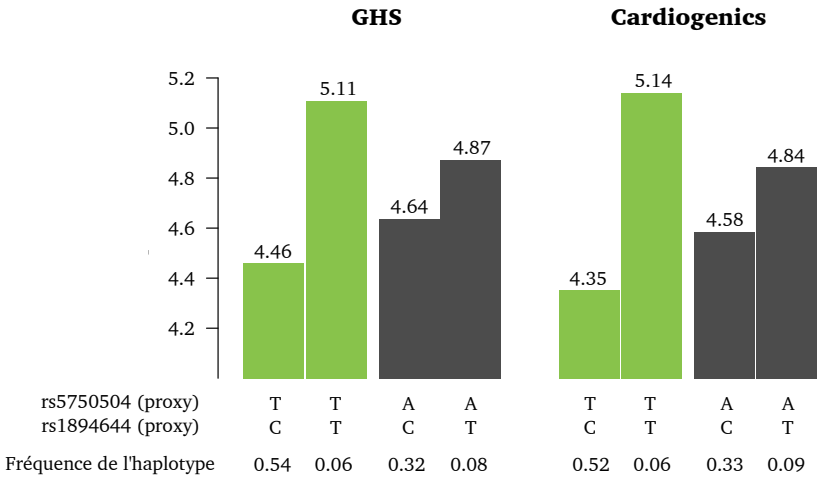


FIGURE 8.13 – Niveaux d'expressions du gène *H1F0* dans le monocyte, selon les haplotypes dérivés du 3utrSNP rs1894644 du gène *H1F0* et du miSNP rs5750504 du microARN hsa-mir-659. La paire de SNPs rs1894644/rs5750504 est représentée par la paire rs763137/rs2899293 dans GHS et par la paire rs1894644/rs6000905 dans Cardiogenics.

miSNP x 3utrSNP	miRNA (CHR)	Gène (CHR)	Probe	Cas			Témoins		
				Proxies	beta[1]	P-value [2]	Proxies	beta[1]	P-value [2]
rs17349873 rs2278768	hsa-mir-3119-1 (1)	ASB1 (2)	ILMN_1683096	rs1330387 rs2278768	0.04	0.83	rs6703198 rs10084192	0.30	0.23
rs107822 rs1042448	hsa-mir-219-1 (6)	HLA-DPB1 (6)	ILMN_1749070	rs213208 rs3128923	-0.25	**$6.6\,10^{-6}$**	rs439205 rs3117222	-0.29	**$8.9\,10^{-9}$**
rs257095 rs2278768	hsa-mir-4636 (5)	ASB1 (2)	ILMN_1683096	rs6555591 rs2278768	0.02	0.89	rs257095 rs10084192	0.07	0.44
rs5750504 rs1894644	hsa-mir-659 (22)	H1F0 (22)	ILMN_1757467	rs2899293 rs763137	-0.26	$1.4\,10^{-4}$	rs6000905 rs1894644	-0.25	**$1.0\,10^{-4}$**
rs6963819 rs10473	Hsa-mir-490 (7)	MXRA7 (7)	ILMN_1743836	rs2350780 rs7221855	0.00	0.98	rs2350780 rs9910052	0.02	0.53
rs262404 rs1044561	hsa-mir-3973 (11)	ASB1 (2)	ILMN_1683096	rs16928224 rs2334004	-0.14	0.37	rs262407 rs10084192	0.11	0.50
rs2284385 rs6060539	hsa-mir-4755 (20)	RBM12 (20)	ILMN_1670841	rs2284390 rs2425125	0.09	0.21	rs2038123 rs6121015	0.10	0.24
rs257095 rs1044561	hsa-mir-4636 (5)	ASB1 (2)	ILMN_1683096	rs6555591 rs2334004	0.02	0.89	rs257095 rs10084192	0.07	0.44

FIGURE 8.14 – Les associations dans Cardiogenics, séparément chez les cas et les témoins, pour les huit interactions significatives dans GHS et réplicables dans Cardiogenics.

Chapitre 9

Discussions et perspectives

Je déteste les discussions : elles vous font parfois changer d'avis.

Oscar Wilde

Les deux chapitres précédents ont présenté les résultats de mes recherches de phénomènes d'interactions entre polymorphismes dans le cadre des maladies multifactorielles. Ce chapitre a pour objectif de discuter ces résultats, de donner de nouvelles perspectives de recherches et de proposer quelques hypothèses pour expliquer ce que nous avons observé.

Rappelons pour commencer que notre objectif n'était pas de tester ou comparer l'ensemble des méthodes de détection d'interaction SNP-SNP disponibles, cela n'aurait de toute façon pas été possible, mais de tenter plusieurs stratégies de recherche et d'appliquer plusieurs méthodes permettant d'en augmenter la puissance.

9.1 Sur la recherche d'interactions entre polymorphismes dans la thrombose veineuse

9.1.1 Rappel des résultats obtenus

Ce travail est à notre connaissance, le première tentative de détection de phénomènes d'interaction associés à la thrombose veineuse à l'échelle du génome entier.

Notre stratégie de recherche ne nous a pas permis d'identifier de nouveaux variants susceptibles de contribuer à la maladie. Elle nous a cependant permis de

tester différentes approches de combinaisons de tests et de pondérations. Elle nous a aussi servi de support pour tester et appliquer nos calculs de puissance, qui ont par ailleurs révélé que l'utilisation de polymorphismes communs était nécessaire à la détection d'interactions avec suffisamment de puissance, dans des études du type de EOVT ou MARTHA. Enfin, en testant certaines des interactions les plus prometteuses sur certains biomarqueurs quantitatifs, nous avons pu mettre en évidence une association significative entre l'interaction SNP-SNP rs9804128-rs4784379 et le niveau plasmatique de facteur VIII.

9.1.2 Analyse et perspectives ouvertes par nos travaux

En revanche les raisons pour lesquelles nous n'avons pas pu détecter d'interaction associée à la maladie thromboembolique veineuse ne sont pas encore identifiées. Il se pourrait d'abord qu'il n'y ait pas d'interaction entre polymorphismes contribuant à la variabilité de la pathologie. Cette hypothèse est cependant en légère contradiction avec les observations de chercheurs travaillant sur la maladie qui suggèrent au contraire que la maladie thromboembolique veineuse pourrait provenir de multiples interactions entre de nombreux facteurs de risque génétiques ou environnementaux [17]. Il est ensuite possible que notre manque de résultats provienne de notre stratégie de recherche pour laquelle nous pourrions imaginer différentes améliorations :

- Nous pourrions tester d'autres méthodes de pondérations comme celle basée sur la *p*-value du test de Levene.
- Nous pourrions aussi ne pas nous restreindre au seul chromosome 20 pour effectuer ces pondérations. Cependant, cela aurait aussi pour conséquence d'augmenter le nombre de tests effectués ce qui n'est pas forcément souhaitable.
- A l'inverse, nous pourrions nous restreindre à des SNPs fonctionnels, situés dans des gènes et dont les différentes formes modifient la séquence protéique induite.
- Une taille d'échantillon plus importante nous permettrait de gagner en puissance et ainsi de pouvoir détecter des effets modestes.

9.1.3 Réflexions liées à nos recherches

Finalement, une autre explication de notre manque de réussite dans la détection d'associations significatives pourrait résider dans la nature des interactions impactant la thrombose veineuse.

Hypothèse de multiples combinaisons concurrentes La thrombose veineuse est une maladie complexe survenant fréquemment des suites d'autres maladies telles que les maladies inflammatoires ou les cancers [19] et dont les facteurs de risques avérés sont très diverses (immobilisation, traumatisme, tabac, pilule contraceptive, etc.). Ce constat serait facilement expliqué par une multitude de mécanismes différents, impliquant des acteurs différents, mais dont la combinaison engendrerait la maladie. Dans un tel scénario, on peut imaginer que la présence de certains facteurs génétiques soient protecteurs pour certaines personnes et au contraire à risque pour d'autres, si bien que sur une large population, il est difficile d'en détecter les effets.

Difficulté de détection de telles interactions Sous cette hypothèse de multiples combinaisons de facteurs concurrents, il faudrait pour détecter ces combinaison rechercher des interactions entre multiples polymorphismes génétiques (pas seulement deux que nous avons fait ici). Une telle entreprise serait cependant difficile car si la détection d'interaction entre deux polymorphismes reste faisable, pour des interactions entre trois, quatre ou plus de polymorphismes, cela devient extrêmement compliqué. Le nombre de combinaisons augmenterait exponentiellement et nécessiterait un très grand nombre d'individus et d'importantes capacités de calculs.

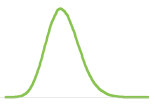
9.2 Sur la recherche de polymorphismes liés aux microARNs et leurs impacts sur l'expression des gènes

9.2.1 Rappel des résultats obtenus

Là encore, ce travail est à notre connaissance le premier à explorer l'ensemble des microARNs à la recherche de SNP qui seraient, en interaction avec d'autres SNPs de leurs régions cibles, associés à l'expression de certains gènes. Il nous a fourni deux résultats très intéressant mais dont les interprétations doivent donner lieu à prudence, notamment en ce qui concerne les implications de microARNs dans les associations détectées.

Le cluster impliquant le microARN hsa-mir-1279

L'analyse d'association simple dans GHS a permis de détecter un cluster de gènes qui pourraient être régulés par le microARN hsa-mir-1279, mais différentiellement selon l'allèle présent au miSNP rs1463335. Les associations identifiées sont fortes et répliquées dans l'étude Cardiogenics ce qui a poussé certains de nos collaborateurs en Allemagne à entamer des analyses fonctionnelles sur ce cluster. De mon point de vue, ces résultats révèlent très probablement la présence d'un phénomène biologique réel important, mais l'implication du microARN hsa-mir-1279 dans ce phénomène



ne me paraît pas forcément évidente. En effet, le microARN en question se trouve dans la séquence du gène *CPSF6*, sous-unité d'un facteur nécessaire notamment à la maturation des région 3'UTR des ARN messager lors de la transcription. Il est possible que hsa-mir-1279 joue un rôle dans le phénomène, mais le gène *CPSF6* semble aussi un bon candidat pour être impliqué dans la régulation du cluster de gènes. De plus, les gènes *LYZ* et *YEATS4* n'étant pas éloignés du microARN, on ne peut exclure que la variation responsable de l'association observée soit située dans une région régulatrice d'un de ces deux gènes et que les associations sur les autres gènes du cluster passent par ce gène. .Nous avons cependant montré que nos associations restaient significatives après ajustements sur ces gènes, ce qui tend à réfuter cette dernière hypothèse.

Les interactions avec les gènes *HLA-DPB1* et *H1F0*

En ce qui concerne les deux résultats d'interaction trouvés associés aux expressions des gènes *HLA-DPB1* et *H1F0*, ma réserve viendrait cette fois de la proximité des miSNPs potentiellement impliqués dans ces associations. En effet, le miSNP rs107822 du microARN hsa-mir-219-1 se trouve à environ 120 000 bases du gène *HLA-DPB1* il n'est pas impossible que le miSNP soit un marqueur pour une variation située dans une séquence régulatrice du gène. De même, le miSNP rs5750504 du microARN hsa-mir-659 se situe à 40 000 bases du gène *H1F0* et l'on peut là aussi imaginer que l'association implique une interaction entre deux SNPs liés au gène. L'hypothèse d'une implication des microARNs n'est pas à exclure mais d'un point de vue statistique, on pourrait se demander quelles étaient les chances que les deux associations détectées (et l'on pourrait rajouter les associations du cluster) impliquent des miSNPs proches des gènes impactés.

9.2.2 Analyse et perspectives ouvertes par nos travaux

Je pense qu'il est important aussi d'avoir un regard critique de notre démarche afin d'avoir des pistes de travail pour de possibles améliorations.

La puce utilisée D'abord, la puce à ADN utilisée dans ce travail était une puce classique d'analyse de SNPs sur le génome. Ce type de puces n'est pas forcément adapté à l'identification de miSNPs. Il existe maintenant des puces spécifiquement dédiées aux identifications de variations dans les microARNs. Il est probable que de telles puces pourraient nous fournir nombre de nouveaux polymorphismes potentiellement impliqués dans l'expression des gènes et notamment des polymorphismes situés dans les séquences seed et mature des microARNs, plus

à même d'avoir des impacts en interaction avec des polymorphismes dans leurs séquences cibles.

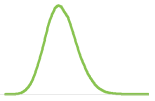
L'identification des miSNPs Du fait de l'absence de base de données de pri-microARNs, nous avons assimilé comme étant des miSNPs tous les SNPs situés à 200 bases d'un microARN. Il est probable que la connaissance précise des séquences des pri-microARNs pourrait être bénéfique à notre travail.

L'identification des 3utrSNPs De même nous avons déclarés 3utrSNPs, tous les SNPs situés dans les régions 3'UTR de gènes pour lesquels nous avions une donnée d'expression. Bien que les algorithmes de prédiction de cible de microARNs ne soient pas forcément parfaits, ils permettent cependant de détecter les régions le plus à même d'être des sites de fixation pour microARNs et nous pourrions sûrement tirer bénéfice de leur utilisation. Par exemple, nous pourrions pondérer nos résultats par certains scores de prédiction de cible de microARNs. Nous pourrions aussi imaginer d'autres types de pondérations, outre celle par la *p*-value de Lévène déjà effectuée. Il est aussi connu que les microARNs ciblent en priorité des séquences d'ARN messager situés dans leur région 3'UTR mais qu'ils peuvent aussi cibler des séquences de leurs régions 5'UTR ou des ORFs [1]. Nous pourrions donc également explorer ces régions.

Phénotype d'intérêt Une autre réflexion que l'on pourrait mener dans le cadre de notre travail serait de se demander si les expressions des gènes sont bien le bon phénotype à étudier pour observer des associations avec des polymorphismes situés dans les séquences des microARNs. En effet, les microARNs régulent les gènes après la transcription et le débat de savoir s'ils ne font que réguler la production de protéines ou s'il peuvent au contraire influencer sur la quantité d'ARN messager reste d'actualité. La tendance semblerait aller vers la seconde solution [51], ce qui justifie notre démarche. Il n'en reste pas moins qu'une partie de la régulation par les microARNs s'effectue au niveau de la traduction de l'ARN messager et n'est donc pas détectable avec notre stratégie.

Type cellulaire étudié Enfin, il est possible que le monocyte ne soit pas le type cellulaire idéal pour la détection de changement d'expression par des miSNPs. Il a été montré que certains microARNs pouvaient être mis en cause dans certains types de cancer du sang et notamment au sein du monocyte pour plusieurs types de leucémies [16] mais nous pourrions envisager d'effectuer des recherches similaires

1. Pour rappel, les ORFs (pour open reading frame) sont les régions d'un gène potentiellement traduites en protéine.



dans le macrophage pour lequel nous avons des données d'expression dans l'étude Cardiogenics.

9.2.3 Réflexions liées à nos recherches

Conservation des microARNs

On s'attendrait a priori à pouvoir détecter facilement des SNPs situés dans les séquences des microARNs. En effet, une variation de séquence dans un microARN mature ou dans sa région seed a le potentiel d'altérer le profil de fixation de ce microARN pour n'importe quelle région de n'importe quel ARN messager susceptible de gagner ou perdre un peu en complémentarité avec sa séquence. Un tel potentiel d' impact a cependant pour probable conséquence une attention particulière de l'organisme pour ce genre de variations. Alors que les polymorphismes situés dans les régions 3'UTR ciblés par les microARNs sont relativement fréquents, ce n'est pas le cas pour les variations situées dans les séquences des microARNs. Les microARNs sont bien conservés entre les espèces proches et les variations au sein de leur séquence mature et de leur région seed sont très rares [20, 100]. Parmi l'ensemble des miSNP identifiés dans GHS, seuls 5 se situaient dans les séquences matures ou seed des microARNs. Cependant, nous n'avons pas trouvé de sur-représentation d'association chez ces miSNPs par rapport à l'ensemble de ceux étudiés.

Interprétation finale des résultats

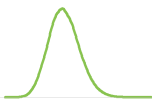
Au final, j'aurai plutôt tendance à penser que notre recherche d'associations entre SNPs liés aux microARNs et expression de gènes a donné un résultat très intéressant qui est que les SNPs liés aux microARNs n'ont peut-être en général qu'un effet modeste sur l'expression des gènes. Si tel était le cas, il ne serait alors pas illogique que nous n'ayons pu trouver d'association avec une réelle implication de SNPs liés aux microARNs. Plusieurs hypothèses de mécanisme d'autorégulation de l'organisme en présence de variations génétiques dans les microARNs peuvent être avancées :

Régulation par les gènes ciblés Les algorithmes de prédictions estiment que chaque microARN pourrait réguler des centaines ou milliers de gènes. Aussi, il est possible que l'ensemble des gènes régulés par un microARN comportent des gènes qui se régulent entre eux. Par exemple, on peut imaginer qu'une variation dans un microARN empêche la régulation d'un gène, mais aussi celle d'un autre qui serait régulateur de ce gène. Ainsi, il y aurait un mécanisme de compensation interne à l'ensemble des gènes ce qui atténuerait les effets d'une variation dans un microARN.

Régulation par les autres microARNs De même, il semble que chaque gène puisse comporter des sites de fixations pour plusieurs microARNs. Il n'apparaîtrait alors pas illogique que notre organisme se soit créé un mécanisme de régulation basé sur la redondance. Plusieurs microARNs ciblent les même gènes et même si l'une des fixations n'est pas possible à cause d'une variation dans un microARN, les autres microARNs continuer de réguler les gènes ciblés par ce microARN ce qui atténue l'effet de la variation [32].

Régulation par les sites de fixation Enfin, une dernière hypothèse pourrait être que contrairement aux estimations effectuées, le nombre de gènes régulés par un microARN soit très réduit, mais que de nombreux gènes puissent cependant accueillir la fixation du microARN, sans que cette fixation n'ait de conséquences particulières. Ainsi, la plupart des sites de fixation aurait simplement pour rôle de réguler l'action du microARN en lui faisant « perdre son temps » ce qui atténuerait là encore les effets des microARN et ainsi des variations qu'ils pourraient contenir [104].

L'hypothèse de l'organisme statisticien Toutes ces hypothèses ne sont pas incompatibles entre elles et il est possible que la réalité consiste en une combinaison de tous ces mécanismes et probablement de bien d'autres. En particulier, je ne peux résister à la tentation d'imaginer notre organisme être un statisticien hors pair en matière de régulation. En multipliant les mécanismes de compensation et la redondance, je l'imagine ne pas empêcher les dysrégulations, mais au contraire, les favoriser, de sorte qu'un mécanisme de régulation déficient à un endroit soit forcément compensé par d'autres dispositifs défaillants à d'autres endroits, pour finalement que l'ensemble ne soit que peu affecté.



Bibliographie

Articles, livres, thèses

1. 3C Study Group. Vascular factors and risk of dementia : design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology* **22,** 316–25 (2003) (cf. p. 79).
2. Abelson, J. F., Kwan, K. Y., O'Roak, B. J. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science (New York, N.Y.)* **310,** 317–20 (2005) (cf. p. 100).
3. Andrieu, N., Dondon, M.-G. & Goldstein, A. M. Increased power to detect gene-environment interaction using siblings controls. *Annals of epidemiology* **15,** 705–11 (2005) (cf. p. 54).
4. Antoni, G. *Identification de facteurs génétiques modulant deux phénotypes intermédiaires de la maladie thromboembolique veineuse : les taux de facteurs VIII et de Von Willebrand* thèse de doct. (Université Paris-Sud, 2012) (cf. p. 92).
5. Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of experimental medicine* **79,** 137–58 (1944) (cf. p. 26).
6. Barbosa-Morais, N. L., Dunning, M. J., Samarajiwa, S. A. *et al.* A re-annotation pipeline for Illumina BeadArrays : improving the interpretation of gene expression data. *Nucleic acids research* **38,** e17 (2010) (cf. p. 82).
7. Barenboim, M., Zoltick, B. J., Guo, Y. *et al.* MicroSNiPer : a web tool for prediction of SNP effects on putative microRNA targets. *Human mutation* **31,** 1223–32 (2010) (cf. p. 109).
8. Bartel, D. P. MicroRNAs : target recognition and regulatory functions. *Cell* **136,** 215–33 (2009) (cf. p. 7).

9. Bateson, W & Punnett, R. On the inter-relations of genetic factors. *Proceedings of the Royal Society of London. Series B* **84,** 3–8 (1911) (cf. p. 26).

10. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 289–300 (1995) (cf. p. 67).

11. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J. *et al.* GenBank. *Nucleic acids research* **39,** D32–7 (2011) (cf. p. 32).

12. Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242 (2000) (cf. p. 32).

13. Bernstein, B. E., Birney, E., Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012) (cf. p. 10).

14. Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8,** 3 –62 (1936) (cf. p. 67).

15. Bostjancic, E., Zidar, N., Stajer, D. *et al.* MicroRNAs miR-1, miR-133a, miR-133b and miR-208 are dysregulated in human myocardial infarction. *Cardiology* **115,** 163–9 (2010) (cf. p. 100).

16. Bousquet, M., Harris, M. H., Zhou, B. *et al.* MicroRNA miR-125b causes leukemia. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 21558–63 (2010) (cf. p. 119).

17. Brouwer, J.-L. P., Veeger, N. J. G. M., Kluin-Nelemans, H. C. *et al.* The pathogenesis of venous thromboembolism : evidence for multiple interrelated causes. *Annals of internal medicine* **145,** 807–15 (2006) (cf. p. 87, 116).

18. Calin, G. A., Ferracin, M., Cimmino, A. *et al.* A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *The New England journal of medicine* **353,** 1793–801 (2005) (cf. p. 100).

19. Carrier, M., Le Gal, G., Wells, P. S. *et al.* Systematic review : the Trousseau syndrome revisited : should we screen extensively for cancer in patients with venous thromboembolism ? *Annals of internal medicine* **149,** 323–33 (2008) (cf. p. 117).

20. Chen, K. & Rajewsky, N. Natural selection on human microRNA binding sites inferred from SNP data. *Nature genetics* **38,** 1452–6 (2006) (cf. p. 120).

21. Cheng, Y. & Zhang, C. MicroRNA-21 in cardiovascular disease. *Journal of cardiovascular translational research* **3,** 251–5 (2010) (cf. p. 100).

22. Chico, T. J. A., Milo, M. & Crossman, D. C. The genetics of cardiovascular disease : new insights from emerging approaches. *The Journal of pathology* **220,** 186–97 (2010) (cf. p. 100).

23. Cohen, S. N. & Chang, A. C. Y. Recircularization and Autonomous Replication of a Sheared R-Factor DNA Segment in Escherichia coli Transformants. *Proceedings of the National Academy of Sciences* **70,** 1293–1297 (1973) (cf. p. 27).

24. Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics* **10,** 392–404 (2009) (cf. p. 53).

25. Cox, D. G., Dostal, L., Hunter, D. J. *et al.* N-acetyltransferase 2 polymorphisms, tobacco smoking, and breast cancer risk in the breast and prostate cancer cohort consortium. *American journal of epidemiology* **174,** 1316–22 (2011) (cf. p. 33).

26. Crick, F & Watson, J. Molecular structure of nucleic acids. *Nature.* (1953) (cf. p. 26).

27. Darwin, C. *The Origin of Species* **2,** 22–79. (John Murray, 1859) (cf. p. 16).

28. De Stefano, V, Martinelli, I, Mannucci, P. M. *et al.* The risk of recurrent deep venous thrombosis among heterozygous carriers of both factor V Leiden and the G20210A prothrombin mutation. *The New England journal of medicine* **341,** 801–6 (1999) (cf. p. 87).

29. Devlin, B & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics.* (1995) (cf. p. 21).

30. Djebali, S., Davis, C. A., Merkel, A. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8 (2012) (cf. p. 10).

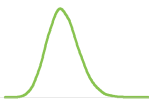
31. Doms, A. & Schroeder, M. GoPubMed : exploring PubMed with the Gene Ontology. *Nucleic acids research* **33,** W783–6 (2005) (cf. p. 100).

32. Dorn, G. W. Decoding the cardiac message : the 2011 Thomas W. Smith Memorial Lecture. *Circulation research* **110,** 755–63 (2012) (cf. p. 100, 121).

33. Dorn, G. W., Matkovich, S. J., Eschenbacher, W. H. *et al.* A human 3' miR-499 mutation alters cardiac mRNA targeting and function. *Circulation research* **110,** 958–67 (2012) (cf. p. 100).

34. Ehret, G. B., Munroe, P. B., Rice, K. M. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478,** 103–9 (2011) (cf. p. 38).

35. Evans, D. M., Marchini, J., Morris, A. P. *et al.* Two-stage two-locus models in genome-wide association. *PLoS genetics* **2,** e157 (2006) (cf. p. 52).



36. Fisher, R. The maximum likelihood method. *Messenger in Mathematics* (1912) (cf. p. 55).

37. Fisher, R. *Statistical Methods for Research Workers* en. **4.** (1925) (cf. p. 73).

38. Friedman, R. C., Farh, K. K.-H., Burge, C. B. *et al.* Most mammalian mRNAs are conserved targets of microRNAs. *Genome research* **19,** 92–105 (2009) (cf. p. 7).

39. Germain, M., Saut, N., Greliche, N. *et al.* Genetics of venous thrombosis : insights from a new genome wide association study. *PloS one* **6,** e25581 (2011) (cf. p. 78, 79, 86, 94).

40. Greliche, N. *Stratégies de Recherches de Phénomènes d'Interactions dans les maladies multifactorielles* thèse de doct. (2013) (cf. p. 14).

41. Greliche, N., Germain, M., Lambert, J.-C. *et al.* A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis (soumis). *BMC medical genetics* (cf. p. 85).

42. Greliche, N., Zeller, T., Wild, P. S. *et al.* Comprehensive Exploration of the Effects of miRNA SNPs on Monocyte Gene Expression. *PloS one* **7,** e45863 (2012) (cf. p. 99).

43. Griffiths-Jones, S., Saini, H. K., van Dongen, S. *et al.* miRBase : tools for microRNA genomics. *Nucleic acids research* **36,** D154–8 (2008) (cf. p. 7, 32, 101).

44. Guo, J. U., Ma, D. K., Mo, H. *et al.* Neuronal activity modifies the DNA methylation landscape in the adult brain. *Nature neuroscience* **14,** 1345–51 (2011) (cf. p. 14).

45. Handford, M. Où est Charlie ? : le voyage fantastique. (1989) (cf. p. 36).

46. Hardy, G. H. Mendelian proportions in a mixed population. *Science* **28,** 49–50 (1908) (cf. p. 45, 61).

47. Heinig, M., Petretto, E., Wallace, C. *et al.* A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* **467,** 460–4 (2010) (cf. p. 82).

48. Hercberg, S., Galan, P., Preziosi, P. *et al.* The SU.VI.MAX Study : a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Archives of internal medicine* **164,** 2335–42 (2004) (cf. p. 78).

49. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics.* (1979) (cf. p. 67).

50. Hua, Y., Zhang, Y. & Ren, J. IGF-1 Deficiency Resists Cardiac Hypertrophy and Myocardial Contractile Dysfunction : Role of microRNA-1 and microRNA-133a. *Journal of cellular and molecular medicine.* (2011) (cf. p. 100).

51. Huntzinger, E. & Izaurralde, E. Gene silencing by microRNAs : contributions of translational repression and mRNA decay. *Nature reviews. Genetics* **12,** 99–110 (2011) (cf. p. 119).

52. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–96 (2003) (cf. p. 75).

53. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–45 (2004) (cf. p. 9, 31, 101).

54. Jacquard, A. Structures génétiques des populations. *Population* **24,** 1155–1160 (1969) (cf. p. 18).

55. Janssens, F. La Théorie de la Chiasmatypie. *La Cellule.* (1909) (cf. p. 26).

56. John, B., Enright, A. J., Aravin, A. *et al.* Human MicroRNA targets. *PLoS biology* **2,** e363 (2004) (cf. p. 7).

57. Johnson, A. D., Handsaker, R. E., Pulit, S. L. *et al.* SNAP : a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics (Oxford, England)* **24,** 2938–9 (2008) (cf. p. 103).

58. Kaprio, J. Twins and the mystery of missing heritability : the contribution of gene-environment interactions. *Journal of internal medicine.* (2012) (cf. p. 39).

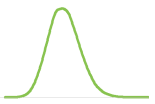
59. Kelly, M., Bagnall, R. D., Peverill, R. E. *et al.* A polymorphic miR-155 binding site in AGTR1 is associated with cardiac hypertrophy in Friedreich ataxia. *Journal of Molecular and Cellular Cardiology* **51,** 848–54 (2011) (cf. p. 101).

60. Kerem, B, Rommens, J. M., Buchanan, J. A. *et al.* Identification of the cystic fibrosis gene : genetic analysis. *Science (New York, N.Y.)* **245,** 1073–80 (1989) (cf. p. 34).

61. Klein, R. J., Zeiss, C., Chew, E. Y. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)* **308,** 385–9 (2005) (cf. p. 34).

62. Koeleman, B., Reitsma, P., Allaart, C. *et al.* Activated protein C resistance as an additional risk factor for thrombosis in protein C-deficient families. *Blood* **84,** 1031–1035 (1994) (cf. p. 87).

63. Krek, A., Grün, D., Poy, M. N. *et al.* Combinatorial microRNA target predictions. *Nature genetics* **37,** 495–500 (2005) (cf. p. 7).



64. Krol, J., Loedige, I. & Filipowicz, W. The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics* **11,** 597–610 (2010) (cf. p. 7).

65. Lango Allen, H., Estrada, K., Lettre, G. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–8 (2010) (cf. p. 38).

66. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75,** 843–54 (1993) (cf. p. 7).

67. Lehman, I. R., Bessman, M. J., Simms, E. S. *et al.* Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli. *The Journal of biological chemistry* **233,** 163–70 (1958) (cf. p. 27).

68. Lelandais, G., Vincens, P., Badel-Chagnon, A. *et al.* Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms. *Bioinformatics (Oxford, England)* **22,** 1359–66 (2006) (cf. p. 16).

69. Levene, H. *Contributions to Probability and Statistics : Essays in Honor of Harold Hotelling : Robust tests for equality of variances* 278–292. (Stanford Univ. Press, Palo Alto, CA, 1960) (cf. p. 45, 61).

70. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005) (cf. p. 7).

71. Lewontin, R. C. The Interaction of Selection and Linkage. I. General Considerations ; Heterotic Models. *Genetics* **49,** 49–67 (1964) (cf. p. 23).

72. Li, J., Harris, R. A., Cheung, S. W. *et al.* Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. *PLoS genetics* **8,** e1002692 (2012) (cf. p. 14).

73. Lu, M., Zhang, Q., Deng, M. *et al.* An analysis of human microRNA and disease associations. *PloS one* **3,** e3420 (2008) (cf. p. 99).

74. Lunetta, K. L., Hayward, L. B., Segal, J. *et al.* Screening large-scale association study data : exploiting interactions using random forests. *BMC genetics* **5,** 32 (2004) (cf. p. 54).

75. Manolio, T. A., Collins, F. S., Cox, N. J. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–53 (2009) (cf. p. 36, 39).

76. Maragkakis, M, Reczko, M, Simossis, V. A. *et al.* DIANA-microT web server : elucidating microRNA functions through target prediction. *Nucleic acids research* **37,** W273–6 (2009) (cf. p. 7).

77. Martinelli, I, Taioli, E, Bucciarelli, P *et al.* Interaction between the G20210A mutation of the prothrombin gene and oral contraceptive use in deep vein thrombosis. *Arteriosclerosis, thrombosis, and vascular biology* **19,** 700–3 (1999) (cf. p. 87).

78. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 560–4 (1977) (cf. p. 30).

79. McVean, G. Evolutionary genetics : what is driving male mutation ? *Current biology : CB* **10,** R834–5 (2000) (cf. p. 14).

80. Mendel, G. *Experiments in Plant Hybridization* (1865) (cf. p. 26, 49).

81. Morange, P. E. & Tregouet, D. A. Lessons from genome-wide association studies in venous thrombosis. *Journal of thrombosis and haemostasis : JTH* **9 Suppl 1,** 258–64 (2011) (cf. p. 86).

82. Morgan, T. The theory of the gene. *American Naturalist.* (1917) (cf. p. 26).

83. Moskvina, V. & Schmidt, K. M. On multiple-testing correction in genome-wide association studies. *Genetic epidemiology* **32,** 567–73 (2008) (cf. p. 76).

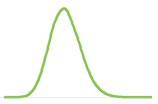
84. Mullis, K, Faloona, F, Scharf, S *et al.* Specific enzymatic amplification of DNA in vitro : the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology* **51 Pt 1,** 263–73 (1986) (cf. p. 27).

85. Nachman, M. W. & Crowell, S. L. Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics* **156,** 297–304 (2000) (cf. p. 14).

86. Nossent, A. Y., Hansen, J. L., Doggen, C. *et al.* SNPs in MicroRNA Binding Sites in 3'-UTRs of RAAS Genes Influence Arterial Blood Pressure and Risk of Myocardial Infarction. *American journal of hypertension.* (2011) (cf. p. 101).

87. Oudot-Mellakh, T., Cohen, W., Germain, M. *et al.* Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway : the MARTHA project. *British journal of haematology* **157,** 230–9 (2012) (cf. p. 79).

88. Paré, G., Cook, N. R., Ridker, P. M. *et al.* On the use of variance per genotype as a tool to identify quantitative trait interaction effects : a report from the Women's Genome Health Study. *PLoS genetics* **6,** e1000981 (2010) (cf. p. 60, 108).



89. Pearson, K. On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that can be reasonably supposed to have arisen from Random Sampling. *Philosophical Magazine* **50,** 157–175 (1900) (cf. p. 45, 62).

90. Pruitt, K. D., Tatusova, T., Brown, G. R. *et al.* NCBI Reference Sequences (RefSeq) : current status, new features and genome annotation policy. *Nucleic acids research* **40,** D130–5 (2012) (cf. p. 101).

91. Purcell, S., Neale, B., Todd-Brown, K. *et al.* PLINK : a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81,** 559–75 (2007) (cf. p. 63).

92. R Development Core Team. R : A language and environment for statistical computing. *R Foundation Statistical Computing.* (2008) (cf. p. 63).

93. Rane, S., He, M., Sayed, D. *et al.* Downregulation of miR-199a derepresses hypoxia-inducible factor-1alpha and Sirtuin 1 and recapitulates hypoxia preconditioning in cardiac myocytes. *Circulation research* **104,** 879–86 (2009) (cf. p. 100).

94. Rao, C. R. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society* **37,** 81–91 (1945) (cf. p. 59).

95. Ren, X.-P., Wu, J., Wang, X. *et al.* MicroRNA-320 is involved in the regulation of cardiac ischemia/reperfusion injury by targeting heat-shock protein 20. *Circulation* **119,** 2357–66 (2009) (cf. p. 100).

96. Ro, S., Park, C., Young, D. *et al.* Tissue-dependent paired expression of miRNAs. *Nucleic acids research* **35,** 5944–5953 (2007) (cf. p. 9).

97. Rosendaal, F. R. Venous thrombosis : a multicausal disease. *Lancet* **353,** 1167–73 (1999) (cf. p. 87).

98. Ryan, B. M., Robles, A. I. & Harris, C. C. Genetic variation in microRNA networks : the implications for cancer research. *Nature reviews. Cancer* **10,** 389–402 (2010) (cf. p. 99).

99. Sanger, F, Nicklen, S & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 5463–7 (1977) (cf. p. 30).

100. Saunders, M. A., Liang, H. & Li, W.-H. Human polymorphism at microRNAs and microRNA target sites. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 3300–5 (2007) (cf. p. 120).

101. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Human molecular genetics* **19,** R227–R240 (2010) (cf. p. 32).

102. Schena, M, Shalon, D, Davis, R. W. *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)* **270,** 467–70 (1995) (cf. p. 28).

103. Schunkert, H., König, I. R., Kathiresan, S. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics* **43,** 333–8 (2011) (cf. p. 82).

104. Seitz, H. Redefining microRNA targets. *Current biology : CB* **19,** 870–3 (2009) (cf. p. 121).

105. Sherry, S. T., Ward, M. H., Kholodov, M *et al.* dbSNP : the NCBI database of genetic variation. *Nucleic acids research* **29,** 308–11 (2001) (cf. p. 32, 102).

106. Shi, D., Li, P., Ma, L. *et al.* A Genetic Variant in pre-miR-27a Is Associated with a Reduced Renal Cell Cancer Risk in a Chinese Population. *PloS one* **7,** e46566 (2012) (cf. p. 100).

107. Sidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association.* (1967) (cf. p. 67).

108. Slaby, O., Bienertova-Vasku, J., Svoboda, M. *et al.* Genetic polymorphisms and MicroRNAs : new direction in molecular epidemiology of solid cancer. *Journal of cellular and molecular medicine.* (2011) (cf. p. 99, 100).

109. Small, E. M. & Olson, E. N. Pervasive roles of microRNAs in cardiovascular biology. *Nature* **469,** 336–42 (2011) (cf. p. 100).

110. Smith, H. & Wilcox, K. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *Journal of molecular biology.* (1970) (cf. p. 27).

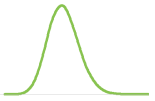
111. Speliotes, E. K., Willer, C. J., Berndt, S. I. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics* **42,** 937–48 (2010) (cf. p. 38).

112. Steen, K. V. Travelling the world of gene-gene interactions. *Briefings in bioinformatics* **13,** 1–19 (2012) (cf. p. 53).

113. Storey, T. A. *Principles of hygiene* (Stanford University Press, 1935) (cf. p. 1).

114. Stouffer, S., Suchman, E. & DeVinney, L. The American soldier : adjustment during army life. (1949) (cf. p. 73).

115. Student. The probable error of a mean. *Biometrika.* (1908) (cf. p. 45).



116. Sturtevant, A. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14,** 43–59 (1913) (cf. p. 26).

117. Sutton, W. The chromosomes in heredity. *The Biological Bulletin.* (1903) (cf. p. 26).

118. The ENCODE Project Consortium. Identification and analysis of functional elements in 1the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007) (cf. p. 9).

119. Tregouet, D. A. & Garelle, V. A new JAVA interface implementation of THESIAS : testing haplotype effects in association studies. *Bioinformatics (Oxford, England)* **23,** 1038–9 (2007) (cf. p. 63).

120. Trégouët, D.-A., Heath, S., Saut, N. *et al.* Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk : results from a GWAS approach. *Blood* **113,** 5298–303 (2009) (cf. p. 78, 86).

121. Van Boven, H., Vandenbroucke, J., Briet, E. *et al.* Gene-Gene and Gene-Environment Interactions Determine Risk of Thrombosis in Families With Inherited Antithrombin Deficiency. *Blood* **94,** 2590–2594 (1999) (cf. p. 87).

122. Van Rooij, E., Sutherland, L. B., Thatcher, J. E. *et al.* Dysregulation of microRNAs after myocardial infarction reveals a role of miR-29 in cardiac fibrosis. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 13027–32 (2008) (cf. p. 100).

123. Vandenbroucke, J. P., Koster, T, Briët, E *et al.* Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet* **344,** 1453–7 (1994) (cf. p. 87).

124. Venter, J. C., Adams, M. D., Myers, E. W. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291,** 1304–51 (2001) (cf. p. 15).

125. Via, M., Gignoux, C. & Burchard, E. G. The 1000 Genomes Project : new opportunities for research and social challenges. *Genome medicine* **2,** 3 (2010) (cf. p. 75).

126. Wald, A. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics.* (1939) (cf. p. 46, 59).

127. Walker, F. O. Huntington's disease. *Lancet* **369,** 218–28 (2007) (cf. p. 28).

128. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–78 (2007) (cf. p. 34).

129. White, R. H. The epidemiology of venous thromboembolism. *Circulation* **107,** I4–8 (2003) (cf. p. 86).

130. Wild, P. S., Zeller, T, Beutel, M *et al.* [The gutenberg health study]. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* **55,** 824–30 (2012) (cf. p. 81).

131. Wilks, S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics.* (1938) (cf. p. 58).

132. Willer, C. J., Li, Y. & Abecasis, G. R. METAL : fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* **26,** 2190–1 (2010) (cf. p. 73).

133. Wojcik, S. E., Rossi, S., Shimizu, M. *et al.* Non-codingRNA sequence variations in human chronic lymphocytic leukemia and colorectal cancer. *Carcinogenesis* **31,** 208–15 (2010) (cf. p. 100).

134. Wu, C, Gong, Y, Sun, A *et al.* The human MTHFR rs4846049 polymorphism increases coronary heart disease risk through modifying miRNA binding. *Nutrition, metabolism, and cardiovascular diseases : NMCD.* (2012) (cf. p. 101).

135. Xu, J., Hu, Z., Xu, Z. *et al.* Functional variant in microRNA-196a2 contributes to the susceptibility of congenital heart disease in a Chinese population. *Human mutation* **30,** 1231–6 (2009) (cf. p. 101).

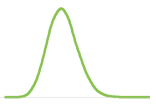
136. Yang, B., Lin, H., Xiao, J. *et al.* The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2. *Nature medicine* **13,** 486–91 (2007) (cf. p. 100).

137. Zee, R. Y. L., Bubes, V., Shrivastava, S. *et al.* Genetic risk factors in recurrent venous thromboembolism : A multilocus, population-based, prospective approach. *Clinica chimica acta ; international journal of clinical chemistry* **402,** 189–92 (2009) (cf. p. 87).

138. Zeller, T., Wild, P., Szymczak, S. *et al.* Genetics and Beyond - The Transcriptome of Human Monocytes and Disease Susceptibility. *PLoS ONE* **5,** e10693 (2010) (cf. p. 103, 112).

139. Zhi, H., Wang, L., Ma, G. *et al.* Polymorphisms of miRNAs genes are associated with the risk and prognosis of coronary artery disease. *Clinical research in cardiology : official journal of the German Cardiac Society* **101,** 289–96 (2012) (cf. p. 101).

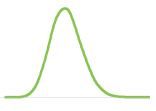
140. Zoller, B, Berntsdotter, A, Garcia de Frutos, P *et al.* Resistance to activated protein C as an additional genetic risk factor in hereditary deficiency of protein S. *Blood* **85,** 3518–3523 (1995) (cf. p. 87).



Sites web, autres

141. *23andMe* : https://www.23andme.com/ (cf. p. 162).
142. *Cardiogenics* : http://www.cardiogenics.org (cf. p. 82).
143. *Consent to Research* : http://weconsent.us/about-us/ (cf. p. 163).
144. *DNA 11* : http://www.dna11.com/ (cf. p. 162).
145. *Dosage des facteurs de la coagulation - Portail santé du ministère de la santé du Luxembourg* : http://www.sante.public.lu/fr/maladies-traitements/020-examens/analyses-biologiques/ (cf. p. 79).
146. *GeneGroove* : http://www.genegroove.com/ (cf. p. 162).
147. *GenePartner* : http://www.genepartner.com/ (cf. p. 162).
148. *Genomes Unzipped* : http://www.genomesunzipped.org/ (cf. p. 163).
149. *GoPubMed* : http://www.gopubmed.org/ (cf. p. 100).
150. *Guinness World Records* : http://www.guinnessworldrecords.com/ (cf. p. 29).
151. *Gutenberg Health Study* : http://www.gutenberghealthstudy.org/ (cf. p. 81).
152. Illumina. *World Personal Genome Registry* : http://www.worldpgr.com/ (cf. p. 161).
153. *Online Mendelian Inheritance in Man, OMIM* : http://omim.org/ (cf. p. 35).
154. *openSNP* : http://opensnp.org (cf. p. 163).
155. *Où est charlie, imitation (Charlie se trouve dans un casier en haut à gauche)* : http://www.nioutaik.fr/images/charlie-ecole.JPG (cf. p. 36).
156. *Personal Genome Project* : http://www.personalgenomes.org/ (cf. p. 163).
157. *Personal Genome Project : Participant profiles* : https://my.personalgenomes.org/users/ (cf. p. 163).
158. *Personal Genome Project Study Guide* : http://www.pgpstudy.org/ (cf. p. 163).
159. *Promethease - SNPedia* : http://snpedia.com/index.php/Promethease (cf. p. 162).
160. *PubMed* : http://www.ncbi.nlm.nih.gov/pubmed (cf. p. 100).
161. *ReMOAT* : http://remoat.sysbiol.cam.ac.uk/ (cf. p. 82).
162. *Warrior Roots* : http://www.warriorroots.com/ (cf. p. 162).

163. *Your DNA Song* : http://www.yourdnasong.com/ (cf. p. 162).



Article 1

A genome-wide search for common SNP x SNP interaction on the risk of venous thrombosis.

En cours de révision dans le journal BMC Medical Genetics

RESEARCH ARTICLE

A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis

Nicolas Greliche[1], Marine Germain[1], Jean-Charles Lambert[2], William Cohen[3], Marion Bertrand[4], Anne-Marie Dupuis[5], Luc Letenneur[6], Mark Lathrop[7], Philippe Amouyel[2,8], Pierre-Emmanuel Morange[3], David-Alexandre Trégouët[1]

Abstract

Background: Venous Thrombosis (VT) is a common multifactorial disease with an estimated heritability between 35% and 60%. Known genetic polymorphisms identified so far only explain ~5% of the genetic variance of the disease. This study was aimed to investigate whether pair-wise interactions between common single nucleotide polymorphisms (SNPs) could exist and modulate the risk of VT.

Methods: A genome-wide SNP x SNP interaction analysis on VT risk was conducted in a French case-control study and the most significant findings were tested for replication in a second independent French case-control sample. The results obtained in the two studies totaling 1,961 cases and 2,338 healthy subjects were combined into a meta-analysis.

Results: The smallest observed p-value for interaction was p = 6.00 10-11 but it did not pass the Bonferroni significance threshold of 1.69 10-12 correcting for the number of investigated interactions that was 2.96 1010. Among the 41 suggestive pair-wise interactions with p-value less than 10-8, one was further shown to involve two SNPs, rs9804128 (IGFS21 locus) and rs4784379 (IRX3 locus) that further demonstrated significant interactive effects (p = 4.83 10-5) on the variability of plasma Factor VIII levels, a quantitative biomarker of VT risk, in a sample of 1,091 VT patients.

Conclusion: This study, the first genome-wide SNP interaction analysis conducted so far on VT risk, suggests that common SNPs are unlikely exerting strong interactive effects on the risk of disease.

Background

Venous Thrombosis (VT) is a common complex disease affecting ~0.2% of individuals a year. VT includes deep vein thrombosis and pulmonary embolism, the latter being characterized by a one year mortality rate of ~10% excluding patients with malignancies [1]. As a complex trait, VT is considered as resulting from the interplay between environmental and genetic factors, that could interact with each other, to modulate VT risk [2, 3]. The recent Genome Wide Association Studies (GWAS) strategy brought great hopes to identify novel susceptibility loci to human diseases and some true successes were obtained in the field of VT genetics. Novel genes recently identified to harbor common susceptibility alleles (i.e with allele frequency > 0.05) for VT include GP6, HIVEP1, KNG1, STAB2, STXBP5 and VWF (reviewed in [4]). However, none of the identified risk alleles demonstrated genetic effects stronger than those of the established VT-associated genes known before the GWAS era, ABO, F2, F5 and FGG [5]. As for most multifactorial diseases, risk alleles for VT identified so far only explain a small proportion of the familial risk of disease [6]. Alternative strategies are needed to identify the army sources that could contribute to the unexplained heritability and these include gene-gene and gene-environment interactions, deep sequencing, transcriptomic analyses and epigenomics [7-10].

In this work, we were interested in assessing whether interaction between common polymorphisms could contribute to VT risk. To our knowledge, studies that have investigated this hypothesis were mainly dedicated to known candidate genes [11, 12] and no attempt has been made to address it without any a priori hypothesis. This is why, we here take advantage

of the large amount of genetic information we have collected through two French GWAS on VT [6, 13] to conduct the first genome-wide search for SNP x SNP interaction with respect to VT risk.

Methods

This work was based on two French GWAS on VT, the Early-Onset Venous Thrombosis (EOVT) and the Marseille Thrombosis Association (MARTHA) studies. These two studies have already been extensively described in [5, 6, 14] for EOVT and in [6, 15-17] for MARTHA.

Studied populations and phenotype measurements

Briefly, in both studies, VT patients were cases, with a documented history of VT and free of well known strong genetic risk factors including antithrombin (AT), protein C (PC) or protein S (PS) deficiency, homozygosity for FV Leiden or F2 20210A mutations and lupus anticoagulant. In EOVT, patients were selected to experience idiopathic VT before the age of 50. Controls were French individuals selected from two healthy populations, SUVIMAX [18] and the Three City Study [19], for EOVT and MARTHA, respectively. The EOVT case-control study included 419 patients and 1,228 healthy subjects, while MARTHA was composed of 1,542 patients and 1,110 healthy subjects, all the individuals being of European origin, with the majority being of French descent.

Several key quantitative biomarkers of VT risk have been measured in MARTHA patients. The detailed description of the corresponding measurements has been previously described in [15] for AT, PC, PS and the agkistrodon contortrix venom (ACV) test that explores the PC pathway, in [17] for Factor VIII (FVIII) and von Willebrand Factor (VWF), and in [16] for Activated Partial Thromboplastin Time (aPTT) and Prothrombin Time (PT).

Genotyping

Individuals participating in the EOVT study were genotyped for 317,139 SNPs using the Illumina Sentrix HumanHap300 Beadchip. The application of the quality control criteria described in [5] led the final selection of 291,872 autosomal SNPs for analysis. As detailed in [6], individuals participating to the MARTHA GWAS were typed with the Illumina Human 610-Quad and Human660W-Quad Beadchips. 481,002 autosomal SNPs remained for analysis after quality control.

Statistical analysis

A two-stage genome wide interaction analysis was carried out. The initial screening for pairwise SNPs interactions was carried out in the EOVT study. The first step of the analysis consisted in reducing redundancy between SNPs by keeping only one SNP out of all SNPs in strong pairwise linkage disequilibrium ($r2>0.90$) within a window of 50kb.

Pairwise SNPs interactions were tested by a logistic regression analysis where both SNPs were coded under an additive model (0,1 and 2 according to the number of rare alleles) and an interaction term was added in the model. All interactions significant at $p < 10\text{-}4$ were further assessed in the larger MARTHA study. When SNPs were not available in the latter sample, the best available proxy in term of r2 , according to the SNAP database [20], was used. The same logistic regression model was applied in the MARTHA study. Results obtained in the two GWAS were then meta-analyzed through a fixed-effect model relying on the inverse-variance weighting as implemented in the METAL software (http://www.sph.umich.edu/csg/abecasis/metal).

Homogeneity of associations across the two GWAS studies was tested using the Mantel-Haenszel method [21].

The most significant interactions were then further assessed in relation to quantitative biomarkers of VT risk in MARTHA patients. For this, standard linear regression analyses were conducted with the same additive allele coding as for the binary trait analysis. Analyses were adjusted for age, sex and ABO blood group. For AT, PC, PS and ACV, individuals under anticoagulant were excluded. The THESIAS software [22] was used to illustrate the detected pairwise SNP interactions.

Results and discussion

We first applied a pairwise tagging approach to discard redundant SNPs using a r2 threshold of 0.90, that led to the final selection of 243,189 SNPs from the EOVT study.

2.96 1010 pairwise SNPs interactions were then tested in EOVT, but none of them reached the Bonferroni corrected p-value of 1.69 10-12. Nevertheless, all interactions with p-value less than 10-4 ($n = 2,126,084$) were further assessed in MARTHA. The smallest observed p-value was 6.73 10-7, but it did not pass the Bonferroni correction ($p < 2.35$ 10-8) for the number of interactions tested at this second step.

The meta-analysis of the results obtained in EOVT and MARTHA led to 41 suggestive interactions with p-values lower than 10-8 and with consistent effects in both studies (Table 1). The smallest one, $p = 6.00$ 10-11, was observed for two SNPs in the vicinity of SURF6 gene that is ~40kb from the ABO locus. After adjusting for the ABO blood group, this interaction vanished ($p = 0.37$) suggesting that this interaction had captured the ABO effect through the linkage disequilibrium extending at this locus.

Despite the lack of study-wise statistical interactions, we could not exclude that some genuine interaction phenomena hide in the list of suggestive interactions (Table 1). We hypothesized that the use of additional biological information on quantitative biomarkers of VT risk could help in digging into this list. We therefore investigated whether the identified interactive SNPs could exert their effect on VT biomarkers available in MARTHA: ACV, aPTT, AT, Fibrinogen, FVIII, PC, PS, PT and VWF. At the Bonferroni threshold of 1.35 10-4 for the number of performed tests (i.e 369 = 41 SNPs x 9 phenotypes), one interaction was statistically significant ($p = 4.82$ 10-5). It involved rs9804128 lying in the promoter region of the IGSF21 gene and the rs4784379

Table 1. Pairwise SNP interactions with suggestive statistical evidence (p < 10) in the meta analysis of two GWAS datasets gathering 1,961 cases and 2,338 controls

						EOVT				MARTHA					
						MAF[2]		Interaction		MAF		Interaction		Combined	
rsID	Allèles[1]	CHR	rsID	Allèles	CHR	SNP1	SNP2	OR[3]	P	SNP1	SNP2	OR	P	OR[4]	P
rs493014	T/G	9	rs886090	G/A	9	0,31	0,33	1,72	1.85×10^{-5}	0,30	0,31	1,60	6.73×10^{-7}	1,64	6.00×10^{-11}
rs1336472	G/A	1	rs4715555	A/G	6	0,41	0,39	1,64	4.10×10^{-5}	0,40	0,38	1,49	2.00×10^{-6}	1,54	4.24×10^{-10}
rs380904	G/A	8	rs8086028	G/A	18	0,30	0,27	1,96	3.76×10^{-6}	0,29	0,31	1,55	1.12×10^{-5}	1,67	4.51×10^{-10}
rs6815916	A/G	4	rs6092326	C/T	20	0,09	0,48	2,37	4.32×10^{-5}	0,09	0,47	1,98	2.95×10^{-6}	2,10	6.84×10^{-10}
rs2282015	T/G	10	rs13050454	G/A	21	0,41	0,43	1,81	3.52×10^{-7}	0,41	0,42	1,37	7.68×10^{-5}	1,50	8.36×10^{-10}
rs7648704	T/G	3	rs4868644	C/T	5	0,33	0,49	1,64	7.36×10^{-5}	0,33	0,49	1,52	2.88×10^{-6}	1,56	9.89×10^{-10}
rs1985317	T/C	9	rs827637	G/A	10	0,39	0,46	0,55	7.13×10^{-7}	0,41	0,46	0,72	7.73×10^{-5}	0,66	1.32×10^{-9}
rs2321744	A/G	13	rs6497540	T/G	16	0,09	0,41	0,43	8.61×10^{-5}	0,10	0,42	0,52	2.98×10^{-6}	0,49	1.38×10^{-9}
rs315122	T/G	12	rs884483	T/C	15	0,29	0,11	2,61	1.92×10^{-5}	0,31	0,12	1,87	7.90×10^{-6}	2,05	1.42×10^{-9}
rs1423386	A/G	5	rs6491679	T/G	13	0,20	0,29	1,92	7.24×10^{-5}	0,20	0,29	1,66	4.17×10^{-6}	1,73	1.63×10^{-9}
rs6491679	T/G	13	rs1423386	A/G	5	0,29	0,20	1,92	7.24×10^{-5}	0,29	0,20	1,66	4.17×10^{-6}	1,73	1.63×10^{-9}
rs7714670	T/C	5	rs12880735	G/A	14	0,44	0,34	1,75	4.59×10^{-6}	0,44	0,36	1,42	3.32×10^{-5}	1,52	1.75×10^{-9}
rs12880735	G/A	14	rs7714670	T/C	5	0,34	0,44	1,75	4.59×10^{-6}	0,36	0,44	1,42	3.32×10^{-5}	1,52	1.75×10^{-9}
rs9392653	C/T	6	rs7780976	A/C	7	0,27	0,18	2,14	2.28×10^{-6}	0,29	0,19	1,57	5.49×10^{-5}	1,74	1.83×10^{-9}
rs9804128	A/G	1	rs4784379	G/A	16	0,27	0,25	1,97	2.73×10^{-5}	0,26	0,24	1,60	9.45×10^{-6}	1,71	1.90×10^{-9}
rs1364505	G/A	7	rs1204660	G/A	20	0,30	0,16	2,14	2.32×10^{-5}	0,33	0,16	1,67	1.11×10^{-5}	1,80	2.10×10^{-9}
rs2288073	A/G	2	rs10771022	G/T	12	0,30	0,34	1,71	7.94×10^{-5}	0,29	0,34	1,55	5.51×10^{-6}	1,60	2.11×10^{-9}
rs1367228	C/A	2	rs3905075	C/T	13	0,43	0,41	1,61	9.44×10^{-5}	0,45	0,40	1,44	4.22×10^{-6}	1,49	2.20×10^{-9}
rs536477	G/A	1	rs1937920	A/G	10	0,43	0,26	0,57	3.27×10^{-5}	0,43	0,27	0,67	1.40×10^{-5}	0,63	2.93×10^{-9}
rs2710201	A/G	7	rs3780293	G/A	9	0,06	0,34	0,35	6.84×10^{-5}	0,06	0,36	0,43	9.92×10^{-6}	0,40	3.30×10^{-9}
rs12541254	G/A	8	rs305009	G/A	15	0,35	0,23	1,99	3.15×10^{-6}	0,34	0,23	1,50	7.63×10^{-5}	1,65	3.33×10^{-9}
rs4507975	A/G	1	rs9914518	G/A	17	0,29	0,46	0,61	9.59×10^{-5}	0,29	0,47	0,67	7.95×10^{-6}	0,65	3.58×10^{-9}
rs2771051	T/G	9	rs827637	G/A	10	0,37	0,46	0,52	9.27×10^{-8}	0,37	0,46	0,75	4.59 x 10-4	0,67	3.82×10^{-9}
rs10516089	T/C	5	rs11072930	T/C	15	0,32	0,28	0,51	2.66×10^{-6}	0,31	0,30	0,69	7.19×10^{-5}	0,63	3.86×10^{-9}
rs10504130	G/A	8	rs2847351	A/G	18	0,15	0,30	2,46	1.04×10^{-5}	0,14	0,32	1,69	3.07×10^{-5}	1,88	4.46×10^{-9}
rs318497	G/A	6	rs7019259	A/G	9	0,49	0,07	0,29	2.56×10^{6}	0,49	0,07	0,51	8.40×10^{5}	0,43	4.54×10^{9}
rs6695223	T/C	1	rs1763510	C/T	6	0,12	0,39	2,49	6.00×10^{-6}	0,13	0,39	1,66	4.31×10^{-5}	1,86	4.70×10^{-9}
rs1336708	A/G	13	rs1423386	A/G	5	0,26	0,20	0,51	6.77×10^{5}	0,25	0,20	0,61	1.20×10^{5}	0,58	4.85×10^{9}
rs1423386	A/G	5	rs1336708	A/G	13	0,20	0,26	0,51	6.77×10^{5}	0,20	0,25	0,61	1.19×10^{5}	0,58	4.85×10^{9}
rs6771316	G/A	3	rs10986432	T/C	9	0,14	0,18	2,41	4.64×10^{5}	0,13	0,17	1,99	2.20×10^{5}	2,13	5.26×10^{9}
rs664910	A/G	3	rs877228	G/A	15	0,30	0,47	1,63	6.05×10^{5}	0,30	0,44	1,44	1.92×10^{5}	1,50	6.63×10^{9}
rs9945428	C/A	18	rs4823535	G/A	22	0,30	0,28	0,58	7.47×10^{5}	0,30	0,26	0,65	1.85×10^{5}	0,62	6.88×10^{9}
rs1910358	T/C	5	rs9981595[5]	T/G	21	0,23	0,12	2,21	9.60×10^{5}	0,23	0,11	1,93	1.63×10^{5}	2,03	7.14×10^{9}
rs6771725	G/T	3	rs10507246	G/T	12	0,26	0,08	2,60	4.02×10^{5}	0,28	0,09	2,04	3.77×10^{5}	2,22	8.60×10^{9}
rs16865717	C/T	2	rs2009579	C/T	20	0,27	0,36	1,90	5.22×10^{-6}	0,29	0,36	1,43	9.59×10^{-5}	1,56	8.82×10^{-9}
rs2028385	A/G	12	rs2038227	A/C	16	0,16	0,39	2,19	3.36×10^{-7}	0,16	0,37	1,47	7.11×10^{-4}	1,69	8.82×10^{-9}
rs10476160	A/G	5	rs1707420	C/T	8	0,21	0,48	0,56	6.35×10^{-5}	0,20	0,48	0,65	2.48×10^{-5}	0,62	9.09×10^{-9}
rs971572	C/A	1	rs10828151	A/C	10	0,32	0,07	0,35	3.43×10^{-5}	0,32	0,07	0,47	4.38×10^{-5}	0,42	9.30×10^{-9}
rs6858430	C/T	4	rs4800250	A/G	18	0,20	0,40	1,86	2.44×10^{-5}	0,21	0,40	1,52	5.16×10^{-5}	1,62	9.67×10^{-9}
rs467650	T/C	5	rs7153749	T/C	14	0,36	0,44	0,59	1.69×10^{-5}	0,37	0,44	0,71	6.00×10^{-5}	0,67	9.91×10^{-9}
rs7153749	T/C	14	rs467650	T/C	5	0,44	0,36	0,59	1.69×10^{-5}	0,44	0,37	0,71	6.00×10^{-5}	0,67	9.91×10^{-9}

[1] Common/minor alleles

[2] Minor Allele Frequency

[3] Odds ratiofor VT associated with the interaction of the two minor alleles under a logistic model assuming additive allelic effect.

[4] Pooled Odds ratio derived from a fixed effect model analysis using the inverse variance method as implemented in METAL. None of the reported interactions demonstrated evidence for heterogeneity across GWAS samples (p>0.05 for all homogeneity test pvalues).

[5] rs2836978 serves as a proxy (r^2=1) for rs9981595 in the discovery GWAS.

mapping 130kb downstream the IRX3 locus, the two SNPs interacting to modulate plasma FVIII levels. As shown in Table 2, carriers of the rs9804128-G and rs4784379-A alleles were associated with the highest plasma FVIII levels compared to the three other alleles combinations. At contrast, these individuals were associated with ~2 fold decreased in VT risk, the frequency of the GA combination being 8.3% in controls and 4.6% in patients (Table 2). Looking deeply to the diplotypes formed by these two SNPs revealed that patients carrying without any ambiguity the GA combination, ie those carrying either the rs9804128-GG genotype and the rs4784379-A allele or the rs9804128-GA genotype and the rs4784379-AA genotype, exhibited the highest plasma FVIII levels (Table 3). Individuals ambiguous for the GA combination, who are those heterozygotes at both rs9804128 and rs4784379, were at intermediate FVIII levels (Table 3).

To our knowledge, this work is the first attempt to investigate, at the genome-wide scale, the presence of interactive effects derived from common SNPs. Despite the use of two large GWAS datasets, this study

Table 2: Interactive effects of the rs9804128 and rs4784379 on the risk of VT and on plasma FVIII levels

		EOVT Fréquence		MARTHA Fréquence		Combinés Fréquence		Patients de MARTHA [1] Moyenne haplotypique attendue pour FVIII	
		Témoins	Cas	Témoins	Cas	Témoins	Cas		
rs9804128	rs4784379	N =1228	N =419	N =1110	N =1542	N =2338	N =1961	Fréquence	[95%CI]
A	G	0,56	0,53	0,58	0,55	0,57	0,55	0,55	68.77 [66.27 - 71.26]
A	A	0,17	0,2	0,17	0,19	0,17	0,19	0,18	62.34 [58.03 - 66.64]
G	G	0,19	0,24	0,17	0,21	0,18	0,22	0,22	62.09 [56.35 - 67.83]
G	A	0,08	0,04	0,09	0,05	0,08	0,05	0,05	91.95 [92.98 - 100.9]
		$p^{(2)}$ =2.73 10^{-5}		$p^{(2)}$=9.45 10^{-6}		$p^{(3)}$ =1.90 10^{-9}			$p^{(4)}$ =6.89 10^{-5}

[1] In MARTHA, 1091 patients were measured for FVIII levels

[2] p-value of the interaction term between the two SNPs in the logistic regression analysis under the assumption of additive allele effects

[3] p-value obtained from the meta-analysis of the EOVT and MARTHA samples using a fixed effect model

[4] p-value of the interaction term between the two SNPs in the linear regression analysis, adjusted for age, sex , ABO blood group and F5/F2 carriers mutations

Table3: Plasma FVIII levels according to the rs9804128 and rs4784379 polymorphisms in 1091 VT patients

	rs4784379		
rs9804128	AA	AG	GG
AA	115.91 (32.80) N =34	132.70 (49.75) N =231	136.16 (51.35) N =321
GA	155.93 (77.17) N =16	141.42 (56.03) N =144	131.76 (47.11) N =266
GG	156.00 (68.98) N =4	150.17 (42.90) N =23	122.90 (60.11) N =52

Mean (SE) are shown.

did not detect interactions that reached the Bonferroni correction for the number of investigated interactions. The absence of such interaction could of course be due to low power, as the power of our second stage interaction analysis was about 50% to detect the most significant observed interactions [23, 24]. There is still no consensus about the most efficiency way to perform a genome-wide search for SNP x SNP interaction. Some people advocate to restrict the search for interaction to the set of most "significant" SNPs observed in single locus analysis. However, in that case, which statistical threshold should be used for selecting SNPs with significant marginal associations ? Nevertheless, we further confined our search for interaction to SNPs with statistical evidence for association in univariate analysis as low as $p < 10^{-3}$ or $p < 0.05$. We did not identify pair-wise significant interaction that were homogeneous between EOVT and MARTHA, and that satisfied the relevant Bonferroni correction (data not shown). Others suggest to use external biological information to refine the research strategy. Pathway-based analysis focusing only on the pairwise interactions between candidate gene SNPs could be such a strategy. By focusing only on SNPs mapping the VT candidate genes listing in the Supplementary Table 1 in [6], we did not detect any Bonferroni-corrected significant interaction that replicate in the EOVT and MARTHA study (data not shown). Another possibly could consist in assessing whether the most promising interactive effects could also be observed on quantitative traits known to be associated with the disease. Doing so, we observed that the rs9804128 and rs4784379 could interact to modulate both the risk of VT and the variability of FVIII levels. The rs9804128 lies in the proximal promoter of the IGFS21 gene and, according to the SNAP database [20], it is not in strong LD ($r2>0.8$) with any other SNP. Conversely, the rs4784379 is in strong LD with several SNPs, all located at least 100kb away from the IRX3 locus. However, the observed interaction could be considered as counterintuitive since the allele combination associated with increased FVIII levels was found less frequent in cases than in controls. This phenomenon could nevertheless be observed in presence of a mortality bias when patients with high levels of FVIII levels are at a higher risk of VT-associated mortality (eg .pulmonary embolism) and then under-represented in the cases sample. Further investigations are needed to replicate this association that involved SNPs at genes on which very little is known with respect to VT .

Conclusion

In conclusion, our work suggests that interactive phenomena between common SNPs are unlikely to contribute much to the risk of the VT.

Competing interests

The authors declare they have no competing interests.

Authors' contribution

NG and DAT carried out statistical analyses.

MG, JCL and WC were responsible for data collection and database management.

AMD, DAT, MB, ML, PA and PEM contributed to the study design whose direct implementation was coordinated by DAT and PEM.

All authors read and approved the final manuscript.

References

1. White RH: The epidemiology of venous thromboembolism. Circulation 2003, 107:I4-8.
2. Rosendaal FR: Venous thrombosis: a multicausal disease. Lancet 1999, 353:1167-1173.
3. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM, Coll I, Felices R, Stone W, Fontcuberta J, Blangero J: Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. Am J Hum Genet 2000, 67:1452-1459.
4. Morange PE, Tregouet DA: Lessons from genome-wide association studies in venous thrombosis. J Thromb Haemost 2011, 9 Suppl 1:258-264.
5. Tregouet DA, Heath S, Saut N, Biron-Andreani C, Schved JF, Pernod G, Galan P, Drouet L, Zelenika D, Juhan-Vague I, et al: Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. Blood 2009, 113:5298-5303.
6. Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, Cohen W, Oudot-Mellakh T, Antoni G, Alessi MC, et al: Genetics of venous thrombosis: insights from a new genome wide association study. PLoS One 2011, 6:e25581.
7. Morange PE, Tregouet DA: Deciphering the molecular basis of venous thromboembolism: where are we and where should we go? Br J Haematol 2010, 148:495-506.
8. Cordell HJ: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 2009, 10:392-404.
9. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al: Finding the missing heritability of complex diseases. Nature 2009, 461:747-753.
10. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 2011, 11:446-450.
11. Auro K, Alanne M, Kristiansson K, Silander K, Kuulasmaa K, Salomaa V, Peltonen L, Perola M: Combined effects of thrombosis pathway gene variants predict cardiovascular events. PLoS Genet 2007, 3:e120.
12. Pomp ER, Doggen CJ, Vos HL, Reitsma PH, Rosendaal FR: Polymorphisms in the protein C gene as risk factor for venous thrombosis. Thromb Haemost 2009, 101:62-67.
13. Tregouet DA, Konig IR, Erdmann J, Munteanu A, Braund PS, Hall AS, Grosshennig A, Linsel-Nitschke P, Perret C, DeSuremain M, et al: Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat Genet 2009, 41:283-285.
14. Smith NL, Heit JA, Tang W, Teichert M, Chasman DI, Morange PE: Genetic variation in F3 (tissue factor) and the risk of incident venous thrombosis: meta-analysis of eight studies. J Thromb Haemost 2012, 10:719-722.
15. Oudot-Mellakh T, Cohen W, Germain M, Saut N, Kallel C, Zelenika D, Lathrop M, Tregouet DA, Morange PE: Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. Br J Haematol 2012, 157:230-239.
16. Tang W, Schwienbacher C, Lopez LM, Ben-Shlomo Y, Oudot-Mellakh T, Johnson AD, Samani NJ, Basu S, Gogele M, Davies G, et al: Genetic Associations for Activated Partial Thromboplastin Time and Prothrombin Time, their Gene Expression Profiles, and Risk of Coronary Artery Disease. Am J Hum Genet 2012, 91:152-162.
17. Antoni G, Oudot-Mellakh T, Dimitromanolakis A, Germain M, Cohen W, Wells P, Lathrop M, Gagnon F, Morange PE, Tregouet DA: Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. BMC Med Genet 2011, 12:102.
18. Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D, Roussel AM, Favier A, Briancon S: The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. Arch Intern Med 2004, 164:2335-2342.
19. 3C Study Group: Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. Neuroepidemiology 2003, 22:316-325.
20. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI: SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008, 24:2938-2939.
21. Mantel N, Haenszel W: Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959, 22:719-748.
22. Tregouet DA, Garelle V: A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies. Bioinformatics 2007, 23:1038-1039.
23. Gauderman WJ: Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 2002, 155:478-484.
24. Demidenko E: Sample size and optimal design for logistic regression with binary interaction. Stat Med 2008, 27:36-46.

Article 2

Comprehensive exploration of the effect of miRNA SNPs on monocyte gene expression.

PLoS One. 2012 ;7(9) :e45863

OPEN ACCESS Freely available online

PLOS ONE

Comprehensive Exploration of the Effects of miRNA SNPs on Monocyte Gene Expression

Nicolas Greliche[1,2], Tanja Zeller[3], Philipp S. Wild[4], Maxime Rotival[1¤], Arne Schillert[5], Andreas Ziegler[5], Panos Deloukas[6], Jeanette Erdmann[7], Christian Hengstenberg[8], Willem H. Ouwehand[6,9], Nilesh J. Samani[10,11], Heribert Schunkert[7], Thomas Munzel[4], Karl J. Lackner[12], François Cambien[1], Alison H. Goodall[10,11], Laurence Tiret[1], Stefan Blankenberg[3], David-Alexandre Trégouët[1,13]*, the Cardiogenics Consortium[¶]

1 INSERM UMR_S 937, Pierre and Marie Curie University (UPMC, Paris 6), Paris, France, **2** Université Paris-Sud, Paris, France, **3** Department of General and Interventional Cardiology, University Heart Center Hamburg, Hamburg, Germany, **4** Departments of Medicine II, University Medical Center, Johannes Gutenberg University Mainz, Mainz, Germany, **5** Institut für Medizinische Biometrie und Statistik, Universität Lübeck, Lübeck, Germany, **6** Human Genetics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom, **7** Universität zu Lübeck, Medizinische Klinik II, Lübeck, Germany, **8** Klinik und Poliklinik für Innere Medizin II, Universität Regensburg, Regensburg, Germany, **9** Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge, United Kingdom, **10** Department of Cardiovascular Sciences, University of Leicester, Leicester, United Kingdom, **11** National Institute for Health Research Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, United Kingdom, **12** Department of Clinical Chemistry, University Medical Center, Johannes Gutenberg University Mainz, Mainz, Germany, **13** ICAN Institute for Cardiometabolism And Nutrition, Pierre and Marie Curie University (UPMC, Paris 6), Paris, France

Abstract

We aimed to assess whether pri-miRNA SNPs (miSNPs) could influence monocyte gene expression, either through marginal association or by interacting with polymorphisms located in 3'UTR regions (3utrSNPs). We then conducted a genome-wide search for marginal miSNPs effects and pairwise miSNPs × 3utrSNPs interactions in a sample of 1,467 individuals for which genome-wide monocyte expression and genotype data were available. Statistical associations that survived multiple testing correction were tested for replication in an independent sample of 758 individuals with both monocyte gene expression and genotype data. In both studies, the hsa-mir-1279 rs1463335 was found to modulate in *cis* the expression of *LYZ* and in *trans* the expression of *CNTN6*, *CTRC*, *COPZ2*, *KRT9*, *LRRFIP1*, *NOD1*, *PCDHA6*, *ST5* and *TRAF3IP2* genes, supporting the role of hsa-mir-1279 as a regulator of several genes in monocytes. In addition, we identified two robust miSNPs × 3utrSNPs interactions, one involving *HLA-DPB1* rs1042448 and hsa-mir-219-1 rs107822, the second the *H1F0* rs1894644 and hsa-mir-659 rs5750504, modulating the expression of the associated genes. As some of the aforementioned genes have previously been reported to reside at disease-associated loci, our findings provide novel arguments supporting the hypothesis that the genetic variability of miRNAs could also contribute to the susceptibility to human diseases.

Citation: Greliche N, Zeller T, Wild PS, Rotival M, Schillert A, et al. (2012) Comprehensive Exploration of the Effects of miRNA SNPs on Monocyte Gene Expression. PLoS ONE 7(9): e45863. doi:10.1371/journal.pone.0045863

Editor: Andrea Vergani, Children's Hospital Boston, United States of America

Received April 20, 2012; **Accepted** August 22, 2012; **Published** September 21, 2012

Funding: The Gutenberg Health 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 Health 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. CARDIOGENICS was funded by the European Union FP6 program (LSHM-CT-2006-037593). NJ Samani holds a Chair supported by the British Heart Foundation. Work described in this paper is part of the research portfolio supported by the Leicester NIHR Biomedical Research Unit in Cardiovascular Disease. Collection of the Cardiogenics controls was part supported through the Cambridge Bioresource, which is funded by the NIHR Cambridge Biomedical Research Centre. Statistical analyses benefit from the C2BIG computing centre funded by the Fondation pour la Recherche Médicale, La Région Ile de France (CODDIM) and the Genomic Network of the Pierre and Marie Curie University (Paris 06). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have the following interests. Part of the Gutenberg Health Study is funded by its contract with Boehringer Ingelheim and PHILIPS Medical Systems including an unrestricted grant for the Gutenberg Health Study. There are no patents, products in development or marketed products to declare. 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.

* E-mail: david.tregouet@upmc.fr

¤ Current address: College London, Hammersmith Hospital, London, United Kingdom,

¶ Membership of the Cardiogenics Consortium is provided in the Acknowledgments.

Introduction

MicroRNAs (miRNAs) represent a class of small (~19–29 nucleotides) non coding RNAs that participate in gene post-transcriptional regulation. By binding to complementary target sites that are mainly located in gene 3'UTR regions, miRNAs inhibit mRNA translation either via mRNA degradation or via repression of mRNA translation [1]. A complete or nearly complete match of the miRNA with its target sequence generally results in a decrease of gene expression while a mismatch lead to a repression of mRNA translation. In general, miRNAs participate in regulating the expression of genes located remote from their

genomic sequence; however when miRNAs are located within gene introns they are highly likely to modulate the expression of the host gene [2,3].

According to the latest miRNA reference database (miRBase release 18, www.mirbase.org) [4], it is estimated that more than 1,500 miRNAs could exist in humans. A given miRNA may have several mRNA targets and participates in the regulation of a network of genes with genomic sequence similarities [5]. Reciprocally, a given mRNA may harbour in its 3'UTR region several different miRNA target sites and then be under the control of a set of miRNAs. It is estimated that, overall, about 50% of the genome would be subject to regulation by miRNAs [6,7], making them one of the most important component of a cell. It is then not surprising to find miRNAs associated with a large number of human diseases (~300 diseases according to the human miRNA disease database [8]) including cardiovascular and metabolic disorders [9–12].

As with any genomic sequence, miRNAs are prone to nucleotide variations that may have non negligible effects. The presence of a single nucleotide polymorphism (SNP) in the long miRNA primary (pri-miRNA) may affect its maturation process, its expression or the binding of the mature form to its target, which would then influence the expression of the target genes [13,14]. This is the case, for example, for rs11614913 located in the pri-miRNA-196. It is hypothesized that this SNP affects miR-196a-2 expression, alters the miRNA–target binding site and influences cancer risks [15,16]. The existence of a SNP in the miRNA genomic sequence may create mature miRNA variants, named isomiRs, whose predicted targets could differ from the original miRNA's targets [17]. In addition, the expression of miRNAs is known to be regulated by transcriptional factors, and by polymorphisms within the transcription factor binding sites, which may then modulate miRNA expression [18]. Finally, the presence of a SNP in the miRNA target sequences could also influence the expression of the targeted mRNAs [19,20]. As an example, the rs58186-C allele located in the 3'UTR region of the *AGTR1* gene has been shown to decrease the efficiency of the binding of miR-155 to this gene. leading to an increase in *AGTR1* expression [20].

In this study, we conducted a genome-wide investigation of the effect of pri-miRNA SNPs (miSNPs) on monocyte gene expression in a large epidemiological study of healthy subjects for whom genome-wide monocyte gene expressions and genotype data have been collected, as part of the Gutenberg Health Study [21–24]. We also conducted a genome-wide search for pair-wise interactions between miSNPs and SNPs located in 3'UTR regions (3utrSNPs). We reasoned that such investigation could help to identify novel miRNA-sensitive regulation of gene expression in a key cell type participating in several disease processes including inflammation, atherosclerosis and immunity [25]. miSNPs effects identified were further validated for replication in a second large monocyte expression dataset, the Cardiogenics Transcriptomic Study (CTS) [26].

Results

The Gutenberg Health Study (GHS) comprised 1,467 individuals (750 men and 717 women) [23]. All these individuals were typed for common SNPs using the Affymetrix Genome-Wide Human SNP Array 6.0 and their monocyte expression profiles were obtained from the Illumina HT-12 v3 Beadchip. Detailed description of these genome-wide expression and genotype data has already been provided elsewhere [21–24].

Probes and SNPs selection

The GRCH37 release of the Human reference genome and the 17th release of the miRNA database [4] were used to identify SNPs located within pri-miRNA sequences and 3'UTR regions. The number of miSNPs genotyped in GHS, or that could be substituted according to the SNAP software [27] by a "proxy" genotyped SNP in strong correlation (when expressed in terms of a pairwise linkage disequilibrium (LD) r^2greater than 0.90) was 294, representing 258 distinct miRNAs.

The pre-processing of the expression data (see Methods) identified 22,004 probes covering 15,786 genes of "perfect" quality score according to ReMOAT [28] and not harboring a SNP in their genomic sequence. These probes were then tested for association with all genotyped miSNPs.

The search for interactions between miSNPs and 3utrSNP was restricted to probes targeting genes known to contain SNPs in their 3'UTR region that were either directly genotyped in GHS, or tagged by genotyped SNPs (r^2>0.90). This led to the selection of a subsample of 8,768 probes characterizing 6,147 genes. In these genes, the total number of 3utrSNPs (or "proxy") that were further studied was 10,783. The distribution of the number of 3utrSNPs per gene is given in Table 1.

Association of miSNPs with gene expression

GHS discovery phase. This analysis can be viewed as an ancillary study of the whole genome-wide association study between all genotyped SNPs and all expressions already conducted in GHS and whose results can be found in a publicly available resource [23]. At the Bonferroni correction level of 7.73×10^{-9} (ie. 0.05/(294×22,004)), fifty-seven associations between miSNPs and gene expression were significant (Table S1). However, forty-eight of these associations implicated miSNPs proxies mapping the genomic region of the genes they were associated with. We interrogated the GHS express database to identify the SNPs showing the strongest association with the associated expression among those with $p<5.50\times10^{-5}$ and located within 1Mb of the probe genomic sequence, thereafter referred to as the best *cis* eSNPs [23]. In six cases, the miSNP proxies were the best *cis* eSNPs. After adjusting for the effect of the best *cis* eSNPs, most miSNPs association vanished and only seven (bold lines in Table S1) remained significant at $p=7.73\times10^{-9}$. Most of these 48 *cis* miSNPs associations are then likely due to LD between miSNPs and "true" *cis* eSNPs. Nevertheless, this must be investigated in greater depth as in several examples the corresponding miRNA was located within an intron of the associated gene, and could therefore participate in the regulation of the host gene.

Of more interest are the nine genome-wide significant associations that involved a miSNP located on a chromosome distinct from the one mapped by the associated gene, so called *trans* associations referring to associations involving SNPs that are located more than 1Mb away, or a distinct chromosome, from the associated probe. As shown in Table 2, the hsa-mir-1279 SNP rs1463335, tagged by the SNP rs317657 ($r^2=1.0$), was associated in *cis* with expression of *LYZ* ($R^2=20.1\%$; $p=1.36\times10^{-76}$) and *YEATS4* ($R^2=13.1\%$; $p=1.32\times10^{-46}$), and in *trans* with expression of *CNTN6* ($R^2=3.3\%$; $p=1.16\times10^{-12}$), *CTRC* ($R^2=3.5\%$; $p=1.39\times10^{-13}$), *COPZ2* ($R^2=3.0\%$; $p=2.33\times10^{-11}$), *KRT9* ($R^2=4.5\%$; $p=1.15\times10^{-15}$), *LRRFIP1* ($R^2=10.0\%$; $p=1.50\times10^{-35}$), *NOD1* ($R^2=2.1\%$; $p=7.25\times10^{-9}$), *PCDHA6* ($R^2=9.2\%$; $p=9.44\times10^{-33}$), *ST5* ($R^2=5.1\%$; $p=2.05\times10^{-18}$) and *TRAF3IP2* ($R^2=4.9\%$; $p=2.74\times10^{-17}$). It is of note that whereas the rs317657-C allele, with minor allele frequency 0.46, was associated with increased expression of *LYZ*, *YEATS4* and *NOD1*, it was associated with decreased levels of *CNTN6, CTRC,*

Table 1 Distribution of the number of 3utrSNPs (or proxy) in the 6,147 studied genes.

# 3utrSNPs per gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	18
# genes	3,435	1,438	670	313	138	80	35	17	7	4	5	1	1	2	1

Note that, in some instances, a genotyped SNP can serve as a proxy ($r^2 > 0.90$) for several 3utr SNPs. This explains why the total number of 3utr proxy SNPs that can be derived from this table (11,353 = 1×3,435 + 2×1,438 + 3*670 + ...) is slightly higher than the number of really studied SNPs (10,783).
doi:10.1371/journal.pone.0045863.t001

COPZ2, *KRT9*, *LRRFIP1*, *PCDHA6*, *ST5* and *TRAF3IP2* expression. After adjusting for the best *LYZ cis* eSNP, the association of rs317657 with *LYZ* expression still retained genome-wide significance ($p = 6.17\times10^{-11}$) while the association with *YEATS4* disappeared ($p = 0.734$) (Table S1). According to the TargetScan bioinformatics tool [5], the position 648 to 654 of the 3'UTR *LYZ* region is predicted to be complementary at 8 bases with the hsa-mir-1279 sequence. This type of matching configuration, called 8mer, is usually considered to be a good prior for predicting potential targets of miRNA. After adjusting for *LYZ* expression, the *trans* association observed with rs317657 were reduced, but remained highly significant, $p = 3.88\times10^{-11}$, $p = 1.15\times10^{-7}$, $p = 2.52\times10^{-6}$, $p = 1.65\times10^{-10}$, $p = 7.16\times10^{-29}$, $p = 2.44\times10^{-5}$, $p = 8.23\times10^{-28}$, $p = 1.81\times10^{-13}$, $p = 5.66\times10^{-10}$ for *CNTN6*, *CTRC*, *COPZ2*, *KRT9*, *LRRFIP1*, *NOD1*, *PCDHA6*, *ST5* and *TRAF3IP2*, respectively. Corresponding p-values for the *trans* associations adjusted for *YEATS4* expression were $p = 1.86\times10^{-9}$, $p = 1.72\times10^{-11}$, $p = 6.45\times10^{-9}$, $p = 9.48\times10^{-12}$, $p = 6.10\times10^{-28}$, $p = 3.76\times10^{-13}$, $p = 1.59\times10^{-28}$, $p = 2.33\times10^{-13}$, $p = 5.10\times10^{-8}$, respectively. When the *trans* associations were adjusted for both *LYZ* and *YEATS4* expressions, they were hardly modified, with p-values ranging between $p = 2.98\times10^{-6}$ (*COPZ2*) to $p = 6.55\times10^{-27}$ (*PCDHA6*). As indicated in Table 3, these nine genes were not strongly correlated with each other, nor with expression of LYZ, the gene in which the rs31757 SNP was located.

Replication in CTS. We focused on the genome-wide significant *trans* associations observed with the hsa-mir-1279 miSNP proxy. These associations were tested for replication in CTS where monocyte expression was measured in a sample of 395 healthy individuals and 363 patients with coronary artery disease [26]. In CTS, the hsa-mir-1279 rs1463335 proxy was the rs998022 ($r^2 = 0.90$). Its pairwise r^2 with the GHS rs317657 proxy was 0.84. The probe tagging the *LYZ* gene expression was not available in CTS, but all other associations were replicable. As indicated in Table 2, they all replicated with consistent pattern of association as in GHS. The rs998022-G allele tagging the rs317657-C allele was associated with increased expression of *YEATS4* ($R^2 = 11.2\%$; $p = 3.21\times10^{-21}$) and *NOD1* ($R^2 = 9.82\%$; $p = 7.83\times10^{-19}$), but with decreased expression of *CNTN6* ($R^2 = 5.9\%$; $p = 7.56\times10^{-12}$), *CTRC* ($R^2 = 8.1\%$; $p = 1.54\times10^{-15}$), *COPZ2* ($R^2 = 9.7\%$; $p = 2.06\times10^{-18}$), *KRT9* ($R^2 = 5.9\%$; $p = 1.11\times10^{-11}$), *LRRFIP1* ($R^2 = 16.7\%$; $p = 6.65\times10^{-32}$), *PCDHA6* ($R^2 = 16.4\%$; $p = 2.67\times10^{-31}$), *ST5* ($R^2 = 17.0\%$; $p = 2.51\times10^{-30}$) and *TRAF3IP2* ($R^2 = 8.9\%$; $p = 5.23\times10^{-17}$). Associations were homogeneously observed in CAD patients and healthy subjects from CTS (Table S2).

Search for miSNP × 3utrSNP interactions

GHS discovery phase. Each 3utrSNP was tested for interaction with all miSNPs with respect to the expression levels of the probes tagging the 3utrSNP-associated gene. Interactions were assessed using a standard linear regression analysis where both SNPs coded as 0,1,2 were included to the model together with the corresponding interaction term. Analyses were adjusted for age and sex. The total number of tested interactions was 4,890,102.

Instead of applying the standard Bonferroni correction to handle multiple testing, we followed the suggestion by Pare et al. [29] and adopted a weighted-Bonferroni correction according to the p-value of the Levene's test. This consists in prioritizing 3utrSNPs according to the significance of the test for a difference in the variance of expressions according to genotypes. This strategy relies on the statistical property that a significant difference in phenotypic variances according to sub-groups could be a marker for interaction phenomena.

Using this weighted-Bonferroni correction, 51 miSNP × 3utrSNP interactions were genome-wide significant at $p < 1.02\times10^{-8}$ (Table 4). Note, only 31 would have been declared significant according the standard Bonferroni procedure (Table 4). Seventeen of the detected interactions involved the *RFPL1* rs13053624 that was found to interact with 17 miSNPs over 16 distinct miRNAs to modulate *RFPL1* expression (probe ILMN_1797383). One of these interacting miRNAs was hsa-mir-3674. Interestingly, according to microSNiPer database [30], *RFPL1* is predicted to harbor a SNP, rs13053817, in a potential target site for hsa-mir-3674 that is, according to the SNAP database, in nearly complete association with the identified rs13053624 ($r^2 = 0.90$). No other strong biological and bioinformatics evidence could be obtained from public databases (miRanda [31], TargetScan [5], DianaMicro [32], PicTar [33], mirBase [4]) in favour of the 30 other genes we identified through our interaction search (Table 4).

Replication in CTS. The fifty-one genome-wide significant interactions were tested for replication in CTS. However, only eight interactions could be replicable, which did not include the aforementioned interaction involving *RFPL1* rs13053624.

Using the same linear regression model, further adjusted for disease status as for the discovery phase, two interactions replicated in CTS at the Bonferroni-corrected level of 6.25×10^{-3} (Table 5).

The first replicated interaction involved the *HLA-DPB1* rs1042448 and hsa-mir-219-1 rs107822 tagged by the rs3128923/rs213208 and rs3117222/rs439205 pairs in GHS and CTS, respectively. These two loci are distant from about 100 kb and the corresponding tag SNPs were in modest linkage disequilibrium (LD), $r^2 = 0.58$ and $r^2 = 0.56$, in GHS and CTS, respectively. In GHS, the haplotype analysis of the rs107822 and rs1042448 proxies revealed that the *HLA-DPB1* rs1042448-A proxy allele (i.e the allele at the proxy SNP that can be used to tag the rs1042448-A allele) was associated with a strong increase in *HLA-DBP1* expression ($\beta = +0.61$, $p = 1.64\times10^{-105}$) when carried on the same haplotype as the hsa-mir-219-1 rs107822-C proxy allele (Figure 1). Conversely, when associated with the hsa-mir-219-1 rs107822-T proxy allele, the increasing effect of the *HLA-DPB1* rs1042448-A proxy allele was significantly reduced

Table 2 *Cis* and *trans*-associations observed with the hsa-mir-1279 rs1463335[(1)].

Associated Gene Expression					GHS			CTS		
Probe	Gene	CHR	Start	End	β[(2)]	SE	P[(3)]	β[(2)]	SE	P[(3)]
ILMN_1748730	CTRC	1	15764937	15773152	−0.03	0.004	$1.39\ 10^{-13}$	−0.06	0.007	$1.54\ 10^{-15}$
ILMN_2252021	LRRFIP1	2	238536223	238690289	−0.05	0.004	$1.50\ 10^{-35}$	−0.12	0.010	$6.65\ 10^{-32}$
ILMN_1699317	CNTN6	3	1134628	1445277	−0.02	0.003	$1.16\ 10^{-12}$	−0.04	0.006	$7.56\ 10^{-12}$
ILMN_1740494	PCDHA6	5	140207649	140391928	−0.04	0.003	$9.44\ 10^{-33}$	−0.10	0.008	$2.67\ 10^{-31}$
ILMN_1663381	TRAF3IP2	6	111880142	111927320	−0.03	0.003	$2.74\ 10^{-17}$	−0.06	0.007	$5.23\ 10^{-17}$
ILMN_2114422	NOD1	7	30464142	30518392	0.05	0.008	$7.25\ 10^{-9}$	0.12	0.013	$7.83\ 10^{-19}$
ILMN_1731063	ST5	11	8714898	8932497	−0.06	0.007	$2.05\ 10^{-18}$	−0.22	0.019	$2.51\ 10^{-30}$
ILMN_1815205	LYZ[(1)]	12	69742133	69748012	0.20	0.010	$1.36\ 10^{-76}$	NA	NA	NA
ILMN_1801387	YEATS4[(1)]	12	69753531	69784575	0.15	0.010	$1.32\ 10^{-46}$	0.19	0.020	$3.27\ 10^{-21}$
ILMN_1792568	KRT9	17	39722092	39728309	−0.04	0.006	$1.15\ 10^{-15}$	−0.11	0.016	$1.11\ 10^{-11}$
ILMN_1667361	COPZ2	17	46103532	46115151	−0.03	0.005	$2.33\ 10^{-11}$	−0.10	0.011	$2.06\ 10^{-18}$

[(1)]The rs1463335 was tagged by the rs317657 and rs998022 in GHS and CTS, respectively. The rs146335 is located on chromosome 12, at position 69,667,075. As a consequence, the association observed with LYZ and YEATS4 are considered as *cis*-associations, the remaining eight as trans-associations.
[(2)]Regression coefficient associated with the rare miSNP allele under an additive effect model, adjusted for age and gender
[(3)]P-value of the association between miSNP and gene expression
doi:10.1371/journal.pone.0045863.t002

($p = 1.88\times10^{-20}$) and became $\beta = +0.18$ ($p = 3.49\times10^{-8}$) illustrating the interaction phenomenon identified through linear regression analysis. This interaction remained significant ($p = 2.81\times10^{-12}$) when the haplotype analysis was further adjusted on the best *cis* eSNP observed for *HLA-DBP1* expression, rs3128963 ($p = 2.30\times10^{-151}$) (see GHS_Express database [23]). The same pattern of associations was observed in CTS (Figure 1). The *HLA-DPB1* rs1042448-A proxy allele was associated with a strong significant increase in *HLA-DPB1* expression ($\beta = +0.63$, $p = 5.24\times10^{-62}$) when carried on the same haplotype as the hsa-mir-219-1 rs107822-C proxy allele. The corresponding increase when the rs1042248-A proxy allele was associated with the hsa-mir-219-1 rs107822-A proxy allele was significantly reduced ($p = 2.68\times10^{-20}$) and did no longer reach significance ($\beta = +0.05$; $p = 0.23$) (Figure 1).

The second replicated interaction involved the *H1F0* rs1894644 and hsa-mir-659 rs5750504 tagged by the rs763137/rs2899293 and rs1894644/rs6000905 pairs in GHS and CTS, respectively (Figure 2). These two loci are distant from about 40 kb and the corresponding tag SNPs were in low LD, $r^2 = 0.15$ and $r^2 = 0.14$, in GHS and CTS, respectively. In GHS and in CTS, the *H1F0* rs1894644-T proxy allele was associated with a strong increase in *H1F0* expression ($\beta = +0.65$, $p = 1.71\times10^{-53}$ and $\beta = +0.79$, $p = 1.36\times10^{-40}$, respectively) when it was on the same haplotype as the rs5750504-T proxy allele. Conversely, when the rs1894644-T proxy allele was on the same haplotype as the rs5750504-A proxy allele, the corresponding increase in *H1F0* expression was lower ($\beta = +0.23$, $p = 9.74\times10^{-13}$ and $\beta = +0.26$, $p = 7.25\times10^{-8}$, respectively). The test for homogeneity of the *H1F0* rs1894644 effect according to the rs5750504 proxy was significant $p = 3.03\times10^{-12}$ and $p = 5.67\times10^{-10}$ in GHS and CTS, respectively, validating the interaction detected through standard linear regression analysis ($p = 2.98\times10^{-10}$ and $p = 1.37\times10^{-8}$, respectively). Note that, in GHS, the rs763137 SNP involved in this interaction was the best *cis* eSNP for *H1F0* ($p = 1.10\times10^{-62}$).

As shown in Table S3, the two replicated interactions were consistent in CAD and healthy subjects composing CTS.

Table 3 Correlation between gene expressions influenced by the rs317657 tagging the hsa-mir-1279 rs1463335.

	CTRC	LRRFIP1	CNTN6	PCDHA6	TRAF3IP2	NOD1	ST5	LYZ	YEATS4	KRT9
LRRFIP1	0.204	1								
CNTN6	0.137	0.237	1							
PCDHA6	0.202	0.449	0.200	1						
TRAF3IP2	0.129	0.271	0.202	0.270	1					
NOD1	0.225	−0.126	0.047	−0.062	0.029	1				
ST5	0.210	0.517	0.192	0.411	0.274	−0.176	1			
LYZ	−0.156	−0.143	−0.070	−0.125	−0.170	0.113	−0.125	1		
YEATS4	−0.079	−0.162	−0.110	−0.113	−0.250	−0.070	−0.140	0.558	1	
KRT9	0.217	0.485	0.168	0.402	0.302	−0.166	0.740	−0.133	−0.121	1
COPZ2	0.188	0.400	0.131	0.341	0.236	−0.140	0.592	−0.143	−0.093	0.590

doi:10.1371/journal.pone.0045863.t003

Table 4 Genome-wide significant ($p<1.02\ 10^{-8}$) interactions between miSNPs and 3utrSNPs on monocyte gene expression in the Gutenberg Health Study.

							GHS				
Gene	**CHR**	**Probe**	**3utrSNP**	**miRNA**	**CHR**	**miSNP**	**miProxy**	**3utrProxy**	**P(1)**	**Levene P-value**	**weighted P(2)**
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-592	7	rs11563750	rs11563505	rs13053817	$1.04\ 10^{-35}$	$3.22\ 10^{-5}$	$1.50\ 10^{-36}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-3920	11	rs12275715	rs12283329	rs13053817	$1.21\ 10^{-26}$	$3.22\ 10^{-5}$	$1.74\ 10^{-27}$
TXNDC5	6	ILMN_1769082	rs8643	hsa-mir-125b-2	21	rs2823897	rs2211981	rs8643	$8.95\ 10^{-18}$	$3.39\ 10^{-1}$	$1.23\ 10^{-17}$
TXNDC5	6	ILMN_1769082	rs1043784	hsa-mir-125b-2	21	rs2823897	rs2211981	rs3734589	$1.26\ 10^{-17}$	$3.18\ 10^{-1}$	$1.64\ 10^{-17}$
LYZ	12	ILMN_1815205	rs710794	hsa-mir-1279	12	rs1463335	rs317657	rs710794	$4.13\ 10^{-15}$	$4.51\ 10^{-23}$	$1.20\ 10^{-16}$
ASB1	2	ILMN_1683096	rs1044561	hsa-mir-125b-2	21	rs2823897	rs2211981	rs2334004	$1.45\ 10^{-16}$	$8.91\ 10^{-1}$	$1.87\ 10^{-15}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4656	7	rs3750013	rs17135110	rs13053817	$2.28\ 10^{-14}$	$3.22\ 10^{-5}$	$3.29\ 10^{-15}$
ASB1	2	ILMN_1683096	rs2278768	hsa-mir-3119-1	1	rs17349873	rs1330387	rs2278768	$3.71\ 10^{-14}$	$1.34\ 10^{-6}$	$4.10\ 10^{-15}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-30c-1	1	rs16827546	rs16827546	rs13053817	$2.89\ 10^{-14}$	$3.22\ 10^{-5}$	$4.16\ 10^{-15}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-1307	10	rs7911488	rs2271751	rs9287035	$2.98\ 10^{-13}$	$9.07\ 10^{-46}$	$4.29\ 10^{-15}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-125b-1	11	rs2081443	rs2081443	rs13053817	$2.40\ 10^{-13}$	$3.22\ 10^{-5}$	$3.47\ 10^{-14}$
PKD1L2	16	ILMN_1742788	rs1901818	hsa-mir-4272	3	rs9868022	rs9868022	rs7198127	$8.92\ 10^{-14}$	$8.80\ 10^{-2}$	$5.47\ 10^{-14}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-4670	9	rs2104533	rs2296666	rs9287035	$5.16\ 10^{-12}$	$9.07\ 10^{-46}$	$7.42\ 10^{-14}$
ASB1	2	ILMN_1683096	rs2278768	hsa-mir-125b-2	21	rs2823897	rs2211981	rs2278768	$5.30\ 10^{-12}$	$1.34\ 10^{-6}$	$5.85\ 10^{-12}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4300	11	rs11603185	rs7944477	rs13053817	$2.02\ 10^{-11}$	$3.22\ 10^{-5}$	$2.92\ 10^{-12}$
SPRY1	4	ILMN_2329914	rs300574	hsa-mir-4666	1	rs16841344	rs4653963	rs300555	$1.52\ 10^{-11}$	$1.16\ 10^{-2}$	$5.10\ 10^{-12}$
HLA-DPB1	6	ILMN_1749070	rs1042448	hsa-mir-219-1	6	rs107822	rs213208	rs3128923	$1.26\ 10^{-10}$	$4.11\ 10^{-8}$	$1.11\ 10^{-11}$
ASB1	2	ILMN_1683096	rs2278768	hsa-mir-4636	5	rs257095	rs6555591	rs2278768	$1.09\ 10^{-10}$	$1.34\ 10^{-6}$	$1.20\ 10^{-11}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4292	9	rs2811749	rs2811749	rs13053817	$1.98\ 10^{-10}$	$3.22\ 10^{-5}$	$2.86\ 10^{-11}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-624	14	rs11156654	rs11156654	rs13053817	$2.20\ 10^{-10}$	$3.22\ 10^{-5}$	$3.18\ 10^{-11}$
GPRC5C	17	ILMN_1724211	rs2706527	hsa-mir-3667	22	rs135771	rs135775	rs2706526	$5.46\ 10^{-9}$	$5.08\ 10^{-79}$	$4.52\ 10^{-11}$
H1F0	22	ILMN_1757467	rs1894644	hsa-mir-659	22	rs5750504	rs2899293	rs763137	$2.98\ 10^{-10}$	$1.30\ 10^{-1}$	$2.18\ 10^{-10}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-548n	7	rs1649215	rs1637670	rs9287035	$1.64\ 10^{-8}$	$9.07\ 10^{-46}$	$2.37\ 10^{-10}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-521-1	19	rs4803178	rs4803178	rs13053817	$2.88\ 10^{-9}$	$3.22\ 10^{-5}$	$4.16\ 10^{-10}$
GPRC5C	17	ILMN_2352090	rs2706527	hsa-mir-3667	22	rs135771	rs135775	rs2706526	$1.06\ 10^{-7}$	$6.63\ 10^{-102}$	$6.80\ 10^{-10}$
GPRC5C	17	ILMN_2352090	rs2706527	hsa-mir-107	10	rs17481096	rs17481096	rs2706526	$1.20\ 10^{-7}$	$6.63\ 10^{-102}$	$7.69\ 10^{-10}$
HLA-DPB1	6	ILMN_1749070	rs1042448	hsa-mir-219-1	6	rs213210	rs213210	rs3128923	$8.98\ 10^{-9}$	$4.11\ 10^{-8}$	$7.88\ 10^{-10}$
MXRA7	17	ILMN_1743836	rs10473	hsa-mir-490	7	rs6963819	rs2350780	rs7221855	$2.66\ 10^{-7}$	$6.10\ 10^{-167}$	$1.04\ 10^{-9}$
SPRY1	4	ILMN_1651610	rs300574	hsa-mir-4666	1	rs16841344	rs4653963	rs300555	$3.82\ 10^{-9}$	$6.28\ 10^{-3}$	$1.12\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-1236	6	rs403569	rs550513	rs13053817	$7.89\ 10^{-9}$	$3.22\ 10^{-5}$	$1.14\ 10^{-9}$
GPRC5C	17	ILMN_2352090	rs2706527	hsa-mir-941-1	20	rs2427555	rs2427554	rs2706526	$2.03\ 10^{-7}$	$6.63\ 10^{-102}$	$1.30\ 10^{-9}$
POGZ	1	ILMN_2329309	rs3811409	hsa-mir-4666	1	rs16841344	rs4653963	rs3811409	$2.24\ 10^{-9}$	$1.12\ 10^{-1}$	$1.53\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4643	6	rs16884450	rs16884450	rs13053817	$1.28\ 10^{-8}$	$3.22\ 10^{-5}$	$1.85\ 10^{-9}$
ASB1	2	ILMN_1683096	rs1044561	hsa-mir-3973	11	rs262404	rs16928224	rs2334004	$1.60\ 10^{-10}$	$8.91\ 10^{-1}$	$2.06\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-3646	20	rs11574730	rs11574730	rs13053817	$1.70\ 10^{-8}$	$3.22\ 10^{-5}$	$2.45\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-4460	5	rs13171514	rs13171514	rs9287035	$2.47\ 10^{-7}$	$9.07\ 10^{-46}$	$3.55\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-3674	8	rs7003112	rs6558541	rs13053817	$2.55\ 10^{-8}$	$3.22\ 10^{-5}$	$3.67\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-1205	8	rs9649959	rs9649959	rs13053817	$2.78\ 10^{-8}$	$3.22\ 10^{-5}$	$4.02\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-4656	7	rs17829969	rs17829969	rs13053817	$2.82\ 10^{-8}$	$3.22\ 10^{-5}$	$4.07\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-4784	2	rs6709245	rs12463867	rs9287035	$3.22\ 10^{-7}$	$9.07\ 10^{-46}$	$4.63\ 10^{-9}$
AAK1	2	ILMN_1880387	rs13427243	hsa-mir-3667	22	rs135771	rs135775	rs13427243	$7.28\ 10^{-9}$	$1.04\ 10^{-1}$	$4.80\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-604	10	rs2368392	rs3758371	rs13053817	$3.69\ 10^{-8}$	$3.22\ 10^{-5}$	$5.32\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-215	1	rs3820455	rs34406824	rs9287035	$3.88\ 10^{-7}$	$9.07\ 10^{-46}$	$5.58\ 10^{-9}$
RBM12	20	ILMN_1670841	rs6060539	hsa-mir-4755	20	rs2284385	rs2284390	rs2425125	$4.06\ 10^{-7}$	$1.65\ 10^{-47}$	$5.62\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-2113	6	rs9375085	rs9375085	rs9287035	$4.02\ 10^{-7}$	$9.07\ 10^{-46}$	$5.79\ 10^{-9}$
RFPL1	22	ILMN_1797383	rs13053624	hsa-mir-1269b	17	rs7210937	rs2240567	rs13053817	$4.93\ 10^{-8}$	$3.22\ 10^{-5}$	$7.10\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-4705	13	rs7337292	rs7337292	rs9287035	$5.10\ 10^{-7}$	$9.07\ 10^{-46}$	$7.33\ 10^{-9}$

Table 4. Cont.

							GHS				
Gene	**CHR**	**Probe**	**3utrSNP**	**miRNA**	**CHR**	**miSNP**	**miProxy**	**3utrProxy**	**P(1)**	**Levene P-value**	**weighted P(2)**
PKD1L2	16	ILMN_1742788	rs1901818	hsa-mir-4473	9	rs16938058	rs16938057	rs7198127	$1.24\ 10^{-8}$	$8.80\ 10^{-2}$	$7.60\ 10^{-9}$
MRPL43	10	ILMN_1678974	rs2295716	hsa-mir-608	10	rs4919510	rs4919510	rs3824783	$3.06\ 10^{-7}$	$9.68\ 10^{-22}$	$9.44\ 10^{-9}$
ECE1	1	ILMN_1672174	rs3026907	hsa-mir-520d	19	rs2217653	rs9304754	rs9287035	$6.62\ 10^{-7}$	$9.07\ 10^{-46}$	$9.52\ 10^{-9}$
ASB1	2	ILMN_1683096	rs1044561	hsa-mir-4636	5	rs257095	rs6555591	rs2334004	$7.57\ 10^{-10}$	$8.91\ 10^{-1}$	$9.74\ 10^{-9}$

(1) P-value of the interaction test derived from the standard linear regression analysis
(2) P-value of the interaction test obtained when the Levene test p-value was used under a weighted-Bonferroni framework.
doi:10.1371/journal.pone.0045863.t004

Discussion

Coupling genome-wide association and expression studies have been an attractive strategy to disentangle the architecture of the genetics of gene expression and to assess whether gene expression dysregulation could mediate the effect of SNPs on disease risk identified through genome-wide association studies [23,34]. To our knowledge, such studies [23,34–37] mainly focused on assessing marginal associations of single SNPs with gene expression. Even if SNP × SNP interactions have often been advocated as a potential source of phenotype variability [38,39], there has been few attempt to assess at the genome-wide scale whether such SNP × SNP interactions could influence gene expression variability. This is likely due to the statistical and computing burdens associated with such investigations characterized by a huge number of tested interactions and the very large sample size required to detect genome-wide significance. We postulated that focusing on plausible "biological" interactions could be one strategy to dig into the complex architecture of SNP × SNP interactions. This is why we undertook what we think is the first systematic and comprehensive search for interactions between SNPs located in the genomic sequence of miRNAs and SNPs located in the 3'UTR gene regions that could participate in monocyte gene expression. This search for interactions was preceded by a genome-wide investigation of miSNPs effect on monocyte expression to assess whether miSNPs could influence gene expression, in particular, through *trans* regulation.

These investigations were conducted in the Gutenberg Health Study where the extensive genome-wide study of marginal SNP associations with monocyte expressions had previously been reported and the results stored in a publicly available resource [23], and we replicated the significant findings in the Cardiogenics study.

Our survey of marginal miSNP effect has pointed out the hsa-mir-1279 miRNA mapping to chromosome 12q15 as a candidate regulator of 10 genes in monocytes. Indeed, we observed that the hsa-mir-1279 rs1463335 tagged by rs317657 or rs1463335 was

Table 5 Replication in Cardiogenics of the miSNPs × 3utrSNPs detected in Gutenberg Health Study.

MiSNP × 3utrSNP	**rs17349873 rs2278768**	**rs107822 rs1042448**	**rs257095 rs2278768**	**rs5750504 rs1894644**	**rs6963819 rs10473**	**rs262404 rs1044561**	**rs2284385 rs6060539**	**rs257095 rs1044561**
miRNA (CHR)	hsa-mir-3119-1 (1)	hsa-mir-219-1 (6)	hsa-mir-4636 (5)	hsa-mir-659 (22)	hsa-mir-490 (7)	hsa-mir-3973 (11)	hsa-mir-4755 (20)	hsa-mir-4636 (5)
Gene (CHR)	ASB1 (2)	HLA-DPB1 (6)	ASB1 (2)	H1F0 (22)	MXRA7 (7)	ASB1 (2)	RBM12 (20)	ASB1 (2)
Probe	ILMN_1683096	ILMN_1749070	ILMN_1683096	ILMN_1757467	ILMN_1743836	ILMN_1683096	ILMN_1670841	ILMN_1683096
Gutenberg Health Study								
Proxies	rs1330387 rs2278768	rs213208 rs3128923	rs6555591 rs2278768	rs2899293 rs763137	rs2350780 rs7221855	rs16928224 rs2334004	rs2284390 rs2425125	rs6555591 rs2334004
β (1)	−0.480	−0.165	−0.233	−0.194	−0.065	0.988	0.164	0.375
Weighted P-value (2)	$4.10\ 10^{-15}$	$1.11\ 10^{-11}$	$1.20\ 10^{-11}$	$2.18\ 10^{-10}$	$1.04\ 10^{-9}$	$2.06\ 10^{-9}$	$5.62\ 10^{-9}$	$9.74\ 10^{-9}$
Cardiogenics Transcriptomic Study								
Proxies	rs6703198 rs10084192	rs439205 rs3117222	rs257095 rs10084192	rs6000905 rs1894644	rs2350780 rs9910052	rs262407 rs10084192	rs2038123 rs6121015	rs257095 rs10084192
β (1)	0.093	−0.274	0.045	−0.268	0.011	−0.025	0.099	0.045
P-value (3)	$4.62\ 10^{-1}$	**$2.03\ 10^{-13}$**	$5.18\ 10^{-1}$	**$1.37\ 10^{-8}$**	$5.98\ 10^{-1}$	$8.29\ 10^{-1}$	$7.22\ 10^{-2}$	$5.18\ 10^{-1}$

(1) Regression coefficient of the interaction term when both miSNP and 3utr proxy SNPs coded 0/1/2 according to the number of carried rare alleles are introduced in a linear regression model together with their interaction term.
(2) P-value of the interaction test obtained in GHS when the Levene test p-value was used under a weighted-Bonferroni framework.
(3) P-value of the interaction test derived from the standard linear regression analysis in Cardiogenics. Bold p-values are significant after Bonferroni correction.
doi:10.1371/journal.pone.0045863.t005

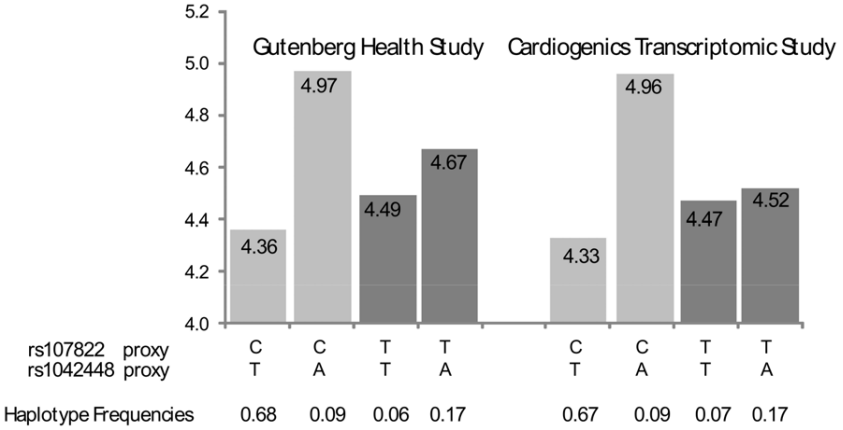


Figure 1. *HLA-DPB1* rs1042448 × *hsa-mir-219-1* rs107822 interaction on *HLA-DPB1* monocyte expression. In the Gutenberg Health Study, the rs1042248/rs107822 pair was tagged by rs3128923/rs213208. In the Cardiogenics Transcriptomic Study, the corresponding tagging pair was rs3117222/rs439205.
doi:10.1371/journal.pone.0045863.g001

robustly associated in *cis* with *LYZ* expression and in *trans* with *CNTN6*, *CTRC*, *COPZ2*, *KRT9*, *LRRFIP1*, *NOD1*, *PCDHA6*, *ST5* and *TRAF3IP2*. The bioinformatics prediction of the *LYZ* gene as a target for hsa-mir-1279 miRNA supports this hypothesis. The lack of strong correlation between the expression of these 10 genes, together with the *trans* association observed after adjusting for *LYZ* expression, could suggest that these nine genes could also be targets for the hsa-mir-1279, despite the absence of such prediction by current bioinformatics tools. However, the observation of positive associations with *LYZ* and *NOD1*, but of negative associations with the other genes, is puzzling as we could have expected, at first sight, a similar pattern of associations if all these genes were target for hsa-mir-1279. Functional experimental work is needed to characterize the role of hsa-mir-1279 in the regulation of these genes in-depth, in particular *TRAF3IP2* as this gene was identified in two independent GWAS as a susceptibility locus for psoriasis [40,41]. Our results, if confirmed, could open therapeutics perspectives as it is possible to use artificial miRNA targets to modify gene expression [42,43]. A *trans* association pattern was also recently observed at the locus 12q15 using an unsupervised gene networks analysis of the same datasets [24]. The rs11177644 located in the 3'UTR region of the *YEATS*4 gene was also found associated in *cis* to *LYZ* and *YEATS4* and in *trans* with a module of 36 genes including the *CNTN6*, *CTRC*, *COPZ2*, *KRT9*, *LRRFIP1*, *NOD1* and *ST5* discussed above. However, unlike what we observed here with hsa-mir-1279 rs1463335, the *trans* associations with rs11177644 had been found mediated by *cis* regulation mechanisms. Using a standard linear regression analysis (see above), we then tested whether these two SNPs could interact to contribute to the identified *trans* associations. We did not observe any strong evidence for such phenomenon as the lowest p-value for interaction was $p = 8.53 \times 10^{-4}$ for *PCDHA6* (data not shown). As the rs11177644 and rs1463335 were in moderate LD ($r^2 = 0.30$ and D' = +0.70), we further conducted an haplotype analysis of the two SNPs (Table 6). This revealed that both SNPs acted additively on *LYZ* expression but, after adjusting for rs11177644, the association of rs1463335 with *YEATS4* was no longer significant ($p = 0.748$). This haplotype analysis also revealed strong *trans* haplotype associations, which were due to a single haplotype, (rs317657_C/rs11177644_A), which was, after adjusting for *LYZ*

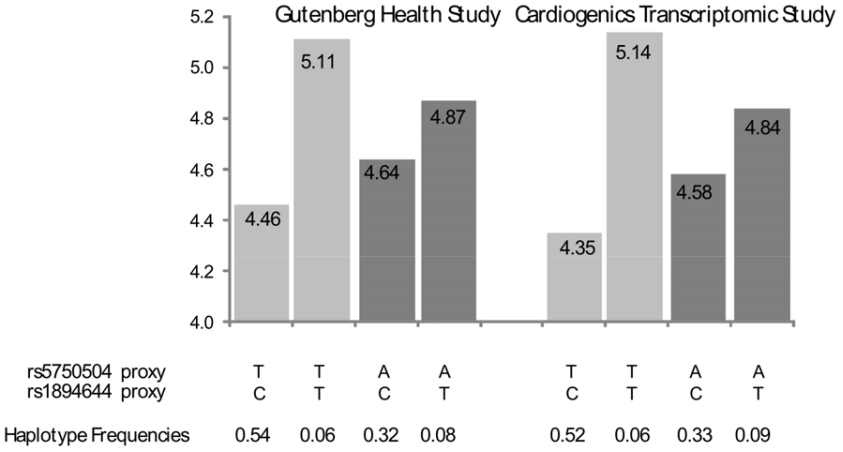


Figure 2. *H1F0* rs1894644 × *hsa-mir-659* rs5750504 interaction on *H1F0* monocyte expression. In the Gutenberg Health Study, the rs1894644/rs5750504 pair was tagged by rs763137/rs2899293. In the Cardiogenics Transcriptomic Study, the corresponding tagging pair was rs1894644/rs6000905.
doi:10.1371/journal.pone.0045863.g002

Table 6 Haplotype effects derived from the rs317657 and rs11177644 at the 12q15 locus in the Gutenberg Health Study (N = 1,467).

Polymorphisms		Haplotype Frequencies	Haplotype effects on Gene Expressions[1]					
rs317657	**rs11177644**		*YEATS4*[2]	*LYZ*[3]	*PCDHA6*	*LRRFIP1*	*ST5*	*KRT9*
C	A	0.399	reference	reference	reference	reference	reference	reference
C	G	0.061	−0.231 [−0.271 – −0.191]	−0.209 [−0.246 – −0.171]	+0.067 [0.051 – 0.083]	+0.074 [0.051 – 0.096]	+0.082 [0.047 – 0.117]	+0.068 [0.041 – 0.096]
T	A	0.155	+0.018 [−0.008 – 0.044]	−0.073 [−0.103 – −0.042]	+0.051 [0.041 – 0.061]	+0.053 [0.041 – 0.065]	+0.058 [0.038 – 0.078]	+0.051 [0.035 – 0.067]
T	G	0.385	−0.258 [−0.277 – −0.240]	−0.281 [−0.301 – −0.260]	+0.061 [0.052 – 0.070]	+0.081 [0.068 – 0.094]	+0.079 [0.060 – 0.098]	+0.054 [0.038 – 0.071]
Haplotypic association			R^2 = 39.3% p = 1.72 10^{-158}	R^2 = 36.6% p = 7.08 10^{-145}	R^2 = 12.3% p = 2.52 10^{-38}	R^2 = 11.4% p = 1.00 10^{-37}	R^2 = 5.20% p = 3.02 10^{-16}	R^2 = 4.69% p = 2.55 10^{-13}

Polymorphisms		Haplotype Frequencies	Haplotype effects on Gene Expressions[1]				
rs317657	**rs11177644**		*NOD1*	*CNTN6*	*CTRC*	*TRAF3IP2*	*COPZ2*
C	A	0.399	reference	reference	reference	reference	reference
C	G	0.061	−0.083 [−0.120 – −0.045]	+0.017 [0.004 – 0.029]	+0.031 [0.011 – 0.051]	+0.022 [0.007 – 0.037]	+0.029 [0.002 – 0.056]
T	A	0.155	−0.033 [−0.056 – −0.010]	+0.019 [0.011 – 0.027]	+0.021 [0.007 – 0.034]	+0.021 [0.011 – 0.030]	+0.035 [0.020 – 0.049]
T	G	0.385	−0.086 [−0.108 – −0.064]	+0.023 [0.016 – 0.031]	+0.039 [0.028 – 0.051]	+0.024 [0.015 – 0.033]	+0.028 [0.014 – 0.042]
Haplotypic association			R^2 = 4.08% p = 8.30 10^{-13}	R^2 = 3.31% p = 3.87 10^{-10}	R^2 = 3.08% p = 6.70 10^{-10}	R^2 = 2.87% p = 3.40 10^{-8}	R^2 = 2.27% p = 2.21 10^{-6}

[1]Haplotype effects were estimated assuming haplotype additive effects after adjusting for age, gender, and *LYZ* and *YEATS4* expressions when appropriate.
[2]After adjusting for rs317657, the rs11177644-G allele was associated with decreased *YEATS4* expression (β = −0.262, p = 1.36 10^{-138}). After adjusting for rs11177644, the rs317657-T allele was no longer associated with *YEATS4* expression (β = −0.003, p = 0.748).
[3]After adjusting for rs317657, the rs11177644-G allele was associated with decreased *LYZ* expression (β = −0.208, p = 9.56 10^{-84}). After adjusting for rs11177644, the rs317657-T allele was associated with decreased *LYZ* expression (β = −0.072, p = 1.24 10^{-10}).
doi:10.1371/journal.pone.0045863.t006

and *YEATS4* expression, strongly associated with increased levels of *NOD1* ($p = 8.30 \times 10^{-13}$), and decreased levels of the eight other genes, with p-values ranging from 2.21×10^{-6} to 2.52×10^{-38} (Table 6). These results suggest that the associations observed at the 12q15 locus are much more complex as initially hypothesized. It appeared that *YEATS4* and *LYZ* expressions could be under the influence of a common *cis* eSNP, but the latter would also be additionally influenced by a miSNP contributing to *trans* associations. As discussed in the following paragraph, further investigating including molecular experiments are required to dissect this complex pattern of association.

Two interactions miSNPs × 3utrSNPs were robustly identified, the first involving *HLA-DPB1* rs1042448 and hsa-mir-219-1 rs107822, the second the *H1F0* rs1894644 and hsa-mir-659 rs5750504. In both cases, the identified 3'UTR rare alleles were found to strongly increase the expression of the associated genes, but these over-expressions were highly reduced in carriers of miSNPs rare alleles. The identified miSNPs are not located within the mature sequence of the associated miRNAs but in their pri-miRNA sequences. These rare alleles could either be associated with increased miRNA expression or could tag for yet-unknown miSNPs within mature sequences leading to the production of isomiRs. It could be speculated that the associated miRNAs or isomiRs would then target the identified 3'UTR regions made sensitive to miRNAs regulation by the identified 3'UTR variants, variants that could create novel motifs for miRNAs' binding and would lead to reduction of the *per se* effect of the 3'UTR variant. Molecular constructs are required to assess such hypothesis. We further checked whether the identified miSNPs could interact with other 3'UTR SNPs located in genes in the vicinity of the *HLA-DBP1* and *H1F0* loci. We did not observe any suggestive evidence ($P<0.05$) for such interaction suggesting that the identified miRNA regulation would be specific to *HLA-DBP1* and *H1F0*.

The identified interactions involved SNPs in modest LD but located within a genomic distance of less than 100 kb. Several examples have already been observed where a given miRNA participates to the regulation of a gene located in its very close vicinity [2,3,44,45]. Nevertheless, one cannot exclude the possibility that the detected interactions are tagging for other complex haplotypic effects spanning a large distance and over several genes, five genes lying between *HLA-DPB1* and hsa-mir-219-1 and three between *H1F0* and hsa-mir-659 (Figure 3). Additional functional experiments would be required to biologically characterize the detected statistical interactions.

Little is known about *H1F0* in human diseases except that it codes for a histone family member protein. Interestingly, hsa-mir-659 has been shown to influence the risk of dementia [46] through a mechanism that could involve histone deacetylation [47,48]. Although speculative, the joint contribution of *H1F0* and hsa-mir-659 on the risk of dementia could deserve further attention. Conversely the *HLA-DPB1* gene has been associated with several complex diseases such as pulmonary hypertension, hepatitis B infection and systemic sclerosis [49–51]. In addition, hsa-mir-219-1 was suggested to play a role in schizophrenia and in N-methyl-D-aspartate (NMDA) glutamate receptor signaling, two pathophysiological mechanisms linked to *HLA-DPB1* [52,53] making our results of valuable information for scientists interested in these pathologies.

Several limitations of this work must be acknowledged. First, because our investigation was conducted on genotyped data of common SNPs, only 258 miRNAs were covered by our study, which represent less than one-quarter of the hypothesized total number of human miRNAs. Second, only one cell type was studied where not all genes are expressed. Therefore not all possible association could be explored. Third, expression were measured using the microarray technology that may be less efficient than emerging mRNA deep-sequencing methods for measuring, especially low abundant, mRNA levels [54,55]. Because a given miRNA can bind several genes and a given 3'UTR can be a target for several miRNAs, compensation

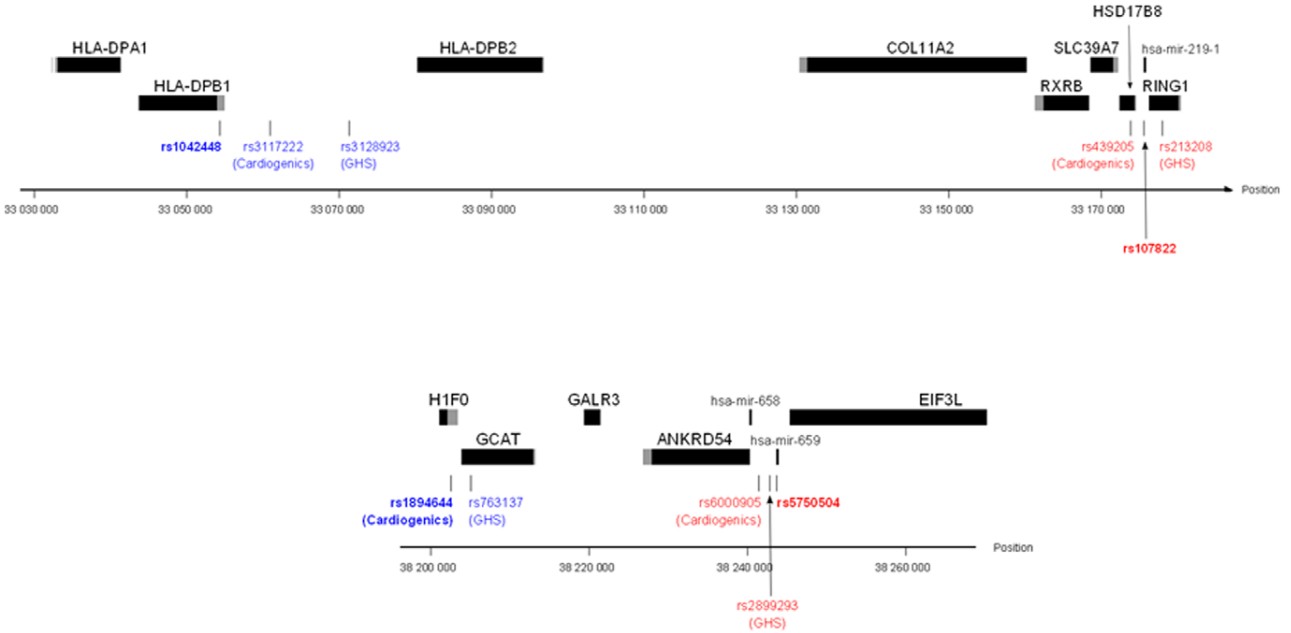


Figure 3. Location of genes, miSNP and 3'UTR SNPs at the two detected interacting loci. Gene are indicated as black rectangles with grey 3'UTR. Bold red and blue SNPs represent miSNPs and 3utrSNPs respectively. Corresponding proxies are non-bold coloured. Top: *HLA-DBP1* locus on chromosome 6; Bottom: *H1F0* locus on chromosome 22.
doi:10.1371/journal.pone.0045863.g003

phenomena are proposed to explain the relative low impact of miRNA regulation on mRNA expression generally observed [56]. Therefore, genetic effects associated with miRNA and 3'UTR SNPs are hypothesized to be a modest size and very large sample size would be required to detect them. Despite having robustly identified two interactions, we cannot then exclude that other interactions with lower magnitude could have been missed due to power considerations, even if the two genome-wide expression datasets used in this work are among the largest collected so far in human epidemiological studies. Third, by discarding from our investigations probes harboring a SNP in their genomic sequence to avoid any bias in the results of the association analyses, some miRNA-sensitive regulatory mechanisms associated to genes tagged by probes matching their 3'UTR region may have been missed. Last, our investigation was conducted in monocytes and results observed may not be portable to other cells or tissues.

Nevertheless, our study illustrates that the proposed strategy searching for interaction between miSNPs and 3'UTR SNPs in genome-wide expression studies could be an alternative to bioinformatics prediction tools to identify miRNA targeted genes.

Materials and Methods

Ethics Statement

This work was based on two genome-wide expression studies, the Gutenberg Health Study (GHS) for the discovery phase and the Cardiogenics Transcriptomic Study (CTS) for the replication stage. Both studies were approved by the Institutional Ethical Committee of each participating center and by the local and federal data safety commissioners (Ethik-Kommission der Landesärztekammer Rheinland-Pfalz) for GHS. These two studies have already been extensively described in [21–23] for GHS and in [26,57] for CTS.

Gutenberg Health Study

This analysis was conducted in a population-based sample of 750 men and 717 women aged 35–74 years, of European descent. Monocytic RNA was isolated from peripheral blood monocytes by negative selection using RosetteSep Monocyte Enrichment Cocktail (StemCell Technologies, Vancouver, Canada), Trizol extraction and purification by silica-based columns. Expression profiles were assessed using the *Illumina* HT-12 v3 BeadChip (Illumina, CA, USA) with ~48,000 probes covering 37,804 genes, and generated data were pre-processed using Beadstudio. Values from probes with $\leq$1 bead were re-imputed using SVD impute from the pcaMethods R package [58]. Data were normalized using quantile normalization and VST transformation as implemented in the lumi R package. To avoid spurious associations due to hybridation difference, probes that contained SNPs or were not annotated to be of "*perfect*" quality according to ReMOAT [28] (Reannotation and Mapping of Oligonucleotide Arrays Technologies, http://remoat.sysbiol.cam.ac.uk) were discarded. Individuals were typed for genome-wide genotype data using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, CA, USA). SNP analysis was restricted to autosomal SNPs with minor allele frequency >0.01, call rate >0.98 and Hardy-Weinberg equilibrium testing p-value $>10^{-4}$.

Cardiogenics Study

The present study included monocyte expression data from 758 individuals from European descent, 363 patients with coronary artery disease and 395 unrelated healthy individuals.

Monocyte RNAs were isolated from whole blood using CD14 micro beads (Miltenyi) and expression profile was processed in a single center using the *Illumina* HumanRef-8 v3 beadchip array (*Illumina* Inc., San Diego, CA) containing 24,516 probes corresponding to 18,311 distinct genes. After hybridization, array images were scanned using the *Illumina* BeadArray Reader and probe intensities were extracted using the Gene expression module (version 3.3.8) of the *Illumina BeadStudio* software (version 3.1.30). Raw intensities were processed in R statistical environment using the Lumi and beadarray packages. All array outliers were excluded and only arrays with high concordance in terms of gene expression measures (pairwise Spearman correlation coefficients within each cell type >0.85) were included in the analyses.

Genomic DNA was extracted from peripheral blood leucocytes by standard procedures (Qiagen). Genome-wide genotyping was carried out using one of two *Illumina* arrays; the Sentrix Human Custom 1.2 M array and the Human 610 Quad Custom array. Data from the two arrays was combined as described in [59]. SNP analysis was restricted to autosomal SNPs with minor allele frequency >0.01, call rate >0.95 and Hardy-Weinberg equilibrium testing p-value $>10^{-5}$.

Statistical analysis

The association of miSNP proxies with probe expression was tested by use of a standard linear regression model under the assumption of additive allele effects (i.e. proxy genotype coded as 0/1/2 according the number of rare alleles). Pair-wise SNPs interactions on probe expression were tested using a standard linear regression model in which both SNP (miSNP and 3utrSNP) genotypes were coded as 0,1,2 together with the corresponding product term for interaction. All analyses were adjusted for age and gender, and additionally for disease status in CTS.

In the Gutenberg Health Study, a weighted-Bonferroni procedure was applied to identify genome-wide significant interactions. Each 3utrSNP was first assessed using the Levene statistic [29] testing the equality of associated-probe expression variance across genotypes. The resulting log(p-value) was then used to weight the interaction p-value obtained from the linear regression analysis. This strategy is expected to be more powerful than a standard Bonferroni correction procedure [60,61] as it gives more weight to interaction involving probes showing higher differences in inter-genotype variance.

For each 3utrSNP u ($u = 1$ to N_{utr}) associated with a Levene test p-value q_u, we define a standardized weight w_u as

$$w_u = (N_{utr} x N_{miSNP}) \log(q_u) \Big/ \sum_{i=1}^{Nutr} N_{miSNP} \log(q_u)$$ such as $\sum_{i=1}^{N} w_i = N$ where N_{utr}, N_{miSNP}, N are the total number of studied 3utrSNPs, miSNPs and interactions, respectively. Each interaction p-value P_i is then weighted by the w_u corresponding to the 3utrSNP that is involved in the interaction, leading to a weighted p-value P_i*. Each Pi* that is then below 0.05/N is then declared genome-wide significant at the 0.05 type I error.

In Cardiogenics, the standard Bonferroni threshold was used to declare significance.

Identified interactions between pairs of SNPs were illustrated through haplotype analyses conducted by the THESIAS software implementing a Stochastic-EM algorithm for haplotype-based association analysis [62]. All other statistical analyses were performed in R v. 2.12.0.

Acknowledgments

Members of the Cardiogenics Consortium not included in the manuscript

Tony Attwood[1], Stephanie Belz[2], Peter Braund[3], Jessy Brocheton[4],Jason Cooper[5], Abi Crisp-Hihn[1], Patrick Diemert (formerly Linsel-Nitschke)[2], Nicola Foad[1], Tiphaine Godefroy[4],Jay Gracey[3], Emma Gray[6], Rhian Gwilliams[6], Susanne Heimerl[7], Jennifer Jolley[1], Unni Krishnan[3], Heather Lloyd-Jones[1], Ulrika Liljedahl[8],Ingrid Lugauer[7], Per Lundmark[8], Seraya Maouche[2,4],Jasbir S Moore[3], Gilles Montalescot[4], David Muir[1], Elizabeth Murray[1], Chris P Nelson[3], Jessica Neudert[9], David Niblett[6], Karen O'Leary[1], Helen Pollard[3], Carole Proust[4], Angela Rankin[1], Augusto Rendon[10], Catherine M Rice[6], Hendrik Sager[2], Jennifer Sambrook[1], Gerd Schmitz[11], Michael Scholz[9], Laura Schroeder[2], Jonathan Stephens[1], Ann-Christine Syvannen[8], Stefanie Tennstedt (formerlyGulde)[2],Chris Wallace[5].

[1]Department of Haematology, University of Cambridge, Long Road, Cambridge, CB2 2PT, UK and National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge, CB2 2PT, UK; [2]Medizinische Klinik 2, Universität zu Lübeck, Lübeck Germany; [3]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester, LE3 9QP, UK; [4] INSERM UMRS 937, Pierre and Marie Curie University (UPMC, Paris 6) and Medical School, 91 Bd de l'Hôpital 75013, Paris, France.

[5]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge, CB2 0XY, UK; [6] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; [7]Klinik und Poliklinik für Innere Medizin II, Universität Regensburg, Germany; [8]Molecular Medicine, Department of Medical Sciences, Uppsala University, Uppsala, Sweden;[9]Trium, Analysis Online GmbH, Hohenlindenerstr. 1, 81677, München, Germany;[10] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK; [10]Institut für KlinischeChemie und Laboratoriums medizin, Universität, Regensburg, D-93053 Regensburg, Germany.

Supporting Information

Table S1 Genome-wide significant ($p<7.7\ 10^{-9}$) associations of miSNPs on monocyte gene expression in the Gutenberg Health Study. (1) SNPs showing the strongest association (with $P<5\ 10^{-5}$) with gene expression within 1Mb of the associated probe. (2) Regression coefficient associated with the rare miSNP allele under an additive effect model, adjusted for age and gender. (3) P-value of the association between miSNP and gene expression. (4) P-value of the association between miSNP and gene expression adjusted for the best *cis* eSNP. (5) Pairwise r2 between miSNP and best *cis* eSNPs in GHS. (6) The best *cis* eSNP and the associated-miSNP coincide.
(XLSX)

Table S2 *Cis* and *trans*-associations observed with the hsa-mir-1279 rs1463335(1) separately in CAD patients and healthy subjects of the Cardiogenics Transcriptomic Study. (1) The rs1463335 was tagged by the rs998022 in CTS. The rs146335 is located on chromosome 12, at position 69,667,075. As a consequence, the association observed with LYZ and YEATS4 are considered as *cis*-associations, the remaining eight as trans-associations. (2) Regression coefficient associated with the rare miSNP allele under an additive effect model, adjusted for age and gender. (3) P-value of the association between miSNP and gene expression.
(DOCX)

Table S3 Patterns of detected miSNPs × 3utrSNPs interaction separately in CAD and healthy subjects of the Cardiogenics Transcriptomic Study. (1) Regression coefficient of the interaction term when both miSNP and 3utr proxy SNPs coded 0/1/2 according to the number of carried rare alleles are introduced in a linear regression model together with their interaction term. (2) P-value of the interaction test derived from the standard linear regression analysis in CTS. Bold p-values correspond to the detected interactions that were significant after Bonferroni correction in the whole CTS.
(DOCX)

Author Contributions

Conceived and designed the experiments: TZ PD JE CH WHO NJS HS TM KJL FC AHG LT SB. Performed the experiments: TZ. Analyzed the data: NG PSW MR AS AZ LT DAT. Contributed reagents/materials/analysis tools: MR AR AZ PD CH WHO NJS HS AHG SB. Wrote the paper: NG FC AHG LT DAT.

References

1. Carthew RW, Sontheimer EJ (2009) Origins and Mechanisms of miRNAs and siRNAs. Cell 136: 642–655.
2. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (2004) Identification of mammalian microRNA host genes and transcription units. Genome Res 14: 1902–1910.
3. Kim YK, Kim VN (2007) Processing of intronic microRNAs. Embo J 26: 775–783.
4. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36: D154–158.
5. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15–20.
6. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136: 215–233.
7. Krol J, Loedige I, Filipowicz W (2010) The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet 11: 597–610.
8. Lu M, Zhang Q, Deng M, Miao J, Guo Y, et al. (2008) An analysis of human microRNA and disease associations. PLoS One 3: e3420.
9. Mishra PK, Tyagi N, Kumar M, Tyagi SC (2009) MicroRNAs as a therapeutic target for cardiovascular diseases. J Cell Mol Med 13: 778–789.
10. Urbich C, Kuehbacher A, Dimmeler S (2008) Role of microRNAs in vascular diseases, inflammation, and angiogenesis. Cardiovasc Res 79: 581–588.
11. Fernandez-Hernando C, Suarez Y, Rayner KJ, Moore KJ (2011) MicroRNAs in lipid metabolism. Curr Opin Lipidol 22: 86–92.
12. Leeper NJ, Cooke JP (2011) MicroRNA and mechanisms of impaired angiogenesis in diabetes mellitus. Circulation 123: 236–238.
13. Slaby O, Bienertova-Vasku J, Svoboda M, Vyzula R (2011) Genetic polymorphisms and MicroRNAs: new direction in molecular epidemiology of solid cancer. J Cell Mol Med.
14. Hughes AE, Bradley DT, Campbell M, Lechner J, Dash DP, et al. (2011) Mutation Altering the miR-184 Seed Region Causes Familial Keratoconus with Cataract. Am J Hum Genet.
15. Tian T, Shu Y, Chen J, Hu Z, Xu L, et al. (2009) A functional genetic variant in microRNA-196a2 is associated with increased susceptibility of lung cancer in Chinese. Cancer Epidemiol Biomarkers Prev 18: 1183–1187.
16. Hu Z, Chen J, Tian T, Zhou X, Gu H, et al. (2008) Genetic variants of miRNA sequences and non-small cell lung cancer survival. J Clin Invest 118: 2600–2608.
17. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, et al. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18: 610–621.
18. Lin Z, Murtaza I, Wang K, Jiao J, Gao J, et al. (2009) miR-23a functions downstream of NFATc3 to regulate cardiac hypertrophy. Proc Natl Acad Sci U S A 106: 12103–12108.

19. Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, et al. (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet 38: 813–818.
20. Martin MM, Buckenberger JA, Jiang J, Malana GE, Nuovo GJ, et al. (2007) The human angiotensin II type 1 receptor +1166 A/C polymorphism attenuates microrna-155 binding. J Biol Chem 282: 24262–24269.
21. Castagne R, Zeller T, Rotival M, Szymczak S, Truong V, et al. (2011) Influence of sex and genetic variability on expression of X-linked genes in human monocytes. Genomics 98: 320–326.
22. Castagne R, Rotival M, Zeller T, Wild PS, Truong V, et al. (2011) The choice of the filtering method in microarrays affects the inference regarding dosage compensation of the active X-chromosome. PLoS One 6: e23956.
23. 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.
24. Rotival M, Zeller T, Wild P, Maouche S, Szymczak S, et al. (2011) Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet doi:10.1371/journal.pgen.1002367.
25. Weber C, Zernecke A, Libby P (2008) The multifaceted contributions of leukocyte subsets to atherosclerosis: lessons from mouse models. Nat Rev Immunol 8: 802–815.
26. 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.
27. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938–2939.
28. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JF, Ritchie ME, et al. (2010) A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res 38: e17.
29. Pare G, Cook NR, Ridker PM, Chasman DI (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS Genet 6: e1000981.
30. Barenboim M, Zoltick BJ, Guo Y, Weinberger DR (2010) MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. Hum Mutat 31: 1223–1232.
31. John B, Enright AJ, Aravin A, Tuschl T, Sander C, et al. (2004) Human MicroRNA targets. PLoS Biol 2: e363.
32. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, et al. (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 37: W273–276.
33. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, et al. (2005) Combinatorial microRNA target predictions. Nat Genet 37: 495–500.
34. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224.
35. 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.
36. 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.
37. 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.
38. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392–404.
39. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, et al. (2011) Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 89: 607–618.
40. Ellinghaus E, Ellinghaus D, Stuart PE, Nair RP, Debrus S, et al. (2010) Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. Nat Genet 42: 991–995.
41. Strange A, Capon F, Spencer CC, Knight J, Weale ME, et al. (2011) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet 42: 985–990.
42. Brown BD, Naldini L (2009) Exploiting and antagonizing microRNA regulation for therapeutic and experimental applications. Nat Rev Genet 10: 578–585.
43. Rayner KJ, Esau CC, Hussain FN, McDaniel AL, Marshall SM, et al. (2011) Inhibition of miR-33a/b in non-human primates raises plasma HDL and lowers VLDL triglycerides. Nature 478: 404–407.
44. Inaoka H, Fukuoka Y, Kohane IS (2007) Evidence of spatially bound gene regulation in Mus musculus: decreased gene expression proximal to microRNA genomic location. Proc Natl Acad Sci U S A 104: 5020–5025.
45. Inaoka H, Fukuoka Y, Kohane IS (2006) Lower expression of genes near microRNA in C. elegans germline. BMC Bioinformatics 7: 112.
46. Rademakers R, Eriksen JL, Baker M, Robinson T, Ahmed Z, et al. (2008) Common variation in the miR-659 binding-site of GRN is a major risk factor for TDP43-positive frontotemporal dementia. Hum Mol Genet 17: 3631–3642.
47. Fiesel FC, Voigt A, Weber SS, Van den Haute C, Waldenmaier A, et al. (2010) Knockdown of transactive response DNA-binding protein (TDP-43) downregulates histone deacetylase 6. Embo J 29: 209–221.
48. Fiesel FC, Schurr C, Weber SS, Kahle PJ (2011) TDP-43 knockdown impairs neurite outgrowth dependent on its target histone deacetylase 6. Mol Neurodegener 6: 64.
49. Kamatani Y, Wattanapokayakit S, Ochi H, Kawaguchi T, Takahashi A, et al. (2009) A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. Nat Genet 41: 591–595.
50. Kominami S, Tanabe N, Ota M, Naruse TK, Katsuyama Y, et al. (2009) HLA-DPB1 and NFKBIL1 may confer the susceptibility to chronic thromboembolic pulmonary hypertension in the absence of deep vein thrombosis. J Hum Genet 54: 108–114.
51. Zhou X, Lee JE, Arnett FC, Xiong M, Park MY, et al. (2009) HLA-DPB1 and DPB2 are genetic loci for systemic sclerosis: a genome-wide association study in Koreans with replication in North Americans. Arthritis Rheum 60: 3807–3814.
52. Verhelst H, Verloo P, Dhondt K, De Paepe B, Menten B, et al. (2011) Anti-NMDA-receptor encephalitis in a 3 year old patient with chromosome 6p21.32 microdeletion including the HLA cluster. Eur J Paediatr Neurol 15: 163–166.
53. Zamani MG, De Hert M, Spaepen M, Hermans M, Marynen P, et al. (1994) Study of the possible association of HLA class II, CD4, and CD3 polymorphisms with schizophrenia. Am J Med Genet 54: 372–377.
54. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321: 956–960.
55. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509–1517.
56. Huntzinger E, Izaurralde E (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet 12: 99–110.
57. Shah S, Nelson CP, Gaunt TR, van der Harst P, Barnes T, et al. (2011) Four Genetic Loci Influencing Electrocardiographic Indices of Left Ventricular Hypertrophy. Circ Cardiovasc Genet.
58. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J (2007) pcaMethods–a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23: 1164–1167.
59. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43: 333–338.
60. Benjamini Y, Hochberg Y (1997) Multiple hypotheses testing with weights. Scand J Stat 24: 407–418.
61. Dalmasso C, Genin E, Tregouet DA (2008) A weighted-Holm procedure accounting for allele frequencies in genomewide association studies. Genetics 180: 697–702.
62. Tregouet DA, Garelle V (2007) A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies. Bioinformatics 23: 1038–1039.

À toi qui m'as feuilleté jusqu'ici… et qui espérais que ce soit fini

Épilogue

Marin Shadok : *Quand on ne sait pas où l'on va, il faut y aller... et le plus vite possible.*

Les Shadoks
http://www.lesshadoks.com/

Vers une disponibilité des données génomiques à la communauté non scientifique

Au chapitre 3, j'ai expliqué comment les avancées technologiques ont pu fournir aux chercheurs les données leur permettant d'identifier certains polymorphismes de prédisposition aux maladies génétiques. Depuis la fin des années 2000, ces avancées permettent désormais également à la communauté non scientifique d'avoir accès à ce genre de données, pour des fins plus ou moins sérieuses.

Séquençage et génotypage personnalisé

Si la communauté scientifique séquence désormais régulièrement des individus afin notamment de rechercher des variants rares pouvant expliquer la survenue de certaines maladies, jusqu'à maintenant, très peu de personnes se sont personnellement faites séquencées. Le registre mondial des génomes personnels en reporte 56 au moment où j'écris ce document dont, hormis les pionniers Craig Venter ou James Watson, quelques célébrités non scientifiques comme Glenn Close, Desmond Tutu ou Henry Louis Gates [152]. En revanche, de plus en plus d'entreprises proposent le génotypage personnel par envoi de Kit de récupération de salive. La plupart de ces sociétés fournissent en même

temps une interprétation des données afin d'informer les individus sur leur généalogie ou certains de leurs risques médicaux. Par exemple, 23andMe, la compagnie leader dans la génomique personnelle aurait déjà génotypé plus de 180 000 personnes, son offre consistant actuellement à un génotypage de plus de 900 000 variants (et leur interprétation) par la puce à ADN Illumina HumanOmniExpress pour 299 €[141]. Certains outils comme Promethease, associé à SNPedia, permettent également aux personnes ayant déjà leurs données, de les interpréter gratuitement[159].

Apparition de nombreux produits dérivés

Surfant sur cet engouement grandissant pour la génomique personnelle, certaines entreprises offrent aussi des services moins scientifiques tels que GenePartner proposant une aide à la recherche du partenaire génotypiquement idéal[147], Warrior Roots qui propose à chacun de découvrir ses ancêtres guerriers et son potentiel athlétique[162], Your DNA Song ou l'application pour iPhone GeneGroove qui créent une musique personnelle à partir d'ADN[146, 163] ou DNA 11 qui propose à ses clients des tableaux artistiques personnalisés à partir leur propre ADN[144](cf. figure 9.1).

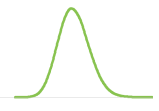
FIGURE 9.1 – L'ADN peut aussi être utilisé pour faire de l'art (en haut à gauche), faire de bonnes rencontres (en haut à droite), connaître ses ancêtres guerriers (en bas à gauche) ou faire de la musique sur iPhone (en bas à droite).

Des données qui deviennent publiques

Il semble aussi que nous nous dirigions peu à peu vers une diffusion des données de génotypage et de séquençage à la communauté scientifique voire le grand public, en témoignent les nombreuses initiatives qui promeuvent l'open data pour ce genre de données. Le Projet Génome Personnel est une longue et large étude dont le but est de séquencer puis de rendre publiques les séquences et informations médicales de 100 000 volontaires qui auront auparavant passé un test permettant de vérifier leurs connaissances génétiques et leur conscience des risques engendrés en rendant ce genre de données disponibles sur internet [156, 158]. Au moment de l'écriture de la thèse l'étude est composée de 2 140 individus, anonymes pour la plupart. Les données génotypiques de 278 d'entre eux sont déjà rendues publiques, tout comme les séquences génétiques complètes de 37 personnes [157]. Consent to Research est un autre projet dont le but est de collecter et rendre publiques des données de volontaires en s'assurant que ceux-ci aient préalablement consenti à les fournir malgré les risques encourus. L'objectif du projet est alors d'emmener les données à ne pas être dédiées à une étude en particulier, mais rendues disponibles à la communauté scientifique[143]. Le projet genomes unzipped, enfin, est un projet mené par 12 personnes dont le but est de tester les risques et bénéfices de l'information génétique en mettant à disposition du public leurs données génotypiques[148].

... peut-être parfois trop ?

OpenSNP est une initiative en léger décalage avec les précédentes. Ce site internet propose aux gens qui le souhaitent de rendre publiques leurs données génotypiques, tout en renseignant librement d'autres informations comme phénotypes. Au contraire des deux premiers projets, ce site internet ne s'assure pas particulièrement que les individus qui soumettent leurs données soient conscients des risques qu'ils prennent. Il permet aux inscrits de créer des phénotypes, de mettre leur nom et leur photo ou de renseigner leurs liens de réseaux sociaux[154]. Il est même possible d'y intégrer les données générées par les produits Fitbit, qui enregistrent par exemple des données sur la qualité du sommeil de l'individu portant l'instrument. Tout cela sous licence creative common zero, licence sans restriction d'utilisation ou de distribution. Au moment de l'écriture de document, 282 individus y avaient mis leurs données de génotype à disposition du public.



Réfléxions

Ces évolutions peuvent pousser certains à se poser des questions. Compte tenu des avancées technologiques prodigieuses en matière de séquençage (cf : figure 9.2) et en extrapolant les avancées de nos connaissances sur le génome humain, il n'est pas inconcevable que dans quelques dizaines d'années, nous soyons en mesure d'avoir des informations sur une personne (mais aussi sur sa famille et sa descendance) que ces derniers préféreraient garder secret. Peut-être cependant que nous ne seront pas capables de tirer beaucoup plus d'informations des données de génotype que nous le pouvons actuellement, que les risques de dérives ne sont pas si importants que certains peuvent le penser et que la société s'adaptera à la diffusion publique de ces nouvelles informations très personnelles. L'avenir nous le dira...

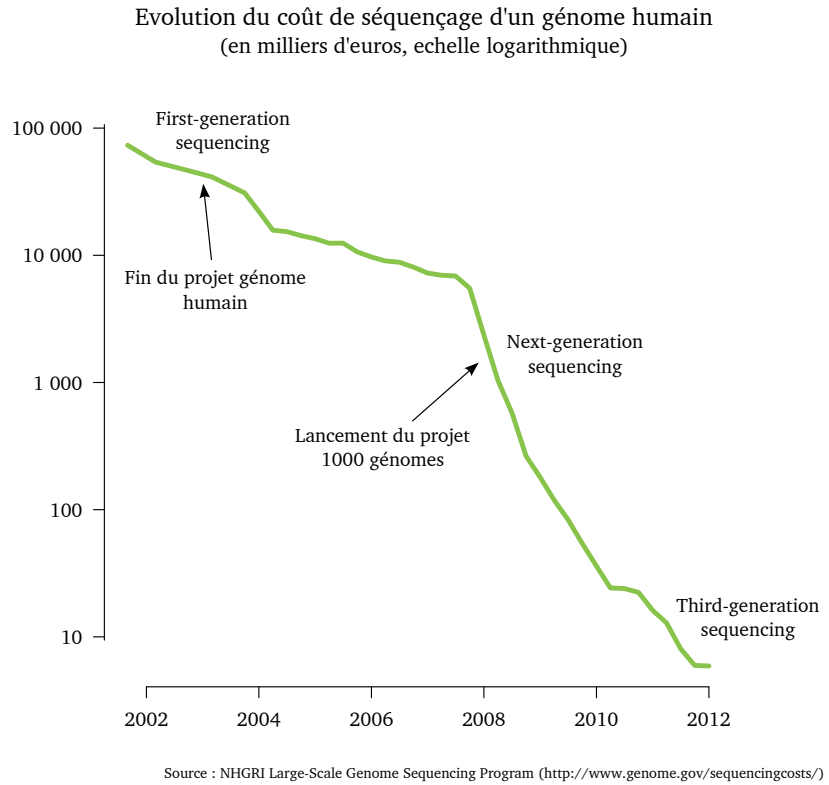


FIGURE 9.2 – Évolution du coût du séquençage humain. L'échelle de coût y est logarithmique.

Title

Research strategies for finding genetic interaction phenomena in multifactorial diseases

Abstract

Recently, Genome-Wide Association Studies (GWAS) have led to the discovery of numerous genetic polymorphisms involved in complex human diseases. However, these polymorphisms contribute only a little to the overall genetic variability of these diseases, suggesting the need for new kind of investigations in order to disentangle the so-called "missing heritability".

The purpose of my PhD project was to investigate how different research strategies relying on statistical and biological considerations could help in determining whether part of this missing heritability could reside in interaction phenomena between genetic polymorphisms.

Firstly, we applied different statistical methodologies and looked for interactions between polymorphisms that could influence the risk of venous thrombosis (VT). Even though this study was based on two large GWAS datasets, we were not able to identify pairwise interactions that survive multiple testing. This work suggests that strong interactive phenomena between common SNPs are unlikely to contribute much to the risk of VT.

Second, by adopting a hypothesis-driven approach relying on biological arguments, we sought for interactions between microRNA related polymorphisms that could alter genetic expression. Using two large GWAS datasets in which genome-wide monocyte expression was also available, we were able to demonstrate the existence of two pairwise interaction phenomena on monocyte expression involving miRNAs polymorphisms: 1/ the expression of HLA-DPB1 was modulated by a polymorphism in its 3'UTR region with a polymorphism in the hsa-mir-219-1 microRNA sequence; 2/ similarly, the expression of H1F0 was influenced by a polymorphism in its 3'UTR region interacting with a polymorphism in the microRNA hsa-mir-659.

Altogether, this project supports for the role of gene x gene interactions in the interindividual variability of biological processes but their identifications remain a tedious task requiring large samples and the development of new research strategies and methodologies.

Keywords

interaction, microRNA, venous thrombosis, monocyte, genetics, GWAS, statistics, power, multiple testing, weighting, heritability, genetics, SNP, complex diseases

Stratégies de recherches de phénomènes d'interactions dans les maladies multifactorielles

Les études d'associations en génome entier ("GWAS") ont récemment permis la découverte de nombreux polymorphismes génétiques impliqués dans la susceptibilité aux maladies multifactorielles. Cependant, ces polymorphismes n'expliquent qu'une faible part de l'héritabilité génétique de ces maladies, nous poussant ainsi à explorer de nouvelles pistes de recherche.

Une des hypothèses envisagées serait qu'une partie de cette héritabilité manquante fasse intervenir des phénomènes d'interactions entre polymorphismes génétiques. L'objectif de cette thèse est d'explorer cette hypothèse en adoptant une stratégie de recherche d'interactions basée sur des critères statistiques et biologiques à partir de données issues de différentes études "GWAS".

Ainsi, en utilisant différentes méthodes statistiques, nous avons commencé par rechercher des interactions entre polymorphismes qui pourraient influencer le risque de thrombose veineuse. Cette recherche n'a malheureusement pas abouti à l'identification de résultats robustes vis à vis du problème des tests multiples.

Dans un deuxième temps, à partir d'hypothèses "plus biologiques", nous avons tenté de mettre en évidence des interactions entre polymorphismes impliqués dans les mécanismes de régulation de l'expression génique associés aux microARNs. Nous avons pu ainsi montrer de manière robuste dans deux populations indépendantes qu'un polymorphisme au sein de la séquence du microARN hsa-mir-219-1 interagissait avec un polymorphisme du gène HLA-DPB1 pour en moduler l'expression monocytaire. Nous avons également montré que l'expression monocytaire du gène H1F0 était influencée par un phénomène d'interaction impliquant un polymorphisme du microARN hsa-mir-659.

En apportant sa propre contribution à l'engouement récent que suscite la recherche d'interactions entre polymorphismes dans les maladies dites complexes, ce travail de thèse illustre clairement la difficulté d'une telle tâche et l'importance de réfléchir à de nouvelles stratégies de recherches.