UNIVERSITE DE NANTES

FACULTE DE MEDECINE

2012 N°36

ANALYSE D'ASSOCIATION GENOME ENTIER DE 3 PATHOLOGIES - LE DIABETE DE TYPE 2, LE SYNDROME DE BRUGADA ET LE PROLAPSUS VALVULAIRE MITRAL: OBSERVATIONS SUR L'ARCHITECTURE GENETIQUE DE TRAITS COMPLEXES

THESE DE DOCTORAT

ECOLE DOCTORALE : BIOLOGIE-SANTE

DISCIPLINE : SCIENCES DE LA VIE ET DE LA SANTE

SPECIALITE : ÉPIDEMIOLOGIE GENETIQUE

Présentée et soutenue publiquement par

CHRISTIAN DINA

Le 22 Octobre 2012, devant le jury ci-dessous

Rapporteurs	**Mme Catherine ANDRE, Chargé de Recherche, CNRS-UMR6061, Rennes**
	Mme Emmanuelle GENIN, Directeur de Recherches, INSERM U1078, BREST
Examinateurs	**M. Philippe MABO, Professeur PU-PH, CHU de Rennes, Rennes**
	M. Ghislain ROCHELEAU, Maître de Conférences, UNIVERSITE LILLE 2, LILLE
Directeur de thèse	**M. Jean-Jacques SCHOTT, Directeur de Recherches, Inserm UMR1087, Nantes**

Le contenu de cette thèse est confidentiel. [16 mois – limite : 28 mars 2015]

Remerciements

Je voudrais tout d'abord et spécifiquement remercier mon directeur de thèse Jean-Jacques Schott. Sa patience et son calme face à un doctorant atypique m'ont été d'une aide immense.

Je voudrais remercier le Professeur Le Marec, directeur de l'unité depuis 2012, pour sa confiance à mon égard et les encouragements qu'il m'a prodigué aussi bien dans les projets de recherche dans le laboratoire que pour la réalisation de mon travail de thèse.

Je voudrais également remercier le Pr. Pierre Pacaud, directeur de l'unité, qui m'a accueilli dans son laboratoire, à l'époque où notre nom était Inserm U915 ainsi que Flavien Charpentier qui dirigeait l'équipe CNRS par laquelle je suis entré dans l'institut du thorax et qui m'a accueilli dans son équipe.

Je souhaiterais également remercier le Docteur Emmanuelle Génin et le Docteur Catherine André d'avoir accepté de juger mon travail ainsi que le Professeur Philippe Mabo et le Docteur Ghislain Rocheleau pour avoir consacré du temps à l'examen de ma thèse.

Je veux également remercier le groupe « clinique » de l'équipe sans qui nos recherches ne pourraient exister. Merci aux Professeurs Vincent Probst et Thierry Le Tourneau, merci à toutes les infirmières de recherche clinique qui ont fourni beaucoup d'énergie dans la constitution de cohorte de patients et dans le recrutement des familles.

Je pense aussi à l'équipe du Pr. Philippe Froguel, à Lille, au sein de laquelle j'ai initié une partie des travaux réalisés pour le travail présenté dans ce document.

Je voudrais remercier tous mes collègues des équipes de génétique et la plateforme de génomique de l'institut du Thorax et particulièrement l'équipe d'analyse qui a aidé puissamment à la bonne réalisation de ce travail de thèse.

Merci à mes parents qui m'ont encouragé sans se lasser afin que je réalise ce projet, à ma famille qui m'a régulièrement encouragé, sans se lasser, ces dernières années.
Merci à Anne, mon épouse, pour la patience dont elle a fait preuve, particulièrement ces dernières semaines, et accessoirement les quelques années précédentes. Merci à Jules et à Victoire pour leur patience, surtout ces derniers week-ends, qui ont été particulièrement difficiles.

Table des Matières:

A. Table des Publications

A.I. Articles reliés au travail de la thèse

the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, Strawbridge RJ, Khan H, Grallert H, Mahajan A, Prokopenko I, Kang HM, **Dina C**, Esko T, Fraser RM, Kanoni S, Kumar A, Lagou V, Langenberg C, Luan J, Lindgren CM, Müller-Nurasyid M, Pechlivanis S, Rayner NW, Scott LJ, Wiltshire S, Yengo L, Kinnunen L, Rossin EJ, Raychaudhuri S, Johnson AD, Dimas AS, Loos RJ, Vedantam S, Chen H, Florez JC, Fox C, Liu CT, Rybin D, Couper DJ, Kao WH, Li M, Cornelis MC, Kraft P, Sun Q, van Dam RM, Stringham HM, Chines PS, Fischer K, Fontanillas P, Holmen OL, Hunt SE, Jackson AU, Kong A, Lawrence R, Meyer J, Perry JR, Platou CG, Potter S, Rehnberg E, Robertson N, Sivapalaratnam S, Stančáková A, Stirrups K, Thorleifsson G, Tikkanen E, Wood AR, Almgren P, Atalay M, Benediktsson R, Bonnycastle LL, Burtt N, Carey J, Charpentier G, Crenshaw AT, Doney AS, Dorkhan M, Edkins S, Emilsson V, Eury E, Forsen T, Gertow K, Gigante B, Grant GB, Groves CJ, Guiducci C, Herder C, Hreidarsson AB, Hui J, James A, Jonsson A, Rathmann W, Klopp N, Kravic J, Krjutškov K, Langford C, Leander K, Lindholm E, Lobbens S, Männistö S, Mirza G, Mühleisen TW, Musk B, Parkin M, Rallidis L, Saramies J, Sennblad B, Shah S, Sigurðsson G, Silveira A, Steinbach G, Thorand B, Trakalo J, Veglia F, Wennauer R, Winckler W, Zabaneh D, Campbell H, van Duijn C, Uitterlinden AG, Hofman A, Sijbrands E, Abecasis GR, Owen KR, Zeggini E, Trip MD, Forouhi NG, Syvänen AC, Eriksson JG, Peltonen L, Nöthen MM, Balkau B, Palmer CN, Lyssenko V, Tuomi T, Isomaa B, Hunter DJ, Qi L; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium, Shuldiner AR, Roden M, Barroso I, Wilsgaard T, Beilby J, Hovingh K, Price JF, Wilson JF, Rauramaa R, Lakka TA, Lind L, Dedoussis G, Njølstad I, Pedersen NL, Khaw KT, Wareham NJ, Keinanen Kiukaanniemi SM, Saaristo TE, Korpi-Hyövälti E, Saltevo J, Laakso M, Kuusisto J, Metspalu A, Collins FS, Mohlke KL, Bergman RN, Tuomilehto J, Boehm BO, Gieger C, Hveem K, Cauchi S, Froguel P, Baldassarre D, Tremoli E, Humphries SE, Saleheen D, Danesh J, Ingelsson E, Ripatti S, Salomaa V, Erbel R, Jöckel KH, Moebus S, Peters A, Illig T, de Faire U, Hamsten A, Morris AD, Donnelly PJ, Frayling TM, Hattersley AT, Boerwinkle E, Melander O, Kathiresan S, Nilsson PM, Deloukas P, Thorsteinsdottir U, Groop LC, Stefansson K, Hu F, Pankow JS, Dupuis J, Meigs JB, Altshuler D, Boehnke M, McCarthy MI. **Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes**. Nat Genet. 2012 Aug 12;44(9):981-990.

Dina C ; Of 508 mice and 40,000 humans. J Mol Cell Cardiol. 2011 Mar;50(3):377-9. Epub 2010 Dec 16.

Voight BF*, Scott LJ*, Steinthorsdottir V*, Morris AP*, **Dina C***, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Boström K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jørgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proença C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparsø T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI; MAGIC investigators; GIANT Consortium. **Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis**. Nat Genet. 2010 Jul;42(7):579-89.

Bezzina CR, Barc J, Mizusawa Y, Remme CA, Gourraud JB, Simonet F, Verkerk AO, Schwartz PJ, Crotti L, Dagradi F, Guicheney P, Fressart V, Leenhardt A, Antzelevitch C, Bartkowiak S, Borggrefe M, Schimpf R, Schulze-Bahr E, Zumhagen S, Behr ER, Bastiaenen R, Tfelt-Hansen J, Olesen MS, Kääb S, Beckmann BM, Weeke P, Watanabe H, Endo N, Minamino T, Horie M, Ohno S, Hasegawa K, Makita N, Nogami A, Shimizu W, Aiba T, Froguel P, Balkau B, Lantieri O, Torchio M, Wiese C, Weber D, Wolswinkel R, Coronel R, Boukens BJ, Bézieau S, Charpentier E, Chatel S, Despres A, Gros F, Kyndt F, Lecointe S, Lindenbaum P, Portero V, Violleau J, Gessler M, Tan HL, Roden DM, Christoffels VM, Le Marec H, Wilde AA, Probst V, Schott JJ, **Dina C**, Redon R. **Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death**. Nat Genet. 2013 Sep;45(9):1044-9. doi: 10.1038/ng.2712. Epub 2013 Jul 21.

A.II. Articles reliés au travail de la thèse mais non-présentés

Bonnefond A, Clément N, Fawcett K, Yengo L, Vaillant E, Guillaume JL, Dechaume A, Payne F, Roussel R, Czernichow S, Hercberg S, Hadjadj S, Balkau B, Marre M, Lantieri O, Langenberg C, Bouatia-Naji N; Meta-Analysis of Glucose and Insulin Related Traits Consortium (MAGIC), Charpentier G, Vaxillaire M, Rocheleau G, Wareham NJ, Sladek R, McCarthy MI, **Dina C**, Barroso I, Jockers R, Froguel P. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat Genet. 2012 Jan 29; 44(3):297-301.

Baruteau AE, Behaghel A, Fouchard S, Mabo P, Schott JJ, **Dina C**, Chatel S, Villain E, Thambo JB, Marçon F, Gournay V, Rouault F, Chantepie A, Guillaumont S, Godart F, Martins RP, Delasalle B, Bonnet C, Fraisse A, Schleich JM, Lusson JR, Dulac Y, Daubert JC, Le Marec H, Probst V. **Parental Electrocardiographic Screening Identifies a High Degree of Inheritance for Congenital and Childhood Non-Immune Isolated Atrio-Ventricular Block**. Circulation. 2012 Aug 16.

Laurent G, Saal S, Amarouch MY, Béziau DM, Marsman RF, Faivre L, Barc J, **Dina C**, Bertaux G, Barthez O, Thauvin-Robinet C, Charron P, Fressart V, Maltret A,Villain E, Baron E, Mérot J, Turpault R, Coudière Y, Charpentier F, Schott JJ, Loussouarn G, Wilde AA, Wolf JE, Baró I, Kyndt F, Probst V. **Multifocal ectopic Purkinje-related premature contractions: a new SCN5A-related cardiac channelopathy**. J Am Coll Cardiol. 2012 Jul 10;60(2):144-56.

A.III. Autres articles

Limou S, Coulonges C, Herbeck JT, van Manen D, An P, Le Clerc S, Delaneau O, Diop G, Taing L, Montes M, van't Wout AB, Gottlieb GS, Therwath A, Rouzioux C, Delfraissy JF, Lelièvre JD, Lévy Y, Hercberg S, **Dina C**, Phair J, Donfield S, Goedert JJ, Buchbinder S, Estaquier J, Schächter F, Gut I, Froguel P, Mullins JI, Schuitemaker H, Winkler C, Zagury JF. **Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS**. J Infect Dis. 2010 Sep 15;202(6):908-15

Isidor B, Lindenbaum P, Pichon O, Bézieau S, **Dina C**, Jacquemont S, Martin-Coignard D, Thauvin-Robinet C, Le Merrer M, Mandel JL, David A, Faivre L, Cormier-Daire V, Redon R, Le Caignec C. **Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis.** Nat Genet. 2011 Mar 6;43(4):306-8.

Postel-Vinay S, Véron AS, Tirode F, Pierron G, Reynaud S, Kovar H, Oberlin O, Lapouble E, Ballet S, Lucchesi C, Kontny U, González-Neira A, Picci P, Alonso J, Patino-Garcia A, de Paillerets BB, Laud K, **Dina C**, Froguel P, Clavel-Chapelon F, Doz F, Michon J, Chanock SJ, Thomas G, Cox DG, Delattre O. **Common variants near TARDBP and EGR2 are associated with susceptibility to Ewing sarcoma.** Nat Genet. 2012 Feb 12;44(3):323-7.

Germain M, Saut N, Greliche N, **Dina C**, Lambert JC, Perret C, Cohen W, Oudot-Mellakh T, Antoni G, Alessi MC, Zelenika D, Cambien F, Tiret L, Bertrand M, Dupuy AM, Letenneur L, Lathrop M, Emmerich J, Amouyel P, Trégouët DA, Morange PE. **Genetics of venous thrombosis: insights from a new genome wide association study**. PLoS One. 2011;6(9):e25581.

Bouatia-Naji N, Bonnefond A, Baerenwald DA, Marchand M, Bugliani M, Marchetti P, Pattou F, Printz RL, Flemming BP, Umunakwe OC, Conley NL, Vaxillaire M, Lantieri O, Balkau B, Marre M, Lévy-Marchal C, Elliott P, Jarvelin MR, Meyre D, **Dina C**, Oeser JK, Froguel P, O'Brien RM. **Genetic and functional assessment of the role of the rs13431652-A and rs573225-A alleles in the G6PC2 promoter that are strongly associated with elevated fasting glucose levels.** Diabetes. 2010 Oct;59(10):2662-71.

Rung J, Cauchi S, Albrechtsen A, Shen L, Rocheleau G, Cavalcanti-Proença C, Bacot F, Balkau B, Belisle A, Borch-Johnsen K, Charpentier G, **Dina C**, Durand E, Elliott P, Hadjadj S, Järvelin MR, Laitinen J, Lauritzen T, Marre M, Mazur A, Meyre D, Montpetit A, Pisinger C, Posner B, Poulsen P, Pouta A, Prentki M, Ribel-Madsen R, Ruokonen A, Sandbaek A, Serre D, Tichet J, Vaxillaire M, Wojtaszewski JF, Vaag A, Hansen T, Polychronakos C, Pedersen O, Froguel P, Sladek R. **Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia.** Nat Genet. 2009 Oct;41(10):1110-5. Epub 2009 Sep 6. Erratum in: Nat Genet. 2009 Oct;41(10):1156.

B. Liste des abréviations

ADN	**A**cide **D**ésoxyribo**n**ucléique
ARN	**A**cide **Ri**bo**n**ucléique
ARNm	**ARN m**essager
bpm	**B**attement **p**ar **m**inute
cM	**c**enti-**M**organ
CNV	**C**opy **N**umber **V**ariation
CR	**C**all **R**ate (pourcentage de succès)
DFE	**D**égénérescence **F**ibro-**E**lastique
DL	**D**éséquilibre de **L**iaison
ECG	**E**lectro**c**ardio**g**ramme
FA	**F**ibrillation **A**uriculaire
FDR	**F**alse **D**iscovery **R**ate
FV	**F**ibrillation **V**entriculaire
FVI	**F**ibrillation **V**entriculaire **I**diopathique
FWER	**F**amily-**W**ise **E**rror **R**ate
Gb	**G**iga**b**ase
GWAS	**G**enome **W**ide **A**ssociation **S**tudy
LD	**L**inkage **D**isequilibrium
Mb	**M**éga**b**ase
ms	**mi**lli**s**econde
MS	**M**ort **S**ubite
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
PA	**P**otentiel d'**A**ction
pb	**p**aire de **b**ase
PVM	**P**rolapsus **V**alvulaire **M**itral
QTc	**QT c**orrigé
RFLP	**R**estriction **F**ragment **L**ength **P**olymorphism
RP	**R**epolarisation **P**récoce
RS	**R**éticulum **S**arcoplasmique
SBr	**S**yndrome de **Br**ugada
SNP	**S**ingle **N**ucleotide **P**olymorphism
TdC	**T**rouble **d**e **C**onduction
TV	**T**achycardie **V**entriculaire
UCSC	**U**niversity of **C**alifornia **S**anta **C**ruz

C. Introduction générale

Les pays industrialisés sont confrontés à de multiples défis afin de maintenir la santé et la qualité de vie de leurs habitants. L'explosion des maladies chroniques touchant les personnes de plus de 60 ans entraîne une augmentation importante des coûts, ce qui pèse dangereusement sur notre capacité à préserver un système de santé viable et efficace pour le plus grand nombre.

Dans un contexte de vieillissement de la population, la prise en charge des maladies fréquentes, qu'elles conduisent à une nécessité de prise en charge lourde ou qu'elles se traduisent par une espérance de survie très faible (en particulier les maladies métaboliques, respiratoires et cardiovasculaires) représente aujourd'hui un enjeu majeur de Santé Publique.

L'objectif est, à terme, de réduire leurs impacts, tant sur la qualité de vie des patients que sur le plan économique.

La transmission des caractères et des maladies dans les familles est une observation courante. Cette ressemblance entre ascendants et descendants, et plus généralement, entre individus apparentés, est portée en partie par un environnement partagé et en partie par les gènes.
La mise en évidence des bases génétiques de certaines maladies humaines, qui apparaît maintenant comme une évidence, est le résultat d'une mise en place conceptuelle longue et fastidieuse, au croisement des approches évolutionnistes, biochimiques et agronomiques.
La stratégie de génétique inverse a été utilisée depuis longtemps pour mettre en évidence les défauts génétiques responsables d'une pathologie, bien sûr, mais encore et surtout de la protéine (dans la majorité des cas), codée par ce gène, qui était identifiée comme importante dans l'étiologie de la pathologie étudiée. Par ailleurs, et de façon aussi importante, à travers la protéine, c'est également la voie biosynthétique dont le rôle dans la survenue de la pathologie étudiée est mis en évidence.

Dans mon travail de thèse, j'ai participé à la recherche des bases moléculaires de trois pathologies, une maladie rare et deux maladies fréquentes, que sont le syndrome de Brugada, le Diabète de Type 2 et le Prolapsus Valvulaire Mitral. Le choix de ces pathologies est en partie dû à mon parcours professionnel mais il représente un panorama très intéressant dans notre connaissance des bases moléculaires des pathologies complexes. Le Diabète de Type 2 est une pathologie fréquente qui a été extrêmement étudiée du point de vue génétique, sur de très grands échantillons, et dont l'architecture génétique est en train d'être mieux comprise. Le Syndrome de Brugada est rare alors que le Prolapsus Valvulaire Mitral est plus fréquent dans la population générale. Pour ces deux pathologies, les études génétiques recherchant des variants génétique fréquents se trouvent à un stade moins avancé. C'est une grande opportunité d'avoir pu travailler sur ces trois maladies simultanément.

D. Épidémiologie Génétique

D.I. Génétique et Statistique

La statistique génétique est la science qui permet d'appliquer des modèles mathématiques aux données génétiques afin de proposer des hypothèses de mécanismes de l'hérédité.

Historiquement, les développements de la génétique et de la statistique sont intimement liés. Dans une expérience devenue classique puisque Francis Galton put à la fois mettre en évidence le lien entre les tailles de deux individus et leur apparentement (génétique) et, simultanément, proposer le nom de régression en statistique (Figure 1, ci-dessous). C'est une des expériences les plus connues de mise en évidence de l'héritabilité [Galton,**1886**].

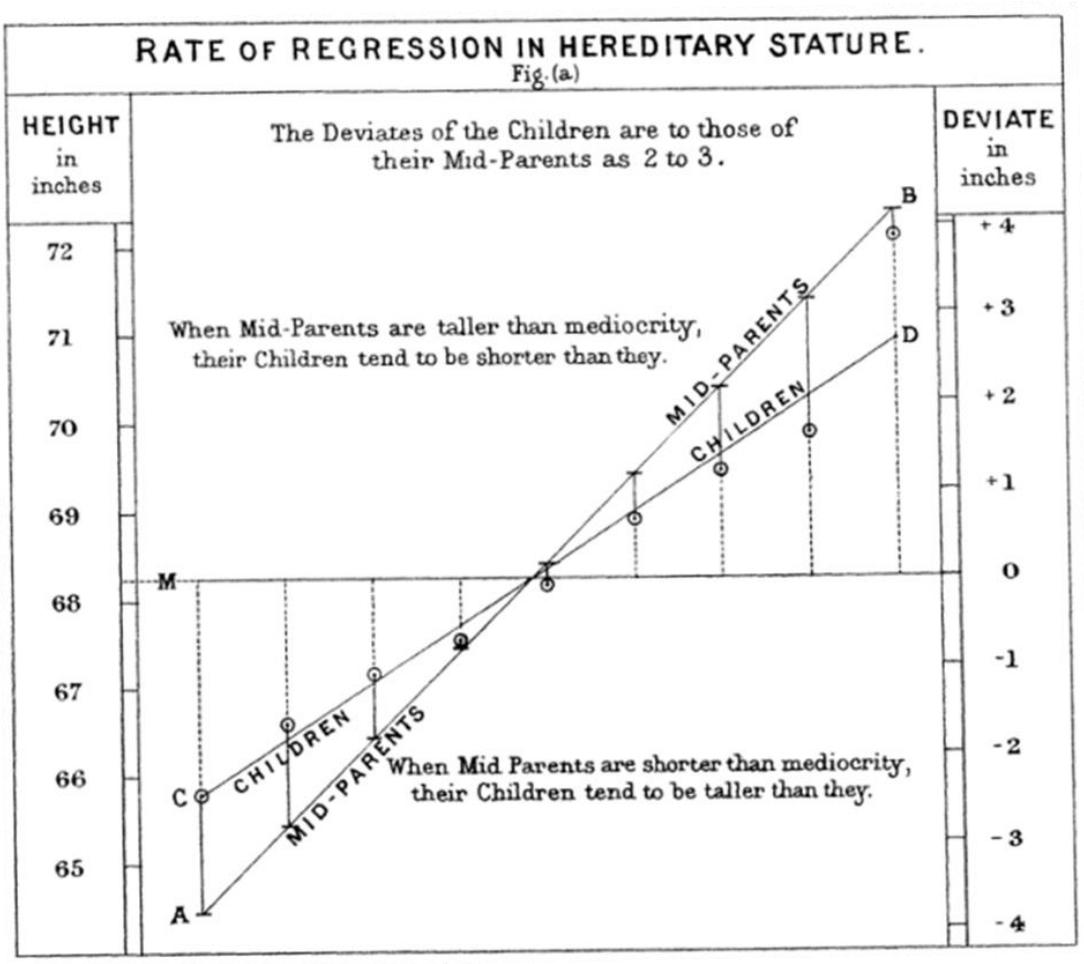


Figure 1 : Expérience de Galton

Galton analysa la relation entre la taille de 954 adultes et celle de leurs parents. Il utilisa la moyenne entre parents et établit qu'il existait une corrélation entre ces deux valeurs. Cette corrélation est liée à l'héritabilité. Il remarque par ailleurs que les enfants de parents présentant une taille extrême ont une tendance à se rapprocher de la moyenne de la population, phénomène qu'il appelle régression et qui, sorti de son contexte, donnera son nom à l'analyse. Plaque IX, sorti de [Galton,**1886**].

Cette expérience conclut à un lien continu entre les valeurs phénotypiques d'individus apparentés.

La première manifestation d'une approche, étonnement moderne, utilisant des méthodes quantitatives pour mettre en évidence une transmission héréditaire d'une pathologie apparaît sous la plume de Pierre Louis Moreau de Maupertuis (1698-1759).

Dans le *Système de la nature,* (1751), ce philosophe, mathématicien, physicien, astronome et naturaliste français de l'époque des Lumières s'appuie sur une étude minutieuse des cas de polydactylie sur plusieurs générations d'une famille berlinoise. Il établit que cette anomalie se transmet aussi bien par le père que par la mère et peut être expliquée par une mutation du patrimoine héréditaire et calcule la probabilité de la retrouver dans le futur chez d'autres membres de la famille. Il a ainsi proposé de tester la probabilité d'observation de la concentration familiale dans le cas où la survenue de cette caractéristique serait « accidentelle », ou, autrement dit « ***il faut voir quelle est la probabilité de cette variété qui est accidentelle dans un premier parent ne se répétera dans ses descendants*** ». La détermination de la prévalence se fait par une recherche systématique du nombre de polydactylies dans la ville de Berlin : ***«..dans une ville de 100 000 habitants j'ai trouvé deux hommes qui avoient cette singularité..*** » et tente d'estimer – de façon empirique - l'imprécision de sa mesure « ***.. et que trois autres m'aient échappé*** ».
Il met en évidence une concentration anormale du nombre de descendants atteints de cette « ***variété*** » puisque si « ***sur 20 000 hommes on puisse compter 1 sexdigitaire : la probabilité que son fils ou sa fille ne naisse pas avec le sexdigitisme est de 20 000 à 1*** » et que « ***la probabilité que cette singularité ne se continueroit pas pendant trois générations consécutives feroit 8 000 000 000 000 à 1*** » (20000^3).
Il remarque le « ***Nombre de fois qui est si grand que la certitude des choses les mieux démontrées en Physique n'approche pas de ces probabilités***» ou, autrement dit, que cette concentration familiale est statistiquement significative et permet de conclure à une distribution non-aléatoire, compatible avec des facteurs « familiaux ». Nous pensons tout de suite à une transmission génétique, même si cette concentration familiale pourrait aussi être la conséquence d'un environnement partagé.

La statistique génétique est la modélisation des transmissions des variations génétiques et le test de différentes hypothèses.

D.II. Génétique et Epidémiologie

L'Epidémiologie est la science qui s'intéresse à la répartition de la fréquence des maladies en fonction de facteurs risques. Cette science peut-être descriptive mais aussi, et principalement, s'appliquer à la recherche de facteurs causant la survenue de ces pathologies [Jean Bouyer, Denis Hémon, Sylvaine Cordier et al.].
L'Epidémiologie Génétique est une discipline qui va étudier la distribution de maladies (ici humaines) en fonction de polymorphismes génétiques ou de structures familiales. Elle vise à comprendre les bases héréditaires de ces maladies, en utilisant les méthodes de la statistique génétique.

D.III. Définitions et Concepts en Génétique

D.III.1. L'ADN

D.III.1.a) *Structure de l'ADN*

L'Acide Désoxyribonucléique est un polymère. Les monomères le constituant s'appellent les nucléotides. Ces nucléotides comprennent une molécule de sucre, le désoxyribose, un groupe phosphate et une partie variable appelée base. Il existe quatre bases classées en deux sous-groupes : les purines comprenant l'adénine et la guanine (notées A et G) d'un côté et les pyrimidines comprenant la cytosine et la thymine (C et T).

L'ordre de ces bases, appelé aussi séquence, contient une bonne partie de l'information génétique transmissible aussi bien de cellule en cellule que de génération en génération.
Chaque base est attachée à la molécule de sucre, le désoxyribose. Chaque molécule de désoxyribose est elle-même attachée au groupe phosphate soit par le carbone 3', soit par le carbone 5'. Enfin, le phosphate d'un nucléotide est attaché au désoxyribose du nucléotide suivant.
L'asymétrie de la molécule de sucre donne une polarité à la molécule d'ADN. D'un côté, c'est le carbone 3' qui se termine par un groupe hydroxyle libre (-OH), d'un autre côté, c'est le carbone 5'. On en tire les noms des deux extrémités d'une molécule d'ADN, l'extrémité 3' et l'extrémité 5'. Connaître cette polarité est important puisque les processus fondamentaux tels que la réplication, la transcription et la traduction se font dans un sens seulement. Il existe deux types de génomes dans la cellule, le génome nucléaire, constitué de 3 200 millions de paires de bases et le génome mitochondrial qui comporte 16 569 paires de bases.
Le génome nucléaire est divisé en paires de chromosomes de tailles inégales (Figure 2, ci-dessous). Il existe deux types de paires de chromosomes – les autosomes et les chromosomes sexuels. Les chromosomes autosomes sont des chromosomes homologues (paires de chromosomes 1 à 22 sur la Figure 2), qui sont identiques en taille et en contenu génique – c'est-à-dire qu'ils ont les mêmes séquences à l'exception des polymorphismes génétiques. Les chromosomes hétérochromosomes, également appelés chromosomes sexuels, sont appelés chromosome X et Y. La femme porte deux chromosomes X, de la même façon qu'un autosome, mais l'homme porte un chromosome X et un chromosome Y. Le chromosome Y a une séquence très différente de celle du X et est beaucoup plus petit mais il possède deux zones homologues (appelée zones pseudo-autosomiques) avec ce chromosome. Le X et le Y forment donc bien une paire chez les individus mâles.

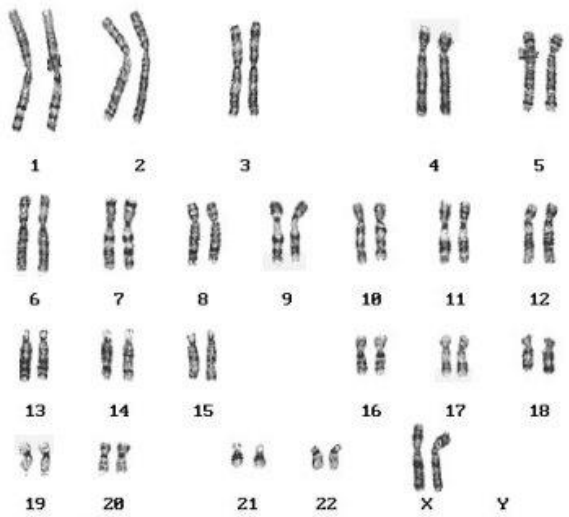


Figure 2 : Caryotype humain : les chromosomes

Caryotype : image des chromosomes au cours de la métaphase de la mitose. Le blocage en métaphase se fait grâce à la colchicine. Le génome humain contient 22 paires de chromosomes. Les chromosomes sont ensuite colorés et font apparaître une alternance de bandes noires et blanches. Ces bandes occupent des positions constantes et définissent des régions fixes du génome permettant d'identifier précisément chacun des chromosomes et aussi de créer une carte de ces chromosomes. Il s'agit d'un caryotype d'un patient féminin, car on observe deux chromosomes X et pas de chromosome Y. (http://lewebpedagogique.com/svt21610/tag/caryotype/)

D.III.1.b) *ADN et hérédité*

Le rôle de l'ADN dans la transmission de l'hérédité a été mis en évidence en plusieurs étapes. D'abord, Sutton [Sutton,**1902**] et Boveri, indépendamment, observèrent une très grande similitude entre les facteurs de Mendel et les chromosomes cellulaires.

Morgan, a confirmé cette observation en établissant le lien entre un trait (les yeux rouges) et un chromosome (chromosome X) chez la mouche drosophile (Figure 3, ci-dessous). Il restait à savoir quelle composante des chromosomes était responsable de la transmission des caractères d'une génération à la suivante. C'est Avery (avec McLeod et McCarthy) et son équipe qui démontrèrent que c'est le passage de l'ADN qui explique la transformation du pneumocoque [Avery et al.,**1944**], puis la transmission du trait dans les générations (de pneumocoques) suivantes.

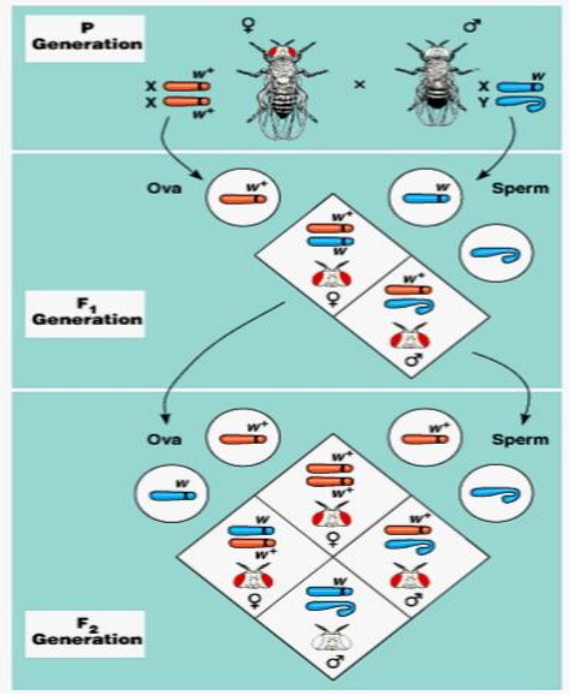


Figure 3 : Morgan et les bases de l'hérédité

En 1908, Thomas H. MORGAN (New-York) étudie le développement d'une petite mouche des fruits, la Drosophile (*Drosophila melanogaster*) ou mouche du vinaigre.
Dans une souche dite "sauvage", il repère un mâle dont les yeux ne sont pas rouges comme ceux des autres individus, mais blancs. Il décide alors d'étudier la transmission de cette mutation appelée "white" en croisant cette mouche avec des femelles de type "sauvage", c'est-à-dire aux yeux rouges (premier croisement). Il réalise ensuite un deuxième croisement où, cette fois-ci, le caractère muté est porté par la femelle (deuxième croisement).
Si les mouches de la première génération (F1) – issues du premier croisement - ont toutes les yeux rouges, les mouches issues du croisement d'individus de F1 présentent des proportions 1/4 :3/4, conformes aux proportions attendues d'après les expériences de Mendel.
Plus inattendu, les porteurs du phénotype « yeux blancs » de la génération F2 sont tous mâles.
Par ailleurs, une étudiante de Morgan (Nettie Stevens) et un chercheur de l'université Columbia découvraient quelques années plus tôt, de façon indépendante, que le chromosome X est lié à la détermination du sexe [Stevens,**1905**][Wilson,**1905**].
Morgan et son équipe en concluent que les chromosomes sont bien les porteurs de l'hérédité : les facteurs décrits par Mendel.
(Pearson Education, Inc. Publishing as Benjamin Cummings)

Enfin, Watson et Crick mirent à jour la structure de l'ADN [WATSON, CRICK,**1953**], constitué de deux brins complémentaires (Figure 4).

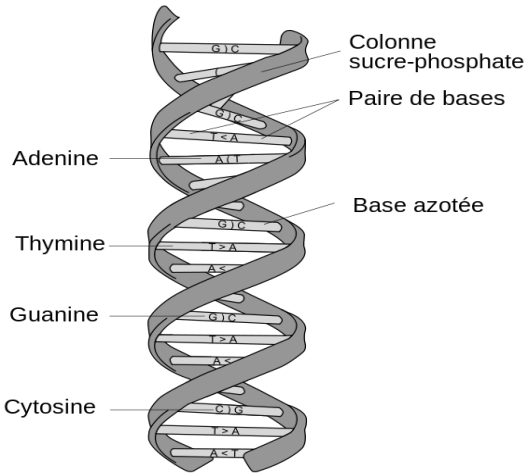


Figure 4 : l'ADN

Chaque brin d'ADN est constitué d'un enchaînement de nucléotides, eux-mêmes composés de bases azotées, de sucres (désoxyribose) et de groupements phosphate. On trouve quatre nucléotides différents dans l'ADN : A, G, C et T, du nom des bases correspondantes (Adénine, Guanine, Cytosine et Thymine). – Source : http://fr.wikipedia.org/wiki/Acide_désoxyribonucléique

D.III.2. Polymorphisme Génétique

Bien que 99.9% du génome soit identique chez tous les humains, il existe une portion non négligeable de variations entre individus. Cette variabilité est à la base de l'analyse génétique dans la mesure où l'on pense qu'elle expliquerait la variabilité des phénotypes humains.

Les localisations sur génome (en paires de bases par exemple) sont appelées des loci (singulier locus). Dans la suite de ce travail, nous emploierons locus (loci) pour désigner des emplacements physiques du génome où il existe une variation entre individus. C'est Bateson, en 1902, qui a proposé le mot d'allèle pour ces différentes versions. Un individu qui possède les deux mêmes versions d'un gène, ou toute autre région polymorphe, est **homozygote**. S'il a deux versions alléliques différentes, il est **hétérozygote**.

C'est toujours William Bateson qui utilisera, en 1905, le mot "**génétique**" pour désigner la science qui étudie l'**hérédité**, c'est-à-dire la transmission des caractères et leurs variations.

D.III.2.a) *SNPs*

D.III.2.a.1) Les substitutions

Les variations les plus simples sont des substitutions de bases. Une base A peut se transformer en C, G ou T (Figure 5). Cette mutation est alors transmise aux molécules d'ADN descendant de la molécule mutée. Sur la Figure 5 (p. 19) on observe deux choses. Suivant le sens de lecture, une variation bi-allélique (consistant en deux allèles) peut-être notée comme « T muté en C » ou « A muté en G ». Il s'agit de la même substitution à cette position. Ensuite, cette figure généralise le cas où il y a plusieurs polymorphismes possibles sur un brin d'ADN. Cela nous permet d'introduire la notion d'haplotype (ensemble d'allèle à différents loci sur un même chromosome).

Nous utilisons l'anglicisme SNP (pour Single Nucleotide Polymorphism) car il est utilisé de façon assez systématique dans la communauté.

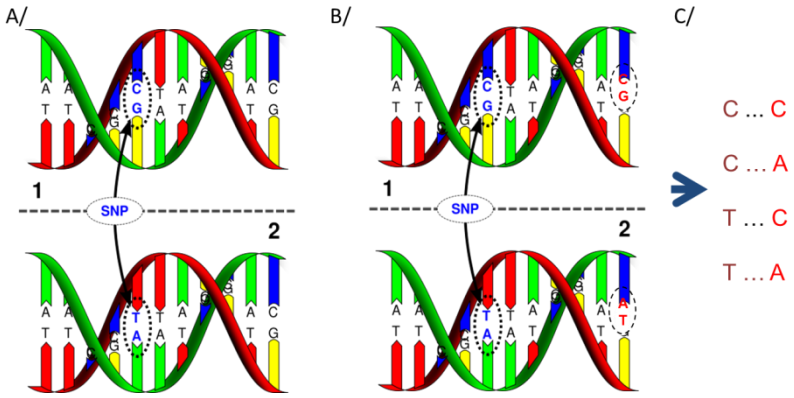


Figure 5 : substitution de base – ou SNP.

A/ Les deux molécules d'ADN diffèrent d'une seule paire de bases (C≡G devient T≡A). Il s'agit d'une substitution de base. Une molécule d'ADN étant formée de deux brins complémentaires, suivant le brin qui est lu en séquençage, le premier allèle est C et le deuxième est T ; ou bien le premier allèle est G et le deuxième est A. Le sens de la lecture est important car une même séquence doit être lue dans le même sens pour deux populations.
B/ Les deux paires de bases diffèrent à deux localisations (deux paires de bases). Il existe une deuxième mutation (en rouge) - C≡G devient A≡T le le brin qui est lu -. On peut dire que T/C (resp. A/G) sont des allèles au locus donné.
C/ Représentation symbolique des haplotypes. Un haplotype est l'ensemble des allèles présents sur un même chromosome (on dit en phase).
Source: Wikipedia Commons. Auteur: David Hall. Adapté.

Il se pose souvent un problème de qualification de ce type de variation. Une modification de la séquence de l'ADN est appelée mutation ou polymorphisme, la différence étant subtile. Bien qu'il n'y ait pas de différence entre une mutation et un polymorphisme, le nom « mutation» est plus souvent utilisé pour définir un changement de séquence ayant un effet pathogène (exemple : la mutation du gène *CFTR*, responsable de la mucoviscidose). Le mot polymorphisme est plutôt réservé aux mutations n'ayant pas d'effet visible. L'emploi du mot mutation est également lié à la fréquence d'une telle variation (voir D.III.5, p.23). Ainsi, des mutations ayant une fréquence supérieure à 5% peuvent indifféremment être appelées mutations ou polymorphismes.

Il est intéressant d'estimer le taux de mutation dans le génome. Cela permet par exemple d'estimer la variabilité attendue entre individus. Avant l'avènement des technologies de séquençage à haut débit, cette estimation reposait sur des approches indirectes. Une des plus intuitives est l'utilisation de la fréquence d'apparition d'une pathologie « génétique » chez des enfants de parents non atteints [Cooper, Krawczak,**1993**]. L'approche indirecte la plus utilisée et sans doute la plus précise est toutefois la comparaison de séquences entre espèces dont on connaît la date de séparation [Nachman, Crowell,**2000**]. Ainsi le taux de mutation dans le génome d'une génération à l'autre a été estimé à 2.5. 10^{-8}.
La capacité de séquencer des génomes entiers a permis d'estimer de façon plus directe ce taux. Il devient en effet possible de comparer la séquence génomique de parents et de leurs enfants. Par ces méthodes, le taux de substitutions est estimé à $2.3.10^{-8}$ [Kong et al.,**2012**][1000 Genomes Project Consortium et al.,**2010**], confirmant les estimations initiales.
Les fréquences des mutations peuvent toutefois varier suivant les régions du génome. L'exemple le plus connu est celui des régions riches en îlot CpG où le taux de mutation est environ 10 fois supérieur.
Deux faits importants sont cependant à noter. D'abord cette fréquence reste très basse quelle que soit le contexte génomique et ensuite, la probabilité de mutations inverses (c'est-à-dire du retour à l'allèle sauvage par mutation) apparaît comme négligeable (puisqu'elle serait de l'ordre de μ^2, soit 10^{-16}).

D.III.2.a.2) Les insertions/délétions

Ces mutations sont des ajouts ou des retraits de bases de la séquence initiale (Figure 6). L'exemple le plus courant est la mutation ΔF508 de la protéine CFTR, une délétion de 3 paires de bases qui enlève une phénylalanine et est responsable de la mucoviscidose. Lorsque la délétion ou l'insertion n'est pas un multiple de 3, les conséquences pour la protéine sont encore plus catastrophiques puisque la protéine synthétisée n'a plus du tout la même séquence, du fait du décalage du cadre de lecture.

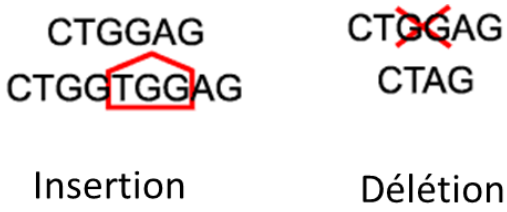


Figure 6 : Insertion et Délétion
Exemple d'insertion de trois bases (TGG) dans une séquence d'ADN ainsi que de délétion de deux bases (GG).
University of California Museum of Paleontology's Understanding Evolution

D.III.2.a.3) Diversité nucléotidique

La diversité nucléotidique moyenne (π) – correspondant au nombre moyen de différences nucléotidiques par site entre paires de séquences prises au hasard chez différents individus de la même espèce - est d'environ 8.10^{-4} [Lander et al.,**2001**][Venter et al.,**2001**][Przeworski et al.,**2000**], représentant un SNP toutes les 1250 paires de bases. Il convient de noter toutefois que π varie fortement d'un chromosome à l'autre et qu'un grand nombre de SNPs rapporté dans les bases de données n'avait pas été confirmé jusqu'à récemment. Les données des 1000 génomes (D.V.3, p. 44) ont permis d'affiner cette estimation.
Il y a cependant une notion importante à souligner lorsqu'on étudie la diversité nucléotidique: cette diversité dépend de la fréquence allélique seuil, et, par conséquent, du nombre d'individus sur lesquels elle est estimée.

D.III.3. Génotypes, Allèles, Fréquences, proportions de Hardy-Weinberg

D.III.3.a) *Génotypes et allèles*

Dans les organismes diploïdes, comme les humains, les allèles se trouvent par paires - à l'exception des polymorphismes des chromosomes X et Y chez l'homme (mâle) et des polymorphismes mitochondriaux.

Pour un marqueur bi-allélique donné, un individu peut présenter trois génotypes possibles (appelées ici AA, Aa et aa de façon générique). Dans une population de N individus, qui comprend N_{AA} individus porteurs du génotype AA, N_{Aa} porteurs du génotype Aa et N_{aa} porteur du génotype aa, on définit trois fréquences génotypiques :

$$G_i = \frac{N_i}{N}$$

Équation 1 : Fréquence génotypique
Fréquence génotypique pour un génotype i, (i = AA, Aa et aa)

Nous sommes souvent intéressés par la fréquence de l'allèle (ne serait-ce qu'en raison d'une manipulation plus aisée). L'unité d'observation est alors le chromosome. Pour un

marqueur génétique autosomal, la taille de la « population » sera 2xN, le nombre de chromosomes.

La fréquence de l'allèle A est le nombre de chromosomes portant l'allèle A divisé par ne nombre total de chromosomes :

$$f_A = \frac{2 \times N_{AA} + N_{Aa}}{2 \times N}$$

Équation 2 : Fréquence allélique

D.III.3.b) *Répartition des allèles dans les génotypes*

Soit un SNP avec 2 allèles : A et a. Il existe dans la population 3 génotypes possibles (AA, Aa et aa) et leurs fréquences seront notées G_{AA}, G_{Aa} et G_{aa}.
Ces fréquences dépendent du lien éventuel entre ces deux allèles ou, autrement dit, de leur indépendance.
Dans une population diploïde de taille supposée infinie où les croisements entre individus se font au hasard (panmixie et pangamie) et où il n'y a ni de mutation, ni de migration (c'est-à-dire pas de nouveaux allèles qui apparaissent), ni de sélection (c'est-à-dire que la probabilité de survie des individus ne dépend pas de leur génotype), les fréquences des génotypes dépendent uniquement des fréquences alléliques et restent stables au cours des générations. C'est la célèbre loi de Hardy-Weinberg découverte indépendamment par le mathématicien anglais Hardy et par le médecin allemand Weinberg en 1908 [Hardy,**1908**][Weinberg,**1908**].
Si on appelle $f_{A,t}$ la fréquence de l'allèle A à la génération t et $f_{a,t}$ celle de l'allèle a ($f_{A,t}$+ $f_{a,t}$ =1), à la génération suivante t+1, les fréquences attendues $G_{AA,t+1}$, $G_{Aa,t+1}$ et $G_{aa,t+1}$ pour les génotypes AA, Aa et aa sont respectivement $f_{A,t}^2$, $2 \times f_{A,t} \times f_{a,t}$ et $f_{a,t}^2$ (proportions de Hardy-Weinberg qui sont obtenues sous l'hypothèse de panmixie). A la génération t+1, la fréquence p_{t+1} de l'allèle A peut alors être obtenue par décompte des allèles comme précédemment :

$$f_{A,t+1} = G_{AA,t+1} + {}^{G_{Aa,t+1}}/_{2} = f_{A,t}^2 + f_{A,t} \times f_{a,t} = f_{A,t}$$

Équation 3 : évolution d'une fréquence allélique

On voit donc que les fréquences alléliques restent inchangées au cours des générations (c'est l'équilibre de Hardy-Weinberg).
Sur des données réelles, on pourra tester si les génotypes observés sont distribués selon les proportions de Hardy-Weinberg en réalisant un test de χ^2 de conformité. Si on reprend l'exemple du paragraphe précédent D.III.3.a), p. 20. On comparera les effectifs observés N_{AA}, N_{Aa} et N_{aa} pour les trois génotypes et les effectifs attendus selon les proportions de Hardy-Weinberg : $N \times f_A^2$, $2 \times f_A \times f_a$, $N \times f_a^2$.

Dans les analyses que j'ai menées, on s'attend à ce que les fréquences génotypiques des marqueurs respectent les proportions de Hardy Weinberg, au moins dans la population générale et/ou chez les témoins.
Dans la population des patients, selon le modèle génétique, on peut s'attendre à ce que les SNPs associés à une pathologie ne suivent pas ces proportions, même lorsque les conditions 1 et 2 sont remplies [Lee,**2003**][Wittke-Thompson et al.,**2005**][Song, Elston,**2006**].
L'expérience des analyses d'association génome entier démontre que toute déviation importante par rapport à cet équilibre est principalement due à des erreurs de génotypage.

Toutefois, un tel déséquilibre, lorsqu'il n'y a pas de problème de génotypage, permet de mettre en évidence des événements d'intérêt, comme un mélange de populations (voir D.III.5, p. 23) ou un effet génétique (et donc une association) lorsqu'il est observé chez les patients.

D.III.4. Évolution des fréquences alléliques

Dans des populations de petite taille, le hasard peut modifier significativement les fréquences alléliques. Cette modification est appelée dérive génétique.
Si quelques individus seulement, et donc un petit nombre de gamètes, sont tirés au hasard pour former la génération suivante, les allèles qu'ils portent ne seront pas exactement dans les mêmes proportions que les allèles du pool génique duquel ils sont issus. Dans une population restreinte, la dérive génétique peut être suffisamment importante pour modifier les fréquences alléliques.

Ainsi, la fréquence des allèles présentant un désavantage sélectif pourrait augmenter grâce à la dérive génique, alors que des allèles avantageux rares pourraient être perdus. Cela peut sembler contre-intuitif mais souligne bien l'effet que peut avoir la variation aléatoire dans une population de taille imitée.

Les deux causes importantes de la dérive génétique sont les goulets d'étranglement et les effets fondateurs.
Goulets d'étranglement : Même les organismes constituant normalement de vastes populations sont susceptibles de traverser occasionnellement des périodes durant lesquelles seul un petit nombre d'individus survit et vont donc refonder une population. Lors de ces goulets d'étranglement, la variation génétique peut-être perdue par l'effet du hasard.

Effet fondateur : Lorsqu'une espèce se répand sur de nouveaux territoires, des populations peuvent être fondées par un petit nombre d'individus pionniers. Ces individus fondateurs ont peu de chance d'avoir la totalité des allèles présents dans leur population d'origine. Même s'ils les ont, leurs fréquences alléliques sont vraisemblablement différentes de celles de la population d'origine.

Dans le cas de la population humaine, le modèle le plus plausible propose une différenciation importante entre les populations il y a environ 100 000 ans, à partir de l'Afrique, où l'homme moderne serait apparu [Lukić, Hey,**2012**]. Cette « sortie de l'Afrique » aurait concerné un nombre réduit d'individus, qui serait à l'origine de toutes les populations non-africaines (Figure 7).

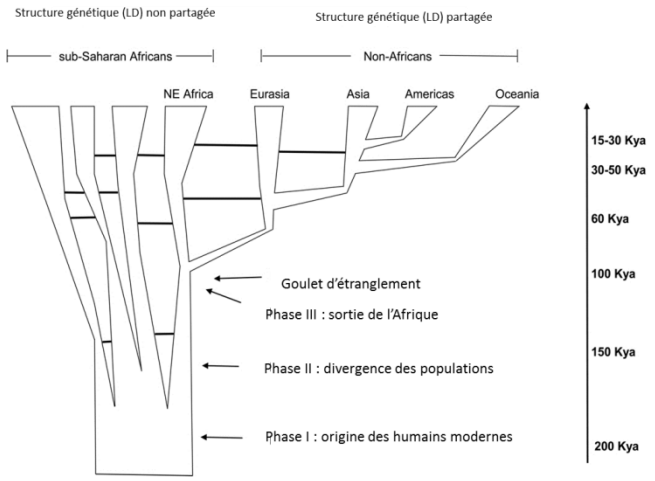


Figure 7 : Histoire génétique de la population humaine

En terme de dérive génétique et de différence de fréquences alléliques entre populations, cela expliquerait la forte variabilité existant entre populations Africaines à l'intérieur de l'Afrique mais également la forte variabilité de fréquence entre populations Africaines et populations Européennes (et/ou Asiatiques).
Par ailleurs, et nous l'expliquerons plus tard, ce scenario expliquerait également la forte différence de structure génétique en terme de déséquilibre de liaison (DL).

D.III.5. Stratification génétique d'une population :

Il est important de noter que la notion de fréquence s'appuie sur la notion de population cible. On peut considérer une fréquence pour toute la population humaine : c'est le nombre de fois où on observerait un allèle donné si on testait tous les êtres humains. En épidémiologie génétique, nous sommes plutôt intéressés par la fréquence d'un allèle dans une population spécifique. Cela peut être, mais pas exclusivement, la population pour laquelle on veut améliorer l'état sanitaire, ou également une population où on s'attend à une certaine homogénéité démographique. En effet, les fréquences alléliques peuvent être très variables en fonction de l'histoire démographique d'un groupe d'individus. Il en résulte la possibilité de très grandes différences entre populations. Si on met à part les phénomènes de sélection, cette différence est principalement due à la dérive aléatoire indépendante d'une fréquence allélique. Cette dérive entraine des trajectoires de fréquences alléliques différentes dans des populations qui ont très peu d'échange gamétique (Figure 8, ci-dessus).

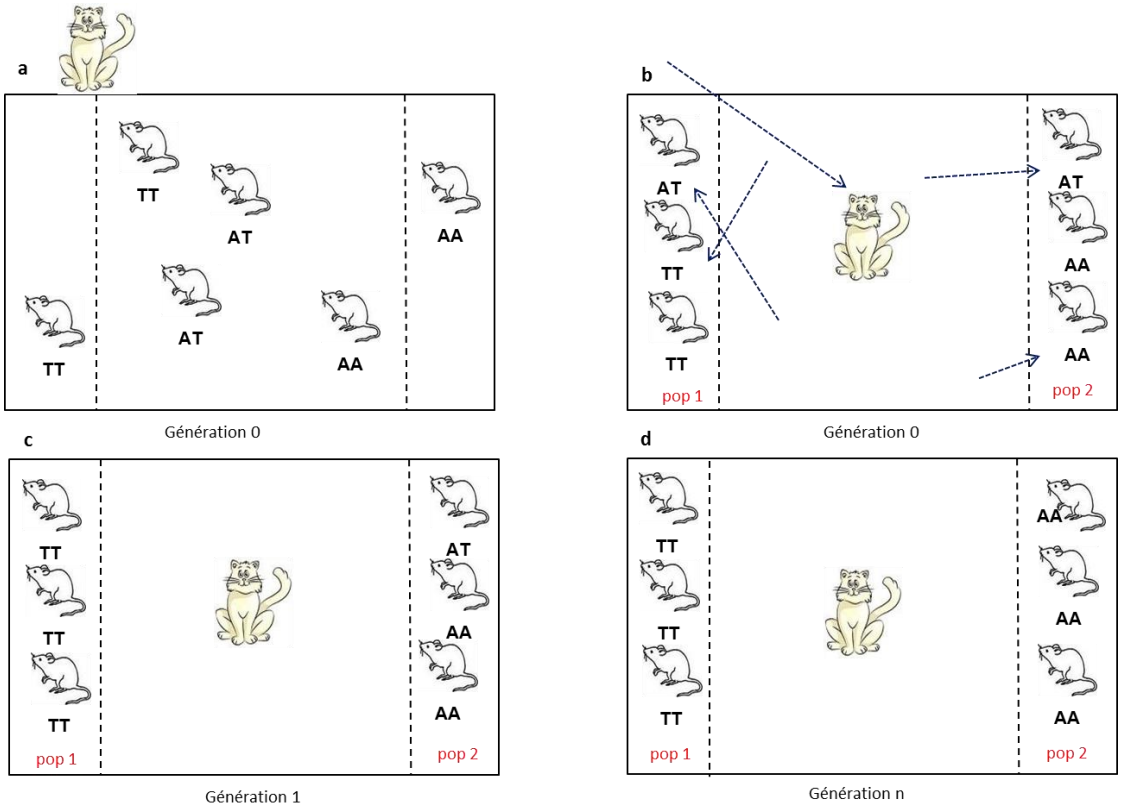


Figure 8 : Sous- Population et Dérive génétique.
La pression extérieure, appliquée sur une population initiale (a) entraîne la création de deux populations (b) qui n'échangent plus de gènes (absence de croisement entre pop 1 et pop 2). Par ailleurs, les fréquences fluctuent aléatoirement dans chaque population de façon indépendante. Du fait de la petite taille de chaque population et d'un déséquilibre de fréquences au moment de la séparation brusque des populations (A est rare dans la pop 1 et T est rare dans la pop 2), la probabilité de fixation de l'allèle T dans la pop 1 et de l'allèle A dans la pop 2 est très grande (c). La conformation visible dans d/ est une des « histoires » possibles – la plus probable – mais n'est pas la seule.

Si ces populations sont petites, on peut considérer qu'il y a un goulet d'étranglement. Certains allèles peuvent se trouver à une fréquence très basse dans une population et disparaître comme l'allèle A en population 1 (Figure 8 – c, ci-dessus) et l'allèle T en population 2 (Figure 8 – d, ci-dessus). Il y a création de différence de fréquence de façon aléatoire.

Par ailleurs, nous observons également que lorsque nous regroupons les individus venant de populations différentes, il peut y avoir un écart important aux proportions de Hardy Weinberg. En effet, il n'existe, à la première génération après ce mélange de deux populations, que des individus homozygotes.
Par contre, à la génération suivante, si l'hypothèse de mariage au hasard (panmixie) est respectée, les proportions seront tout de suite retrouvées (avec un test des proportions de Hardy-Weinberg non significatif).

D.III.6. Recombinaison génétique

Lors de la première division réductionnelle de la division méiotique ont lieu des événements de recombinaison homologue appelés « *crossing-over* » par un appariement des chromosomes homologues et échange de chromatides non sœurs. Ces recombinaisons

contribuent au brassage du génome. La **fraction de recombinaison (θ)** correspond à la probabilité de recombinaison entre deux loci. Cette probabilité dépend de la distance entre ces deux loci.
Les deux valeurs extrêmes sont :
θ =0 : aucune recombinaison entre les deux loci. Ces deux loci sont sans doute extrêmement proches sur le chromosome.
θ =0,5 : correspond à 50% de gamètes recombinés sur l'ensemble des gamètes transmis et signifie donc une non-liaison entre les deux loci (très éloignés sur le même chromosome, ou sur des chromosomes différents).
La mesure de la distance génétique entre deux loci est le **centiMorgan (cM)**, 1 cM correspondant à une fréquence de recombinaison de 1% pendant la méiose.
Ce nom de mesure rend hommage à une des premières expériences de cartographie génétique réalisée par Morgan et son étudiant, Sturtevant.

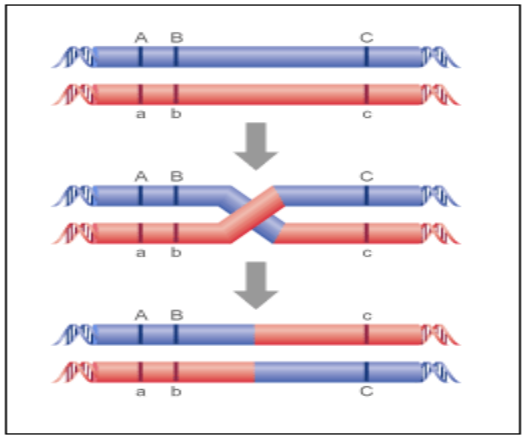


Figure 9 : La recombinaison
En bleu : Chromosome d'origine paternelle.
En rouge : Chromosome d'origine maternelle.
La recombinaison implique un phénomène physique appelé « crossing-over » entre les deux chromosomes homologues d'une paire. Les allèles paternels et maternels peuvent être échangés. Il y a plus de probabilité de recombinaison entre A et C qu'entre A et B, en raison de la position physique sur le chromosome (la deuxième chromatide a été omise pour plus de clarté). D'après http://genome.wellcome.ac.uk/.

Le taux de recombinaison n'est pas lié de façon linéaire à la distance. En effet, un faisceau convergent d'observations et de modèles indique que ce taux est discontinu (alternance de régions à haute probabilité de recombinaison, appelées « hot spots » et de régions à basse probabilité, « cold spots » [Jeffreys et al.,**2001**][Arnheim et al.,**2003**].

Par ailleurs, il existe des variations du taux de recombinaison non seulement entre mâles et femelles mais également entre individus [Broman et al.,**1998**][Cheung et al.,**2007**][Coop et al.,**2008**]. Ces différences peuvent être d'ailleurs gouvernées par certains loci du génome humain [Kong et al.,**2008**][Chowdhury et al.,**2009**].

D.III.7. Déséquilibre de liaison (DL) et haplotypes

Le DL est défini comme l'association non aléatoire entre allèles à différents locus. Si la notion de DL date du début du siècle (voir par exemple Jennings (1917)), la première mesure couramment utilisée fut introduite par Lewontin (1964) il y a environ 40 ans. Reprenons la notation en deux allèles A et B.
Considérons deux locus bi-alléliques A/a et B/b. Soit f_i la fréquence de l'allèle i (i= A,B,a,b)

De manière similaire, p_{ij} désigne la fréquence de la présence simultanée des allèles i et j aux deux loci dans la population. Lewontin [Lewontin,**1964**] proposa alors de mesurer le DL par la statistique :

$$D = p_{AB} - f_A \times f_B$$

Équation 4 : DL de Lewontin

Il y a déséquilibre de liaison entre les locus si *D* diffère significativement de zéro. On note que D est la covariance de deux variables aléatoires (V.A.) qui prendraient les valeurs 0 ou 1 suivant l'allèle observé sur un chromosome (on peut choisir arbitrairement V.A.=0 pour l'allèle A et 1 pour l'allèle B). Cet écart à l'indépendance peut être testé par un simple test du χ^2:

$$x^2 = N_c \times r^2 = N_c \times D^2/(f_A \times f_a \times f_B \times f_b)$$

On note que le test est moins simple en pratique car dans le cas d'individus diploïdes, il est impossible d'avoir un compte exact des « haplotypes ».

Il est toutefois difficile de comparer la mesure D pour différentes paires de marqueurs car elle dépend des fréquences alléliques.

Il est possible de normaliser cette valeur de plusieurs façons. Nous pouvons soit chercher le D_{max}, afin de normaliser le *D* comme l'a proposé Lewontin [Lewontin,**1964**] :

$$D_{max} = \min(f_A \times f_b, f_a \times f_B)\,; si\ D > 0$$

$$D_{max} = \min(f_A \times f_B, f_a \times f_b)\,; si\ D < 0$$

La normalisation s'exprime sous la forme :

$$D' = {}^{D}/_{D_{max}}$$

Il est également possible de calculer une valeur s'approchant de la corrélation statistique :

$$r^2 = \frac{D^2}{f_A \times f_a \times f_B \times f_b}$$

En effet, nous avons vu que D est une valeur similaire à la covariance entre A et B alors que f_A x f_a et f_B x f_b sont les variances de ces variables aléatoires. Ce r^2 est donc assimilable à un coefficient de corrélation.
D'autres mesures existent mais ne seront pas discutées ni utilisées ici.

Le DL entre deux marqueurs diminue en fonction de la distance génétique (taux de recombinaison) d'une génération à la suivante. En effet, la recombinaison génétique réduit la corrélation en redistribuant les allèles sur les chromosomes (Figure 10).

Puisqu'il y a possibilité de recombinaison à chaque génération, le LD à la génération n suit la loi suivante :

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

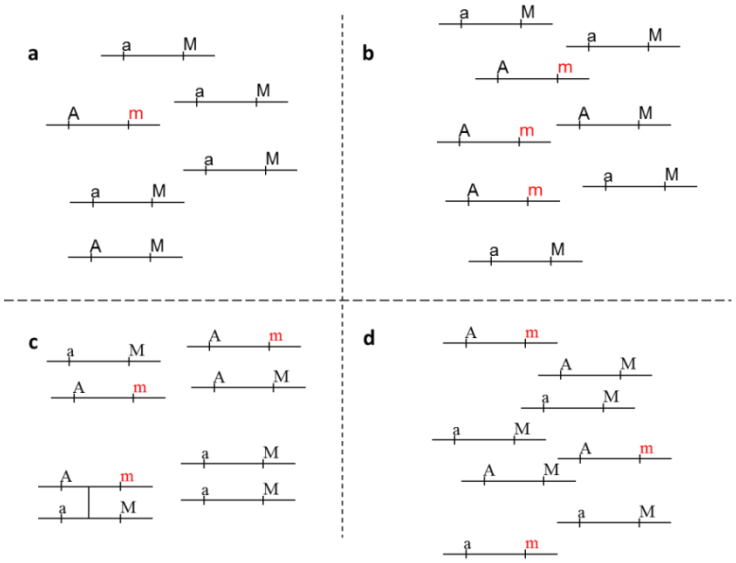


Figure 10 : Déclin du Déséquilibre de liaison

À l'apparition d'un polymorphisme (**a**), il se crée immédiatement un déséquilibre de liaison entre ce polymorphisme et les voisins. L'évolution aléatoire peut entraîner une augmentation de la fréquence du nouvel haplotype dans la population (**b**). En fonction de la distance entre les deux polymorphismes, il peut arriver qu'il y ait échange de matériel génétique entre deux chromosomes au moment de la méïose (**c**). À la suite de la recombinaison, la diversité haplotypique est augmentée. (4 haplotypes en **d** existant au lieu de trois dans **a**, **b** et **c**).

Les travaux de Daly [Daly et al.,**2001**] sur une séquence de 500 kb provenant de la région 5q31 chez 256 patients atteints de la maladie de Crohn ont révélé le faible nombre d'haplotypes différents à l'intérieur des blocs. Par exemple, pour 7 blocs s'étendant sur une distance de 92 kb et contenant 31 SNP, seulement 4 des 231 haplotypes possibles ont été observés pour 94% des chromosomes. Cette structure en bloc a depuis été largement confirmée.

D.III.8. Modèle biométrique et effets des allèles

Nous nous intéressons, en épidémiologie génétique, à la distribution d'un caractère dans la population. Notons que parler de distribution souligne bien qu'il y a des différences entre individus. Ces caractères peuvent être la taille, le poids, la couleur des cheveux ou une pathologie comme le diabète de type 2 ou le Syndrome de Brugada. On appelle ces caractères des **phénotypes** ou des **traits**. C'est un aspect anatomique, physiologique, moléculaire ou comportemental d'un organisme vivant, qui peut être analysé.

D.III.8.a) *Décomposition de la variance d'un trait*

Nous prenons le cas particulier de l'effet d'un SNP bi-allélique sur un trait quantitatif que nous appellerons L.
Ce trait L est décomposé en deux composants que sont la part génétique G et la part environnementale E.

$$L = G + E$$

Chacun de ces composants suit une loi statistique que nous allons décrire.

Nous nous plaçons dans un premier temps dans le cas où la part génétique n'est due qu'à une mutation (ou un polymorphisme) bi-allélique – 3 génotypes (AA, Aa et aa). Pour se conformer aux notations classiques dans le domaine, nous allons définir une nouvelle notation : f_A=p et f_a=q.

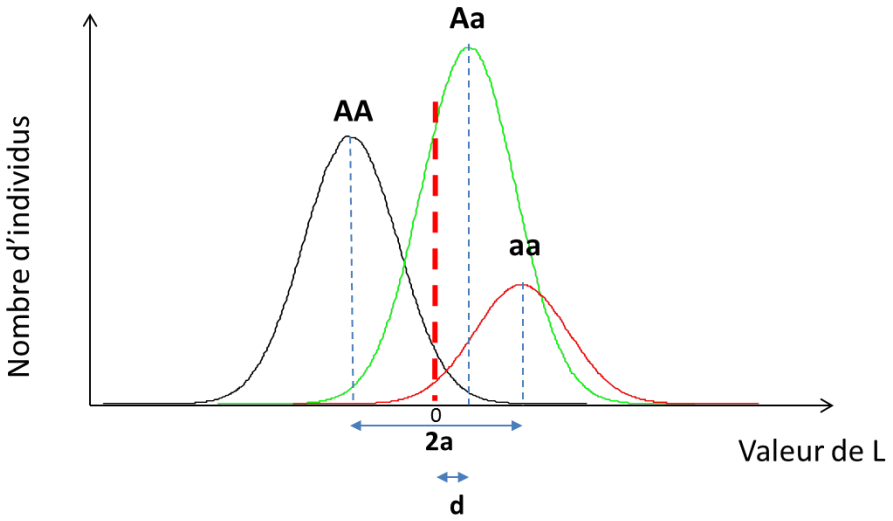


Figure 11 : Distribution d'un trait quantitatif dans la population.
Distribution d'un trait quantitatif en fonction de leur génotype à un locus bi-allélique. La distribution des individus porteurs d'un génotype AA est représentée en noir, celle des porteurs d'un génotype Aa est représentée en vert et enfin celle des porteurs du génotype aa est représentée en rouge. L'aire sous la courbe pour chaque distribution (noire, verte et rouge) est proportionnelle à la fréquence du génotype en population. q^2 en noir, 2*p*q en vert, p^2 en rouge. Chaque génotype a une moyenne: μ_{AA}=-a ; μ_{Aa} = d ; μ_{aa} =+a.

La distribution de la valeur génomique dans la population suit donc une loi multinomiale de paramètre P_g (vecteur des fréquences génotypiques). Par ailleurs, dans ce modèle, il y a une valeur fixe par génotype, égale pour tous les individus porteurs du même génotype à ce locus.
La part environnementale, E, suit une loi normale centrée sur 0 et dont la variance σ_E^2 est la variance environnementale (non génétique).

	AA	Aa	aa	Population
Fréquence (*g*)	q^2	$2pq$	p^2	
Moyenne / Génotype	-a	d	+a	
	μ_{AA}	μ_{Aa}	μ_{aa}	
Contribution à la moyenne	$-aq^2$	$2dpq$	ap^2	$\boldsymbol{M = a(p-q) + 2pqd}$
Contribution à la variance	$(-a^2)q^2$	$2pqd^2$	$(a^2)p^2$	$\boldsymbol{V = 2pq[a + d(p-q)]^2 + [2pqd]^2}$

Tableau 1 : modélisation d'un phénotype

Puisque (p^2-q^2) = (p-q) x (p+q) et que p+q=1 par définition, nous avons une moyenne en population du trait, *M* :

$$\boldsymbol{M = a \times (p-q) + 2 \times p \times q \times d}$$

Équation 5 : Moyenne d'un trait en population

La variance du trait s'obtient également facilement comme $\sum_{i=AA}^{aa} g_i \times (\mu_i - M)^2$ ou également sous la forme $\sum_{i=AA}^{aa} g_i \times (\mu_i)^2 - M^2$. C'est sous cette dernière forme que la variance est calculée dans le Tableau 1 (p. 28).

Il faut se rappeler que ce modèle a été mis en place à un moment où l'on imaginait que chaque maladie était due à un gène (implicitement on pensait à une seule mutation), plutôt fréquente dans la population : on l'appelait le gène majeur. Le reste de la variabilité interindividuelle étant due à une variation environnementale et à la composante polygénique. Aujourd'hui, pour beaucoup de pathologies fréquentes, la seule composante polygénique suffirait à expliquer une bonne partie de la variabilité.

D.III.8.b) *Composante polygénique*

Dans ce contexte, la valeur de *L* est égale à la somme des contributions (en principe faibles) d'un grand nombre (*J*) de polymorphismes (effet *ge*).

$$L = \mu + \sum_{j=1}^{J} ge_j + \epsilon$$

Équation 6 : décomposition en effets polygéniques

Bien que pour chaque combinaison d'allèles aux différents locus, la valeur génotypique soit fixe, le grand nombre de loci entraîne une distribution quasi-continue de L (Figure 12, ci-dessous).

Pour simplifier on définit la composante $G = \sum_{j=1}^{J} ge_j$. G suit une loi normale $N(\mu_G, \sigma_G^2)$. , σ_G^2 est la variance génétique polygénique.

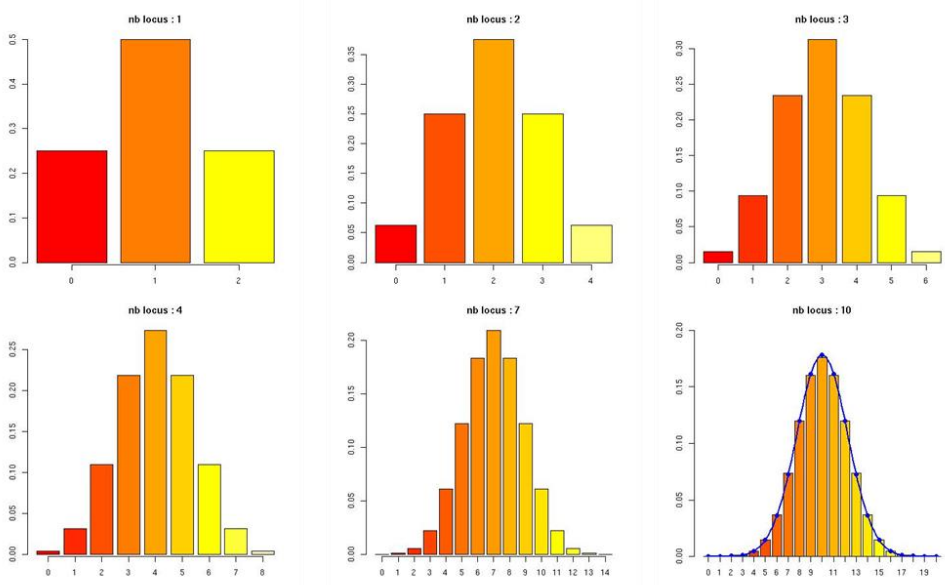


Figure 12 : Composante polygénique

Histogramme de la distribution des différentes classes de phénotype attribuable à la composante génétique seule. Pour simplifier la figure, on fait l'hypothèse que tous les polymorphismes génétiques ont la même fréquence et le même effet a. Pour un seul locus, il y a trois valeurs possibles (-a, 0 et a). Pour deux loci, il y a cinq valeurs possibles (-2a, -a, 0, a, 2a) chaque valeur étant rencontrée avec une fréquence proportionnelles au produit des fréquences alléliques qui la constituent : la valeur -2a est ainsi observée avec une fréquence p^4. Lorsque le nombre de loci en jeu devient important (modèle oligo-génique puis polygénique), la distribution discrète prend un caractère quasi-continu et s'approche d'une loi normale.

La transmission de cette composante polygénique se modélise en généralisant les lois de Mendel à un grand nombre d'allèles. Chaque parent transmet la moitié de son génome (en moyenne) et la valeur polygénique du descendant direct d'un couple est tirée d'une loi normale dont la moyenne est la moyenne des valeurs polygéniques des parents et la variance est σ_G^2, la variance polygénique.

La moyenne phénotypique, dans ce contexte polygénique, devient :

$$\boldsymbol{M = \sum_{i=1}^{J} a_i \times (p_i - q_i) + 2 \times \sum_{i=1}^{J} p_i \times q_i \times d_i}$$

Équation 7 : Moyenne en population (cas polygénique)

On notera enfin que ce modèle, classique, n'intègre pas encore la transmission non génétique des caractères (effets épi-génétiques). Il existe toutefois une part de la variance, que je ne vais pas traiter ici, qui est la variance environnementale partagée. Bien que non traitée dans ce travail, elle est disponible dans une grande partie des logiciels de décomposition de la variance et pourrait sans doute modéliser cette source de ressemblance familiale non génétique de façon satisfaisante.

La variance génétique dans l'hypothèse polygénique est estimée comme dans le chapitre D.III.8.a), p. 27, dans le cas où il n'y a pas d'interaction entre loci (épistasie).

D.III.8.c) *Héritabilité :*

L'héritabilité, au sens large, est la proportion de la variance d'un phénotype due à l'ensemble de facteurs génétiques. Elle inclut la variance de dominance. L'héritabilité au sens étroit est la variance phénotypique due aux facteurs génétiques additifs. C'est ainsi la proportion de la variance transmise d'une génération à la génération suivante. En effet, chaque parent ne transmet que la moitié de son génome.

D.III.8.d) *Cas d'un phénotype binaire*

D.III.8.d.1) Susceptibilité et seuils

Le modèle génétique présenté dans les chapitres précédents représentait un trait quantitatif. L'extension de ce modèle à un phénotype binaire (malade / non malade) se fait par l'introduction d'une valeur seuil pour L sur l'axe des abscisses et sera défini en fonction de la prévalence de la maladie. Si le phénotype L suit une loi normale de moyenne 0 et variance 1, et pour une pathologie avec une prévalence (fréquence dans la population) de Π, on définit un seuil Z tel que $Z=\Phi^{-1}(\Pi)$ (Figure 13, ci-dessous). Un individu t dont $L_t < Z$ aura un phénotype binaire *Ph* = 0. Un individu t dont $L_t \geq Z$ aura un phénotype binaire *Ph* = 1.

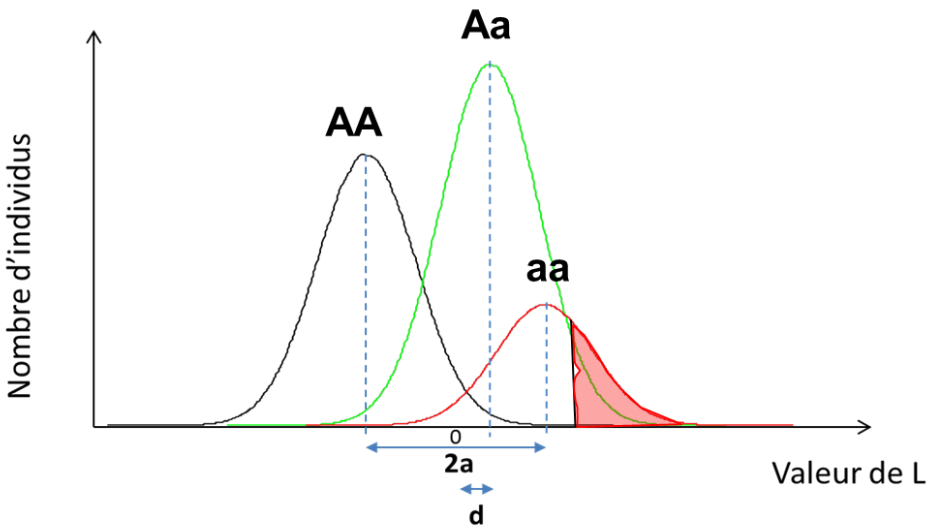


Figure 13 : modèle biométrique pour phénotype binaire

On introduit un seuil Z pour lequel, l'aire sous la courbe de distribution des valeurs génotypiques à droite (surface en rouge) est égale à la prévalence Π. Ce seuil peut varier en fonction de covariables (sexe par exemple) et pour différentes tranches d'âge. Cette prévalence (ou ensemble de prévalences) est en général estimée par des études épidémiologiques préalables.

Pour un génotype donné, la probabilité d'être atteint pour un génotype i est égale à $\emptyset(Z, m = \mu_i,\ \sigma^2 = 1 - V_a)$, soit la probabilité, sous une loi normale de moyenne μ_i et de variance égale à la variance environnementale plus la variance polygénique. Cette valeur est la **pénétrance** de chaque génotype i (notée P_i).

$$P_i = P({}^{Ph\,=\,1}/_{G\,=\,i})$$

Équation 8 : pénétrance

Par ailleurs, le risque de survenue d'une pathologie est aussi exprimé en termes de risque relatif (RR).
Les risques relatifs des génotypes Aa et aa par rapport au génotype de base (choisi de façon arbitraire) sont :

$$RR_{Aa} = \frac{\emptyset(Z, m = \mu_{iAa}, \sigma = 1 - V_a)}{\emptyset(Z, m = \mu_{iAA}, \sigma = 1 - V_a)}$$

Équation 9 : Risque relatif de l'hétérozygote

Et

$$RR_{aa} = \frac{\emptyset(Z, m = \mu_{iaa}, \sigma = 1 - V_a)}{\emptyset(Z, m = \mu_{iAA}, \sigma = 1 - V_a)}$$

Équation 10 : Risque relatif de l'homozygote muté

Les risques relatifs sont très utilisés en épidémiologie mais ils ne sont pas toujours facilement estimables, surtout dans un contexte d'association cas-témoins.
La mesure la plus utilisée est le rapport de cotes (plus communément appelé, même en français, Odds Ratio).
Si on définit p_x la probabilité d'être atteint lorsqu'on est exposé et $p_{\bar{x}}$ la probabilité d'être atteint lorsqu'on n'est pas exposé, alors l'Odds Ratio s'écrit :

$$OR = \frac{p_x \times (1 - p_{\bar{x}})}{p_{\bar{x}} \times (1 - p_x)}$$

ce qui dans notre modélisation se transforme, pour le génotype Aa, par exemple en :

$$OR = \frac{\emptyset(Z, m = \mu_{iAa}, \sigma = 1 - V_a) \times (1 - \emptyset(Z, m = \mu_{iAA}, \sigma = 1 - V_a))}{\emptyset(Z, m = \mu_{iAA}, \sigma = 1 - V_a) \times (1 - \emptyset(Z, m = \mu_{iAa}, \sigma = 1 - V_a))}$$

L'Odds Ratio est en général appelé OR pour simplification. Lorsque la prévalence de la maladie est basse (< 1%), l'OR et le RR sont proches.
Enfin, l'OR est directement estimable dans la régression logistique, qui est la technique statistique de recherche de liaison que nous allons employer.

D.III.9. Mesures de concentrations familiales

Historiquement, il était important de mesurer l'héritabilité d'un trait afin d'estimer l'opportunité de poursuivre des études génétiques pour expliquer sa variabilité en population. L'héritabilité est la proportion de la variance du trait due à l'ensemble des variations génétiques. On peut également définir l'héritabilité attribuable à un (ou plusieurs) locus, c'est-à-dire à un sous-ensemble seulement de toutes les variations génétiques.

Dans le cas d'un phénotype binaire, il est également courant, et en général plus simple, d'estimer l'augmentation de risque pour l'apparenté d'un individu atteint (λ_R). R représente le degré d'apparentement par rapport à un individu proposant. Le risque de l'apparenté de niveau R (frère-sœur, cousin, oncle-neveu …) est comparé au risque en population générale, c'est-à-dire à la prévalence. Les estimations de ces rapports en fonction des degrés de parentés R ont été développés par James [James,**1971**] et intégrés pleinement dans la recherche de variants génétiques responsables de pathologie complexes par Risch [Risch,**1990**]. La valeur la plus utilisée car la plus accessible dans des études épidémiologiques est le λ_S (le S est utilisé pour « Sib », qui signifie germain), qui est la probabilité pour un individu d'être atteint si son germain (son frère ou sa sœur) est atteint.

Il est à noter que ce rapport est plutôt un témoin de la concentration familiale au sens large car il est difficile de différencier l'effet génétique de l'effet environnemental partagé – surtout

avec des paires de germains qui, en principe, passent un partie non négligeable de leur existence dans des conditions similaires.
Malgré cela, le λ_s et l'héritabilité sont deux mesures de référence dans la littérature lorsqu'on essaye d'estimer le poids global des polymorphismes génétiques sur le risque d'une pathologie mais également la proportion de ce poids qui est expliquée par les variants déjà mis en évidence.

D.III.9.a) *Estimation empirique du λ_S*

Nous recherchons la probabilité pour un individu d'être atteint si son germain est atteint et nous comparons cette probabilité avec la prévalence en population (Figure 14, ci-dessous).

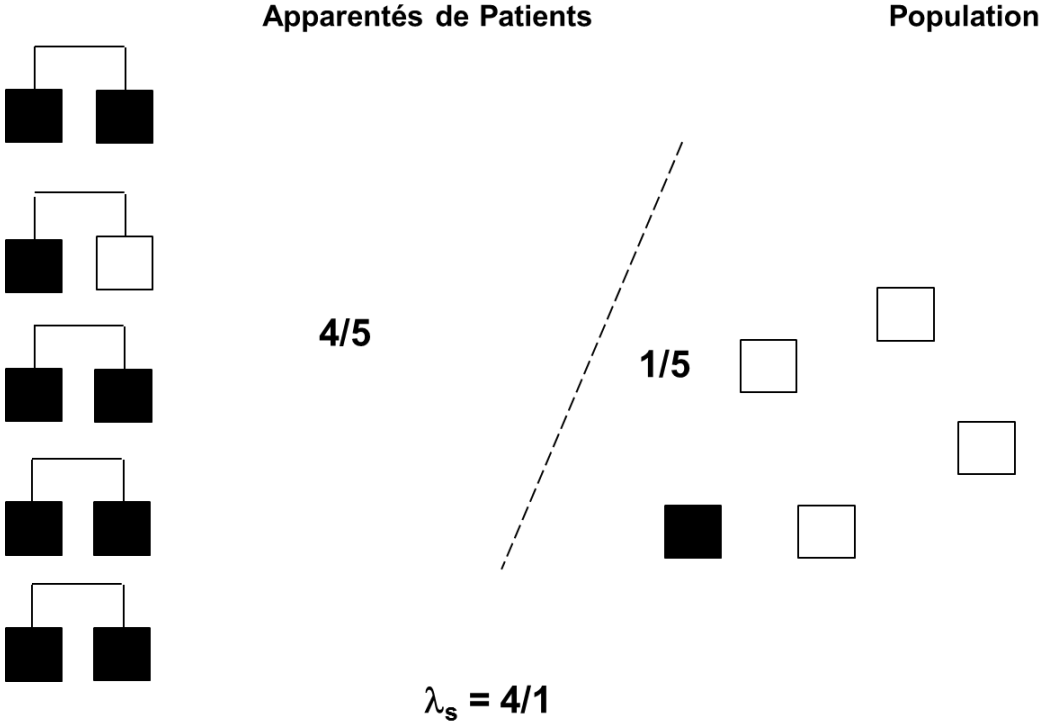


Figure 14 : calcul λ_S
Exemple de calcul de risque relatif pour des apparentés de type germain (frères ou sœurs). La prévalence de la pathologie est estimée chez les apparentés de degré R d'individus atteints, qu'on appelle proposant (ici 4 atteints sur 5 possibles) et également dans la population générale (ici 1 atteint sur 5 possibles).

Cette mesure permet d'estimer le taux de concentration familiale d'une pathologie. Son objectif est d'identifier les maladies ayant une forte probabilité d'avoir des bases génétiques même si la concentration familiale peut également être due à des facteurs environnementaux partagés.

D.III.9.b) *Relation entre λ_S dans un cas multi-loci*

Pour un ensemble de N SNPs associés à une pathologie, si p_j et OR_j sont le RAF (fréquence de l'allèle augmentant le risque) et l'Odds Ratio (OR) correspondant pour le SNP $_j$ [Risch, Merikangas,**1996**][Risch,**1990**], le λ_s du à ces SNPs s'obtient par l'équation suivante :

$$\lambda_S = \prod_{j=1}^{N} \left[1 + \frac{p_j(1-p_j)(OR_j - 1)^2}{2\left[(1-p_j) + p_j OR_j\right]^2} \right]^2$$

Équation 11 : Risque relatif pour un germain – cas multi-locus

Il s'agit d'un modèle simplifié, faisant l'hypothèse d'un effet multiplicatif des polymorphismes génétiques. Les relations entre le λ_s et l'odds-ratio sont discutées de façon plus étendue dans [Rybicki, Elston,**2000**].
En général, on compare le λ_S dû à un ensemble de facteurs de risque identifiés (dans notre cas les génotypes) au λ_S estimé en population.

D.III.9.c) *Héritabilité en fonction du λ_s :*

Si l'on revient à la notion de variable de susceptibilité latente L décrite en (D.III.8.d.1), nous nous intéressons à la proportion de la variance (où autrement dit l'héritabilité) expliquée par l'ensemble de N SNPs déjà mis en évidence. Cette valeur est directement liée à l'augmentation de risque λ_S due à cet ensemble de SNPs.
Soit K la prévalence, la variance de la variable de susceptibilité L expliquée par un ensemble de SNPs, à partir de leur λs est [Wray et al.,**2010**][Reich et al.,**1972**] :

$$h_L^2 = \frac{2[T - T_1\sqrt{1-(T^2 - T_1^2)(1-T/\omega)}\,]}{\omega + T_1^2(\omega - T)}$$

Équation 12 : Héritabilité par λ_s

Dans cette expression, $T=\boldsymbol{\Phi}^{-1}(1-K)$, $T_1=\boldsymbol{\Phi}^{-1}(1-\lambda_S K)$, et $\omega=z/K$, où z est la hauteur d'une densité Gaussienne standard à la valeur T.

D.IV. Génétique Inverse

Pour identifier les causes des pathologies humaines, la stratégie de génétique inverse propose de rechercher les variations génétiques liées à une pathologie - principalement à travers une analyse de liaison (D.IV.1, p. 34) ou une analyse d'association (D.IV.2, p. 38) - dont la protéine défectueuse est inconnue. Elle s'oppose à une stratégie de génétique classique qui visait à identifier le gène responsable d'une pathologie par l'observation d'un phénotype et des hypothèses fonctionnelles en découlant. La mise en évidence du rôle de l'Insuline dans la survenue du diabète en est un très bon exemple. La purification d'extraits de pancréas (organe dont l'importance dans la survenue du diabète a été découverte fortuitement) a permis l'identification de l'insuline. L'effet de cette protéine sur la survenue du diabète a été testé par injection à un chien diabétique. On a alors observé une disparition des symptômes de cette maladie. On en a déduit que des mutations du gène de l'Insuline pourraient avoir un effet sur la survenue du diabète.
La même découverte selon une approche de génétique inverse, telle que je la définis ici, serait advenue par la mise en évidence d'une mutation dans le gène de l'insuline, puis d'une recherche de la séquence de la protéine responsable, de sa localisation dans le pancréas, et enfin par la recherche de son rôle physiologique.
Le terme génétique inverse est assez ancien, datant de l'époque où les méthodes de biologie moléculaire ne permettaient pas la détermination d'un grand nombre de génotypes pour un grand nombre d'individus.
L'approche « gène candidat », qui visait à utiliser des approches comme l'analyse de liaison ou l'analyse d'association près et dans des gènes identifiés comme candidats à partir des hypothèses fonctionnelles des chercheurs, basée sur les caractéristiques d'une pathologie, peut être considérée comme un mélange des deux approches.

D.IV.1. Analyse de Liaison génétique

L'analyse de liaison utilise des marqueurs génétiques répartis sur les chromosomes pour localiser les gènes qui influent sur l'expression d'un phénotype, comme la survenue d'une

maladie [Bernstein,**1931**][Haldane,**1936**]. Leur position sur le génome est connue alors que celle des gènes (et donc de la mutation de ce gène responsable de la pathologie) qui nous intéressent ne l'est pas, la plupart du temps. Pour identifier la position d'une telle mutation, nous testons la transmission conjointe entre les génotypes des marqueurs génétiques et les génotypes de cette mutation (inférés comme expliqués dans le paragraphe D.IV.1.a), p. 35, dans une famille, ou, plus généralement, dans un ensemble d'individus dont nous connaissons les relations familiales (paires d'individus apparentés). Les allèles des marqueurs permettent de définir des régions chromosomiques transmises en même temps que la maladie.

Classiquement, on distingue l'analyse de liaison paramétrique et l'analyse de liaison non paramétrique.

D.IV.1.a) *Analyse de liaison paramétrique*

Dans l'analyse de liaison paramétrique, nous sommes amenés à calculer explicitement le taux de recombinaison (D.III.6, p. 24) entre la mutation (locus) responsables de la pathologie et les marqueurs génétiques. Nous recherchons les marqueurs génétiques qui présentent la recombinaison minimale avec cette mutation.
Une quantité, le Lod-score, est créée, qui compare la vraisemblance (notion expliquée Figure 15) des données observées (phénotypes des individus d'un pedigree et génotypes au niveau des marqueurs) et inférées (génotypes inférés pour la mutation causale) sous H0 (pas de liaison, soit $\theta = 0.5$) et sous H1(liaison, soit $0.5 > \theta_1 \gg 0$).

$$Z_i(\theta_i) = log_{10}\left[\frac{L_i(\theta=\theta_1)}{L_i(\theta=0.5)}\right]$$

Équation 13 : Lod-Score

En pratique, on calcule ce Lod-Score pour différentes valeurs de θ et on choisit celle qui maximise le Lod-score.

Une des difficultés de cette méthode est la nécessité d'utiliser le génotype de la mutation causant la pathologie alors que nous voulons précisément identifier cette mutation. Dans l'analyse paramétrique, nous allons, à partir d'un modèle génétique (défini en D.III.8.d) déduire la probabilité pour un individu de la famille d'être porteur d'un génotype donné (Figure 15). La connaissance de ce modèle provient, classiquement, de l'analyse de ségrégation [MORTON,**1958**]. Dans l'exemple de la Figure 15, p. 36, nous sommes en mesure d'inférer de façon quasi certaine le génotype au niveau de la mutation, en raison du caractère extrême du modèle. Pour des modèles plus complexes, les méthodes classiques d'analyse chercheront toutes les combinaisons génotypiques possibles (en fonction du modèle et du phénotype) et le Lod-score sera une somme des Lod-scores pour chaque configuration familiale pondérée par la probabilité de chaque configuration.

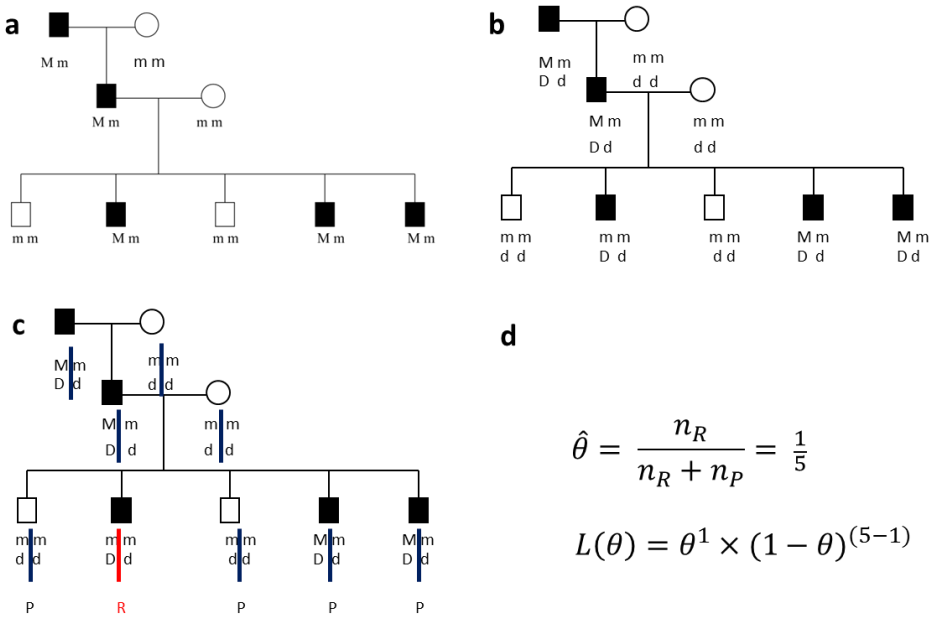


Figure 15 : Liaison, Recombinaison et Vraisemblance

a/ Nous disposons d'un pedigree avec les statuts (noir pour patient atteint, blanc pour membre de la famille non atteint) et d'un marqueur (allèle M et m) génotypé pour tous les individus. Par ailleurs nous avons un modèle M supposant l'effet d'une mutation (d'allèles D et d) avec trois pénétrances, une fréquence de l'allèle délétère et (a). M <- {q=0.0001, P_{DD}=0, P_{Dd}=1, P_{dd}=1}. **b/** Ce modèle, dont on suppose qu'il a été identifié dans des études épidémiologiques, nous permet d'inférer de façon (quasi) certaine le génotype de la mutation chez tous les membres de la famille. Les individus non atteints sont forcément dd. Par ailleurs, du fait de la rareté de l'allèle délétère en population, il est pratiquement impossible que les individus atteints soient double porteurs DD. Il est donc possible, sous ce modèle extrême (pour raisons didactiques), d'assigner les génotypes au locus de la mutation.
c/ Ayant les données sur trois générations, il est aisé de savoir quels allèles ont été hérités de la grand-mère et du grand-père. Pour la mère, on connaît les haplotypes car les deux loci sont homozygotes. Pour le père, il est également facile d'identifier les allèles grand-paternels et grand-maternels et donc la phase. Ceci permet de voir, chez chacun des 5 enfants, s'il y a eu recombinaison ou non des chromosomes paternels. Les chromosomes recombinants sont marqués R et les chromosomes parentaux (en grand probabilité non recombinants) marqués P. Ici, 4 des 5 chromosomes paternels transmis n'ont pas eu de recombinaisons. L'estimation du taux de recombinaison (en **d**) est évidente (1/5). La vraisemblance $L(\theta)$ des données (nombre de recombinaisons estimées, statuts des membres du pedigree, modèle de transmission génétique) en fonction du taux de recombinaison θ peut alors être estimée **(d)**. On estime la probabilité d'observer 1 recombinant (R) et 4 non-recombinants pour un ensemble de valeurs de θ.

Le calcul de la vraisemblance donné dans la Figure 15, (ci-dessus), permet de calculer les Lod-scores et de choisir le θ pour lequel il est maximum (Tableau 2).

	H0 : θ = 0.5	**H1 : θ = 0.3**	**H1 : θ = 0.2**	**H1 : θ = 0.1**	**H1 : θ = 0**
L(θ)	0.03125	0.07203	0.08192	0.06561	0
Lod-Score (Équation 13)		0.3626634	0.4185399	0.32212	$-\infty$

Tableau 2: Lod-score

La vraisemblance est maximale pour θ=0.2. Cela n'est pas étonnant car l'estimation directe de θ est égale à 0.5 (Figure 15). Le Lod-score est, de façon automatique, maximum pour cette valeur. Nous observons également que la vraisemblance est nulle pour θ=0 et Lod-score est infiniment petit. En effet, l'hypothèse d'un taux de recombinaison nul est très exigeante et donc très sensible à des erreurs de modèle et de génotypage. De plus, nous avons observé au moins une recombinaison, ce qui rend impossible un θ égal à 0.

En fait, le plus souvent, on ne connaît pas le mode de transmission de la pathologie, ni la fréquence allélique de l'allèle délétère et encore moins les pénétrances. L'ignorance de ce modèle peut mener à une incapacité à identifier la liaison lorsqu'elle existe [Clerget-Darpoux et al.,**1986**]. Plusieurs alternatives ont été proposées, dont les plus connues sont la maximisation sur un ensemble de modèles possibles – méthode MOD-Score [Risch,**1984**][Clerget-Darpoux et al.,**1986**] et l'analyse de liaison non-paramétrique [Penrose,**1935**].

D.IV.1.a) *Analyse de liaison non-paramétrique*

Les méthodes non paramétriques recherchent également une corrélation entre la ressemblance au niveau du marqueur génétique et la ressemblance au niveau du phénotype, pour des individus apparentés. L'idée d'observer cette corrélation directement, sans passer par l'utilisation d'un modèle est presque aussi ancienne que l'analyse de liaison elle-même [Penrose,**1935**].
La ressemblance génétique est mesurée, en général pour une paire d'individus apparentés, par l'Identité par descendance (IPD) qui est appelée en anglais «Identity By Descent (IBD) ». Pour une paire d'individus, le nombre d'allèles identiques par descendance à un marqueur donné est le nombre d'allèle que ces individus ont reçu en commun de leur ancêtre commun (Figure 16).

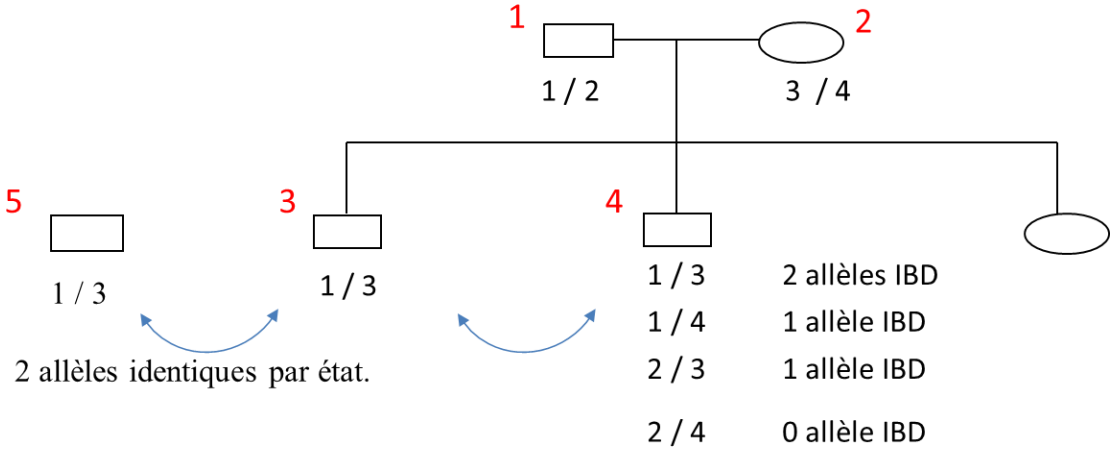


Figure 16 : Statut IBD

La transmission de ce marqueur à 4 allèles est suivie à travers cette famille nucléaire (parents 1 et 2 et enfants 3 et 4). Si l'on suppose que l'enfant 3 a reçu les allèles 1 et 3 de ses parents 1 et 2 respectivement, il existe trois statuts d'identité par descendance possibles à partir de quatre génotypes possibles transmis. Le germain 4 peut recevoir les mêmes allèles, 1 et 3, et la paire présente deux allèles identiques par descendance. Si le germain 4 reçoit les allèles 1/4 ou 2/3, la paire alors ne présente plus qu'un seul allèle identique par descendance (respectivement l'allèle 1 du père ou l'allèle 3 de la mère). Enfin, si le germain 4 reçoit les allèles 2 / 4, la paire n'a aucun allèle identique par descendance.
Les 4 configurations IBD sont équiprobables et il en résulte que P(IBD=2)=1/4, P(IBD=0)=1/4 et P(IBD=1)=2x1/4=1/2.
Enfin, il est important de distinguer le statut d'identité par descendance (IBD) du statut d'identité par état. Ici, la paire constituée des individus 3 et 5 a deux allèles semblables (identique par état) mais en l'absence d'un lien familial, on ne peut pas déterminer s'ils sont hérités d'un ancêtre commun et donc s'ils sont identiques par descendance.

La ressemblance au niveau du phénotype peut prendre diverses formes. Dans le papier original de Penrose (1935), il s'agissait de tester le partage d'allèles pour des paires de germains atteints. La ressemblance est une fonction qui prend la valeur 1 lorsque les deux membres d'une paire sont atteints et 0 – donc non compté – pour toutes les autres paires. Le test le plus connu portait sur le carré de la différence entre les phénotypes des individus d'une paire d'apparentés [Haseman, Elston,**1972**].

Un grand nombre de mesures de ressemblance et de tests statistiques sont proposés et leurs propriétés respectives en termes de puissance et de robustesse sont l'objet de recherches poussées [Holmans,**1993**].

Dans le cas d'un phénotype binaire, la ressemblance entre individus apparentés se traduit par trois statuts de paires : paires concordantes atteintes, paires discordantes et paires concordante non-atteintes. La stratégie la plus populaire a longtemps été l'analyse des paires d'apparentés atteints et, principalement, des germains atteints (nommée « Affected Sib Pairs – ASP ») [Fishman et al.,**1978**], qui est celle utilisée dans le papier original de Penrose. Sous l'hypothèse de non liaison (H_0), nous connaissons les proportions attendues des trois états IBD (1/4,1/2,1/4) (légende de la Figure 16). Il est alors possible d'estimer cette distribution dans des paires de germains atteints et de comparer ces deux distributions par un χ^2 de conformité. Certaines méthodes s'intéressent directement à la proportion d'allèles identiques par descendance, qui se déduit des trois probabilités d'IBD et est d'un maniement plus pratique puisqu'il s'agit d'une seule valeur (Équation 14).

$$\pi = 0 \times P(IBD = 0) + \frac{1}{2} \times P(IBD = 1) + 1 \times P(IBD = 2)$$

Équation 14 : Proportion moyenne d'allèles identiques par descendance

Il est important de noter que dans l'analyse dite non-paramétrique, nous sommes en fait amenés à faire l'hypothèse que tous les individus sont porteurs de la mutation causale et que les non porteurs ne sont pas porteurs [Göring, Terwilliger,**2000**]. De façon paradoxale, ce test paramétrique peut-être en final interprété comme un test paramétrique extrêmement contraignant.

L'analyse de liaison est pleinement efficace lorsqu'une mutation dans un seul et même gène de susceptibilité ségrége dans toutes les familles de l'échantillon, et qu'il est la cause nécessaire et suffisante de la survenue de la maladie. Elle reste efficace s'il y a hétérogénéité à l'intérieur d'un gène (allèles différents entre familles) car ce qui est testé est le partage de zones chromosomiques à l'intérieur de chaque famille. Cette conformation n'est cependant sans doute pas le cas des maladies multifactorielles [Martin et al.,**1992**] où l'on s'attend à une hétérogénéité importante et impliquant plusieurs loci.
Il est important de noter que la méthode d'identité par descendance est non seulement extensible à des paires d'apparentés plus éloignés [Weeks, Lange,**1988**] – initialement proposé en utilisant le statut IBS, d'identité par état - mais peut également être utilisée pour des apparentés dont on ne connaît pas nécessairement de façon explicite le degré d' apparentement. Cette idée est à la base de stratégies d'analyse de liaison en population développées récemment [Purcell et al.,**2007**][Gusev et al.,**2011**][Browning, Thompson,**2012**].

D.IV.2. Analyse d'Association Génétique

D.IV.2.a) *Principe*

L'analyse d'association teste, comme l'analyse de liaison, une corrélation entre un marqueur et un phénotype d'intérêt.
Le principe des études d'association génétique consiste à comparer les distributions alléliques (ou génotypiques) au niveau d'un marqueur génétique en fonction du phénotype d'un groupe d'individus. Par exemple, si nous avons des patients et des témoins, nous rechercherons des marqueurs où les fréquences alléliques sont significativement différentes entre ces deux groupes. Si ces distributions sont significativement différentes, alors on conclut qu'il existe une association entre l'allèle (ou le génotype) le plus fréquent chez les cas et la maladie étudiée [Balding,**2006**].

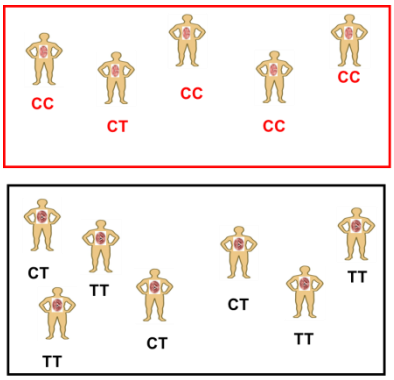


Figure 17 : Principe de l'analyse cas témoins génétique

Pour un marqueur génétique donné, on estime les fréquences alléliques chez les patients (groupe rouge) et les témoins (groupe noir). Dans l'analyse cas-témoins nous définissons une statistique comme une fonction F de la différence de fréquences d'un allèle dans les deux groupes.

Je présente l'analyse d'association pour des marqueurs bi-alléliques de type SNP car ce sont pratiquement les seuls marqueurs étudiés dans cette thèse. L'analyse d'association est parfaitement utilisable avec de marqueurs plus complexes (multi-alléliques) et les mêmes modèles (avec quelques variantes) peuvent être appliqués.

D.IV.2.b) *Modélisation statistique d'un effet génétique*

La façon la plus directe et naturelle de résumer le contraste entre génotypes (D.IV.2.a), p. 38) est d'estimer la différence de fréquences génotypiques entre ces deux groupes. La statistique résumant cette différence suivra une loi de distribution de probabilité théorique prédéterminée. La comparaison à cette distribution permet d'évaluer la force de cette association, par exemple par une p-valeur.

La table peut s'écrire sous la forme suivante, pour une analyse avec N1. témoins et N2. cas, pour un SNP donné :

	TT	**CT**	**CC**	
Témoins	N_{11}	N_{12}	N_{13}	$N_{1.}$
Cas	N_{21}	N_{22}	N_{23}	$N_{2.}$
	$N_{.1}$	$N_{.2}$	$N_{.3}$	N

Tableau 3 : répartition cas - témoins

En cas d'indépendance, l'effectif attendu pour une case donnée *ij* est $E_{ij=}\frac{p_i \times p_j}{N}$.

Ce test génotypique n'impose pas de contraintes sur le modèle, c'est-à-dire que les augmentations de risque des génotypes CT et TT sont indépendantes. La statistique évaluant la force de l'association entre trait et génotypes est :

$$S_1 = \sum_i \sum_j \frac{(N_{ij} - E_{ij})^2}{E_{ij}} \sim X^2\ (2\ ddl)$$

La façon la plus classique de tester si S est suffisamment grand pour conclure à une association est de calculer sa p-valeur (D.IV.2.c), p. 42) .*S* suit une suit une loi du χ^2 avec 2 degrés de liberté.

En général, on peut faire des hypothèses sur les effets des allèles afin de simplifier ce modèle. Par exemple, on suppose que le risque augmente constamment en fonction du nombre d'allèle (ou diminue constamment en fonction du nombre d'allèles complémentaires). Dans ce cas, il est possible de tester un modèle moins général et ainsi de s'assurer une certaine puissance, dans le cas où le modèle est proche de la réalité.

Nous allons principalement utiliser une extension de cette méthode, la régression logistique. Si M est une variable aléatoire notée 1 si un individu est malade et 0 s'il est témoin :

$$\ln\left(\frac{P(M=1/X,G)}{1-P(M=1/X,G)}\right) = \alpha + B_g \times G + B_c \times X$$

Équation 15 : modèle logistique

X est une matrice de covariables non génétiques (âge, sexe, statut fumeur ...) et G un codage du génotype au SNP testé. En cas de non association, B_g est égale à 0. Son estimation n'est donc pas significativement différente de cette valeur.

$$P\left(M=1/_{X,G}\right) = \frac{e^{\alpha+B_g\times G+B_c\times X}}{1+e^{\alpha+B_g\times G+B_c\times X}}$$

Équation 16

Dans le test le plus simple, G est égal au nombre d'allèles de référence (la référence étant arbitraire) chez un individu.

	TT	**CT**	**CC**	**TT**		**CT**		**CC**	
				G_1	G_2	G_1	G_2	G_1	G_2
Général	1	RR_1	RR_2	0	0	1	1	2	0
Multiplicatif	1	RR_1	RR_1^2	0	0	1	0	2	0
Dominant	1	RR_1	RR_1	0	0	1	0	1	0
Récessif	1	1	RR_1	0	0	0	0	1	0

Tableau 4 : codage du génotype sous différents modèles

Suivant le modèle génétique attendu, on code de façon différente, dans la régression logistique, le génotype d'un individu. Dans la partie gauche du tableau, nous montrons l'augmentation (ou diminution) du risque par rapport au génotype de base TT (résumé par le risque relatif). Dans la partie droite, pour chaque génotype, il y a deux variables G1 et G2 qui résument le génotype à un SNP. Dans le cas du modèle « Général », la variable G1 est le nombre d'allèles de référence (ici **C** choisi arbitrairement) et la variable G2 est un indicateur d'hétérozygotie. Les modèles dominants et récessifs font l'hypothèse qu'il y a domination de l'effet d'un des deux allèles. Il n'y a plus qu'un seul paramètre (ou risque relatif).

Dans le Tableau 4 où est présenté un sous ensemble des modèles génétique existants, le modèle «Général» est l'équivalent du test de *S* sans aucune contrainte, présenté précédemment.

Nous allons tester si le (ou les) coefficient(s) β est significativement différent de 0.
Ces coefficients sont obtenus par maximisation de la vraisemblance. La contribution d'un individu i (Y_i pouvant prendre les valeurs 0 ou 1 – par exemple cas ou témoin) à la vraisemblance est la probabilité pour cet individu de faire partie d'une catégorie (Équation 17) :

$$\prod_i P(Y_i = 1/X,G)^{Y_i} \times [1 - P(Y_i = 1/X,G)]^{1-Y_i}$$

Équation 17 : vraisemblance de la loi logistique (1 individu i)

La vraisemblance totale pour un groupe de N individus s'écrit donc comme le produit de vraisemblances individuelles (Équation 18) :

$$V = \prod_{i=1..N} P(Y_i = 1/\ X, G)^{Y_i} \times [1 - P(Y_i = 1/X, G)]^{1-Y_i}$$

Équation 18 : vraisemblance de la loi logistique (1 échantillon)

La vraisemblance dépend donc des valeurs des coefficients de régression B_g et B_c (Équation 16), qui sont les paramètres à estimer pour trouver le maximum de vraisemblance.
Dans la pratique, les logiciels utilisent une procédure approchée pour obtenir une solution satisfaisante de la maximisation. Les résultats dépendent de l'algorithme utilisé et de la précision adoptée lors du paramétrage du calcul mais pour tester l'association, il est important d'avoir la vraisemblance sous l'hypothèse nulle et, dans la plupart des cas, le maximum de la vraisemblance (et les paramètres pour lesquels cette vraisemblance est maximale) – (Figure 18).

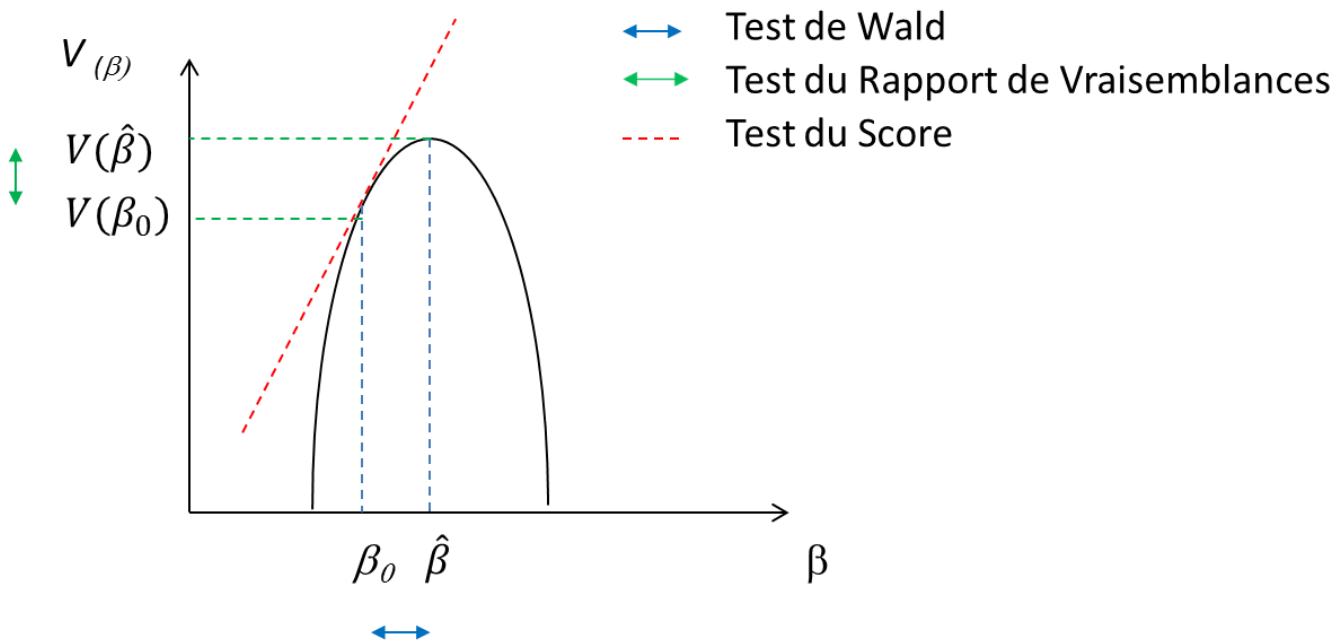


Figure 18 : Différents tests à partir de la vraisemblance
Représentation de la fonction de vraisemblance (Équation 18) lorsqu'il n'y a qu'un seul paramètre à estimer (la courbe est en une dimension). Dans ce cas simple, on peut balayer le spectre des valeurs de β et trouver la valeur de β qui permet de maximiser la vraisemblance ($\hat{\beta}$). L'association génétique peut alors être testée de trois façons : en comparant β à β_0 (test de Wald), en comparant $V(\beta)$ à $V(\beta_0)$ (test du Rapport de Vraisemblance) et enfin en comparant l'angle de la tangente à la fonction V au point $\beta=\beta_0$ à 0 (test du score). Dans le test du score, si $\hat{\beta}$ est près de β_0, la tangente est horizontale.

Dans le contexte de mon travail de thèse, je me suis principalement intéressé à l'association d'un variant sous un modèle multiplicatif (Tableau 4), l'effet des autres variables étant sans importance (même si ces variables doivent être inclues pour prendre en compte les possibles co-variables). J'utiliserai donc principalement le test de Wald sur le coefficient de régression du génotype:

$$S_2 = \frac{\hat{\beta}_g}{\sigma(\widehat{\beta_g})} \sim N(0,1)$$

Équation 19 : Z-score ou statistique d'association d'un marqueur (test de Wald)

Lorsqu'on veut tester plusieurs effets de façon simultanée (par exemple pour tester l'effet simultané de deux génotypes), on applique un test de rapport de vraisemblance. Il suffit de comparer $V(\hat{\beta})$ et $V(\beta_0)$ (Équation 20).

$$W = 2 \times (\log(V(\hat{\beta}) - \log(V(\beta_0)) \sim x_n^2$$

Équation 20 : statistiques d'association – rapport de vraisemblance

Le rapport de vraisemblances W suit une loi de χ^2 avec un nombre de degrés de liberté n. n est la différence entre le nombre de paramètres estimés sous H0 et sous H1. En pratique pour notre test d'association génétique, il s'agit du nombre de SNPs pour lesquels on teste l'association de façon simultanée.

D.IV.2.c) *Significativité statistique*

Cette notion sera reprise plus en détail en D.VI.4, p. 52, mais pour chaque statistique *S* mise en place, qui mesure l'association entre un phénotype et un génotype, nous cherchons à obtenir la probabilité d'observer *S* sous l'hypothèse de non association : nous l'appellerons ici p-valeur.

Il y a plusieurs possibilités pour obtenir cette p-valeur : de façon asymptotique, de façon exacte ou par permutation. En règle générale, toute statistique *S* suit une loi de probabilités, c'est-à-dire que pour une valeur de **s** donnée parmi les valeurs possibles de S, on saura sa probabilité de survenue ainsi que la probabilité d'observer une valeur supérieure à **s** (et bien sûr la probabilité d'observer une valeur inférieure). Si nous connaissons cette loi, il est alors possible d'en tirer sur des tables ces probabilités (D.IV.2.c.1), p. 42). Si ce n'est pas le cas, alors nous pourrons estimer sa loi de probabilités de façon empirique (D.IV.2.c.2), p. 42).

D.IV.2.c.1) Test asymptotique

Le test asymptotique est possible lorsque la statistique qu'on a choisie suit une loi de probabilité connue. Ainsi, comme expliqué précédemment, la statistique *S* suit une loi du χ^2 à deux degrés de liberté. La statistique issue de l'analyse de régression logistique suit une loi normale à 1 degré de liberté.

D.IV.2.c.2) Les permutations

Lorsqu'on ne connaît pas la loi d'une statistique, il est possible de l'estimer empiriquement par des méthodes d'échantillonnage de type Monte-Carlo. Dans les études cas-témoins, on peut réaliser chaque simulation en permutant à chaque fois les statuts cas-témoins de façon à simuler sous l'hypothèse d'indépendance. Cette stratégie s'avère très intéressante car elle permet d'étudier le comportement de la distribution entière des statistiques spécifiquement pour notre étude. Elle permet par exemple d'estimer la distribution de la p-value la plus petite de notre expérience de façon adaptée au nombre de phénotypes, des marqueurs et aux corrélations entre marqueurs et phénotypes.

Dans le cas d'analyse d'association génome entier qui nous occupe ici, nous pouvons permuter le statut malade/non malade (ou le vecteur des statuts pour chaque individu).

D.V. Les projets Génomiques

D.V.1. Carte génomique

Cette approche a été rendue possible à l'échelle du génome entier grâce au projet HapMap. Il a permis le développement de cartes denses de SNPs et l'obtention d'informations sur leur fréquence, leur répartition physique, et sur les groupes de SNPs en déséquilibre de liaison [Ardlie et al.,**2002**][Frazer et al.,**2007**][The International HapMap Consortium,**2005**]. Cette structure avait été mise en évidence sur des études pilotes [Abecasis et al.,**2001**][Tishkoff, Verrelli,**2003**][Chakravarti et al.,**1984**], et d'abord par une étude initiale sur 500 individus et s'est confirmée depuis.

Ce phénomène de déséquilibre de liaison, permet de dire que si un SNP varie dans un bloc, les autres SNPs du bloc varient également. Ainsi, le génotypage (onéreux) de l'ensemble des SNPs d'un bloc peut être évité. La plupart des SNPs étant localisés dans des blocs clairement définis grâce au projet HapMap, le génotypage (onéreux) de l'ensemble des SNPs d'un bloc peut être évité ce qui permet de limiter le nombre de SNPs étudiés («TagSNP») dans les GWAS.

Le but de l'étude d'association est de détecter le ou les variants génétiques prédisposant à une pathologie en génotypant des marqueurs génétiques (SNP) en déséquilibre de liaison avec ces variants, et dont la position chromosomique est connue.

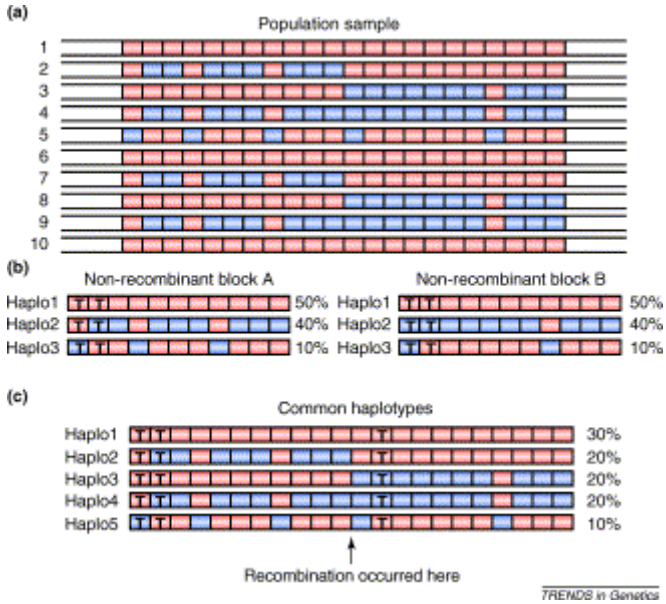


Figure 19 : Blocs haplotypiques

a/ Une « population » d'haplotypes (par exemple identifée dans un échantillon d'individus de HapMap) constitués de 22 SNPs (un carré représente un SNP). Nous voyons que les haplotypes 1,6 et 9 sont identiques ; 2 et 7 sont identiques ; 3 et 8 sont identiques ; 4 et 9 sont identiques. Sur toutes les combinaisons possibles (2^{22}), il y a donc très peu de variantes haplotypiques observées. **b/** De plus, il y a un niveau de régularité encore plus bas puisqu'en coupant après le 11^{e} SNP, pour chaque « bloc », on n'observe que 3 haplotypes possibles. Il n'y a pas de trace d'une possible recombinaison historique à l'intérieur de chaque bloc. Le déséquilibre de liaison dans chaque bloc est donc très fort. Il y a eu par contre recombinaison à un moment donné entre le dernier SNP du Bloc 1 et le premier SNP du bloc 2. Toutefois, **c/**, on n'observe pas d'haplotypes à 22 SNPs représentant toutes les combinaisons d'haplotypes à 11 SNP. On n'en observe que 5 sur 9. Cela signifie que le nombre de recombinaisons entre ces deux blocs est bas également et donc que le déséquilibre de liaison entre les deux blocs est aussi important (bien qu'il soit moins fort que celui observé à l'intérieur des blocs). Adapté de [Cardon, Abecasis,**2003**]

On peut représenter cette structure de plusieurs façons. L'option la plus courante est de montrer une matrice où la mesure du LD est représentée par un niveau de couleur (Figure 20, p. 14).

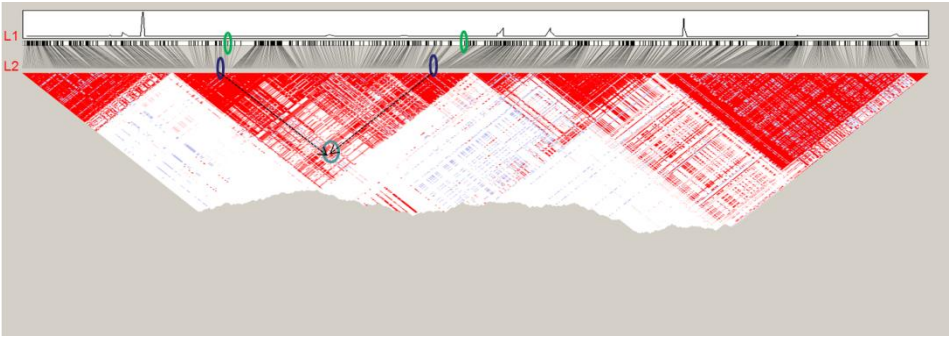


Figure 20 : Représentation du Déséquilibre de Liaison

Dans cette région, sur la ligne L1, un trait représente la position d'un SNP (entourés en vert). Les SNPs sont représentés ensuite sur la ligne L2 de façon à ce qu'ils soient espacés de façon régulière (pour les besoins de la représentation – entourés en bleu foncé). Le DL, deux à deux, des SNPs est représenté sous la forme d'une demi-matrice. Un très fort DL est représenté par un rouge très fort. Sur la figure, le point entouré en bleu clair et pointé par les deux flèches représenté le DL entre les deux SNPs choisis. Adapté à partir du site HapMap (www.hapmap.org).

D.V.2. Le projet Génome Humain :

Le Projet Génome Humain a débuté en 1990. Son objectif était le séquençage de 3 milliards de bases du génome pour 2005 afin de faciliter le clonage de gènes impliqués dans les pathologies héréditaires, et de comprendre l'association entre génotypes et phénotypes.
Le séquençage complet du génome humain a été entrepris simultanément par un consortium public international piloté par le National Institute of Health et la compagnie privée Celera. Près de 300 séquenceurs automatiques ont été utilisés chez Celera et environ 600 pour le projet public répartis dans plusieurs laboratoires du monde entier. Les objectifs fixés au commencement ont été atteints avec 2 ans d'avance, la séquence initiale du génome ayant été publiée en 2001 [Venter et al.,2001][Lander et al.,2001] et la séquence complète en 2003 (2,86 milliards de bases) [International Human Genome Sequencing Consortium,2004].
Les grandes découvertes issues de ce projet ont été le faible nombre de gènes humains (26 000 seulement face aux 300 000 attendus) et la faible quantité de variation du génome entre 2 individus (0.1%).
Ce projet a été le point de départ de la caractérisation du génome humain qui se poursuit aujourd'hui avec le projet 1 000 génomes, mais il a également ouvert le séquençage aux génomes d'espèces différentes. On peut désormais dénombrer près de 2 500 génomes (559 eucaryotes et 1944 procaryotes) séquencés ou en cours de séquençage (données issues du « NCBI Entrez Genome Project »).
Dans le cadre de mon travail de thèse, le Projet Génome Humain a permis d'identifier les positions de toutes les séquences d'ADN chez l'Homme et ainsi de pouvoir positionner sur le génome les variations génétiques qui ont été découvertes par la suite et estimer leurs relations en termes de déséquilibre de liaison, notamment dans le projet HapMap/1000 Genomes.

D.V.3. Le projet HapMap.

Le projet HapMap visait à établir un catalogue des variations génétiques les plus fréquentes chez l'humain et décrire leurs positions, leur distribution et fréquence allélique dans les différentes populations humaines ainsi que leur déséquilibre de liaisons par paires.
Un des objectifs principaux était de permettre aux chercheurs, à partir de ces informations, de mener des études d'association, principalement à l'échelle du génome, afin de mettre en évidence des liens entre variations génétiques et les risques de maladies humaines.

Dans ce projet, 270 individus originaires de trois continents (Amérique du Nord, Afrique, Asie) ont été caractérisés génétiquement. Les données Nord-Américaines et Africaines consistaient en trio (deux parents et leur enfant). Il s'agit de 30 trios de résidents des Etats-Unis originaires de l'Europe du Nord et de l'Ouest et disponibles à travers le Centre d'Etudes du Polymorphisme Humain et de 30 trios issus de la population Yoruba d'Ibadan au Nigéria.
Les deux autres échantillons sont composés de 45 individus non apparentés originaires de de la région de Tokyo ainsi que 45 individus également non apparentés originaires de la région de Pékin.

Même s'il est clair que les populations respectives (codée CEU pour les Européens, YRI pour les Yorubas et ASN pour l'ensemble chinois/japonais) ne sont pas strictement représentatives des populations de leurs continents, ils ont été souvent utilisés dans ce but.

Dans la première étape de l'analyse (HapMap I), 1.3 million de SNPs ont été caractérisés et génotypés [The International HapMap Consortium,**2005**]. Dans une deuxième étape (HapMap II), le projet présentait 3.4 millions de SNPs génotypés en total [International HapMap Consortium et al.,**2007**]. Le déséquilibre de liaison entre ces marqueurs génétiques a été calculé. La connaissance de la structure du déséquilibre de liaison a permis d'identifier les groupes de SNPs en haut déséquilibre de liaison (dont les génotypes sont fortement corrélés) et plus généralement de réaliser une carte génomique dans chaque population (D.V.1). Dans ces groupes, il était possible de sélectionner un seul SNP représentant les SNPs de son groupe. Ces SNPs sont appelés, en anglais, des « tagSNPs ».
Pour les deux premières versions de ce projet (arrivant à au total de 3.4 millions de SNPs), les SNPs à génotyper étaient choisis à partir de bases de données existantes (principalement dbSNP) et devaient être fréquents (MAF > 5%). Leur représentativité par rapport à l'ensemble de SNPs fréquents du génome (non connu) n'était pas garantie même si elle était très probable.

Il est à noter qu'une phase additionnelle de ce projet a été développée - HapMap 3 [International HapMap 3 Consortium et al.,**2010**]. Dans cette phase, 1.6 millions de SNPs, sélectionnés à partir de puces commercialisées, ont été génotypés dans 1184 individus issus de 11 populations différentes. L'intérêt était double : il s'agissait de comparer la structure de déséquilibre de liaison entre ces différentes populations et de disposer d'un panel de référence plus grand dans chaque ensemble cohérent de populations (Européens, Africains, Asiatiques) afin d'augmenter la précision de l'estimation de ce déséquilibre de liaison.

D.V.1. Le projet 1000 génomes.

Dans le cadre des études basées sur les nouvelles méthodes de séquençage (« Next Generation Sequencing » appelé « NGS »), qui réduisent considérablement le prix du séquençage direct du génome, il est nécessaire de pouvoir différencier les variants fréquents (présents dans les bases de données) des variants rares. Or les bases de données existantes (dbSNP - http://www.ncbi.nlm.nih.gov/SNP/) sont riches en information mais ne sont pas exhaustives. Nous avons d'ailleurs vu que même les variations fréquentes n'étaient pas toutes connues et on ne pouvait exclure le fait que certains SNPs fréquents ne soient pas bien corrélés aux SNP déjà existants.
C'est pourquoi, entre autres, en janvier 2008, un consortium international a annoncé le lancement du projet 1000 génomes (http://www.1000genomes.org). L'objectif de cette collaboration est de produire un catalogue complet des variations génétiques humaines (SNP et variations du nombre de copie d'ADN dans leur contexte haplotypique) de fréquences égales ou supérieures à 1%. Le projet a débuté avec 3 projets pilotes qui vont guider la stratégie de l'étude finale et les méthodes d'analyse. Ces projets combinent la lecture du génome entier de près de 200 individus, le séquençage de 2 trios (parents-enfant) et une étude complète de tous les exons de 900 génomes. Le plan pour le projet entier est

de séquencer le génome d'environ 2 000 individus avec une profondeur de lecture de 4X. La première partie, qui est presque achevée, inclut 1101 échantillons provenant de 12 populations différentes (Europe, Asie de l'Est et du Sud, Afrique de l'Ouest, Amérique).

Ce projet des 1000 génomes devrait permettre d'accroître considérablement l'efficacité des études de génétique médicale. Ils devraient aider à la recherche de l'effet de variants rares dans les pathologies humaines mais également d'augmenter, par rapport au projet HapMap, l'information apportée aux études d'association à l'échelle du génome entier («Genome-Wide Association Study»).
Dans le cadre de ma thèse, les données issues de ce projet ont été utilisées pour l'imputation génétique mais également pour pouvoir mieux estimer l'origine géographique des individus et enfin comparer des fréquences alléliques de la population générale avec mes patients issus des mêmes régions.
En se conformant aux procédures d'éthique les plus complètes, le Projet 1000 Génomes utilisera des échantillons venant de donneurs volontaires.

Code	Population	Continent	HM 1 - 2	HM 3	1000 G
ASW	Africains Américains, Etats-Unis	Afr		Oui	Oui
CEU	Européens de l'Utah, Etats-Unis	Eur	Oui	Oui	Oui
CHB	Chinois à Pékin, Chine	Asi	Oui	Oui	Oui
CHD	Chinois à Denver, Etats-Unis	Asi		Oui	Oui
GIH	Indiens Gujarati à Houston, Etats-Unis	Asi		Oui	Oui
JPT	Japonais à Tokyo, Japon	Asi	Oui	Oui	Oui
LWK	Luhya à Webuye, Kenya	Afr		Oui	Oui
MXL	Mexican à Los Angeles	Afr		Oui	Oui
MKK	Massai à Kinyawa, Kenya	Afr		Oui	Oui
TSI	Toscans d'Italie	Eur		Oui	Oui
YRI	Yoruba à Ibadan, Nigeria	Afr	Oui	Oui	Oui
PEL	Péruviens à Lima	Amr			Oui
PUR	Porto-Ricains, Porto-Rico	Amr			Oui
CLM	Colombiens de Medellin, Colombie	Amr			Oui
GBR	Britanniques, Grande-Bretagne	Eur			Oui
FIN	Finlandais, Finlande	Eur			Oui
IBS	Ibériques Espagnols, Espagne	Eur			Oui

Tableau 5 : Populations utilisées dans HapMap et 1000 génomes
Il y a actuellement 17 populations utilisées dans le projet 1000 génomes. Onze populations parmi ces dix-sept ont été utilisées dans HapMap 3 et quatre dans HapMap1 et 2. Les différentes populations (**Population**) représentent 4 grands groupes (Continent) que sont les Européens (Eur), les Africains, les Asiatiques (Asn) et les Américains (Amr - en fait sud-américains).

D.VI. Les Analyses d'association génome entier

D.VI.1. Description

L'analyse d'association génome entier est une stratégie visant à chercher de façon exhaustive, sur tout le génome, les régions présentant une forte association avec le phénotype d'intérêt. Pour cela, nous disposons d'outils permettant de caractériser le génotype d'un groupe d'individus (patients et témoins dans notre cas) pour un grand nombre de marqueurs (typiquement > 100 000).

En premier lieu, on procède à une simple analyse d'association (D.IV.2) sur tous les marqueurs tour à tour.

D.VI.1.a) *Couverture du génome*

Pour réaliser une analyse du génome la plus exhaustive possible, c'est-à-dire en testant (idéalement) tous les SNPs (fréquents), tout en maintenant un prix réaliste, les généticiens ont très tôt utilisé la propriété de corrélation par blocs observé sur le génome (D.V.1, p. 43) pour identifier un sous-ensemble de SNPs qui puisse résumer la variabilité à tous les SNPs (fréquents). Ces SNPs sont connus sous leur dénomination anglophone de « tag-SNPs ».

D.VI.1.b) *QQ-plot et Manhattan Plot*

Le QQ-plot (quantile-quantile plot) est la distribution des p-valeurs observées par rapport à ce qui est attendu sous l'hypothèse nulle (sous l'hypothèse nulle nous nous attendons à une distribution uniforme des p-valeurs).

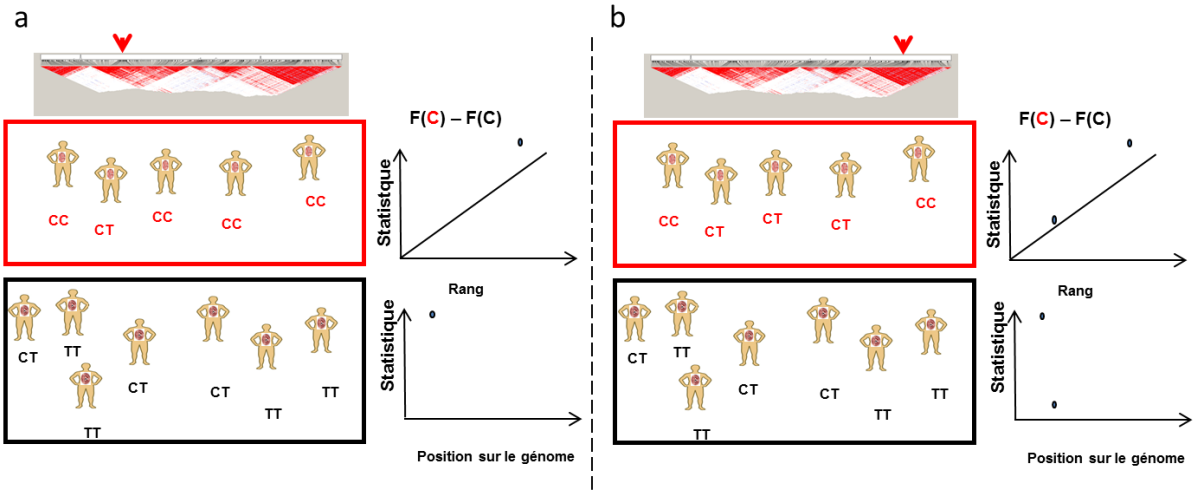


Figure 21 : QQplot et Manhattan Plot

Construction d'un QQ–plot et d'un Manhattan-plot. Tous les SNPs sont analysés à tour de rôle – ici je donne l'exemple de deux SNPs. La position des SNPs et leur DL respectif est représenté comme montré précédemment (**Erreur ! Source du renvoi ntrouvable.**). Le premier SNP, localisé dans le deuxième bloc (flèche rouge) en partant de la gauche. Dans le carré rouge nous avons des patients et dans le cadre noir, des témoins. Chaque individu a son génotype pour le SNP correspondant
On voit qu'il y a une différence de fréquences entre les deux groupes. La statistique correspondante est placée d'abord sur un graphique (bas) en fonction de sa position dans le génome. Elle est également placée sur un graphique en fonction de son rang – du plus bas au plus haut - (ici la statistique est haute donc son rang se trouve assez haut). On voit que cette statistique de l'exemple **a** est plus haute (au-dessus de la ligne diagonale) que sa valeur attendue pour le rang donné. En b, un deuxième SNP, dans un bloc différent, ne montre pas de différence de fréquence entre cas et témoins.

Le Manhattan plot permet de visualiser les résultats (toujours le log10 de la p-valeur) en fonction de la position sur le génome (chromosome-position). Cette figure est très utile pour tester si des signaux d'association sont bien regroupés et donc que des SNPs adjacents ont des résultats corrélés. Le principe des deux graphiques est expliqué sur la Figure 21. Classiquement, tous les SNPs sont enregistrés sur le QQplot et le Manhattan plot afin de synthétiser les résultats sur le génome entier (Figure 22, p. 48).

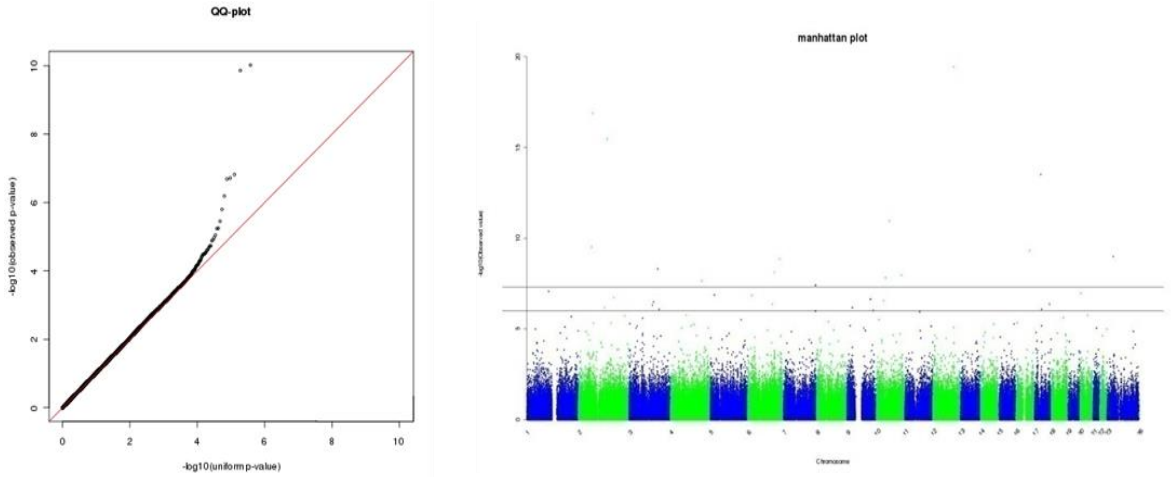


Figure 22 : QQ plot et Manhattan plot en situation

D.VI.2. Imputation et Meta-Analyses

D.VI.2.a) *Imputation : principe*

L'ambition des analyses d'association génome entier est de tester tous les variants fréquents du génome (ensemble $\in_T$) afin d'évaluer leur degré d'association à la maladie. A l'heure actuelle, cependant, seul un sous ensemble ($\in_S$) de tous les SNPs existants peut-être interrogé par des puces de génotypage. Ce sera le cas tant que le séquençage génome-entier ne sera pas accessible au prix du génotypage. Du fait de la structure du génome en blocs haplotypiques [Daly et al.,**2001**], qui induit une corrélation entre SNPs, il est possible d'utiliser le sous ensemble $\in_S$ de SNPs afin d'obtenir une information sur $\in_T$.

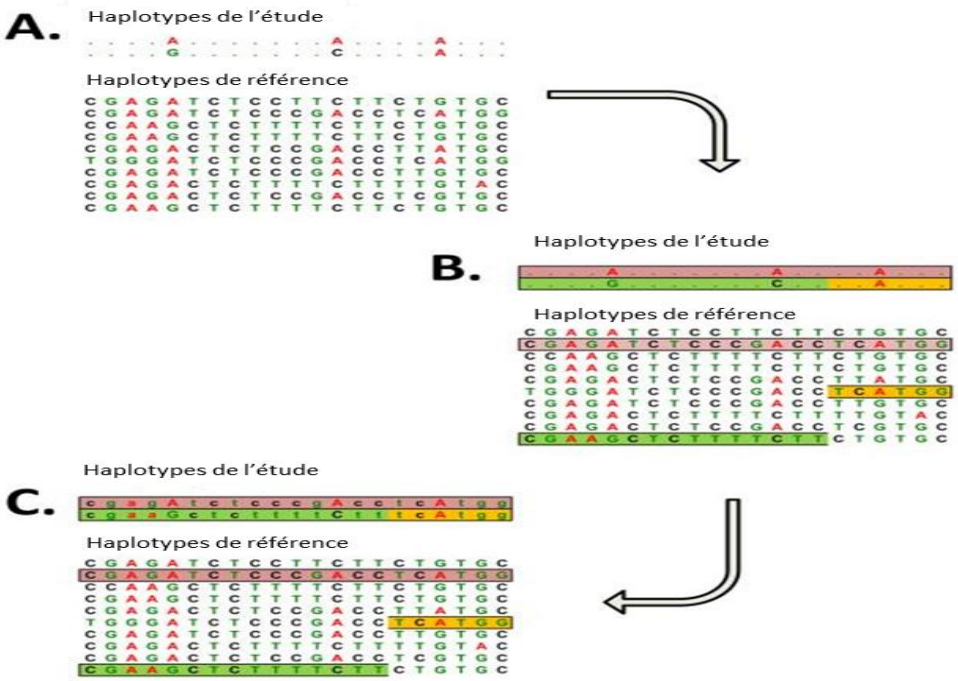


Figure 23 : Imputation des génotypes.

A – Les haplotypes de l'étude (contenant $\in_S$ des SNPs) sont alignés sur les halplotypes du panel de référence (contenant $\in_T$).
B – Les haplotypes communs entre l'étude et le panel de référence sont identifiés (violet, vert et orange).
C – Les génotypes aux SNPs non génotypés dans notre étude sont imputés
Adapté de [Li et al.,**2009**]

Chaque haplotype de l'étude est représenté comme une mosaïque des haplotypes existant dans le panel de référence. Cette méthode repose donc sur l'hypothèse *qu'un nombre limité d'haplotypes ségrègent dans la population*. Cette hypothèse est équivalente à l'existence de blocs haplotypiques (ou forte corrélation entre SNPs sur des régions discrètes).
Il semble important que la population de référence soit suffisamment proche de la population d'étude. De nombreuses études ont été menées afin de s'assurer de la robustesse de cette méthode en présence d'une différence entre la population de référence et la population étudiée [Mueller et al.,**2005**][Bakker et al.,**2006**] et il a été démontré que cette estimation est satisfaisant dans le cas où les populations sont suffisamment proches. Il est également très important que la population de référence soit la plus grande possible afin de permettre une meilleure estimation de la corrélation entre SNPs. Il a été montré que l'estimation de r2 est biaisée vers une surestimation lorsque la population de référence est réduite [Terwilliger, Hiekkalinna,**2006**].
Cette contradiction dans le choix du panel de référence – entre une grande population et une population démographiquement homogène – pourrait être résolue par une nouvelle méthode récemment proposée.
En effet, Marchini et ses collègues [Howie et al.,**2011**] proposent d'utiliser une population de panel la plus large possible, y compris en ajoutant des individus très éloignés démographiquement (Européens vs Africains par exemple). Cette méthode n'abandonne pas l'idée de ressemblance entre les deux populations mais l'applique plutôt aux segments de chromosome que sur le génome entier. Pour chaque région du chromosome en cours d'imputation (définie par une position début et une position de fin), le programme IMPUTE cherche, dans la population de référence, tous les segments (haplotypes) aux mêmes positions qui sont suffisamment proches (au sens d'identité par état) de ceux présents dans la population.

En pratique, l'imputation est assurée par des programmes librement distribués et le chercheur influe principalement sur les paramètres. Cela a été le cas dans ce travail de thèse [Marchini et al.,**2007**][Browning, Browning,**2007**][Li et al.,**2010**].

L'imputation nécessite une étape de phasage des données étudiées suivie d'une comparaison entre les haplotypes observes dans l'étude et ceux observés dans le panel. C'est cette dernière étape qui permet l'imputation proprement dite. Si ce sont deux étapes, elles ne sont pas indépendantes dans la mesure : en effet, à chaque itération, il y a phasage des données puis imputation à partir des haplotypes estimés. La répétition de ces opérations en bloc permet de mieux prendre en compte l'incertitude de la phase haplotypique estimée.
La première étape est la plus longue et doit être refaite à chaque nouvelle version du panel de référence. Il a donc été proposé de séparer ces deux étapes et de n'effectuer le phasage qu'une seule fois dans les données. Le prix à payer pour l'accélération de la procédure semble évident. Nous utilisons un seul résultat de phasage (le plus probable). Il en résulte que l'incertitude du phasage n'est plus prise en compte dans l'étape d'imputation. Les équipes qui ont proposé cette approche ont étudié la perte d'efficacité de l'imputation et l'estiment minime [Howie et al.,**2012**]. Ils ont par ailleurs conseillé de relancer la procédure complète (étapes 1 et 2, ensemble) sur les régions qui sortent significatives.
Nous allons appliquer cette stratégie, en utilisant le programme SHAPEIT [Delaneau et al.,**2012**] pour le phasage et le programme IMPUTE [Marchini et al.,**2007**] pour l'imputation.

D.VI.2.b) *La méta-analyse*

La Meta-Analyse est la méthode qui permet de combiner les résultats de plusieurs analyses testant une même association. Dans notre cadre il s'agit du lien marqueur maladie. Le but principal est d'augmenter la puissance statistique de découverte d'un locus impliqué dans la pathologie d'intérêt.
Cette méthode revient à faire une moyenne pondérée des effets individuels de chaque étude. Cette combinaison doit, idéalement, suivre une loi statistique bien identifiée afin de pouvoir être testée.
Cette combinaison peut se faire pour plusieurs valeurs. Classiquement nous utilisons une combinaison des statistiques suivant une loi normale [Stouffer et al.,**1949**]. Pour cela, nous pouvons utiliser les p-valeurs.
Si nous avons J études avec j=1..J portant sur I SNPs communs avec i=1..I.

Soit p_{ij} la p-valeur d'un test d'association pour le SNP_i dans l'étude J. Soit S un poids de valeurs (-1,1) selon la direction de l'effet du SNP.
Nous construisons une statistique :

$$f(p_{i.}) = \sum_{1}^{J} S_{ij} \times \omega_{ij} \times \emptyset^{-1}(1 - \frac{p_i}{2})$$

Équation 21 : Statistique de la Loi Normale Inverse

Les ω_{ij} sont choisis de telle façon que $\sum_1^J {\omega_{ij}}^2 = 1$. De cette façon, la statistique $f(p_{i.})$ suit une loi normale de moyenne 0 et de variance 1. Les poids sont en général construits à partir des tailles d'échantillons. Dans le cas d'études cas témoins, si le nombre de cas et de témoins diffèrent de façon importante (typiquement un grand nombre de témoins par rapport aux cas), il convient de prendre en compte ce facteur. La construction de ce poids prête le flanc à l'objection de l'instabilité des résultats en fonction du poids. Dans le cas de nos études, il n'y a pas de grande différence de rapport entre cas et témoins et l'utilisation de la taille des groupes comme pondération dans la statistique est raisonnable.
Par ailleurs, cette méthode ne s'occupe pas de l'homogénéité des effets entre études.

L'autre méthode classique est la création d'une statistique à partir des effets des SNPs dans chaque échantillon – ici les Odds-Ratios. En pratique, on estime une moyenne des

logarithmes des Odds-Ratios de chaque étude, pondérés par leur écart-type. Nous pouvons également estimer la déviation standard

Par ailleurs, on peut ajouter classiquement au corpus des méthodes de la Meta-Analyse le test d'homogénéité entre les études. En cas d'hétérogénéité entre études, on introduit une pénalité (en termes de nombre de degrés de liberté et d'estimation de la variance de l'effet) : il s'agit d'une méta-analyse à effets aléatoires.
Il est à noter que le but des méta-analyses en génétique est d'abord d'établir une relation statistique entre un statut et un marqueur génétique en utilisant le plus grand nombre possible de cas et de témoins. Ainsi, l'estimation de l'effet est moins importante dans la phase de découverte.
Enfin, la façon traditionnelle d'estimer l'hétérogénéité, et surtout de tester l'association en cas d'hétérogénéité peut mener à une perte de puissance [Han, Eskin,**2011**].

D.VI.3. Contrôles de Qualité du génotypage

D.VI.3.a) *Artefacts expérimentaux*

Pour obtenir les génotypes de chaque individu à chaque marqueur (SNP), nous étudions l'intensité de chaque allèle (allèle A et allèle B). Chaque allèle émet une intensité de couleur différente qui est transformée en une mesure appelée Log-Ratio. La procédure de génotypage est expliquée en D.VII.3, p. 58.

La plateforme de génotypage Axiom MC et l'instrument GeneTitan MC d'Affymetrix, permettent d'analyser 96 échantillons à la fois (regroupés en plaques) et de tester 572 000 marqueurs génétiques par échantillon.
Les différences entre plaques (« batch effects »), dues à des variations techniques, peuvent être importantes et entrainer des fausses associations positives.

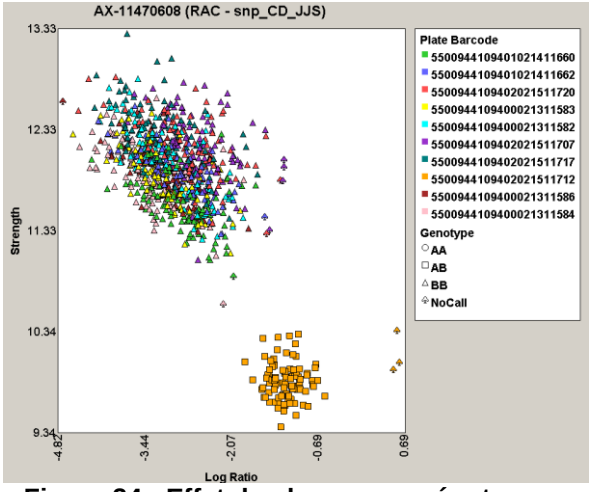


Figure 24 : Effet de plaques en génotypage
Les intensités des génotypes des individus de la plaque 1511712 (en jaune foncé) sont plus basses. Il y a donc formation d'un groupe à part et assignation d'un même génotype – hétérozygote – pour toute la plaque (jaune foncé). L'algorithme de génotypage automatique le détecte comme représentant une classe génotypique d'hétérozygotes. [Affymetrix Analysis Guide]

Les différences de fréquences génétiques entre plaques (chaque plaque étant testée par rapport aux autres) sont visualisées par des graphes de type « QQplots », (Figure 25, ci-dessous).

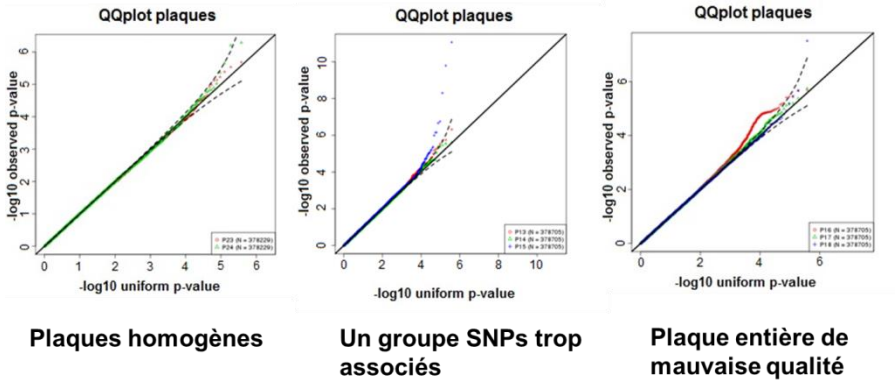


Figure 25 : Artefact expérimental lors du génotypage

Nous observons la distribution des statistiques mesurant les différences de fréquence alléliques entre chaque plaque et le reste des individus. Nous retirons les SNPs qui montrent une différence de fréquences significative pour au moins une plaque.

Nous avons décidé de retirer les SNPs montrant une trop grande différence entre plaques. Toutefois, comme nous le verrons dans d'autres processus de contrôle qualité, il faut choisir un seuil – par exemple la p-value de la différence entre une plaque et toutes les autres. Dans un premier temps j'ai décidé d'exclure les SNPs montrant une différence significative à l'échelle du génome ($5x10^{-8}$).

D.VI.4. Tests Multiples et seuils de significativité

D.VI.4.a) *Quantités utilisées :*

Lors d'une analyse d'association classique, SNP par SNP, on teste l'hypothèse H0 : "Il n'y pas d'association entre le marqueur et le phénotype" contre l'hypothèse H1: "Le marqueur est associé au phénotype". Ce test doit se faire en contrôlant l'erreur de première espèce à un niveau α (nous prendrons ici l'exemple α = 5%). En analyse d'association sur un ensemble de n tests SNP par SNP, cela signifie qu'on s'attend à avoir 5% de nos tests qui rejettent l'hypothèse nulle (et concluent à une vraie association).

Ainsi dans une analyse SNP par SNP avec 1 000 000 de marqueurs, on fera n = 1 000 000 tests au niveau 5% et on s'attend ainsi à avoir 5000 faux positifs ! Ce nombre de faux positifs sera très largement supérieur au nombre de vrais positifs (*Vp*). La question du contrôle des erreurs dans les tests multiples est donc essentielle pour ces analyses.

	H_0 acceptée	*H_0 rejetée*	
H_0 fausse	*Fn*	*Vp*	V
H_0 vraie	*Vn*	*Fp*	F
	n-R	R	n

Tableau 6 : Hypothèses et nombre de tests

Vn = nombre de vrais négatifs ; Vp= nombre de vrais positifs ; Fn = nombre de faux négatifs ; Fp = nombre de faux positifs. V = nombre de tests sous l'hypothèse alternative H1 ; F =

Nous nous intéressons à la proportion de vrais positifs identifiés dans une étude et nous voulons quantifier, pour chaque marqueur, la probabilité qu'il a d'appartenir à l'échantillon des vrais positifs (Tableau 6, ci-dessus).

D.VI.4.b) *Seuil de décision :*

La décision de rejeter l'hypothèse nulle est prise, comme dans toute étude statistique, lorsque la probabilité d'observer une statistique sous H0 est suffisamment petite. En général, la communauté en Epidémiologie-Génétique suit la règle de décision classique qui veut qu'on rejette l'hypothèse nulle si le risque alpha (α) est inférieur à 5%. C'est-à-dire si la probabilité d'obtenir le résultat statistique effectivement observé est inférieure à 5%, nous rejetons l'hypothèse nulle. Dans la suite du document, nous appellerons p-valeur cette probabilité d'observer une statistique donnée si l'hypothèse nulle est vraie.
Ce seuil est conventionnel et, en fait, pour établir une règle de décision, c'est plutôt la stratégie suivie et l'impact du résultat final (en termes de santé publique, de coût, de temps de travails …) qui peut orienter le choix des seuils. Le seuil de 5% peut par exemple convenir par exemple pour des essais cliniques où il est très important d'être convaincu qu'un nouveau médicament est véritablement efficace, entre autres par rapport aux médicaments déjà existants.

En revanche dans le cas d'une étude génétique, en général nous allons tenter de répliquer un ensemble de marqueurs, il y a moins de risques à conclure à tort à une association. En effet, rejeter l'hypothèse nulle pour un marqueur génétique, alors qu'elle est vraie aurait pour conséquence de tester ce marqueur à perte dans une étude de réplication.

En général, ce qui nous intéresse dans la première phase d'une étude d'association génétique (ou toute phase intermédiaire) est d'augmenter le nombre de vrais positifs, en acceptant d'avoir une proportion plus ou moins importante (à décider par l'utilisateur) de faux positifs.
Il est à noter qu'en fin de parcours, le critère pour déclarer une association reste ce risque alpha de 5% (bien qu'il soit corrigé par le nombre de tests indépendants).

D.VI.4.c) *Tests Multiples :*

D.VI.4.c.1) FWER (Family Wise Error Rate)

Le FWER est défini comme la probabilité de rejeter à tort au moins une fois l'hypothèse nulle alors que l'hypothèse nulle est vraie pour tous les tests (on définit n le nombre de tests).

$$FWER = 1 - P(Fp = 0) = 1 - (1 - \alpha)^n$$

Équation 22 : FWER, formule de Sidak

L'approximation la plus connue de cette probabilité a été donnée par Bonferroni et est égale à :

$$FWER = 1 - P(Fp = 0) = 1 - n\alpha$$

Équation 23 : FWER, approximation de Bonferroni

Cette approximation est valable lorsque α est très petit (ce qui est en général le cas en analyse d'association génétique).

Elle est valide pour des tests indépendants et devient très conservative en cas de dépendance. Or nous avons vu que les marqueurs génétiques sont en général dépendants en blocs (D.V.1, p. 43). Cet inconvénient peut être contourné soit en utilisant des permutations (D.IV.2.c.2), p. 42), soit en estimant le vrai nombre de tests, prenant en compte la corrélation entre marqueurs génétiques (D.VI.4.c.2), p. 54).

D.VI.4.c.2) Nombre de tests

Lors d'une analyse d'association de type Etude d'Association Génome Entier (EAGE, appelée également GWAS, de l'anglais Genome Wide Association Study), nous testons tour à tour N marqueurs génétiques avec deux allèles. En général, nous faisons un test par marqueur. Il est donc important de prendre en compte le nombre de tests effectués lors de la détermination du seuil de significativité.

La définition du nombre de tests indépendants dans une analyse d'association génome entier, et par conséquent la p-valeur nécessaire pour un SNP afin qu'il soit déclaré comme associé à la pathologie n'est pas triviale et a fait l'objet d'investigations. Il s'agissait principalement de trouver le nombre de composantes indépendantes expliquant un grand pourcentage de la variabilité génotypique par une méthode proche de l'analyse en composantes principales, expliquée dans (D.VI.6, p. 55). Il s'agit, en simplifiant à l'extrême, d'identifier les groupes de marqueurs fortement corrélés entre eux.

En principe, la p-valeur seuil devrait dépendre de la population d'origine puisque le nombre de composantes résumant l'information génétique dépend du déséquilibre de liaison et donc de l'histoire démographique de cette population. Cette histoire est potentiellement différente d'une population à l'autre.

Ainsi, la p-valeur seuil était estimée à 7.10^{-08}[Dudbridge, Gusnanto,**2008**] pour un ensemble infiniment dense de SNPs, et à 10^{-07} pour les variants fréquents (MAF > 5%) [Pe'er et al.,**2008**] dans une population Européenne.

Même si l'estimation du nombre de tests indépendants devrait dépendre de la population sur laquelle on effectue le test, il est maintenant universellement accepté de prendre le seuil de $1.\ 5.10^{-08}$ comme seuil de détection (pour un nombre de tests indépendants de 1 M). Pour des populations Européennes, et en se concentrant sur des variants fréquents, nous pouvons en déduire que ce seuil est légèrement conservatif. Toutefois, nous pouvons également considérer qu'un risque α de 5% laisse une part trop importante à l'incertitude. Encore une fois, il s'agit d'un choix de l'investigateur qui doit dépendre des conséquences de l'analyse. Ainsi, si l'on compte mener des études longues et coûteuses sur un gène identifié par ce type d'analyse, il est sans doute mieux d'obtenir une significativité beaucoup plus grande, donc une p-valeur plus petite que 5%, même corrigée à l'échelle du génome.

Enfin, il est à noter que la permutation des données permettrait de donner la meilleure estimation du risque de la p-valeur pour un échantillon donné. Cette approche avait été adoptée dans la première analyse génome entier du Diabète de Type 2 à laquelle j'ai contribué [Sladek et al.,**2007**] (position de 4^{e} - 2^{e} ex-aequo) . Une permutation des statuts pour le phénotype (ou le vecteur de phénotypes) suivi de l'application des analyses effectuées dans la première partie permet en effet d'obtenir des p-valeurs adaptées à l'étude.

Par ailleurs, il est même possible de mimer un scénario à deux phases (initiale et réplication) par permutation, en divisant simplement l'échantillon initial [Dudbridge,**2006**].

D.VI.5. Taux de Fausse Découvertes (False Discovery Rate - FDR).

Classiquement, on présente la FDR (je garderai cette expression anglophone du fait de son utilisation *quasi*-systématique dans le domaine) comme une alternative à la p-valeur (qu'on appelle aussi FWER). La FWER est très exigeante alors que la FDR peut permettre d'augmenter la puissance d'une analyse.
La FDR est la proportion attendue de faux positifs parmi les positifs. On note que nous définissons nous-mêmes les positifs puisque nous devons choisir une p-valeur seuil. Le nombre de positifs, dans notre Tableau 6 est *R*. La proportion attendue de fausses découvertes est E(F_p/R). Cette proportion dépend du niveau individuel d'association (p-valeur) ainsi que de la proportion de tests sous H_0 (F) et sous H_1 (V).
L'estimation du taux de fausses découvertes proposé par Benjamini et Hochberg est la méthode la plus connue et sans doute la plus utilisée dans notre domaine [Benjamini, Hochberg,**1995**]. Cette méthode s'appuie sur l'hypothèse que le nombre de tests sous H_1 (V) est petit, et donc négligeable par rapport au nombre de tests sous H_0 (F). Il en résulte que $\hat{V}=n\alpha$. Les auteurs proposent donc d'utiliser la majoration :

$$FDR \leq \max(n \times \alpha / R(\alpha); 1)$$

R(α) le nombre de positifs observés à un niveau α donné.
Soit une série de P-valeur ordonnées $P_1 < \ldots P_i \ldots < P_N$. On recherche le nombre k maximal pour lequel :

$$P(k) \leq \frac{k}{N} \times \alpha$$

et on déclare significatifs les tests H_i : i =1,…, k.

Elle n'est également valide que lorsqu'il y a des SNPs indépendants. La situation se complique dans nos études par le fait que nous avons une grande corrélation entre les tests. Des adaptations de la procédure ont été proposées pour prendre en compte la dépendance [Benjamini,**2001**] mais ces corrections sont plutôt conservatives (c'est-à-dire qu'elles ont tendance à ne pas conclure au rejet de H_0 alors qu'H_0 est faux, ou autrement dit n'identifient pas des associations vraies). Il n'est pas certain qu'elles s'appliquent parfaitement à la structure de dépendance des données génotypiques. Par ailleurs, il est important de noter que la procédure initiale de Benjamini et Hochberg est malgré tout valide même dans un certain nombre de cas de corrélation [Benjamini,**2001**].
Dans ce contexte, il serait peut-être plus intéressant d'explorer des approches prenant en compte la dépendance entre tests, même si elles ont été développées dans le cadre de l'analyse d'expression [Leek, Storey,**2008**][Friguet, Causeur,**2011**][Blum et al.,**2010**], donc pour un nombre de variables moins important (et pour des corrélations moins fortes) que dans nos études.

D.VI.6. Biais du à la Stratification génétique et Analyse en Composantes Principale

Nous avons vu (D.III.5, p. 23) que la population humaine est divisée en ensemble de sous-populations et que le flux de gènes à l'intérieur est supérieur au flux de gènes entre ces sous-populations. Il y a donc, par dérive génétique, des différences de fréquence allélique entre sous-populations.

Dans les analyses cas témoins, surtout lorsque les cas et les témoins sont issus de modes de recrutement différents (ce qui est le cas dans ma thèse), il est possible que l'origine ethnique des deux groupes soit différente.
Le risque d'un tel déséquilibre est d'autant plus grand dans le cas où certaines pathologies sont plus fréquentes dans certains groupes de populations.

Un déséquilibre populationnel entre cas et témoins entrainera une différence de fréquence pour des marqueurs (ici des SNPs) seulement en raison de leur différence entre populations et non de leur effet sur la maladie et donc à une différence de fréquences entre malades et non-malades (Figure 26).

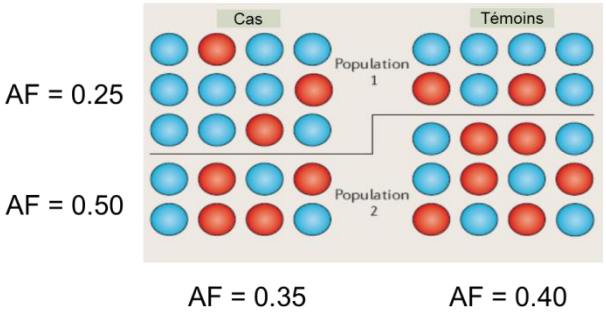


Figure 26 : Stratification et Association

Chaque point représente un chromosome porteur d'un allèle (bleu ou rouge). Les cas et les témoins comprennent deux populations (1 et 2) en proportions différentes – 60% chez le cas, 40% chez les contrôles d'individus issus de la population 1. Le SNP ne montre pas de différence entre cas et contrôles dans chaque population. Par contre, l'allèle rouge a une fréquence allélique de 25% dans la population 1 et une fréquence de 50% dans la population 2. Il en résulte une fausse association lorsqu'on compare cas et témoins. Modifié de [Balding,**2006**].

Les analystes ont été très tôt sensibles à ce problème, qui s'est toutefois trouvé amplifié dans le contexte de grandes études d'association génome entier incluant des individus non apparentés. Dans le cadre de la première étude d'association sur le Diabète de Type 2 [Sladek et al.,**2007**] – article non inclus dans ma thèse -, nous avons proposé d'utiliser une Analyse en Composantes Principales (ACP) afin d'extraire l'information géographique. Cette approche était inspirée par la méthodologie proposée par l'équipe de Luigi Luca Cavalli-Sforza dans les années 70 [Menozzi et al.,**1978**]. Les unités observées étaient les fréquences alléliques par sous-population. Dans leur papier initial, Menozzi et ses collègues ont démontré que la première composante expliquait environ 30% de la diversité géographique et était corrélée à un gradient Nord Sud (Figure 27, ci-dessous).

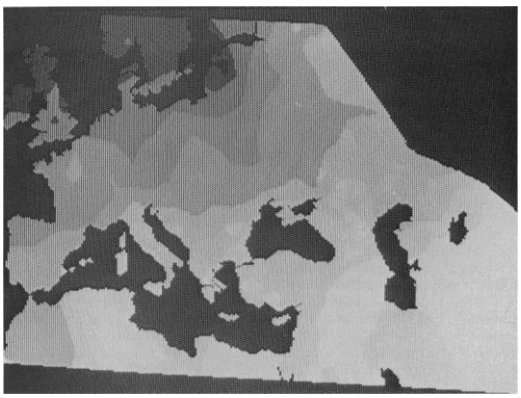


Figure 27 : ACP et Clines

(issu de **[Menozzi et al.,**1978**]**). La première composante principale établie à partir de 38 loci indépendants (groupe sanguin ABO, groupes sanguins Rh, groupes sanguins MNS, Lewis - Le, Duffy - Fy, Haptoglobine - Hp, phosphomutoglutase - PGM1, HLA-A, et HLA-B dans 67 populations. Les marqueurs non issus du locus HLA ont été sélectionnés à partir des tables issues de **[Mourant,**1954**]**.
L'intensité du grisé indique la valeur, pour chaque localisation géographique. Cette première composante représente 27% de la variabilité totale et semble suivre un gradient Nord / Sud-Est.

Notre idée était d'utiliser, comme unité de mesure, non plus une sous population mais l'individu lui-même. Pour chaque marqueur, un individu a une fréquence de 0,0.5 ou 1. Le principe de cette analyse est illustré dans la Figure 28 (ci-dessous).
Dans un deuxième temps, nous nous proposions d'intégrer la position des individus sur une ou plusieurs composantes dans l'analyse d'association. Cela pouvait se faire soit en créant des sous-groupes d'individus proches sur les différentes composantes – avec un nombre minimal de cas et de contrôles dans chaque sous-groupe, soit en utilisant les positions sur une ou plusieurs composantes (qui restent à définir) en tant que covariable quantitative.

Une méthode très similaire et basée sur l'ACP a été publiée sous le nom de EIGENSTRAT [Price et al.,**2006**].

Le but est de trouver, pour chaque individu, sa position « moyenne » représentant tout le génome. Pour cela nous avons bien sélectionné des SNPs en (relatif) équilibre de liaison. En utilisant la stratégie proposée dans le logiciel PLINK, nous avons utilisé les distances génétiques entre individus. À partir de ces données, nous utilisons une technique très proche de l'ACP, le MDS (Multi-Dimensional Scaling).
Plus récemment, plusieurs équipes ont proposé d'appliquer des méthodes de modèles mixtes. Cela revient à utiliser la matrice des différences génétiques entre individus [Kang et al.,**2010**].

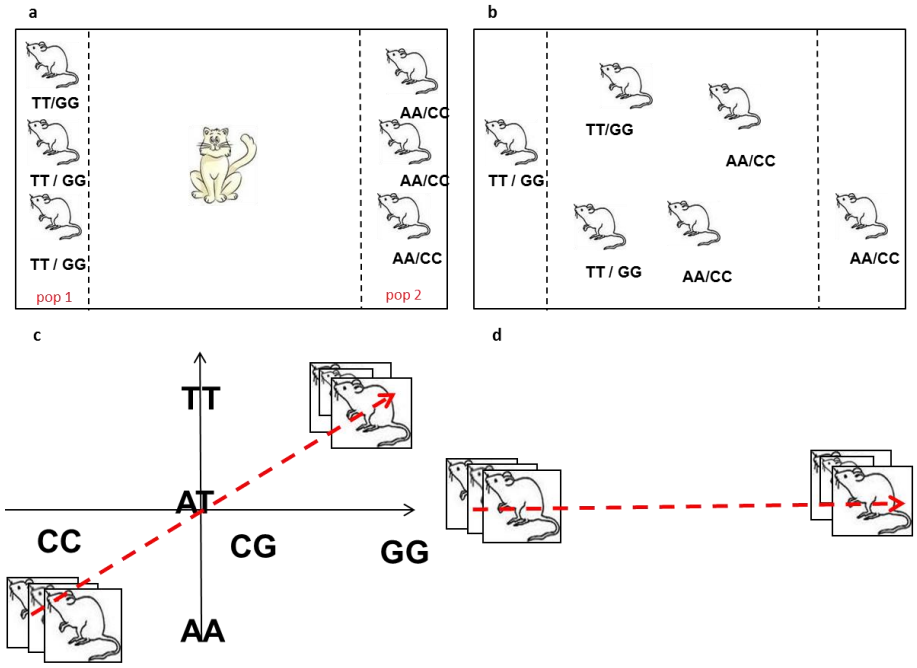


Figure 28 : Principe graphique de l'Analyse en Composantes Principales
a/ Dans les deux populations (pop1 et pop2) de la Figure 8 , nous avons indiqué le génotype à un 2e locus. Ce locus a également été mené à fixation avec un allèle différent par le même processus aléatoire que le premier. **b/** Lorsque la barrière disparaît, il peut y avoir mélange des populations. Pour identifier les populations d'origine, les individus sont classés selon leur génotype. **c/** Sur un graphe qui a autant de dimensions que de SNPs, chaque individu est placé aux coordonnées correspondant à leur génotype (3 souris sont AA/CC et 3 autres souris sont TT/GG). Nous remarquons deux choses : 1/ les souris sont séparées assez facilement sur ce graphe en deux groupes (en fait les populations 1 et 2) ; 2/ Il suffit d'un graphe en 1 dimension (ligne rouge) pour avoir la même information. **d/** La dimension a été réduite (passage de 2 à 1) pratiquement sans perte d'information. Nous avons une seule **composante** qui résume parfaitement la différence entre les deux populations. Si nous devions générer une représentation géographique semblable à celle présentée Figure 27, il suffirait de prendre la position de chaque souris sur la composante (dans **d**) et de la transformer en intensité de couleur sur leur position géographique (dans **b**). Nous aurions des valeurs très foncées à droite et très claires à gauche. C'est le principe de l'analyse en composantes principales appliquée à la génétique.

D.VII. Technologie de Génotypage

D.VII.1. Principe

Le principe du génotypage d'individus avec des SNPs organisés sur des puces à ADN met en relation les principes d'hybridation entres brins d'ADN par complémentarité des bases, de fluorescence en microscopie et de capture d'ADN sur des surfaces solides. Les principaux composants d'une puce sont : (i) le support sur lequel est hybridé l'ADN cible, (ii) les ADN sondes et (iii) un système de détection qui enregistre et interprète le signal d'hybridation. Deux types de puces ont été élaborés par les industriels Affymetrix et Illumina. Utilisant tous deux le principe de fluorescence, ils divergent principalement sur la nature du support d'hybridation : pour la technologie Affymetrix, l'ADN cible est directement synthétisé sur des puces, alors que pour la technologie Illumina, l'hybridation se réalise sur des billes.
Les méthodes d'analyse de puce à ADN permettant de déterminer la configuration génotypique d'un SNP d'un individu, reposent avant tout sur la probabilité pour le signal de fluorescence résultant, de correspondre à tel ou tel génotype.
En général, les individus à génotyper sont regroupés par plaques de différentes tailles (typiquement 24, 48 et 96). Dans nos études, nous avons choisi les plaques de 96 individus. Ce regroupement est important car la conséquence sera que des groupes d'individus seront traités en même temps, dans une même expérience, dans des conditions qui, malgré les efforts de standardisation, seront différentes des autres groupes de 96 individus. Le risque est un effet de plaque (« batch effect ») qui peut induire de fausses associations, si les patients et les témoins sont groupés sur des plaques différentes.

D.VII.2. Marqueurs SNPs et puces ADN

La puce ADN Axiom Affymetrix peut interroger 567000 SNPs par l'intermédiaire de plus de 1,2 millions de sondes. Le choix de ces SNPs a été établi à partir des données HapMap (456 000 SNPs), mais également à partir des données du projet 1000 Genomes (81 000), ainsi que 25 000 sondes tirées d'autres sources, non-dévoilées par Affymetrix®.

6500 petites insertions ou délétions, appelées « indels », sont également testées sur la puce. C'est une spécificité de cette puce qui n'est pas observable encore sur d'autres puces.

La distribution des sondes SNPs est pangénomique, privilégiée dans les régions non-géniques (329 000 régions non géniques / 238 000 géniques). Ils peuvent être situés dans des régions régulatrices de gènes, « *enhancer* », promoteur, « *silencers* », etc.

D.VII.3. Graphe des résultats de génotypage

Comme expliqué précédemment, la détermination (ou assignation) d'un génotype se fait en fonction du rapport entre l'intensité de l'hybridation sonde « spécifique » de l'allèle A et celle de la sonde de l'allèle B (Figure 29, ci-dessous).
Dans un monde parfait, le rapport d'intensité entre l'allèle A et l'allèle B serait de 0, 50% et 100%. Le logarithme du rapport B/A serait respectivement -2, 0 et 2. Par ailleurs, toujours dans des conditions idéales, pour un génotype donné, la variation autour de ces valeurs serait très basse, sinon nulle.

Dans les conditions réelles de génotypage, nous assistons évidemment à des variations dues à un ensemble de causes allant de la qualité de l'ADN à l'affinité de la sonde pour sa cible et à la technologie de détection.

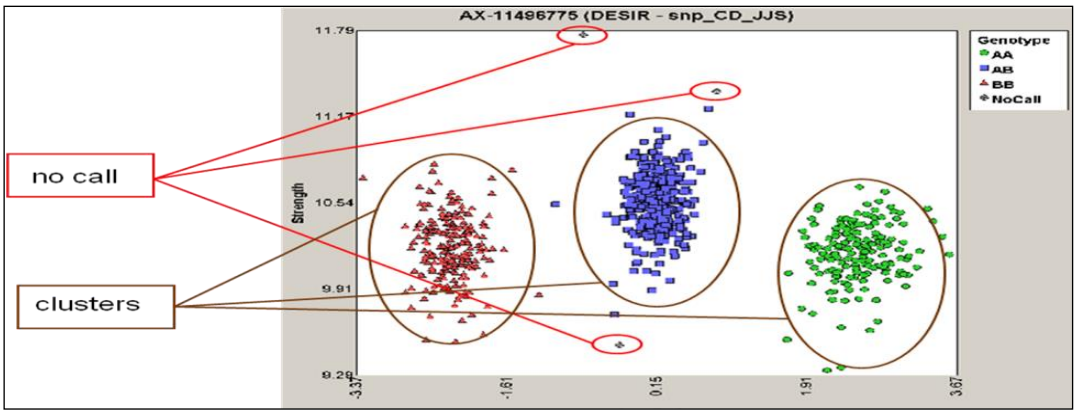


Figure 29 : Exemple d'un « cluster graph (population "DESIR")

Chaque point = 1 génotype d'un individu. Génotype AA : rond vert, AB : carré bleu, BB triangle rouge et « No Call » : trèfle gris
En abscisse : Log Ratio de chaque génotype, mesuré en fonction des intensités de fluorescences entre les deux canaux, AT et GC
En ordonnée : Intensité totale (somme des deux intensité) du génotypage.
Extrait de [Affymetrix Analysis Guide].

Les résultats de ces expériences sont plus ou moins facilement interprétables suivant la qualité de l'ADN, la spécificité de la sonde ou le pourcentage de séquences GC. Les SNPs de mauvaise qualité ne permettent pas une assignation très sûre des génotypes du fait d'une mauvaise séparation entre les classes génotypiques (Figure 30, ci-dessous).

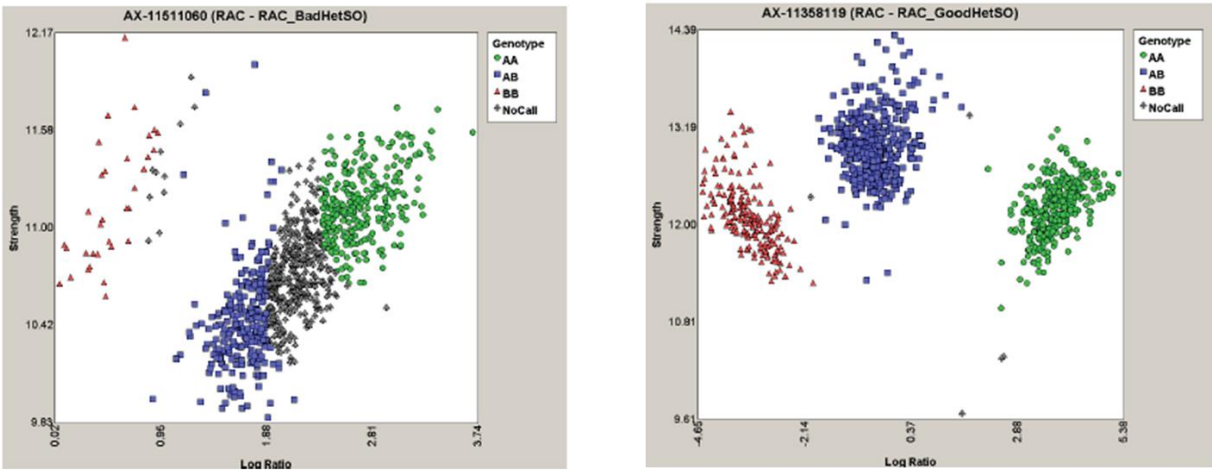


Figure 30 : Exemple de graphes montrant différentes qualités de génotypage

Représentation des Intensité total (Strength) et des différences d'intensité (Log Ratio) des allèles A et B. Dans le cas du « SNP de mauvaise qualité », il y a une difficulté de l'algorithme à séparer les génotypes bleus et verts. Sur la figure de droite, la séparation se fait parfaitement. Extrait de [Affymetrix Analysis Guide]

Étant donné qu'on ne peut inspecter visuellement chaque SNPs, il existe des procédures de vérification automatique de la qualité des assignations de génotypes.
Ces valeurs sont intégrées au système de génotypage apt (Affymetrix Proprietary Tools).

La distance entre les clusters les plus proches (en valeur absolue) est calculée par la valeur « Fisher Linear Discriminant ». Cette valeur permet d'identifier les SNPs pour lesquels les regroupements sont beaucoup trop proches (Figure 31, ci-dessous).

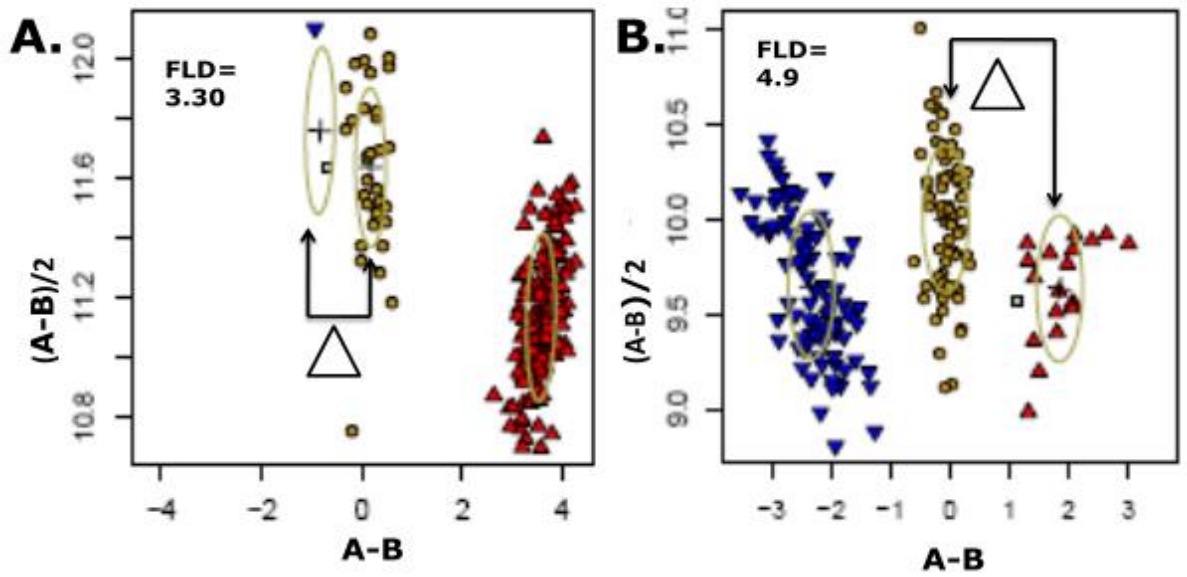


Figure 31 : Valeur de qualité de génotypage : Fisher Linear Discriminant

Le Fisher Linear Discriminant (FLD) est utilisé pour reconnaitre la qualité de séparation des trois différents groupes de génotype (AA, AB et BB). Un FLD grand indiquera une bonne séparation des groupes de génotypes et donc une bonne qualité de génotypage. Il est recommandé, dans le manuel utilisateur de GeneTitan, de supprimer de l'analyse les marqueurs ayant un FLD inférieur à 3.6. Extrait de [Affymetrix Analysis Guide].

La deuxième mesure intégrée dans le système Affymetrix Axiom est le décalage de l'hétérozygote (Figure 32, ci-dessus).

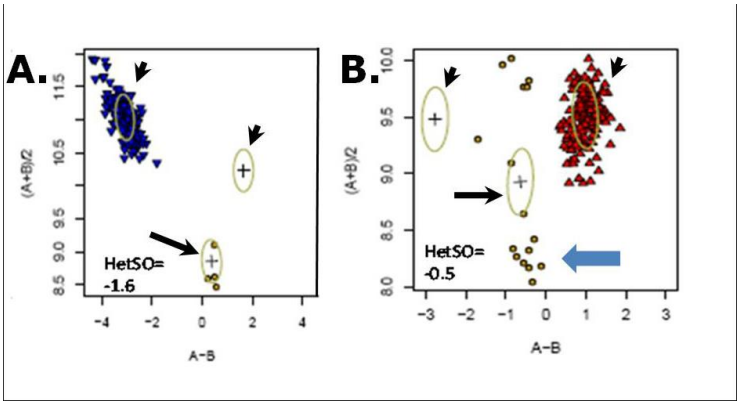


Figure 32 : Valeur de qualité de génotypage : HetSO

Le HetSo est défini comme la distance verticale entre le centre du cluster hétérozygote et la ligne reliant les centres des groupes des homozygotes. Il mesure donc le décalage du signal moyen des hétérozygotes par rapport au signal moyen homozygote. Les SNP ayant un HetSO inférieur à -0.5 sont considérés de mauvaises qualité et doivent être supprimés. Extrait de [Affymetrix Analysis Guide]

Enfin, la dernière mesure qui est commune aux outils Affymetrix et aux procédures classiques d'analyse est le pourcentage d'assignation réussie qu'on appelle « call-rate » (Figure 33, ci-dessous).

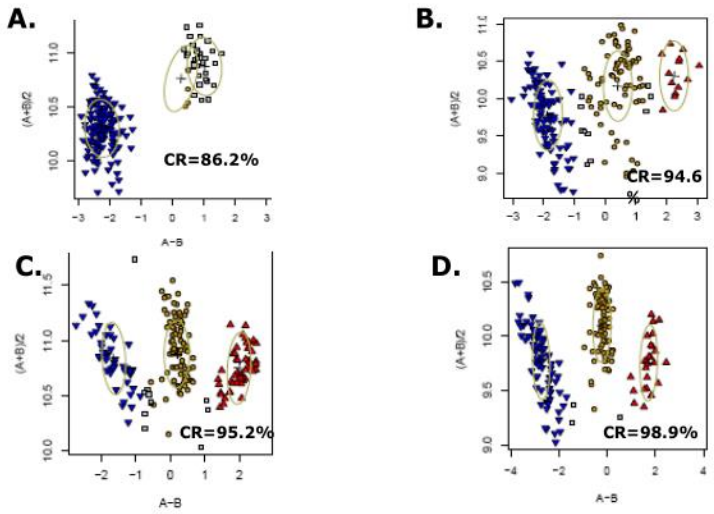


Figure 33 : Pourcentage de succès

Cluster graphes de 4 SNPs dans un ordre décroissant de qualité. Le SNP A ne montre que deux groupes alors que trois sont attendus –le pourcentage d'assignation (Call Rate) est très bas (~86%). Les SNPs B, C et D présentent des qualités d'assignation de plus en plus haute, avec des groupes génotypiques de mieux en mieux séparés. Les carrés blancs représentent les individus pour lesquels on n'a pas pu définir les génotypes. Extrait de [Affymetrix Analysis Guide].

D.VII.4. Impact de la qualité des génotypes

La qualité des génotypes est très importante dans une analyse d'association car des erreus, surtout si elles affectent un sous-groupe d'individus (seulement les cas, ou seulement les témoins par exemple) peuvent entraîner une fausse association et amener à conclure à tort à l'effet d'un gène.

Alors qu'il serait très important de génotyper des cas et des témoins sur chaque plaque afin de pouvoir les apparier, dans la réalité, cette approche est rarement adoptée. Il arrive très souvent que les cas et les témoins soient génotypés à part, par exemple en raison d'un décalage de la disponibilité des données.

D.VIII. Gènes et pathologies : historique

D.VIII.1. Garrod et alcaptonurie

La première relation entre un gène et un enzyme est établie en 1902 par Archibald Garrod (St Bartholomew's Hospital, Londres), à partir d'une observation portant sur une anomalie métabolique chez l'homme : l'alcaptonurie. L'alcaptonurie est une anomalie d'excrétion, affectant le métabolisme de la tyrosine et de la phénylalanine. Les sujets atteints souffrent d'arthrite débilitante. C'est une maladie rare, dont l'incidence est estimée à 1/250000 (OMIM 203500).

Cette maladie se manifeste par le noircissement des urines lorsqu'elles sont exposées à l'air. Le noircissement est dû à la présence dans les urines d'acide homogentisique, qui est un produit intermédiaire de la dégradation de la tyrosine et de la phénylalanine. Cette substance est dégradée chez les individus normaux, mais pas chez les ceux souffrant d'alcaptonurie, chez lesquels elle s'accumule. Le sérum des premiers contient l'enzyme capable de la métaboliser (l'homogentisate 1,2 desoxygénase), mais cet enzyme n'est pas présent dans le sérum des patients.

Garrod observe que la transmission de cette anomalie s'effectue chez l'homme en strict accord avec les lois de Mendel (selon un mode autosomal récessif), ce qui lui suggère qu'elle est due à un gène unique. Il propose donc que la déficience enzymatique soit due à une anomalie du gène responsable de la synthèse de cet enzyme. Garrod publie ces observations en 1909 dans *Les erreurs innées du métabolisme*, livre où il avance une explication similaire pour plusieurs maladies génétiques (albinisme, cystinurie, pentosurie). Plus généralement, il propose que chaque enzyme serait le fruit de l'activité d'un gène.

D.VIII.2. Pathologies expliquées historiquement

Les années 80 ont vu les premières identifications de régions chromosomiques associées à des maladies génétiques telles que la Chorée de Huntington [Gusella et al.,**1983**][Macdonald,**1993**], la mucoviscidose [Riordan et al.,**1989**][Rommens et al.,**1989**] ou encore la myopathie de Duchenne [Monaco et al.,**1986**]. En septembre 2008, la base de données OMIM® (Online Mendelian Inheritance in Man®) comptabilisait 6487 entrées avec une description phénotypique et l'on disposait des bases moléculaires pour 3770 (58%) d'entre elles [Amberger et al.,**2009**]. La base de données OMIM® recense environ 70 nouvelles entrées par mois et estime à plus de 1500 le nombre de pathologies génétiques sans base moléculaire identifiée.

Ces différents succès, dont je ne présente que les plus connus, ont rapidement motivé la recherche systématique, par cartographie génétique, de gènes responsables de pathologies humaines.

E. Résultats

E.I. Objectif et présentation de mes travaux :

L'objectif de ma thèse était de participer, par l'analyse d'association génome entier, à l'élucidation des bases moléculaires de trois pathologies cardio-métaboliques (Diabète de Type 2, Syndrome de Brugada et Prolpasus Valvulaire Mitral), chacune correspondant à un projet. Les travaux d'association génome-entier sont par nature collaboratifs et impliquent un grand nombre de partenaires. Dans ces trois études, j'ai participé de façon importante aux analyses statistiques et à l'élaboration des hypothèses. Dans le Projet 1 (étude du Diabète de Type 2), j'ai participé au Comité d'Analyse et au Comité d'écriture du papier, réalisé les analyses sur nos données cas-témoins et participé aux méta-analyses. Dans le Projet 2 (étude du Syndrome de Brugada) j'ai coordonné les analyses d'association génome entier et y ai fortement participé. Dans le Projet 3, j'ai coordonné les analyses statistiques préalables à l'imputation et à la méta-analyse et j'ai réalisé directement ces deux dernières.

E.II. État des lieux des analyses d'association génome-entier :

E.II.1. De la théorie des variants communs (Common Disease / Common Variant)…

Le début de ma thèse se situe à une période où les analyses d'association génome entier battaient leur plein. Un grand nombre de maladies, en général fréquentes dans la population, étaient étudiées par cette approche.

La stratégie de recherche de variants fréquents, qui était à l'origine de ces études, reposait sur un corps d'hypothèses basés sur des modèles de génétique des populations mais également de l'absence de résultats clairement significatifs des analyses de liaison pour les pathologies fréquentes. Un des papiers que l'on peut considérer comme fondateur de cette orientation a été écrit par Eric Lander, à partir des arguments que j'ai évoqués, sous la forme de « Policy Forums », dans la revue Science [Lander,**1996**]. Ce papier évoquait la possibilité de recruter « … quelques douzaines de patients et de témoins … » à la place de « milliers de familles humaines ... ». Cette vision apparaît rétrospectivement comme très optimiste. Il est intéressant de noter que Lander pensait surtout utiliser l'association génétique pour augmenter notre capacité à identifier les variants causaux dans des grandes régions préalablement mises en évidence par liaison, que ce soit sur des famille humaines ou des croisements animaux [Lander,**1996**]. Par ailleurs, et peut-être de façon plus importante, Lander cite plusieurs exemples « frappants » (« tantalizing examples ») d'effet important d'un allèle fréquent – l'effet des allèles de l'Apo-lipoprotéine 4 dans la maladie d'Alzheimer ainsi que celui du Facteur 4 dans la thrombose veineuse.

La justification du recours à l'analyse d'association comme alternative est historiquement attribué au papier de Risch et Merikangas qui comparait la puissance respective de ces deux approches [Risch, Merikangas,**1996**]. Ce papier signifiait clairement que si les allèles à risque étaient fréquents dans la population, et d'effet modeste, l'analyse d'association aurait une plus grande puissance pour les détecter que l'analyse de liaison.

Par ailleurs, il est important de rappeler également que la théorie soutenant le recours aux analyses d'association génome entier (dans leur forme actuelle) reposait sur l'hypothèse d'un nombre infini d'allèles possibles et donc de la rareté des mutations récurrentes [KIMURA, CROW,**1964**][Kimura,**1969**]. En effet, la corrélation entre allèles physiquement proches sur le génome (DL voir D.III.7, p. 25) est maintenue lorsque le taux de recombinaison θ est faible mais également s'il n'y a pas apparition récurrente d'un même allèle à un même locus. Dans ce dernier cas, en effet, il deviendrait difficile de détecter l'effet d'un SNP par analyse des SNPs voisins, du fait qu'ils ne seraient plus corrélés. Notons toutefois que cela n'empêcherait pas de détecter une association si on génotype le variant causal lui-même.

Enfin, Chakravarti, dans une revue parue dans le journal *Nature Genetics* évoque la difficulté à identifier des variants à effet modeste par la stratégie de clonage positionnel, c'est-à-dire par analyse de liaison [Chakravarti,**1999**]. Par rapport aux arguments précédents, il ajoute également l'idée qu'en cas d'effet épistatique, où la présence simultanée des allèles à risque à l'état homozygote pour plusieurs SNPs est nécessaire à la survenue de la pathologie, les allèles pathogènes sont forcément fréquents.
Il semble donc qu'en dehors de l'existence d'exemples de variants délétères (causant la pathologie) fréquents dans quelques cas ainsi que de la difficulté à identifier des signaux de liaison robustes dans les pathologies communes, il n'y ait pas eu d'expression claire d'une théorie de variants communs pour des pathologies fréquentes.
Cela n'empêcha pas le développement et l'application de cette stratégie avec, comme première étude, l'analyse d'association génome entier recherchant les gènes de la dégénérescence maculaire liée à l'âge [Klein et al.,**2005**]. Une association a été découverte en étudiant seulement 96 patients et 50 témoins, le SNP associé montrant un Odds Ratio de 4.7 et se trouvait près du gène codant pour le Facteur H du complément. Il semblait donc que les prédictions se réalisaient et que la recherche de polymorphismes fréquents, à effet fort, pour des maladies fréquentes et complexes allait déboucher sur des résultats concluants.
Et il est vrai que des associations significatives à l'échelle du génome avaient été identifiées pour un très grand nombre de phénotypes et de pathologies (Figure 34, p. 65, ci-dessous). Cependant, dans la plupart de cas, les effets génétiques étaient beaucoup plus faibles.

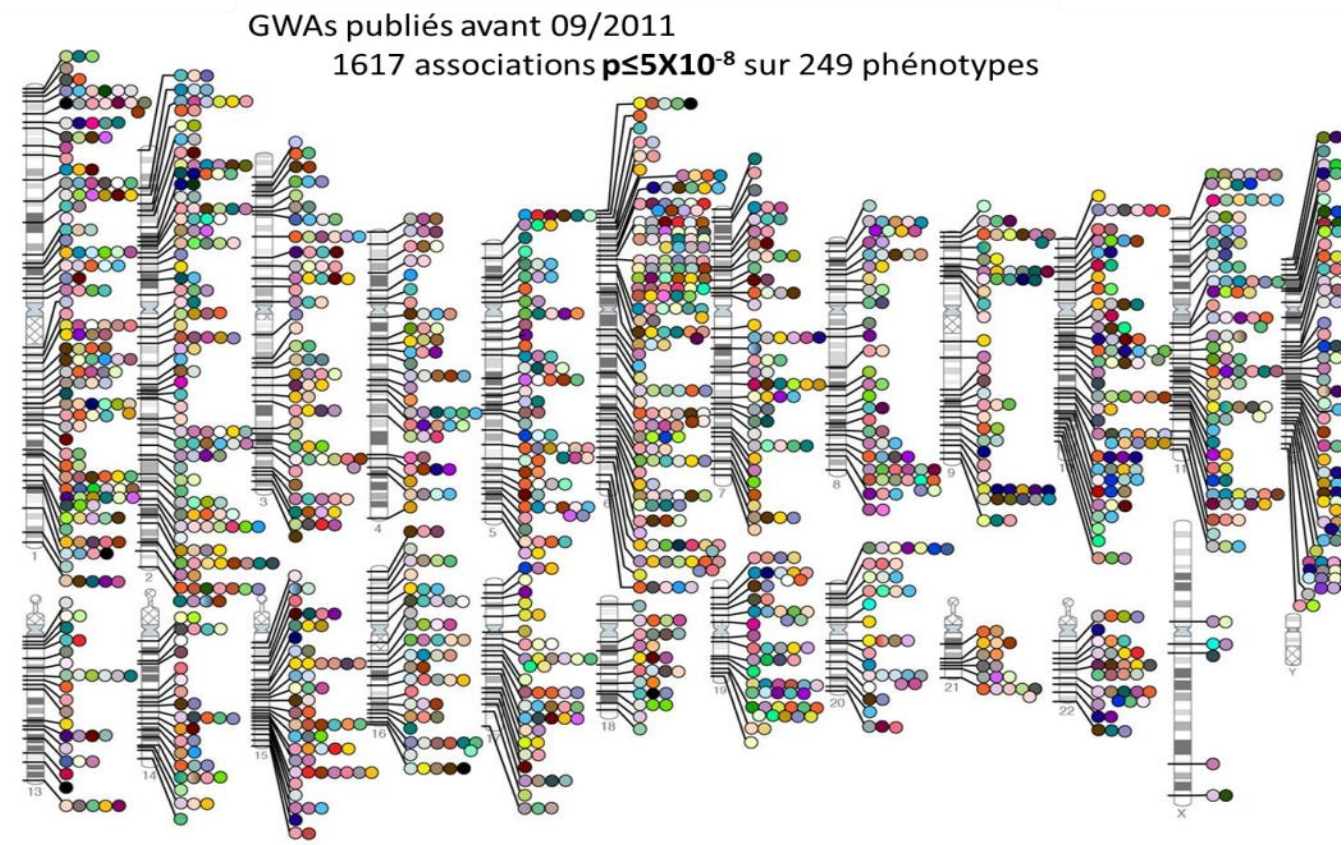


Figure 34 : Résultats significatifs d'Etudes d'Association Génome Entier en Septembre 2011
Localisation des SNPs (et régions de SNPs) significativement associés à une pathologie humaine ou à un phénotype en population générale. Chaque couleur représente un phénotype particulier (pathologie, trait quantitatif …). NHGRI GWAS catalog (http://www.genome.gov/26525384).

E.II.2. A l'héritabilité manquante …

Après une période de grand espoir quant à la capacité de cette stratégie à mettre en évidence les mécanismes moléculaires en jeu dans les maladies complexes, les chercheurs adoptaient une attitude plus critique. Il apparaissait que beaucoup de polymorphismes génétiques fréquents associés aux pathologies étudiées ne présentaient qu'un effet individuel faible et, ensemble, n'expliquaient qu'une part réduite de l'héritabilité attendue (Figure 35, ci-dessous).

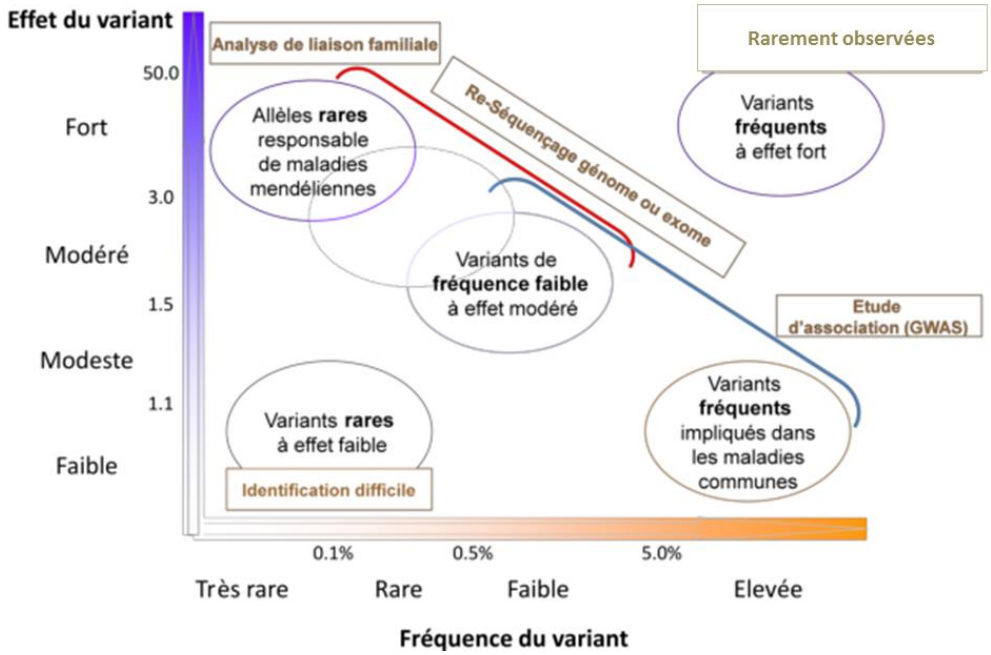


Figure 35 : fréquence de l'allèle à risque et de l'intensité de son effet (odds ratio)

En abscisse : Fréquence du variant dans la population générale.
En ordonnée : Effet du variant. Les variants identifiés à ce jour sont très rarement des variants fréquents à effet fort. Ces découvertes ont par ailleurs été faites avant la mise en œuvre des GWAs. Actuellement, des stratégies complémentaires de séquençage du génome dans des cohortes phénotypiquement très homogènes et de génotypage dans des cohortes plus importantes sont combinées pour identifier des polymorphismes causaux dans tout le spectre des effets et des fréquences.
Modifié d'après (Manolio et al. 2009; Tsuji, 2010)

L'exemple de la maladie d'Alzheimer [Saunders et al.,**1993**][Poirier et al.,**1993**][Corder et al.,**1993**], ou de la dégénérescence maculaire [Klein et al.,**2005**][Haines et al.,**2005**][Edwards et al.,**2005**] qui représentaient des archétypes des allèles qu'on s'attendait à trouver dans ces nouvelles études semblaient en fait isolés. Et ce d'autant plus que la découverte de l'effet de l'allèle Apo Eε4 avait été faite bien avant l'ère des GWAS.

En dehors de ces exemples, les effets de la plus grande partie des marqueurs associés à une pathologie ou à un phénotype quantitatif montrent un effet très faible (voir D.III.8 pour une définition des effets). Les Odds-ratio (pour les traits binaires) ou les coefficients de régression (pour les traits quantitatifs) étaient beaucoup plus modestes qu'attendu et qu'observés dans l'étude de la Dégénérescence Maculaire Liée à l'Age (DMLA).

Un des cas les plus évidents concerne la taille. La corrélation entre parents et enfant pour ce phénotype facilement mesurable et les estimations de l'héritabilité faites par un grand nombre d'études sont autour de 80% [Perola et al.,**2007**][Visscher et al.,**2006**]. En 2009, au moins trois études ont identifié des variants associés à la taille [Gudbjartsson et al.,**2008**][Lettre et al.,**2008**][Weedon et al.,**2008**]. Or, il en ressortait de façon évidente que les 54 variants génétiques mis en évidence expliquaient une part infime de l'héritabilité

[Maher,**2008**], moins de 5 %. Dans son commentaire, Maher proposa le terme d'héritabilité manquante («missing heritability»), qui fut aussitôt repris dans [Manolio et al.,**2009**]. On a parlé aussi, de façon plus informelle, de «dark matter», en référence à la matière noire des astronomes et des cosmologues, qui désigne une catégorie de matière hypothétique jusqu'à présent non détectée, invoquée pour rendre compte d'observations, comme par exemple les estimations de masse des galaxies, des amas de galaxies et les propriétés des fluctuations du fond cosmologique. De la même façon nous savons qu'il existe des mutations et des polymorphismes génétiques non encore mis en évidence par les méthodes actuelles.

En parallèle, certains chercheurs mettaient également en cause l'effet des variants fréquents même quand ils étaient associés à une pathologie. La théorie de l'effet synthétique de variants rares [Dickson et al.,**2010**] fait l'hypothèse que les polymorphismes fréquents associés ne reflètent pas l'effet d'un autre SNP également fréquent (et assez proche physiquement sur le chromosome) mais plutôt celui d'un ensemble de variants rares. Ces variants rares, étant récents dans l'évolution, peuvent être situés très loin physiquement des variants fréquents associés, car le nombre de générations n'était pas suffisamment grand pour permettre des recombinaisons (voir D.III.7, p. 25).

F. Projet 1 : Diabète de Type 2

F.I. L'homéostasie du Glucose et le Diabète de Type 2

Le glucose joue un rôle capital dans l'organisme : c'est un substrat catabolique servant (entre autres) au fonctionnement de l'ensemble des cellules de l'organisme dont les muscles, le cerveau et les hématies. La régulation de la glycémie est contrôlée pour maintenir un apport énergétique constant à tous les organes. Elle est régulée par l'insuline principalement ainsi que par ses antagonistes, le glucagon, l'adrénaline, le cortisol en période de stress, et l'hormone de croissance.
Ces hormones sont des messagers primaires qui se fixent sur leur récepteur et activent, par l'intermédiaire de diverses cascades de transduction, les voies métaboliques impliquées dans la régulation de la glycémie (catabolisme et anabolisme).
La régulation de la glycémie met en jeu le système hormonal, ainsi que plusieurs organes (pancréas, foie et rein principalement). Cette régulation fait partie des processus de maintien de l'homéostasie au sein de l'organisme.
La glycémie à jeun *normale* chez l'homme est statistiquement comprise entre 0,80 et 1,10 g/L[Grimaldi,**2005**].
Le glucose est donc le meilleur des éléments lorsqu'il apporte l'énergie nécessaire au bon fonctionnement de l'organisme mais lorsque, mal régulé, il atteint des concentrations trop hautes dans le sang, il devient un élément adverse. Les complications peuvent alors avoir des conséquences graves pour la santé du patient : elles peuvent être d'ordre macro-vasculaire entraînant un risque d'accidents cardio-vasculaires et des accidents vasculaires cérébraux ; d'ordre micro-vasculaire entrainant des rétinopathies, néphropathies et neuropathies.
L'homéostasie du glucose est maintenue par un jeu d'équilibre entre la concentration du glucose dans le sang et son entrée dans les cellules, favorisée par l'insuline et son récepteur, afin de le stocker sous forme de glycogène (Figure 36, ci-dessous).
Dans ce contexte, le diabète de type 2 est une maladie métabolique caractérisée par l'hyperglycémie. Cette hyper-glycémie peut résulter, en général, soit de dysfonctionnements de la sécrétion d'insuline par les cellules béta pancréatiques, soit d'un défaut de l'action de l'insuline.
Le Diabète de Type 2 est donc un trouble de l'assimilation, de l'utilisation et/ou du stockage du glucose. Son diagnostic repose sur le calcul du taux de glycémie à jeun et deux heures après une Hyper-Glycémie à jeun Provoquée par voie Orale (HGPO). Selon les critères de l'Organisation Mondiale de la Santé, un individu est déclaré diabétique lorsque la Glycémie à jeun dépasse 7 mmol/l et/ou lorsque la Glycémie 2 heures après HGPO est supérieure à 11.1 mmol/l.

Cette maladie est d'autant plus grave qu'un excès chronique de sucres dans le sang, comme nous l'avons vu, s'associe à long terme à des dommages touchant divers organes incluant plus particulièrement les yeux, les reins, les nerfs, le cœur et les vaisseaux sanguins.

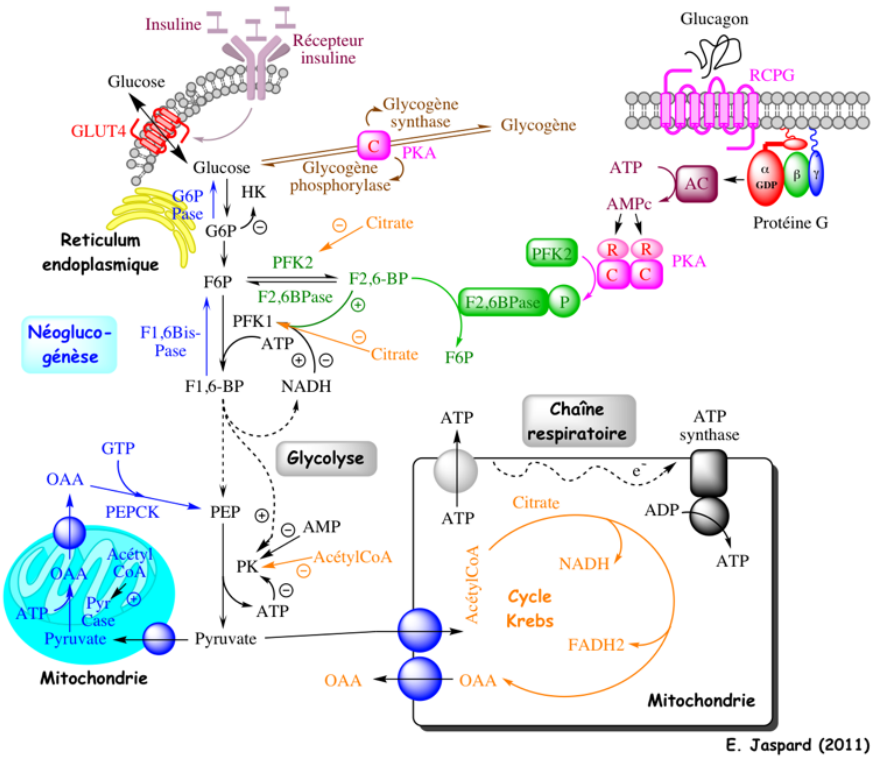


Figure 36 : le Glucose et son cycle intracellulaire dans le foie

Le Glucose rentre dans la cellule par le transporteur GLUT4, qui est activé par l'Insuline à travers le récepteur à l'insuline. Le glucose est alors stocké en Glycogène. Inversement, le Glucagon, également par l'action de son récepteur, induit la néo-synthèse de glucose par Glycogénolyse. Source http://ead.univ-angers.fr/~jaspard/ , tous droits de reproduction sont réservés et strictement limités.

Lorsque la concentration en insuline est faible, le transporteur du glucose GLUT4 est localisé dans des vésicules de stockage des cellules hépatiques et musculaires.
Quand le niveau de glucose circulant est élevé, l'insuline facilite l'entrée du glucose dans les cellules. Cette accélération se fait par augmentation de la synthèse et de la translocation de GLUT4 des compartiments de l'endosome vers la membrane plasmique. L'absorption du glucose augmente. Dans le sens inverse, le glucagon induit la glycogénolyse et la libération du glucose dans le sang.

F.I.1. Insuline et Glucagon

Ces deux protéines sont les principaux régulateurs de l'entrée du glucose dans les cellules. Elles sont secrétées dans la partie endocrine du pancréas : les îlots de Langerhans [Langerhans, Paul,**1869**]. Un pancréas humain peut contenir jusqu'à 10^6 îlots et chaque îlot peut contenir jusqu'à 5000 cellules. Ces cellules se répartissent en cinq types :
- les cellules Bêta (~60% des cellules endocrines d'un îlot de Langerhans humain adulte) qui sécrètent l'insuline;
- les cellules alpha (20-30%) qui sécrètent le glucagon;
- les cellules delta (~10%) qui sécrètent la somatostatine ;
- les cellules PP (<5%) qui sécrètent le polypeptide pancréatique ;
- les cellules epsilon (~1%) qui sécrètent la ghréline.

Les îlots de Langerhans sont très fortement vascularisés, avec un système de capillaires très fins, et innervés et pouvant ainsi répondre de façon très adaptée aux différents stimuli physiologiques.

L'insuline est secrétée dans les cellules Bêta des ilôts de Langerhans du pancréas, en réponse à la concentration élevée du glucose. Par ailleurs, ces cellules stockent également une masse importante de molécules d'insuline afin de pouvoir réagir rapidement aux stimuli extérieurs.

F.II. Description de la pathologie

Le diabète de type 2 est une pathologie caractérisée par un excès de sucre dans le sang, due à une absence de régulation appropriée de l'homéostasie glucidique. Depuis quelques années, la maladie touche également des individus plus jeunes, particulièrement s'ils sont en surpoids. Le diabète de type 2 représente environ 90 % de tous les diabètes.
Le nom vient du grec *Diabetes* qui signifie « siphon » et qui rend compte de l'idée que les fluides ne restent pas dans le corps mais passent juste au travers. Il s'agit en fait d'une conséquence de la forte concentration de sucre dans le sang qui entraîne une évacuation rapide et intensive de l'eau.

Les cellules de l'organisme ne reconnaissent plus l'insuline, en général en raison de stimuli extérieurs comme l'excès de poids et la sédentarité, et le pancréas se voit contraint d'en produire toujours davantage pour compenser le déficit en insuline active. Au cours du temps, le pancréas se fatigue et réduit sa production d'insuline, qui peut alors complètement cesser. Il en résulte un manque d'insuline et une hyperglycémie.

Cette maladie résulte en fait de l'interaction de deux mécanismes : tout d'abord, un dysfonctionnement des cellules bêta du pancréas dont on a vu qu'elles fabriquent l'insuline. On a, par exemple observé que chez les patients diabétiques (Type 2), la masse des cellules bêta est fortement réduite (30 à 60%) [Marchetti et al.,**2004**][Butler et al.,**2003**][Guerra et al.,**2005**].
Le deuxième mécanisme pourrait être la résistance des cellules à l'insuline, ou insulino-résistance : l'insuline peine à faire entrer le sucre dans les cellules pour leur apporter leur énergie. La production de cette hormone augmente pour essayer de « forcer » l'entrée du sucre dans les cellules afin de maintenir une glycémie normale.
Lorsque cette première mesure ne suffit plus, le pancréas s'épuise et le sucre s'accumule dans le sang et c'est le Diabète de Type2.

Cette résistance à l'insuline apparaît normalement avec le vieillissement, mais elle est surtout provoquée par de mauvaises conditions hygiéno-diététiques : l'excès calorique et la sédentarité. Si le muscle ne travaille pas, c'est le cas de la sédentarité, ou si les réserves en graisses sont très importantes et utilisées préférentiellement comme dans l'obésité, les cellules deviennent moins sensibles à l'insuline.

F.III. Épidémiologie

Si Arétée de Cappadoce décrit le diabète comme une maladie «... heureusement plutôt rare …» en son temps (haute Antiquité), il n'en est plus du tout de même de nos jours. Le diabète a en effet aujourd'hui atteint des proportions épidémiques : 240 millions d'individus sont atteints de diabète de par le monde et il fait autant de victimes que le sida (3,8 millions de décès estimés dans le monde en 2007) (http://www.diabetesatlas.org/).
En 50 ans, l' « épidémie » s'est développée dans les populations dont la façon de vivre a radicalement changé : passant d'un modèle traditionnel imprégné d'agriculture à un mode de vie «occidentalisé ». La prévalence du diabète dans la population adulte des pays industrialisés est de l'ordre de 6% [Stumvoll et al.,**2008**][Stumvoll et al.,**2007**]. Le nombre de personnes diabétiques atteigne 380 millions (80% habitant dans les pays en voie de développement) , soit plus de 7% de la population adulte mondiale [Stumvoll et al.,**2008**][Stumvoll et al.,**2005**].

La prévalence du diabète traité a augmenté, en France métropolitaine de 2,6 à 3,6% entre 2000 et 2005. En 2007, la prévalence du diabète traité en France (incluant les régions d'outre-mer) était de 3,95%, soit 2,5 millions de personnes atteintes. Il existait un risque accru chez les hommes ayant un âge supérieur à 40 ans et près d'un homme sur cinq était diabétique à 75 ans.
Le nombre d'individus souffrant de diabète en France, en 2016 est estimé à 2,8 Millions, ce qui correspondrait à une prévalence de 4,5 [Kusnik-Joinville et al.,**2008**][Bonaldi et al.,**2011**].

Les conséquences en termes de santé publique sont dévastatrices. Selon le site de l'Association Française des Diabétiques (http://www.afd.asso.fr/sites/default/files/dp-2013-AFD-bat-web.pdf), en France, en 2011, les coûts pour l'Assurance Maladie des personnes traitées pour le diabète s'élevaient à 17,7 milliards d'euros (10 % des dépenses de santé). En France, le nombre de personnes diabétiques a augmenté de 90 % en 10 ans 1 et l'on compte aujourd'hui plus de 3 millions de personnes diabétiques, soit près de 5 % de la population. Selon les prévisions, un Français sur dix sera touché par le diabète d'ici 10 ans si l'on ne fait rien.
Par ailleurs, les complications entrainées par le diabète ont des conséquences fortement délétères sur la qualité et l'espérance de vie de patients et ainsi sur les frais de prise en charge et les systèmes de santé publique en général.

Le 20 décembre 2006, l'Assemblée générale des Nations Unies votait la Résolution 61/225, reconnaissant le diabète en tant que «maladie chronique, invalidante et coûteuse, associée à de lourdes complications qui représentent des risques graves pour les familles, les pays et le monde entier».
La découverte des bases étiologique de la survenue de cette maladie apparaît donc cruciale.

F.IV. Génétique

Le Diabète de Type 2 et la génétique ont une longue histoire de recherche commune. Du fait de sa prévalence, cette pathologie se prêtait assez facilement aux analyses de jumeaux et de ségrégation permettant d'estimer l'héritabilité de cette pathologie et notre capacité à identifier des gènes responsables de sa survenue. Par ailleurs, l'existence et la facilité (relative) d'accès à des phénotypes métaboliques intermédiaires (glycémie et insulinémie à jeun, à 2 heures etc …) permettait d'étendre l'estimation de l'héritabilité de cette pathologie par comparaison avec des phénotypes très corrélés dans différentes populations. Par ailleurs, le grand nombre de personnes atteintes en France et dans le monde était un gage de réussite des études génétiques, du fait qu'elles permettaient la constitution de grandes séries de cas et de témoins.

F.IV.1. Études de jumeaux

Les études de jumeaux sont fréquentes dans la recherche sur les maladies métaboliques. La plupart de ces études concluent à une héritabilité assez forte du diabète de Type 2 ou de l'index de glycémie. Il a par exemple été observé que la concordance de survenue du Diabète de Type 2 pour des jumeaux monozygotes allait de 50 à 90%, supérieure à la concordance entre jumeaux dizygotes – 37%[Beck-Nielsen et al.,**2003**][Florez et al.,**2003**][Poulsen et al.,**1999**]. Enfin, l'héritabilité proprement dite du diabète de Type 2 est de 26% [Poulsen et al.,**1999**].

F.IV.2. Études familiales

La plupart des études familiales d'héritabilité ont été effectuées sur les variables de l'homéostasie du glucose (fonction des cellules Bêta, glycémie à jeun ..) et concluaient à une forte héritabilité[Hanson et al.,**2001**][Poulsen et al.,**1999**][Mills et al.,**2004**].
Les preuves de l'existence d'une composante polygénique ont fait l'objet d'une revue très complète par Permutt et Cox [Permutt et al.,**2005**].

F.IV.3. Gènes identifiés

Au début de mon travail de thèse, en 2009, plusieurs études avaient déjà mis en évidence l'existence de plusieurs associations génétiques, dont un travail auquel j'ai participé en tant que co-auteur (2e rang à égalité)[Sladek et al.,**2007**] – hors période de thèse.
À ce moment, une vingtaine de gènes était mise en évidence par l'analyse d'association (Figure 37).
Parmi les gènes candidats on retrouvait le gène du récepteur nucléaire *PPARG* [Deeb et al.,**1998**][Altshuler et al.,**2000**] ainsi que le gène codant pour le canal potassique *KCNJ11* (E23K) [Gloyn et al.,**2003**][Laukkanen et al.,**2004**].
Cinq analyses d'association génome entier (incluant celle à laquelle j'ai participé [Sladek et al.,**2007**] avaient ensuite identifié 18 localisations génomiques (plutôt que des gènes à proprement parler)[Zeggini et al.,**2007**][Steinthorsdottir et al.,**2007**][Scott et al.,**2007**][Zeggini et al.,**2008**]. La plupart de ces associations correspondent à des loci contenant des gènes qui semblent impliquer la voie des cellules Bêta [Perry, Frayling,**2008**].

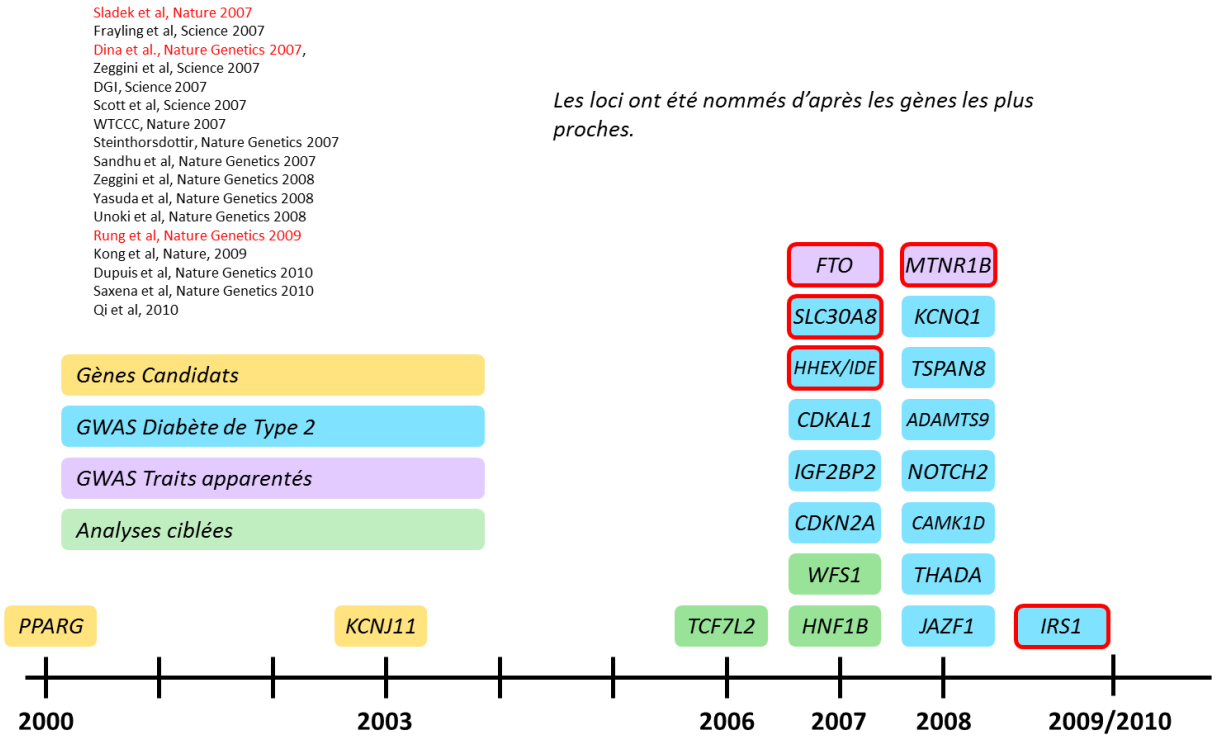


Figure 37 : Loci associés au Diabète de Type 2 en 2009
Les gènes entourés en rouge proviennent d'études auxquelles j'ai participé (hors thèse). Les analyses ciblées sont soit des analyses de liaison suivies de recherche de gènes, soit des analyses bio-informatiques suivies de tests d'association.

La recherche de nouveaux gènes impliqués dans la survenue du diabète de type 2 ainsi que la mise en évidence de l'architecture génétique étaient initiés depuis longtemps et très avancées au moment où j'ai réalisé mes travaux.

F.V. Description de l'étude

DIAGRAM + est un Consortium transatlantique dont le but est d'élucider, dans un effort commun, les bases génétiques du Diabète de Type 2 (Figure 38).
Il regroupe principalement des équipes ayant collecté des échantillons cas-contrôles (ou cas-population) et qui ont génotypé ces individus. Certaines équipes avaient déjà publié leurs analyses d'association génome entier individuellement.

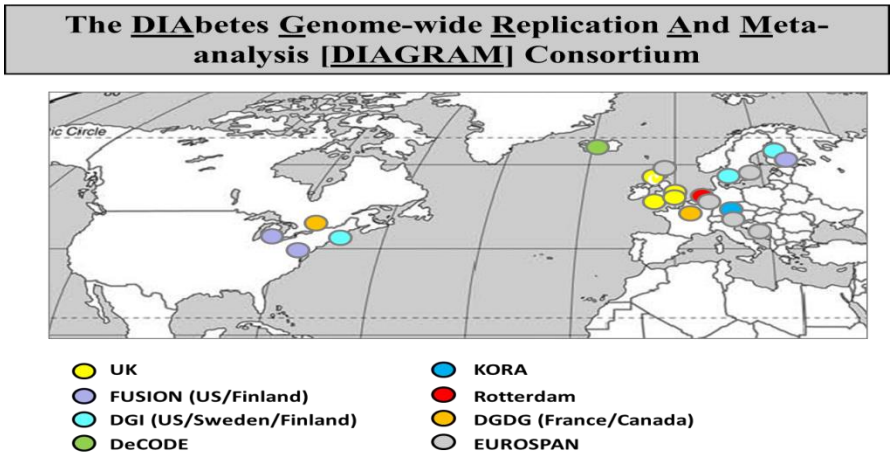


Figure 38 : Équipes et populations dans le Consortium DIAGRAM+
Chaque couleur représente un membre du Consortium, chaque membre peut être lui-même une fédération d'équipes

Je présente ici l'article principal de ce travail (Voight, Scott, Steinthorsdottir, Morris, **Dina** et al., 2010 – que j'appelle DIA1) où j'ai été co-auteur en première position à égalité, et le papier suivant, où je suis également co-auteur (Morris et al. 2012 – DIA2), et qui est la suite de cette étude. Dans ce deuxième papier, j'ai participé de façon moins importante que dans le premier, à la suite de l'étude initiée dans DIA1.

Ce travail portait sur l'analyse d'un très grand nombre de cas et de témoins en meta-analyse portant sur 43 338 cas (dont 8 130 dans la phase initiale) et 101 150 contrôles (dont 39 000 dans la phase initiale).

F.VI. Résultats

F.VI.1. Nouvelles régions géniques

Nous avons identifié, dans l'analyse DIA1, par méta-analyse, 11 nouvelles associations dépassant le seuil de $5x10^{-8}$ en p-valeur.

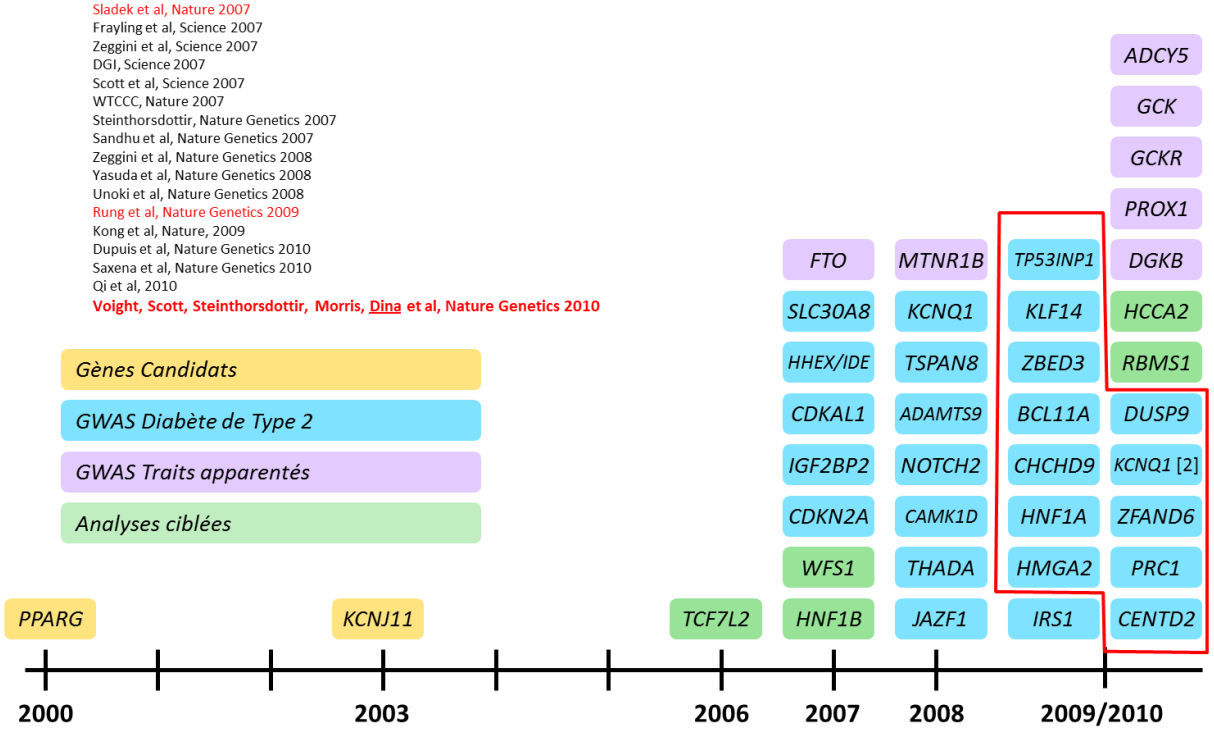


Figure 39 : Loci associés au Diabète de Type 2 en 2010
Entourés de rouge figurent les gènes identifiés par l'étude Voight et al. 2010 ; qui fait l'objet d'une partie de mon travail de thèse.

Entre autres, nous avons identifié le premier gène sur le chromosome X, DUSP9 (Figure 39).

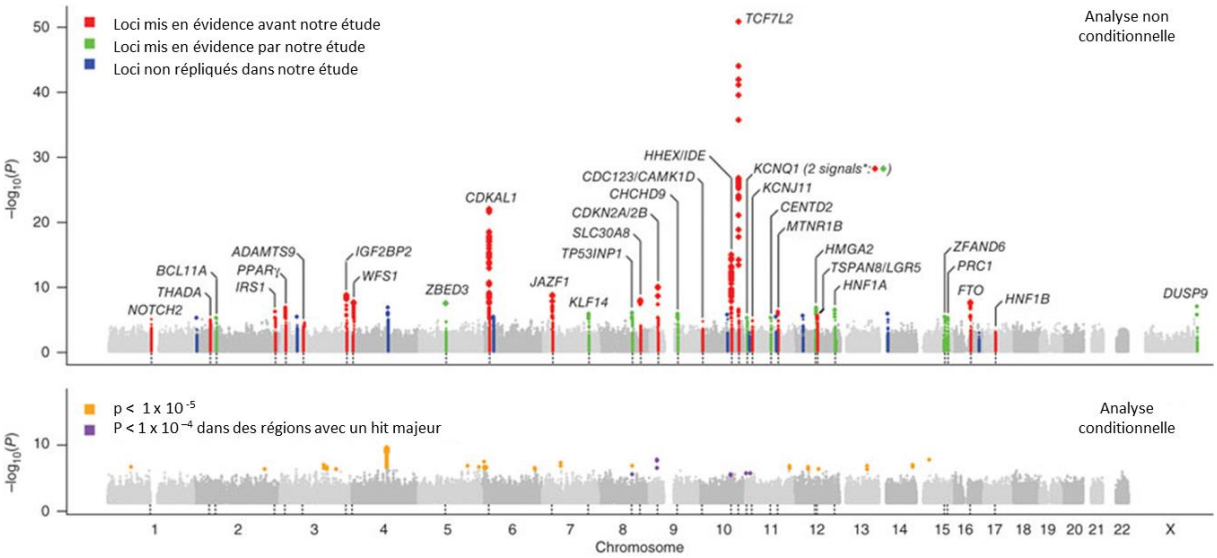


Figure 40 Manhattan Plot des résultats de la Meta-Analyse DIAGRAM+

Ce Manhattan plot montre l'état de la recherche au moment de la publication de [Voight et al.,**2010**], premier article de ma thèse. L'analyse non conditionnelle est l'analyse d'association classique des SNPs. Les ensembles de SNPs mis en rouge représentent les loci mis en évidence avant la parution de notre étude. Les groupes de SNPs en vert représentent les groupes de SNPs mis en évidence par note étude. Dans l'analyse conditionnelle, nous avons testé l'association pour chaque SNP du génome après correction sur le génotype des SNPs les plus associés dans chacun des 39 loci.

La Figure 40 souligne bien que dans notre étude, qui est le troisième effort d'analyse génome entier, les signaux d'association (points en vert) sont de moins en moins forts. Les effets («Odds Ratios») sont de plus en plus petits et la proportion de la variance expliquée diminue également (Tableau 7).

SNP	Chr	Position B36 (Mb)	Allèle référence[b]	Allèle à risque[b]	f	Gène proche[c]	Etape 1[d]		Etape 2[d]		Etape 1 et 2[d]	
							OR (95%CI)	*p*- value	OR (95%CI)	*p*- value	OR (95%CI)	*p*- value
							max 8,130 patients et 38 987 témoins		max 35 240 patients et 62 163 témoins		max 43,370 cases et 101 150 témoins	
Nouveaux loci												
rs155222	11	72110746	C	A	0,88	*CENTD2*	1.13 (1.07-1.19)	7.0×10^{-6}	1.14 (1.11-1.18)	3.2×10^{-18}	1.14 (1.11-1.17)	1.4×10^{-22}
rs243021	2	60438323	G	A	0,46	*BCL11A*	1.09 (1.05-1.13)	8.1×10^{-6}	1.08 (1.06-1.10)	6.2×10^{-11}	1.08 (1.06-1.10)	2.9×10^{-15}
rs231362	11	2648047	A	G	0,52	*KCNQ1*	1.11 (1.06-1.16)	6.4×10^{-6}	1.07 (1.05-1.09)	3.2×10^{-9}	1.08 (1.06-1.10)	2.8×10^{-13}
rs445705	5	76460705	A	G	0,26	*ZBED3*	1.16 (1.10-1.23)	4.2×10^{-8}	1.07 (1.04-1.10)	2.7×10^{-7}	1.08 (1.06-1.11)	2.8×10^{-12}
rs972283	7	130117394	A	G	0,55	*KLF14*	1.10 (1.06-1.15)	1.8×10^{-6}	1.06 (1.03-1.09)	6.4×10^{-6}	1.07 (1.05-1.10)	2.2×10^{-10}
rs804268	15	89322341	C	A	0,22	*PRC1*	1.10 (1.06-1.15)	8.2×10^{-6}	1.06 (1.03-1.08)	1.6×10^{-6}	1.07 (1.05-1.09)	2.4×10^{-10}
rs896854	8	96029687	C	T	0,48	*TP53INP1*	1.10 (1.06-1.15)	1.2×10^{-6}	1.05 (1.03-1.08)	2.2×10^{-5}	1.06 (1.04-1.09)	9.9×10^{-10}
rs116343	15	78219277	A	G	0,60	*ZFAND6*	1.11 (1.06-1.16)	5.1×10^{-6}	1.05 (1.03-1.08)	1.2×10^{-5}	1.06 (1.04-1.08)	2.4×10^{-9}
rs153134	12	64461161	G	C	0,10	*HMGA2*	1.20 (1.12-1.29)	1.7×10^{-7}	1.08 (1.04-1.12)	1.1×10^{-4}	1.10 (1.07-1.14)	3.6×10^{-9}
rs795719	12	119945069	A	T	0,85	*HNF1A*	1.14 (1.08-1.19)	4.6×10^{-7}	1.05 (1.02-1.09)	4.6×10^{-4}	1.07 (1.05-1.10)	2.4×10^{-8}
rs132921	9	81141948	T	C	0,93	*CHCHD9*	1.20 (1.11-1.29)	1.5×10^{-6}	1.08 (1.04-1.13)	2.4×10^{-4}	1.11 (1.07-1.15)	2.8×10^{-8}
rs594532	X	152553116	A	G	0,79	*DUSP9*	1.25 (1.14-1.37)	2.3×10^{-6}	1.32 (1.16-1.49)	2.3×10^{-5}	1.27 (1.18-1.37)	3.0×10^{-10}
Loci connus												
rs757832	2	226728897	G	A	0,64	*IRS1*	1.12 (1.07-1.17)	8.7×10^{-7}	1.10 (1.08-1.13)	2.2×10^{-15}	1.11 (1.08-1.13)	5.4×10^{-20}
rs138715	11	92313476	C	T	0,28	*MTNR1B*	1.12 (1.07-1.17)	1.0×10^{-6}	1.08 (1.05-1.10)	4.4×10^{-10}	1.09 (1.06-1.11)	7.8×10^{-15}

Tableau 7 : résultats d'association des étapes 1 + 2 (P valeur < 5 × 10 −8)

Analyses d'association et méta-analyse. La méta-analyse est effectuée selon la méthode d'inverse de la normale. Les « loci connus » sont des loci qui ont été mis en évidence par d'autres études pendant que nous menions la méta-analyse.
[a] – des SNPs en DL ont été utilisés dans certaines études. [b] - . [c] - Le gène indiqué correspond au gène le plus proche du SNP associé où au meilleur candidate dans la région d'association. [d] – p-valeurs sont bilatérales.

Les analyses conditionnelles ne mettent pas en évidence d'association extrêmement forte.
Il apparaît donc, qu'au-delà de l'importance relative de chaque nouveau locus (et gène) mis en évidence, le but de ces analyses est plutôt de chercher une information physiologique à partir de l'ensemble des gènes mis en évidence.

L'autre objectif est d'identifier un ensemble de marqueurs modérément associés mais n'atteignant pas le niveau d'association génome-entier afin de créer une puce de réplication appelée Cardio-Metabochip [Voight et al.,**2012**]. Ce projet est né de cette observation que les effets étant de plus en plus petits, la détection de nouvelles associations nécessiterait de génotyper un nombre d'individus de plus en plus grand (de l'ordre de centaines de milliers). Génotyper ces nouveaux individus à l'échelle du génome nécessitait un investissement financier trop grand. L'alternative proposée dans le projet Cardio-Metabochip (Figure 41) était de créer une puce avec un nombre réduit de SNPs et donc d'un coût réduit, afin que tous les laboratoires ayant à disposition des cas et des témoins puissent contribuer à la recherche génétique des bases des moléculaire des phénotypes cardiovasculaires et métaboliques.

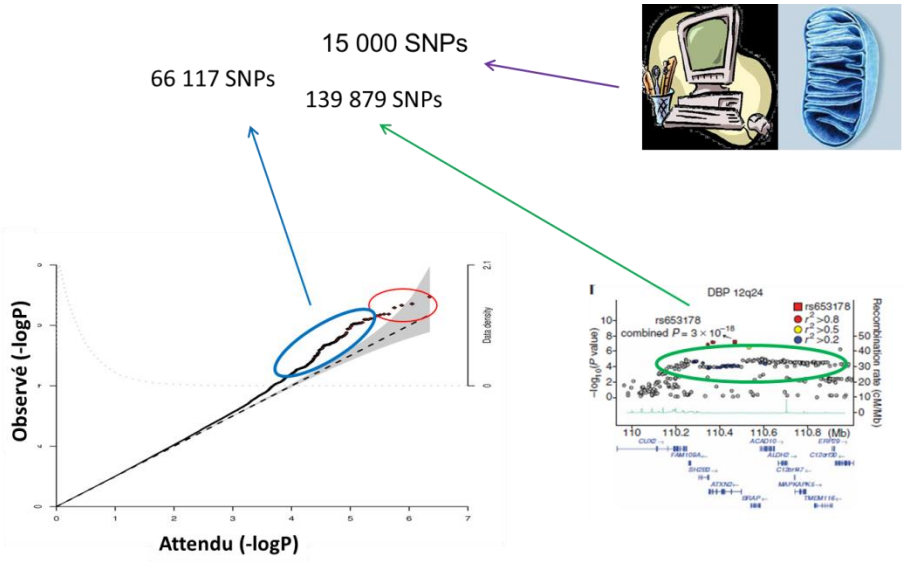


Figure 41 : Cardio-Metabo Chip (220996 SNPs)

En rouge – SNPs associés p < $5x10^{-8}$, En bleu – SNPs avec association suggérée $1x10^{-5}$ > p > $5x10^{-8}$, et entourés en vert - SNPs sélectionnés dans les régions d'association (<$5x10^{-8}$) afin d'étudier le signal (recherche des SNPs causaux).
Environ 200 000 SNPs ont été sélectionnés pour satisfaire trois stratégies différentes qui sont : le test de SNPs non encore associés à $5x10^{-8}$, la localisation, dans une zone d'association déjà définie, des variants les plus associés et potentiellement causaux et le test d'un petit nombre de SNPs afin de vérifier des hypothèses biologiques souvent peu disponibles dans les puces classiques (comme les SNPs du génome mitochondrial ou des SNPs identifiés par analyse *in silico*).

L'utilisation de cette puce à ADN, en réplication sur 34 840 patients et 114 981 témoins issus de 25 études cas-témoins, a permis de détecter, dans un deuxième temps (3e papier de ma thèse) ; [Morris et al.,**2012**], 8 nouvelles associations.

F.VI.2. Nouvelles voies biosynthétiques

Deux stratégies ont été adoptées et permis d'affiner nos connaissances des bases moléculaires et physiologiques du diabète de type 2. Une première analyse a été effectuée dans le premier papier [Voight et al.,**2010**]. Ces analyses ont été étendues dans la deuxième

étude [Morris et al.,**2012**] où un plus grand nombre de gènes ont été mis en évidence. Toutefois, je présenterai plus spécifiquement les résultats du premier article [Voight et al.,**2010**] où ma participation a été plus importante, en décrivant brièvement ce que le deuxième papier a apporté.
Dans ce paragraphe, j'appellerai la liste de gènes se retrouvant dans les zones d'association T2D_genes. Chaque zone d'association sera appelée REGION_i, i étant le numéro de la région. Le numéro de la région est donné dans l'ordre en fonction de la force d'association du SNP le plus associé.

F.VI.2.a) *Association des SNPs avec les données quantitatives*

Pour connaître le mode d'action des polymorphismes et gènes associés au Diabète de Type 2, nous avons testé l'effet des SNPs associés à l'échelle du génome sur des phénotypes liés au diabète de Type 2, dans une population générale. Les deux phénotypes que nous avons étudiés mesurent l'Insulino-résistance et la sensibilité à l'insuline. Il s'agit d'une combinaison des mesures de l'insulinémie et de la glycémie à jeun [Matthews et al.,**1985**]. La variable HOMA-B (mesurant l'activité des cellules Bêta) est essentiellement un rapport entre insulinémie et glycémie. A l'opposé, la variable HOMA-IR (mesurant l'Insulino-résistance) est essentiellement un produit entre ces deux valeurs. L'efficacité de ces mesures pour évaluer l'activité des cellules bêta et l'insulino-résistance ont été extensivement testées avec succès [Wallace et al.,**2004**].

À côté du Consortium DIAGRAM+, il existe le consortium MAGIC (http://www.magicinvestigators.org/) dont le but est de mettre en évidence des associations avec des traits de la glycémie, de l'insulino-résistance et des traits métaboliques en général - par exemple [Saxena et al.,**2010**]. Pour les SNPs associés au diabète de type 2, l'association avec phénotypes liés à l'homéostasie du glucose aussi bien dans ce Consortium MAGIC que dans l'effort « Metabo-Chip » a été utilisée afin de générer des hypothèses sur l'effet des gènes que nous avons mis en évidence.
Dans chaque étude nos collaborateurs ont estimé l'association de tous les marqueurs à l'échelle du génome et ensuite combiné les p-valeurs en utilisant la méthode de l'inverse normale (D.VI.2.b), exactement comme dans l'analyse principale sur le diabète de Type 2. Les résultats pour les 38 SNPs permettent de rechercher les modes d'action des gènes causant le diabète de Type 2, en fonction de leur association avec la fonction des cellules bêta, l'insulino-sensibilité ou avec les deux (Figure 42).

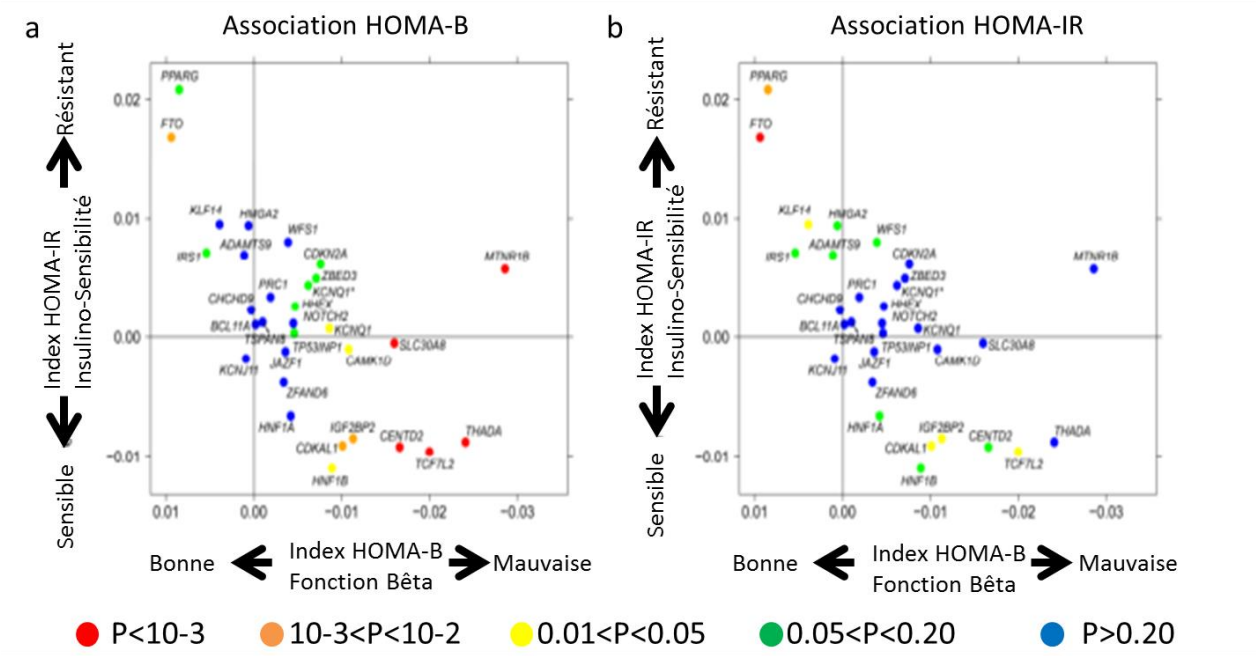


Figure 42 : effet des SNPs T2D sur la fonction des cellules bêta et l'insulino-résistance

Chaque point représente un SNP associé ($p< 5x10^{-8}$) au diabète de Type 2 dans la méta-analyse DIAGRAM+ [Voight et al.,**2010**]. L'association de ces SNPs avec les traits quantitatifs HOMA-B et HOMA-IR a été testée dans les cohortes du Consortium MAGIC et nous avons conservé les effets (bs) du SNP à risque (celui qui augmente le risque du diabète de Type 2) ainsi que les p-values d'association. Les noms des gènes ont été choisis en raison de leur proximité avec le SNPs montrant la plus forte association dans la région. Les coordonnées correspondent aux effets (β issu de la régression linéaire) de chaque SNPs sur les traits (sur HOMA-B en abscisse et HOMA-IR en ordonnée). Les codes couleurs indiquent le degré d'association entre le SNP et le trait quantitatif : HOMA-B en **a/** et HOMA-IR en **b/**.

Il apparaît que pour un nombre de SNPs important, l'allèle augmentant le risque de diabète de Type 2 est également associé à une diminution de la fonction des cellules Bêta.
A l'inverse, il ne semble pas que ces SNPs soient associés à une insulino-résistance particulière (à l'exception du SNP proche du gène *FTO*).
Nous avons conclu, dans ce premier papier [Voight et al.,**2010**], que les bases génétiques du diabète de Type 2 semblaient liées à une dégradation de la fonction des cellules bêta. Cette observation est confirmée dans le deuxième papier [Morris et al.,**2012**].

F.VI.2.b) *Voies biologiques des gènes associés*

Nous avons ensuite testé les liens entre les gènes proches des SNPs les plus associés afin d'identifier une surreprésentation de ces gènes dans une voie biologique particulière ou un excès d'interactions de ces gènes entre eux par la méthode GRAIL [Raychaudhuri,**2011**].
Cette méthode teste la « connectivité » des gènes se trouvant dans les régions d'association (REGION_i).

Les résultats peuvent être représentés sous la forme d'un graphique de connexion (Figure 43).

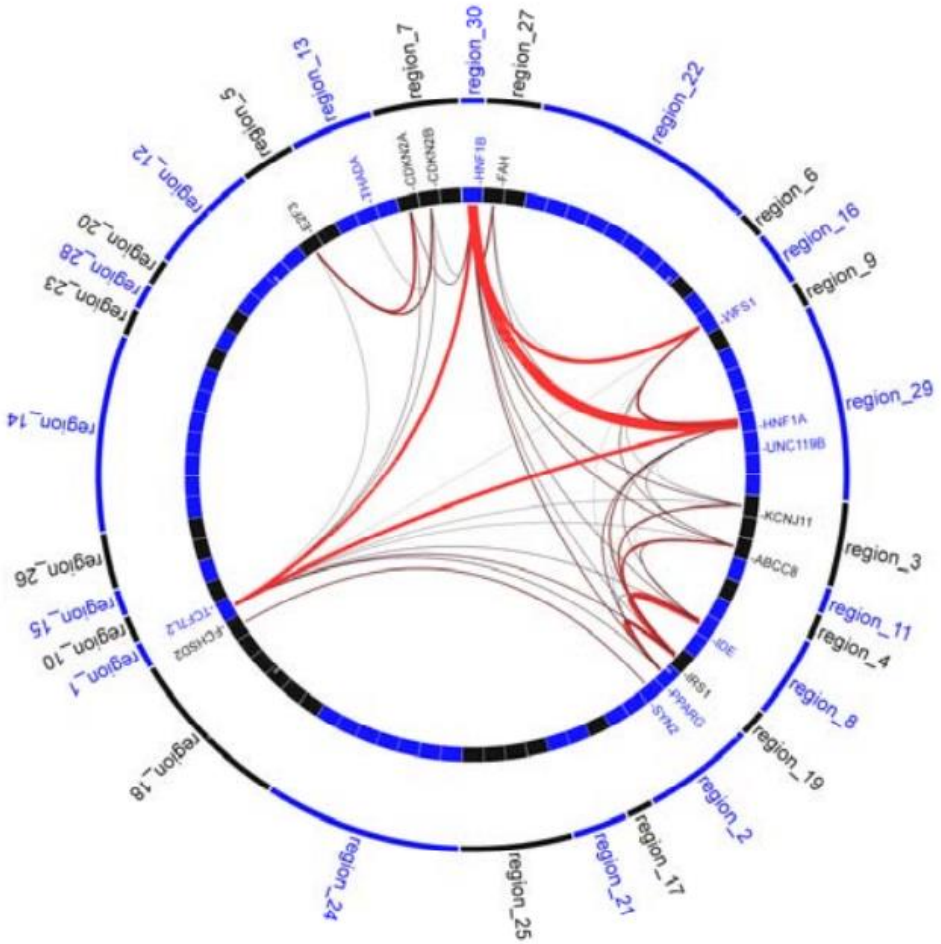


Figure 43 : graphe circulaire de connectivité des gènes du diabète de Type 2

Ce graphique représente le degré de connectivité entre les 30 loci (REGION_i ; i allant de 1 à 30) associés au diabète de Type 2. La connectivité est testée sur les articles publiés avant 2006 – afin d'éviter des redondances de résultats issus d'analyses EAGE précédentes. La significativité de la connexion est traduite par l'épaisseur de l'arc qui joint deux gènes. Les gènes dont le nom apparaît sont ceux dont la connectivité est significative à un niveau de 5% (p-valeur < 0.05).

Une connectivité importante est apparue, mais elle repose principalement sur des gènes dits mono-géniques, c'est à dire des gènes où l'on a trouvé des mutations ayant un effet très fort sur le diabète de Type 2. Le détail des résultats est par ailleurs montré dans le Tableau 8. Le terme le plus cité y est « diabète », reflétant l'importance des gènes déjà découverts en lien avec le diabète de Type 2 avant l'avènement des EAGE (et qu'on appelle gènes mono-géniques dans l'article). Nous retrouvons également les termes « bêta » et « pancreatic », soulignant l'importance des défauts de développement des cellules Bêta dans la pathogénèse du diabète de Type 2.

Région	p-valeur	Gène	Terme
region_29	1.2E-04	HNF1A	'diabetes'
region_30	4.2E-04	HNF1B	'insulin'
region_8	2.4E-03	IDE	'cyclin'
region_13	3.7E-03	THADA	'tumors'
region_3	6.0E-03	ABCC8	'beta'
region_19	0.016	IRS1	'pancreatic'
region_7	0.020	CDKN2B	'glucose'
region_16	0.025	WFS1	'syndrome'
region_1	0.045	TCF7L2	'catenin'
region_9	0.054	SLC30A8	'cytogenetic'
region_15	0.056	ADAMTS9	'mutation'
region_27	0.075	FAH	'benign'
region_5	0.078	E2F3	'degrading'
region_17	0.093	KCNQ1	'cases'
region_18	0.098	FCHSD2	'patients'
region_11	0.11	TSPAN8	'tumor'
region_2	0.11	PPARG	'ppargamma'
region_10	0.15	JAZF1	'zinc'
region_21	0.17	AGGF1	'mutations'
region_20	0.17	BCL11A	'hepatocyte'
region_23	0.19	HMGA2	
region_25	0.21	CCNE2	
region_28	0.27	TLE4	
region_6	0.37	IGF2BP2	
region_14	0.39	PPIAL4	
region_22	0.45	FURIN	
region_24	0.54	CPA5	
region_12	0.64	SEC61A2	
region_4	0.65	FTO	
region_26	0.66	FAT3	

Tableau 8 : détail des résultats GRAIL.

Les régions (**Région**) identifiées dans l'analyse d'association génome entier sont rangées dans l'ordre de leur p-valeur GRAIL (**p-valeur**). Dans chaque région nous présentons le gène le plus connecté (**Gène**). La dernière colonne (**Terme**) inclut les mots qui sont surreprésentés dans les résumés d'articles reliant les gènes les plus connectés. Attention : les termes ne sont pas spécifiques à un gène donné !

Nous avons également utilisé la base de données PANTHER (http://www.pantherdb.org/)[Thomas et al.,**2003**][Mi et al.,**2005**] afin de rechercher des processus biologiques où les gènes se trouvant aux loci du diabète de Type 2 soient statistiquement surreprésentés. Nos gènes ont donc été comparés à une liste de 25 431 gènes humains de référence, dont 14 110 ayant un processus d'annotation disponible. Nous avons testé, pour chaque processus biologique, si le nombre de gènes issus de la liste T2D_genes et présents dans ce processus était statistiquement supérieur au nombre attendu.
La significativité a été testée par simulation : nous avons généré 20 000 listes aléatoires de gènes (imitant la liste originale) et testant à chaque fois la surreprésentation de gènes dans chacun des 242 processus biologiques.

Nous n'avons pas identifié de processus biologique avec un excès significatif de gènes issus de T2D. Les processus les plus positifs (mais non significatifs) étaient la « Signalisation cellulaire par adhésion » ("Cell adhesion-mediated signaling"), la « Transduction du signal » ("Signal transduction") et le « contrôle du cycle cellulaire » ("Cell cycle control").

F.VI.2.c) *Conclusion*

Il apparait que nous n'avons pas réussi à mettre en évidence de façon claire des bases physiologiques impliquées dans l'hérédité du diabète de Type 2. Nous pouvons raisonnablement inférer de nos résultats que l'effet des variants génétiques sur le diabète de Type 2 passe par une diminution de la fonction des cellules Bêta, plutôt que par l'insulino-résistance. Il semble qu'il y ait une grande hétérogénéité des causes du diabète de type 2 ou alors que les processus importants pour le développement de cette maladie soient peu représentés dans les bases de données existantes. Cette observation contredit les résultats obtenus sur d'autres pathologies comme la Maladie de Crohn où les SNPs les plus associés sont proches de gènes codant pour les cytokines de la famille de l'IL-12/IL-13[Wang et al.,**2009**].

F.VI.3. Architecture génétique du Diabète de Type 2

Les deux articles de ma thèse ont permis d'avancer sur plusieurs aspects de la connaissance de l'architecture génétique du diabète de Type 2. Par architecture génétique j'entends le nombre de marqueurs associés au diabète de type 2, leurs fréquences alléliques (appelé spectre des fréquences alléliques), leurs effets (par exemple en terme d'Odds-Ratio) et par conséquent en terme d'héritabilité (notion vue D.III.8.c).

F.VI.3.a) *Risque familial, héritabilité*

Le risque de récurrence familiale λ_R d'un ensemble de génotypes (décrit en D.III.9.a) est calculé en faisant l'hypothèse que les effets des SNPs sont indépendants. Il s'agit d'un produit de λ_{Ri}, i étant le $i^{ème}$ SNP. Le risque de récurrence, estimé comme dans l'Équation 11, est de 1.094 dans le premier article [Voight et al.,**2010**]. Il passe à 1.10 en ajoutant les nouveaux gènes de l'article [Morris et al.,**2012**].
A partir du λ_R, nous pouvons estimer l'héritabilité due aux loci sélectionnés (Équation 12). Elle a été estimée à 8%.

F.VI.3.b) *Nombre de SNPs associés au Diabète de Type 2 :*

Par ailleurs, et c'est là sans doute la plus grande originalité de cette « méta-étude », une estimation de la proportion de variants fréquents impliqués dans le diabète de type 2 a été menée. La méthodologie est assez intuitive. Nous faisons l'hypothèse que, dans l'ensemble des SNPs associés au diabète de type 2 (Vrais Positifs) mais qui ne sont pas détectés (à cause de leur faible effet au vu de la taille de l'échantillon), une large proportion montrera des effets communs entre l'étape 1 et l'étape 2.
Pour chaque locus on définit l'allèle à risque dans l'étape 1 (analyse GWA). Nous étudions en phase 2 la distribution du z-score de l'effet allélique (tel que défini précédemment dans (D.IV.2.b) - Équation 19). S'il existe un sous-ensemble de SNPs associés au diabète de type 2, nous aurons un mélange de distributions, avec une majorité de SNPs appartenant à la distribution de moyenne 0 et un ensemble de SNPs appartenant à la distribution alternative. En final, nous obtenons plutôt pour chaque SNP une probabilité d'appartenir à chacune de ces distributions (Figure 44).

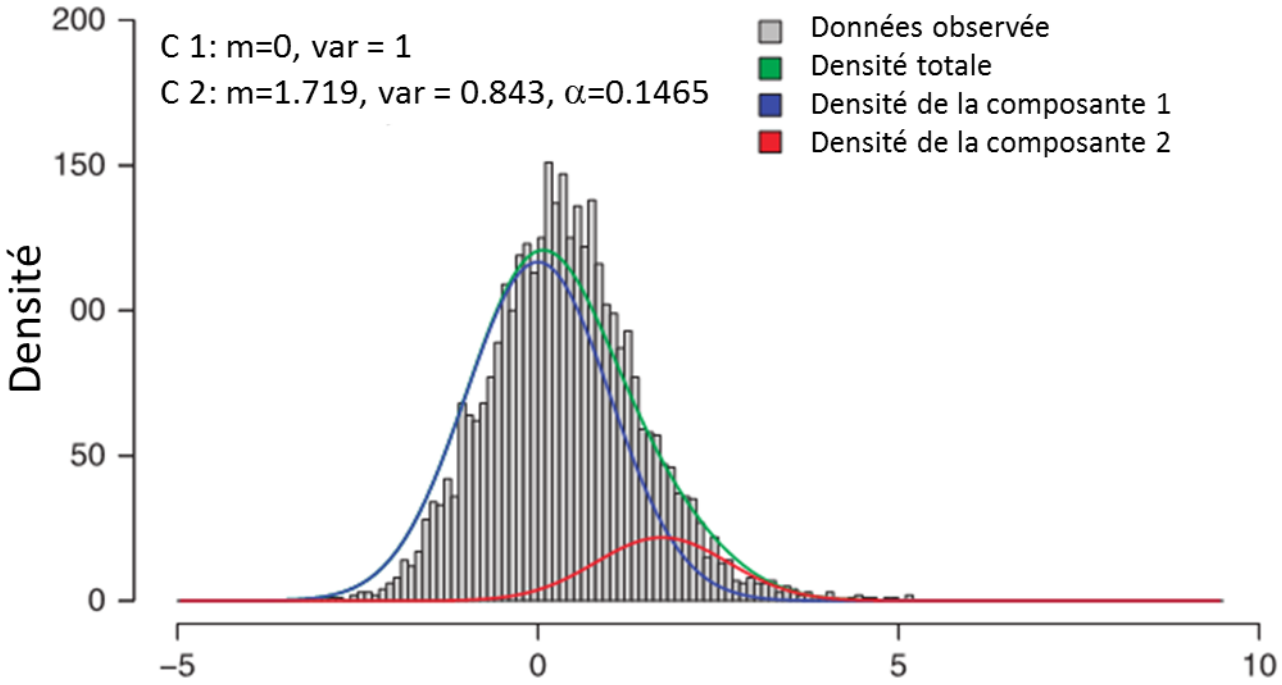


Figure 44 : distribution des effets génétiques dans le diabète de Type 2 (Morris et al. 2012)

La distribution des effets des allèles à risque, définis dans les analyses GWAs utilisées pour sélectionner les SNPs de la MetaboChip est présentée sous la forme d'un histogramme. Comme attendu, nous observons une majorité d'effets positifs. La distribution de ces effets est compatible avec un mélange de deux classes de SNPs, les vrais positifs (distribution rouge) et les SNPs neutres (en bleu).

Cette proportion est intéressante pour faire progresser notre connaissance des bases moléculaires de cette pathologie, mais elle permet également d'estimer la proportion de vrais positifs pour une pathologie complexe, fréquente et sans gène majeur.

En final, nous estimons qu'en plus des 64 loci déjà mis en évidence avec une significativité dépassant le seuil à l'échelle du génome, 488 (456–521 dans un intervalle de confiance à 95%) signaux indépendants sont associés au diabète de Type 2.

F.VI.4. Co-localisation d'association à différents phénotypes

Nous avons par ailleurs observé (dans [Voight et al.,**2010**]) que certaines régions montrant une association avec le diabète de type 2 montraient également une association avec d'autres traits (13 sur 30). Ces « co-associations » avaient lieu plus souvent que ce qui était attendu par hasard. La significativité de ces co-localisations a été faite par simulations.
Il s'agissait de sélectionner au hasard une liste de loci de même taille que les loci identifiés et de tester le nombre de SNPs déjà identifiés comme associés dans une analyse GWAS déjà publiée (identifiés dans la liste disponible sur le site http://www.genome.gov/gwastudies/ [Hindorff et al.,**2009**]). La probabilité d'observer ce nombre de co-localisations dans une fenêtre de 500 kb est de $1.6.10^{-3}$, et elle passe à 7.10^{-5} si nous testons des fenêtres de 100 kb (8 observées). Il s'agit à chaque fois de signaux indépendants (déséquilibre de liaison très bas < 0.1).
La co-localisation la plus intéressante dans le cadre de mon travail est la double association de la région KCNQ1 avec le Diabète de Type 2 et l'intervalle QT de l'électrocardiogramme.

F.VII. Papier n° 1 Voight et al.

F.VIII. Papier n°2 : Morris et al.

G. Projet 2 : Pathologies Cardiaques – Les troubles du rythme

- Recherche des bases génétique du Syndrome de Brugada par Analyse d'association génome entier.

G.I. Le Cœur

Le cœur est un organe essentiel qui permet, par son action de pompe, la circulation du sang dans le corps, aussi bien en direction de tous les muscles et organes (circulation systémique ou grande circulation) qu'en direction des poumons (circulation pulmonaire ou petite circulation.

G.I.1. Description- Anatomie du coeur

Le cœur (**Figure 45**) est un organe comprenant un "sac" séreux, *le péricarde*, qui entoure un muscle, *le myocarde.* Ce muscle est tapissé à l'intérieur par *l'endocarde* qui est une membrane fine. Il est également entouré par des artères dite coronaires qui permettent une bonne irrigation.

Cet organe est caractérisé par quatre cavités : deux oreillettes, droite et gauche, ainsi que deux ventricules, également droite et gauche (**Figure 45**, ci-dessous). Une cloison, le septum, sépare les ensembles (oreillette et ventricule) droits et gauches. Le sang appauvri en oxygène entre dans l'oreillette droite et se trouve ensuite éjecté dans le ventricule droit. Le sang continue son parcours par l'artère pulmonaire pour atteindre les poumons. Le sang oxygéné revient ensuite dans l'oreillette gauche, est éjecté vers le ventricule gauche. Enfin le sang est renvoyé dans l'aorte pour rejoindre la circulation sanguine générale.
Il est très important que le passage du sang entre les différentes cavités (oreillettes vers ventricules et ventricules vers les circulations pulmonaire ou systémique. Un système de valves empêche le sang de refluer des ventricules vers les oreillettes, assurant ainsi une circulation normale. Ces valves et leurs positions sont décrites de façon plus détaillées dans le chapitre H.I, p 158, et dans la Figure 69, p 158.

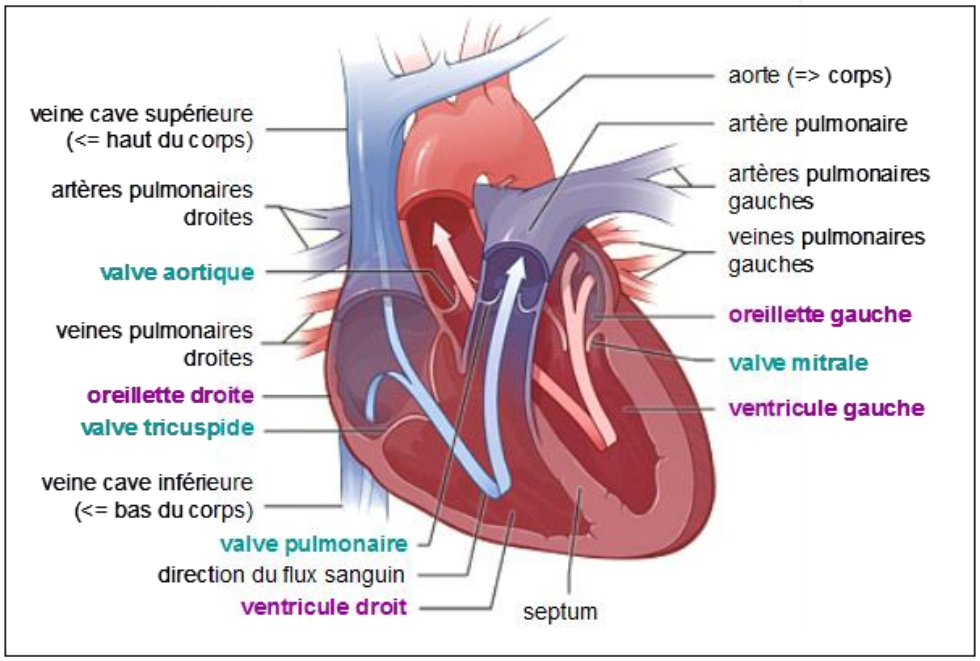


Figure 45 : Structure du Cœur
Le trajet du sang non oxygéné est représenté en bleu et celui du sang oxygéné est représenté en rouge. La paroi musculaire épaisse du ventricule gauche permet la propulsion du sang oxygéné dans l'aorte lors de la systole. Modifié d'après http://www.nhlbi.nih.gov/health.

Le circuit de la circulation systémique, qui distribue le sang oxygéné à tous les organes du corps est beaucoup plus long que le circuit pulmonaire. Le ventricule gauche est ainsi plus puissant et volumineux que le ventricule droit et le muscle le composant est plus épais. Cette différence correspond au besoin d'envoyer un volume de sang bien plus important.

G.I.2. Le cycle cardiaque

L'ensemble des phénomènes électriques et mécaniques du cœur constituent le cycle cardiaque. Ce cycle est décomposable en une phase de contraction (Systole), qui touche d'abord les oreillettes (Systole auriculaire), puis les ventricules (systole ventriculaire), et d'une phase de décontraction (Diastole) touchant aussi bien les oreillettes que les ventricules (**Figure 46**).

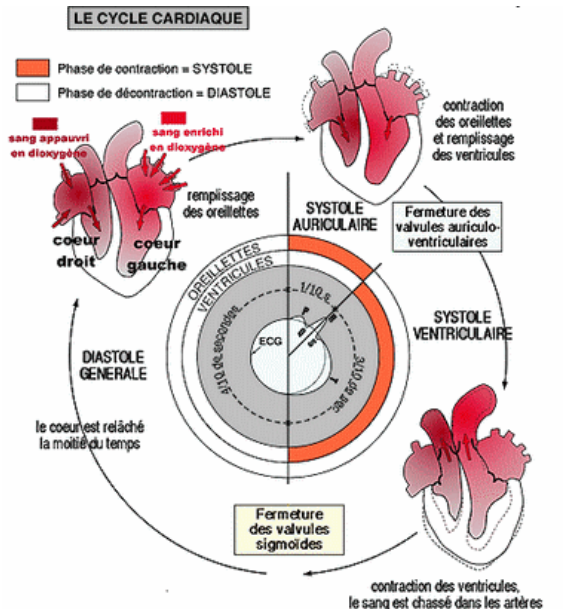


Figure 46 : Cycle cardiaque
Le cycle cardiaque se décompose en une phase systolique au cours de laquelle le sang est envoyé des oreillettes vers les ventricules puis des ventricules vers le système circulatoire, et une phase diastolique de décontraction où le sang entre de façon passive dans les ventricules. L'électrocardiogramme rapporte l'activité électrique du cœur qui permet les contractions et les valves assurent une circulation du sang dans le bon sens en se refermant lors de la systole ventriculaire. le site du service de réanimation du centre hospitalier de la Région d'Annecy : (http://www.reannecy.org/PAGES/espace%20paramedical/cardio/physio_cardiaque.html)

La durée d'un cycle complet dure 0.92 seconde, 0.27 seconde pour la systole ventriculaire et 0.65 seconde pour la diastole ventriculaire.
Lors de la systole auriculaire, il y a contraction des oreillettes, ce qui permet le refoulement du sang vers les ventricules. On parle de remplissage actif. Les valves auriculo-ventriculaires se referment ensuite, évitant un reflux du sang vers les oreillettes.
Lors de systole ventriculaire, la contraction des ventricules entraîne l'expulsion du sang vers le système circulatoire. Comme lors de la systole auriculaire, deux valves dites sigmoïdes (la valve pulmonaire et la valve aortique) s'ouvrent pour laisser passer le sang et se ferment ensuite, empêchant tout reflux du sang vers le ventricule.

Dans un dernier temps, toutes les parties du cœur se relâchent et permettent le remplissage passif des ventricules. C'est la diastole, où le remplissage du cœur se fait à plus de 80% dans les conditions normales. Le cœur en activité normale passe 1/3 du temps en systole et 2/3 en diastole.

G.II. Le système électrique cardiaque

G.II.1. Activité électrique cardiaque

Le mouvement du muscle cardiaque permettant la circulation du sang est le résultat d'un stimulus électrique autonome généré par le tissu cardiaque lui-même. Une partie des cellules cardiaques, formant le tissu nodal, est en effet capable de s'auto-exciter. Ce stimulus parcourt le cœur depuis le nœud sinusal jusqu'à l'extrémité inférieure (Figure 47, ci-dessous) sous la forme d'un Potentiel d'Action (P.A.). Il déclenche la contraction des différentes chambres du cœur.
Ce stimulus est reflété par l'électrocardiogramme (E.C.G.), un tracé papier, qui est le résultat de la mesure de potentiel, identifié à partir de plusieurs électrodes à la surface de la peau (G.II.2, p. 118).

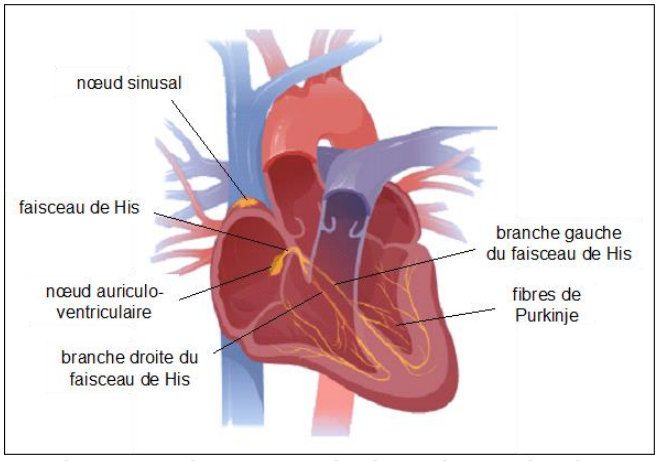


Figure 47 : Le cœur et le tissu de conduction

L'influx électrique prend naissance dans le Nœud sinusal. Il s'étend aux deux oreillettes et au Nœud auriculo-ventriculaire. Le Nœud auriculo-ventriculaire permet d'assurer un rôle de ralentisseur de l'onde de dépolarisation et de relais de cette onde vers les ventricules. Le passage vers les ventricules se fait à travers le réseau de His. L'influx est ensuite distribué suivant les branches du faisceau de His (ou Tawara) et aboutit au réseau de Purkinje. Modifié d'après http://www.nhlbi.nih.gov/health

G.II.1.a) *Bases cellulaires du potentiel d'action (P.A)*

Le mécanisme de contraction du myocarde est classique : la dépolarisation des cellules musculaires entraîne la propagation d'une impulsion électrique le long des fibres musculaires. Cette impulsion entraîne la contraction du muscle.

La dépolarisation se fait par le passage d'ions à travers la membrane des cellules du myocarde grâce aux canaux ioniques. Elle entraîne un potentiel d'action (PA), qui est transmis de cellule en cellule. Une trentaine de courants ioniques différents agissent au niveau cardiaque pour la genèse et le maintien du PA chez les vertébrés [Boyett et al.,**1996**].

Les canaux ioniques sont des protéines membranaires permettant le passage d'ions à travers la membrane. Ils sont à l'origine du potentiel d'action et donc de la contraction cardiaque. Les canaux ioniques sont présents au niveau de la membrane sous différents états (fermé ou ouvert) selon l'effet de divers stimuli (potentiel d'action, ligands, étirement).

Les différences de concentrations en ions (Na+, K+, Ca2+, Cl-) entre le milieu extracellulaire et le cytoplasme des myocytes créent une différence de potentiel (DP) électrique.
Au repos, cette différence de concentrations détermine la différence de potentiel intracellulaire (négatif) et extracellulaire (positif), qui est de l'ordre de -90 mV.

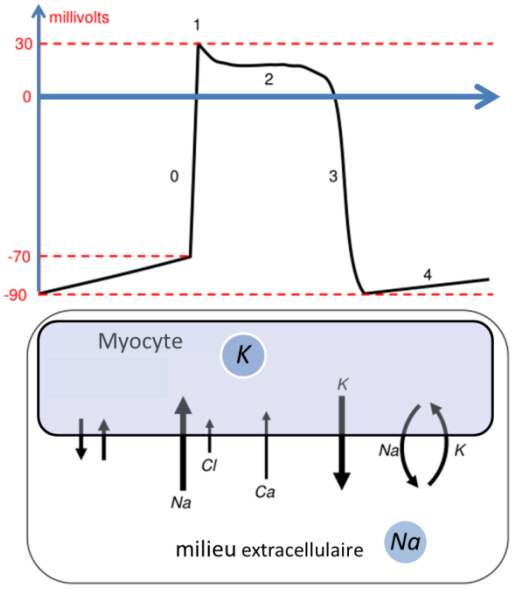


Figure 48 : Potentiel d'Action
*Les cinq phases du potentiel d'action d'une cellule du myocarde et les échanges ioniques au niveau cellulaire sont : respectivement, la **phase 0** correspondant à la dépolarisation de la cellule, la **phase 1** au début de la repolarisation, la **phase 2** à la repolarisation lente, la **phase 3** à la repolarisation rapide, et la **phase 4** au repos. Dans le milieu extracellulaire, le Sodium (Na) est majoritaire ; dans le milieu intracellulaire, c'est le Potassium (K). (adapté de* http://www.dematice.org/ressources/DCEM1/pharmacologie/D1_phar_02/res/resource_01.png)

Lors de la contraction du myocarde, des échanges ioniques se déroulent et définissent ainsi le potentiel d'action, qui comprend 5 phases successives (Figure 48) :
• La phase 0 ou dépolarisation rapide : après une excitation électrique au-dessus du seuil d'activation de la cellule, un afflux rapide d'ions Na+ rentre dans la cellule et inverse rapidement la polarité de la cellule.
• La phase 1 ou début de repolarisation : elle est caractérisée par une repolarisation rapide et de courte durée, due à l'inactivation des canaux Na+ et au flux sortant d'ions potassium K+.
• La phase 2 ou plateau : elle correspond à la phase de repolarisation lente. Elle est due à l'entrée lente des ions Ca2+ dans la cellule, qui atténue l'influence des canaux K+ où les ions continuent à sortir, ralentissant ainsi la phase de repolarisation.

• La phase 3 ou repolarisation : elle correspond à la phase de repolarisation finale, et se caractérise par la fermeture des canaux ioniques Na+ et Ca2+ spécifiques, qui ramène la cellule au potentiel de repos originel. Durant cette phase, les ions K+ sont toujours sortants tandis que le potentiel de la cellule tend vers son seuil de repos.
• La phase 4 : elle correspond au potentiel de repos, où la cellule est plus facilement excitable.
Il existe une période durant laquelle toute stimulation externe serait incapable de générer un nouveau PA : la période réfractaire absolue (PRA). Cet intervalle de temps se situe entre le début du PA et la moitié de la phase 3 (aux alentours de −50 mV), pendant lequel la cellule est inexcitable. Cette période est suivie par la période réfractaire relative (PRR) pendant laquelle un début de réponse commence progressivement à apparaître en réponse à des intensités de stimulation très élevées seulement. Ces périodes réfractaires sont dues aux états d'inactivation par lesquels passent les canaux sodiques et calciques avant de retrouver leur état de disponibilité initial.
Il est également à noter que le PA est légèrement différent suivant les tissus, et donc la localisation dans le cœur (Figure 49).

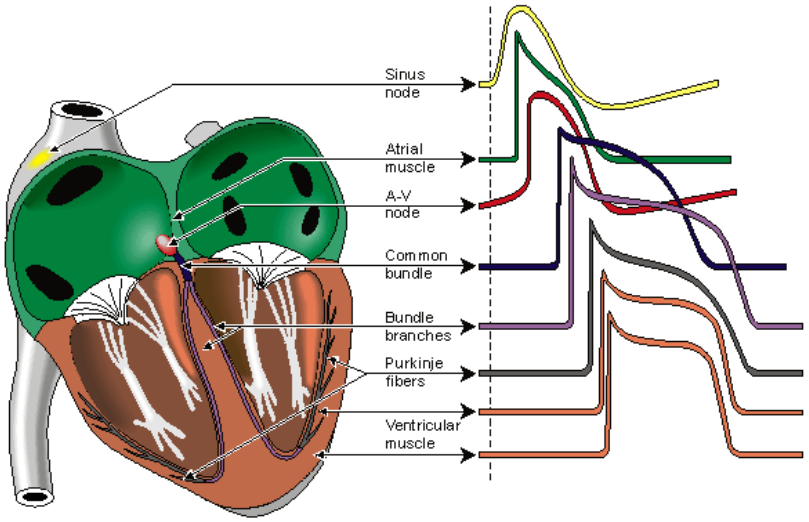


Figure 49 : Propagation du potentiel d'action dans le cœur
Profil du potentiel d'action dans les différents tissus. En raison de propriétés de conductivité différentes, le potentiel d'action n'a pas exactement le même profil suivant le tissu. Extrait de [Jaakko Malmivuo, Robert Plonsey]

G.II.1.b) *Système Nerveux Autonome*

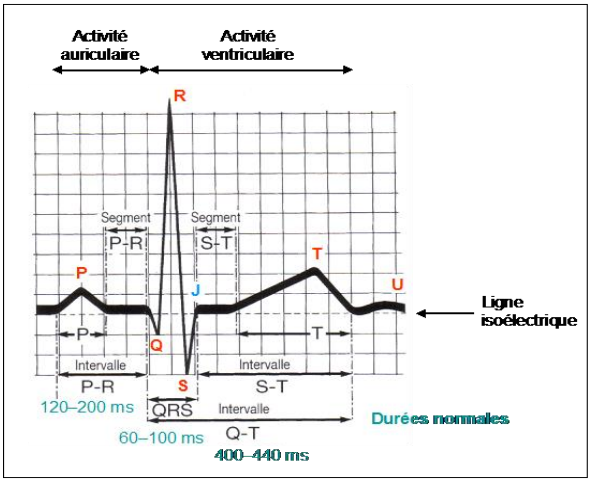
Le muscle cardiaque est unique car il est capable de créer et de maintenir son propre rythme (G.II.1.a), ci-dessus). Cependant, la fréquence cardiaque peut être très rapidement modifiée par l'activité des nerfs qui innervent le cœur et par les substances chimiques en circulation.
C'est le Système Nerveux Autonome (SNA) qui contrôle et modifie le rythme cardiaque, notamment par envoi d'influx au niveau du nœud sinusal.

G.II.2. Électrocardiogramme (ECG)

L'ECG est un enregistrement de l'activité électrique du cœur. Cette activité est mesurée grâce à des électrodes placées à la surface de la peau. Un électrocardiographe amplifie le signal électrique.

L'ECG standard est enregistré sur 12 dérivations (six dérivations des membres et six précordiales), avec une vitesse de déroulement du papier à 25 mm par seconde et une amplitude de 10 mm pour 1 mV.

a



b

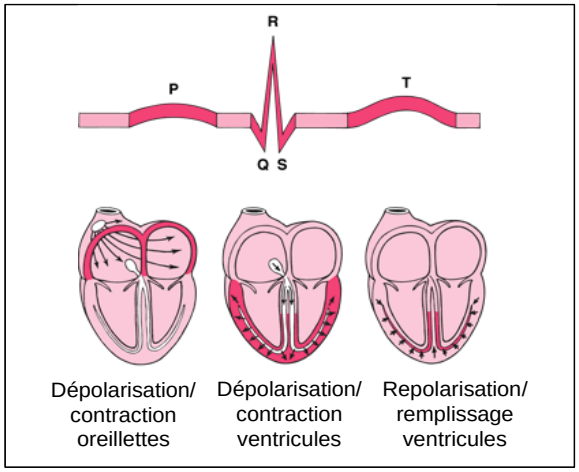


Figure 50 : Principaux paramètres de l'ECG

Les paramètres importants de l'ECG (Figure 50, ci-dessus), pour lesquels ont été cherché des gènes de susceptibilités sont :

Intervalle RR : fréquence des battements cardiaques.

L'intervalle RR correspond au délai entre deux dépolarisations des ventricules. C'est cet intervalle qui permet de calculer la fréquence cardiaque.

Intervalle PP : période de polarisation des oreillettes.

L'intervalle PP correspond au délai entre deux dépolarisations des oreillettes.

Segment PR : pause au niveau du nœud auriculo-ventriculaire (AV).

Le segment PR correspond au délai entre la fin de la dépolarisation des oreillettes et le début de celle des ventricules. C'est le temps pendant lequel l'onde de dépolarisation est bloquée au niveau du nœud AV.

Intervalle PR : durée de conduction auriculo-ventriculaire.

L'intervalle PR correspond au délai entre le début de la dépolarisation des oreillettes et celle des ventricules. C'est le temps de propagation de l'onde de dépolarisation jusqu'aux cellules myocardiques ventriculaires.

Intervalle QT : durée de systole ventriculaire

Cet intervalle correspond au temps de systole ventriculaire, qui va du début de l'excitation des ventricules jusqu'à la fin de leur relaxation.

Segment ST : durée de stimulation complète des ventricules

Le segment ST correspond à la phase pendant laquelle les cellules ventriculaires sont toutes dépolarisées, le segment est alors isoélectrique.

Onde P : dépolarisation des oreillettes depuis le nœud sinusal jusqu'au nœud auriculo-ventriculaire.

Complexe QRS : dépolarisation des ventricules.

Onde T : repolarisation des ventricules.

Onde U : parfois observée après l'onde T et de faible amplitude. Sa signification est discutée mais l'hypothèse la plus probable lorsqu'elle est physiologique est une repolarisation tardive du réseau de Purkinje.

Ces différentes composantes renseignent également sur l'état électrique des différentes parties du cœur (Figure 50 b).

G.II.2.a) *La Mort Subite cardiaque*

La mort subite cardiaque est un problème de santé publique majeur puisqu'elle touche environ un individu sur mille par an dans les pays développés (40 000 cas par an en France) [Jouven, Escande,**2006**][Zipes, Wellens,**1998**][Huikuri et al.,**2001**]. Elle est définie comme

une mort inattendue et brutale survenant dans l'heure suivant l'apparition des premiers symptômes. Elle est liée à un épisode de fibrillation ventriculaire. Le taux de survie, est inférieur à 3% en France, on parle alors de « mort subite récupérée ». Le mécanisme létal est dans 80% des cas un trouble du rythme ventriculaire : torsades de pointes ou tachycardie qui peut dégénérer en fibrillation (Figure 51, ci-dessous).

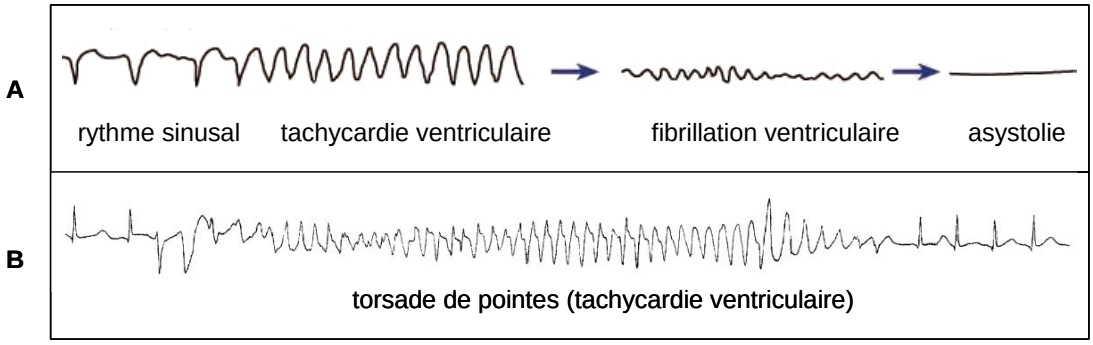


Figure 51 : Mécanismes de la mort subite cardiaque par arythmie ventriculaire

A : La tachycardie ventriculaire (TV monomorphe ou polymorphe, complexes QRS larges, fréquence cardiaque rapide) peut dégénérer en fibrillation ventriculaire (FV, perte de toute activité électrique organisée des ventricules) et causer la mort par asystolie. Modifié d'après *[Huikuri et al.,**2001**]*

B : Les torsades de pointes décrivent des torsions autour de la ligne isoélectrique. Elles sont liées à un phénomène de réentrée ventriculaire (désynchronisation des périodes réfractaires des cellules myocardiques), causée par une bradycardie importante ou un allongement de l'intervalle QT comme dans cet exemple chez un garçon de 15 ans. Modifié d'après *[Napolitano, Priori,**2002**]*

Dans 90% des cas, la mort subite est la conséquence d'une cardiopathie sous-jacente méconnue, le plus fréquemment une maladie coronarienne (suivie d'un angor par exemple) ou une cardiomyopathie dilatée ou hypertrophique. Toutefois, il arrive dans un grand nombre de cas que ces épisodes de fibrillation ventriculaire surviennent sur des cœurs sains.
Environ 5 à 10% des cas de mort subite sont attribués aux troubles du rythme sans cardiopathie sous-jacente. Elle survient surtout chez les jeunes [Sarkozy, Brugada,**2005**].
Pour ces troubles du rythme liés à la mort subite cardiaque, nous soupçonnons l'existence d'une forte composante génétique.
Ainsi, une étude clinique et génétique sur 43 familles où au moins un membre avait fait une mort subite avant 40 ans a démontré qu'il existait, chez les apparentés, une arythmie, primaire ou liée à une maladie structurale cardiaque, dans 40% des cas (17 familles sur 43). Pour la majorité de ces familles, une origine moléculaire a été identifiée sous la forme d'une mutation dans un gène responsable d'une arythmie [Tan et al.,**2005**].
On distingue un grand nombre de troubles du rythme cardiaque qu'on appelle primaires, sans cardiopathie sous-jacente (le syndrome du QT long congénital, le syndrome du QT court congénital, le syndrome de Brugada, la forme héréditaire de la maladie de Lenègre, les formes héréditaires de dysfonction sinusale, la tachycardie ventriculaire polymorphe catécholaminergique, la dysplasie ventriculaire droite arythmogène et les formes familiales de fibrillation auriculaire). Ces troubles du rythme semblent être un substrat important de la mort subite sur cœur sain [Roberts, Brugada,**2003**]. La plupart de ces syndromes semblent être des « canalopathies », c'est à dire des dysfonctionnements des canaux ioniques [Cerrone, Priori,**2011**].
Notre laboratoire s'est intéressé à ces différentes pathologies du rythme cardiaque dans le but de mieux stratifier le risque de mort subite et par ce biais identifier des polymorphismes génétiques augmentant le risque de survenue de cette affection. Leur survenue en l'absence de toute dégradation du myocarde ou du système coronaire cardiaque entraîne l'hypothèse d'une origine génétique. On a d'ailleurs parlé de maladies arrythmogéniques héritées [Cerrone, Priori,**2011**].
C'est dans ce contexte que j'ai été amené à travailler sur le Syndrome de Brugada.

G.II.2.b) *Génétique des troubles du rythme*

Du fait de la place centrale que jouent les canaux ioniques dans la transmission du potentiel d'action, les gènes codant pour ces canaux ioniques ont été immédiatement des candidats idéaux pour la recherche de mutations délétères. Les différents canaux qui se trouvent à toutes les étapes de la transmission du potentiel d'action (Figure 52) ont été testés par analyse de liaison et/ou séquençage du gène chez les patients atteints de ces pathologies.

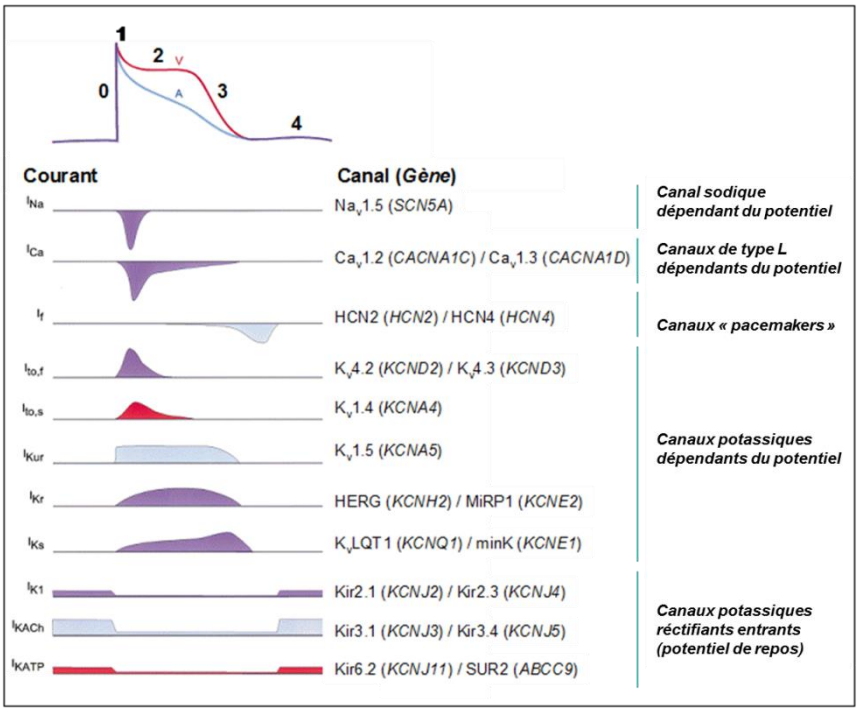


Figure 52 : Représentation des principaux courants ioniques responsables des potentiels d'action auriculaires et ventriculaires humains et des canaux ioniques associés

V *:* PA ventriculaire, A *:* PA auriculaire. En rouge : les courants enregistrés uniquement dans le ventricule ; en bleu *:* uniquement dans l'oreillette, en violet : dans les deux compartiments. Vers le bas les courants entrants et vers le haut les courants sortants. Kv : Canaux potassiques dépendants du voltage, Kir : Canaux potassiques rectifiants entrants (inward rectifier). I signifie courant. Modifié d'après **[Pond, Nerbonne,**2001**]**

Cette hypothèse s'est avérée très féconde puisqu'en 2011, 21 gènes avaient été trouvés liés à au moins un de ces syndromes, dont 12 codaient directement pour des protéines canales (Tableau 9). Une partie des autres gènes codent pour des protéines impliquées dans l'adressage de ces canaux.

Gene (symbole)	Locus chromosomique	Protéine	Phenotype
KCNQ1	11p15.5	sous unité α IKs canal potassique (KvLQT1)	QT long
KCNH2	7q35–q36	sous unité α IKr canal potassique (HERG)	QT long
SCN5A	3p21	sous unité α canal sodique (Nav 1.5)	QT long
KCNJ2	17q23.1–q24.2	*IK1 canal potassique (Kir2.1)*	QT long
KCNE1	21q22.1–q22.2	sous unité β (IK)s canal potassique (MinK)	QT long
KCNE2	21q22.1–q22.2	sous unité IKr canal potassique β- (MiRP)	QT long
ANK2	4q25–q27	Ankyrine B, protéine d'ancrage	QT long
CACNB2b	10p12	sous unité α- canal calcique	Syndrome de Brugada et QT court
Cav3	3p24	Caveoline	QT long
SCN4b	11q23.3	sous unité β-4 du canal sodique	QT long
AKAP9	7q21–q22	A-kinase-anchoring protein	QT long
SNTA1	20q11.2	α1-syntrophine	QT long
KCNJ5	11q23.3–24.3.	Kir3.4 sous unité of IKAch channel	QT long
GPD1-L	3p22.3	Glycerol-3-phosphate dehydrogenase 1 like	Syndrome de Brugada
SCN1b	19q13.1	sous unitéβ-1 canal sodique	Syndrome de Brugada
KCNE3	11q13–q14	sous unité β des canaux IKs et Ito	Syndrome de Brugada
SCN3B	11q23.3	sous unité β-3 du canal sodique	Syndrome de Brugada
RyR2	1q42–43	recepteur Cardiac Ryanodine	CPVT
CASQ2	1p13.3–p11	calsequestrine cardiaque	CPVT

Tableau 9 : liste des gènes impliqués dans des troubles du rythme primaires.
Cette liste n'est pas exhaustive. En particulier, de nouveaux gènes impliqués dans le Sydnrome de Brugada seront mis en évidence plus tard (Tableau 11). Adapté de [Napolitano et al.,**2012**].

L'implication majoritaire de protéines canaux ioniques est plus visible sur la Figure 53.

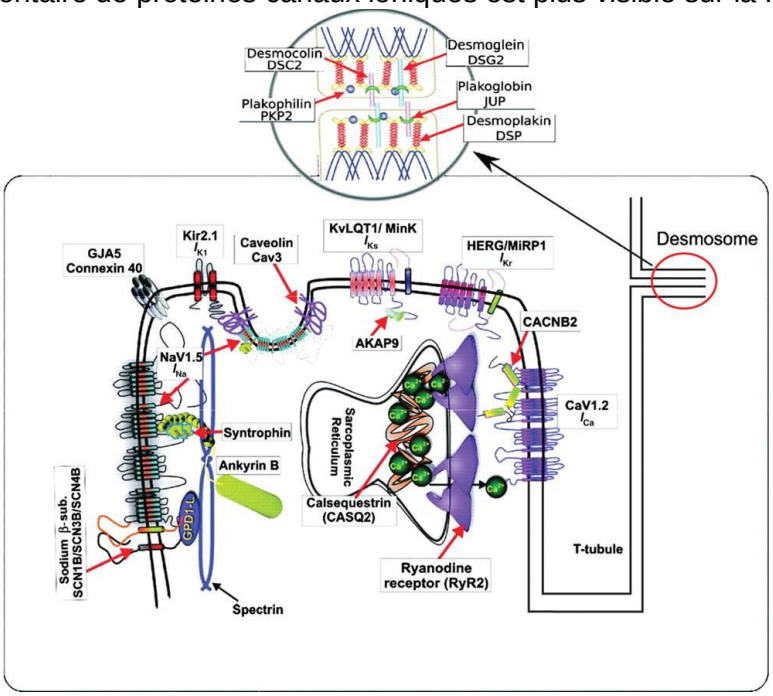


Figure 53 : protéines impliquées dans la survenue de troubles du rythme primaires [Priori,2010**]**
Cette liste n'est pas exhaustive. En particulier, de nouveaux gènes impliqués dans le Sydnrome de Brugada seront mis en évidence plus tard (Tableau 11).

Ces troubles du rythme sont donc appelés souvent des « canalopathies ».

Par ailleurs, les bases génétiques de la mort subite [Arking et al.,**2011**] ainsi que de la fibrillation ventriculaire [Bezzina et al.,**2010**] et fibrillation auriculaire [Gudbjartsson et

al.,**2007**][Gudbjartsson et al.,**2007**] ont été explorées, par le biais d'analyses d'association génome entier. Si l'architecture génétique de la Fibrillation Auriculaire semblait assez proche de celle observée pour des phénotypes complexes classiques, les résultats des études sur la mort subite et la fibrillation ventriculaire n'ont mis en évidence qu'une seule association chacun. Le gène *BAZ2B* semblait associé avec le phénotype de Mort Subite [Arking et al.,**2011**], à travers un SNP rare en population générale (~1%). De façon similaire, le gène *CXADR* a été trouvé associé à la fibrillation ventriculaire chez des individus ayant subi un Infarctus du Myocarde. L'association repose sur un SNP fréquent en population cette fois, ce qui renforce la confiance qu'on peut avoir dans le résultat. Toutefois, pour ces deux dernières études, il sera important de répliquer les résultats dans d'autres échantillons.
La différence dans la force des résultats entre ces deux types de fibrillations peut tenir au fait que la Fibrillation Auriculaire, de par sa prévalence plus haute, a été étudiée dans un plus grand nombre d'études (cas témoins ou cohortes) et sur de plus grands échantillons de patients et de témoins. Il est également possible qu'il n'y ait pas de polymorphisme génétique fréquent impliqué dans le phénotype de fibrillation ventriculaire, en raison, par exemple, du désavantage sélectif qu'un tel phénotype peut induire (décès avant l'âge de la reproduction).

En parallèle à ces recherches dirigées sur le phénotype pathologique, plusieurs équipes ont mis en place un programme de recherche de gènes expliquant la variabilité interindividuelle des traits quantitatifs de l'ECG. Il s'agissait de grandes études génome entier portant sur plusieurs cohortes comprenant chacune plusieurs centaines d'individus.
Après une première étude d'association génome entier qui a permis d'identifier un SNP associé à l'intervalle QT [Arking et al.,**2006**], des études récentes, basées sur des population allant de 13000 à 40 000 individus, ont mis en évidence 13 SNPs indépendants associés avec ce même phénotype [Pfeufer et al.,**2009**][Newton-Cheh et al.,**2009**]. De telles études ont été également menées avec succès sur d'autres traits et identifiés 22 SNPs associés avec la durée du complexe QRS [Holm et al.,**2010**][Chambers et al.,**2010**][Sotoodehnia et al.,**2010**], 10 avec l'intervalle PR [Holm et al.,**2010**][Pfeufer et al.,**2010**], et 9 avec le rythme cardiaque [Holm et al.,**2010**][Cho et al.,**2009**][Eijgelsheim et al.,**2010**].
Par ailleurs, des SNPs associés à l'intervalle QT ont également été trouvés significativement associés au risque de mort subite [Albert et al.,**2010**][Kao et al.,**2009**]. Cette association reste modeste et l'influence de ces polymorphismes associés aux paramètres ECG dans la population sur des conditions pathologiques qui semblent être les extrêmes des distributions de ces paramètres, tels que le Syndrome de Brugada, reste à démontrer.
Un état des lieux des phénotypes de l'ECG a été présenté dans l'éditorial que j'ai écrit dans Journal of Molecular and Cellular Cardiology [Dina,**2011**], qui est présenté dans cette thèse (Annexe 3).

G.III. Le Syndrome de Brugada

G.III.1. Description

En 1992, Pedro et Josep Brugada décrivent les caractéristiques électrocardiographiques de 8 patients. Ces caractéristiques forment une nouvelle entité clinique qui va devenir donc un syndrome. Ce syndrome associe un aspect typique sur l'ECG, des troubles du rythme et de la conduction, un risque de mort subite par tachycardie ventriculaire (TV) polymorphe et une présomption d'hérédité [Brugada, Brugada,**1992**].
Au niveau de l'ECG, les caractéristiques principales semblent être un sus-décalage du segment ST persistant visible dans les dérivations précordiales droites (V1 à V3) supérieur à 0,1 mV, ainsi qu'un bloc de branche droit sans allongement de l'intervalle QT (Figure 54).

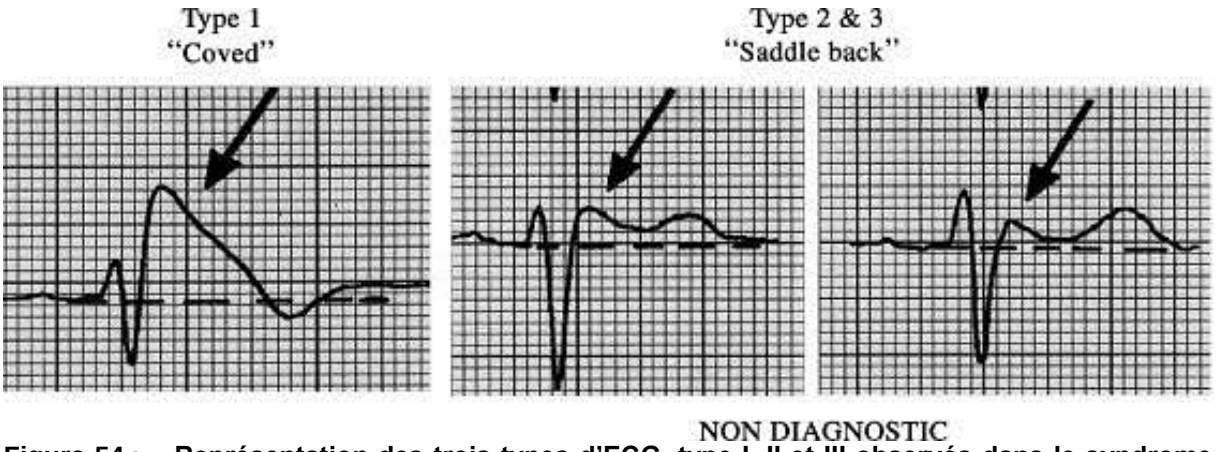


Figure 54 : – Représentation des trois types d'ECG, type I, II et III observés dans le syndrome de Brugada.

Les flèches pointent le point J. D'après [Brugada et al.,**2006**]

Tableau 10 – Caractéristiques électrocardiographiques des ECG de type I, II et III du syndrome de Brugada

	Type 1	**Type 2**	**Type 3**
Point J	≥2 mm	≥2 mm	≥2 mm
Onde T	Négative	Positive ou biphasique	Positive
Aspect ST-T	Convexe	En selle	En selle ou convexe
segment ST (portion terminale)	Décroissance progressive	≥1 mm	< 1 mm

Modifié d'après **[Antzelevitch et al.,**2005**]**

Il est très important de souligner que le diagnostic d'un syndrome de Brugada doit être fait sur cœur sain.

G.III.2. Épidémiologie

Le syndrome de Brugada (SBr) toucherait environ 1 caucasien sur 2000 (0,05% de la population générale). Cette prévalence peut monter jusqu'à 0,12% [Gallagher et al.,**2008**] (individus ayant un SBr de type I) voire 0,19% à 0,58% en incluant les aspects de type II et III.

Cette imprécision dans l'estimation de la prévalence vient du fait que le phénotype ne s'exprime pas toujours spontanément. On peut alors démasquer le syndrome avec des outils pharmacologiques comme les bloqueurs de canaux sodiques : l'ajmaline et la flécaïnide. Ces bloqueurs font partie des anti-arythmiques de classe Ic et sont administrés en perfusion continue sur 10 min, à raison de 1mg/kg pour l'ajmaline et 2 mg/kg pour la flécaïnide.

La proportion de patients asymptomatiques a été estimée en s'appuyant sur le caractère génétique des arythmies, et, plus particulièrement sur l'effet majeur du gène *SCN5A* (Tableau 9), qui sera décrit en G.III.3.a). Hong et ses collègues ont testé par produit anti-arythmique (ici Ajmaline) des porteurs de mutation *SCN5A* , qui avaient donc une probabilité *a priori* forte d'être atteints [Hong et al.,**2004**]. Sur les 61 porteurs de mutations *SCN5A,* 20 étaient spontanés. Sur les 41 restants et testés à l'Ajmaline, 6 étaient inconnus et 28 ont

présenté le sus-décalage du segment ST. Le test à l'Ajmaline permet de doubler (au minimum) le nombre d'atteints. Bien que ces données soient obtenues sur un substrat à haut risque génétique, cela donne une idée de la sous-estimation possible de la prévalence. Il est à noter que sur les 41 porteurs de mutation et non symptomatiques, seulement deux ont réagi à l'Ajmaline.
Dans les deux cas, bien que de façon très différente, on en conclut à une sous-estimation de la prévalence de ces conditions.
La pénétrance des mutations *SCN5A* passe, elle, de 32 à 79% [Hong et al.,**2004**].
L'âge moyen d'apparition des premiers symptômes est en moyenne de 40 ans avec une distribution très hétérogène (de 2 jours à 84 ans)[Antzelevitch et al.,**2005**]. Le SBr touchent préférentiellement les hommes : 72 à 80% [Eckardt et al.,**2005**][Benito et al.,**2008**].

Il a été estimé que le syndrome de Brugada est responsable d'au moins 4% des morts subites cardiaques et de 12% des morts subites sur cœur sain [Antzelevitch et al.,**2005**].
Ses conséquences en termes de Santé Publique sont donc loin d'être négligeables.

G.III.3. Bases Génétiques du Syndrome de Brugada

Le syndrome de Brugada étant une pathologie plus rare et récemment décrite, la recherche de ses bases génétique est moins riche en études estimant son héritabilité. On peut se baser sur les études démontrant l'héritabilité des phénotypes ECG, et plus particulièrement de l'intervalle PR [Havlik et al.,**1980**], estimée à 34% ainsi que du QRS [Li et al.,**2009**][Mutikainen et al.,**2009**], même si pour le QRS, des rapports contradictoires semblent indiquer une héritabilité plus faible, voire nulle [Russell et al.,**1998**][Havlik et al.,**1980**].
Il est à noter que le caractère héritable fait partie de la définition initiale du syndrome. En effet, ce syndrome, comme la plupart des arythmies primaires (G.II.2.a) sont diagnostiquées sur des cœurs sains, sans aucun malformation visible ni antécédents vasculaires, ce qui privilégie des causes génétiques.
Par ailleurs, au commencement des travaux qui font l'objet d'un chapitre de cette thèse, un certain nombre de mutations impliquées dans la survenue de ce syndrome étaient connues. Du fait de l'importance des canaux ioniques dans le maintien d'un potentiel d'action régulier, il n'est pas étonnant que les premiers gènes impliqués dans cette pathologie codent pour ce type de protéine.

G.III.3.a) *Études familiales :*

Dix gènes ont été identifiés à ce jour (Tableau 11). En 1998, l'équipe de Chen a mis en évidence 3 types de mutations : faux-sens, substitution d'un acide nucléique dans un site d'épissage et un décalage du cadre de lecture dans le gène *SCN5A* par une approche gène candidat [Chen et al.,**1998**]. Ces variants ont été identifiés dans 2 petites familles (6 atteints dans l'une et 2 dans l'autre) et chez un cas sporadique souffrant de fibrillation ventriculaire idiopathiques associées à un aspect ECG compatible avec les critères du SBr. La ségrégation génotype/phénotype démontre le caractère autosomique dominant de cette pathologie. Une mutation du gène *SCN5A* est retrouvée chez environ 18 à 30% des patients [Antzelevitch et al.,**2005**].
En 2010, l'étude rétrospective de Kapplinger et al. recense près de 300 mutations dans le canal sodique cardiaque Nav1.5 (codé par *SCN5A*) associées au Syndrome de Brugada [Kapplinger et al.,**2010**].
Des études fonctionnelles dans des systèmes de réexpression ont montré une perte de fonction du courant sodique (INa). Cette perte de fonction se manifeste par plusieurs mécanismes comme une diminution de l'expression du canal ou par une dysfonction intrinsèque du canal.
Des troubles de la conduction sont observés chez les patients atteints du SBr. Il a été décrit que les paramètres tels que l'intervalle PQ, la durée du complexe QRS et l'intervalle HV sont

prolongés chez les patients porteurs d'une mutation *SCN5A* par rapport aux non porteurs [Smits et al., 2002; Yokokawa et al., 2007]

Variant	Gene	Courant Ionique	Effet Fonctionnel	Mode de transmission	% de porteurs	1ère référence
BrS1	*SCN5A*	I_{Na}	P. d F.	A D	11–18%	[Chen et al.,**1998**]
BrS2	*GPD1-L*	I_{Na}	P. d F.	A D	<1%	[London et al.,**2007**]
BrS3	*CACNA1c*	I_{Ca}	P. d F.	A D	<1%	[Antzelevitch et al.,**2007**]
BrS4	*CACNB2*	I_{Ca}	P. d F.	A D	<1%	[Antzelevitch et al.,**2007**]
BrS5	*SCN1B*	I_{Na}	P. d F.	A D	<1%	[Watanabe et al.,**2008**]
BrS6	*KCNE3*	I_{to}	G. d F.	A D	<1%	[Delpon et al.,**2008**]
BrS7	*SCN3B*	I_{Na}	P. d F.	A D	<1%	[Hu et al.,**2009**]
BrS8	*MOG1*	I_{Na}	P. d F.	A D	<1%	[Kattygnarath et al.,**2011**]
BrS9	*KCNE5*	I_{to}	G. d F.	A D	<1%	[Ohno et al.,**2011**]
BRS10	*KCND3*	I_{to}	G. d F.	A D	<1%	[Giudicessi et al.,**2011**]

Tableau 11 : Liste des gènes impliqués dans le syndrome de Brugada
Adapté de [Berne, Brugada,**2012**]. L. o F. = Perte de Fonction. G. d F. = Gain de Fonction. A D = Autosomique dominant

Une étude récente de notre équipe montre des défauts de ségrégation des mutations *SCN5A* au sein de 13 familles. La pénétrance de ces mutations dans cette population est de 18% à l'état de base et de 61% après un test pharmacologique. Douze de ces familles comptent au moins un individu non pénétrant, et cinq comportent des individus atteints non porteurs de la mutation familiale [Probst et al.,**2009**]. Les variants retrouvés dans le gène *SCN5A* ne semblent pas suffisants pour expliquer le SBr. Ce travail rétrospectif montre que le modèle mono-génique, initialement décrit, n'est pas entièrement responsable de la survenue de cette maladie. D'autres facteurs (génétiques) doivent intervenir dans l'expressivité phénotypique du SBr.
Une approche gène–candidat portant sur les gènes codant pour des canaux ioniques (Sodium mais aussi Calcium et Potassium) ont été menées depuis la mise en évidence initiale de *SCN5A* et ont abouti à l'implication de 10 gènes décrits dans le Tableau 11. Il est visible que le principal gène responsable du phénotype SBr reste, pour l'instant, *SCN5A*.
Les mutations dans le gène *SCN5A* restent de loin les plus fréquemment identifiées dans le SBr avec une prévalence d'environ 18-30%. Cependant aucune relation génotype/phénotype parfaite n'a été observée au sein d'une famille suffisamment informative.
La faible pénétrance (32%) conduit à considérer, dans les études familiales, un nombre élevé d'individus non pénétrants. L'expressivité d'une même mutation est variable tant au niveau électrocardiographique qu'au niveau symptomatique. Enfin, l'utilisation des bloqueurs pharmacologiques permettant de démasquer l'aspect ECG caractéristique n'apporte pas une réponse fiable à 100%. Toutes ces limites freinent les études pan-génomiques familiales qui pourraient mener à l'identification de nouveaux marqueurs moléculaires.

Il est à noter qu'une étude récente a mis en évidence un haplotype, présent dans environ 20% de la population asiatique, associé à l'allongement des intervalles PR, QRS, à l'élévation du segment ST et à la réduction de l'expression du canal sodique in vitro. Cet haplotype est spécifique de la population asiatique ce qui corrèle avec la forte incidence de SBr dans cette région du monde [Bezzina et al.,**2006**]. C'est un premier indice d'un effet d'un variant fréquent dans la survenue du syndrome de Brugada.

Une étude très récente testant 12 gènes potentiellement impliqués dans la survenue du syndrome de Brugada. [Crotti et al.,**2012**] conclut à une fréquence très importante de

mutations dans *SCN5A*, principalement. Il apparaît que les mutations dans d'autres gènes décrits comme liés au SBr sont très rares.

G.III.4. Description de l'analyse

Mon projet 2 portait sur la recherche de polymorphismes fréquents dans le Syndrome de Brugada dans une étude multicentrique par approche d'association génome entier.

G.III.4.a) *Groupe patients*

L'unité U1087 a joué un rôle important dans la mise en place d'un consortium ayant pour objectif l'étude des Arythmies, principalement la Mort Subite cardiaque et le Syndrome de Brugada. Ce consortium était soutenu par un financement « Alliance Against Sudden Cardiac Death » de la fondation Leducq (http://www.allianceagainstscd.org/). Pour la réalisation de l'analyse génome entier, nous avons mobilisé 8 centres à travers l'Europe (Figure 55).
Nous avons éliminé, dans cette première phase, les individus Asymptomatiques/Induits. Il s'agissait d'enrichir l'échantillon de formes plus « sévères ».

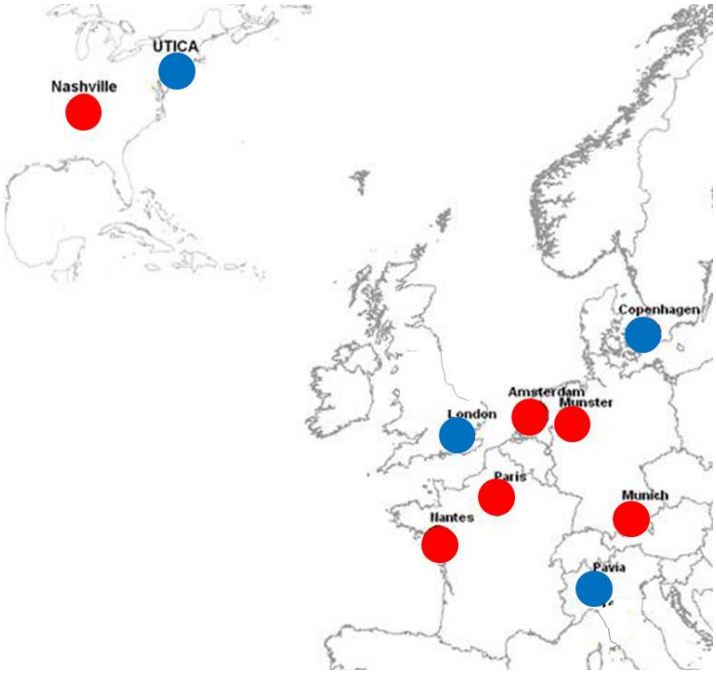


Figure 55 : Consortium pour l'analyse EAGE Mort Subite
Liste des centres impliqués dans l'étude d'analyse d'association génome entier du syndrome de Brugada (GWAs SBr). En rouge, les équipe impliqués dans le Réseau Leducq Mort Subite participant au projet GWAs SBr. En bleu, les centres ayant rejoint l'initiative GWAs SBr. **Nashville** : Vanderbilt Medical Center ; **Amsterdam** : Academic Medical Center, Cardiology Department of Clinical and Experimental Cardiology ; **Munich** : Ludwigs Maximilian Universität München ; **Münster** : Universitätsklinikum Münster ; **Paris** : Université Paris 6 ; **Nantes** : Institut du Thorax

Les patients venaient de plusieurs centres Européens et Américains, induisant ainsi une grande hétérogénéité démographique (Figure 56).

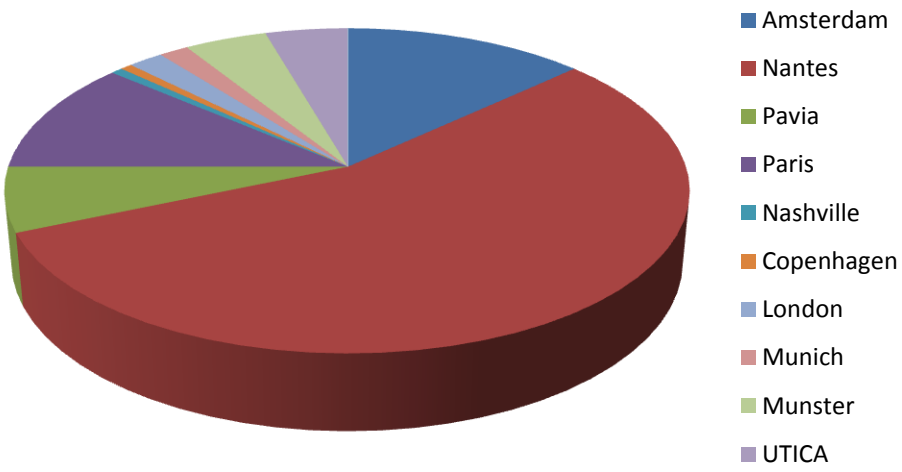


Figure 56 : Origine des patients atteints du syndrome de Brugada
Les patients étudiés dans l'analyse d'association génome entier sont recrutés dans différents centres Européens, induisant une grande hétérogénéité démographique et également génétique. Toutefois, la forte proportion des patients recrutés à Nantes permet de créer une étude (en adjoignant les individus recrutés à Paris) génétiquement homogène avec les témoins français de l'étude DESIR (décrite en G.III.4.b).

Par ailleurs, l'étude est organisée en deux phases, qui comprennent la phase d'analyse d'association proprement dite (GWAs) suivie de deux tentatives de réplication Européennes (Réplication 1 et Réplication 2) ainsi que d'une population japonaise. Les caractéristiques de l'ECG sont résumées dans le Tableau 12.

	GWAs	Réplication 1	Réplication 2	Population japonaise
Sexe	261/51	241/56	219/82	198/20
Age à l'ECG	47 (15)	46 (13)	45 (13)	46 (15)
Fibrillation ventriculaire	144/30	103/22	NA	46/32
Symptômes (Non/Oui)	138/172	128/162	298/0	127/91
Spontanés (Non/Oui)	102/210	89/208	302/0	119/92
QRS	106.66 (20)	102.47(17)	103.90 (23)	98.63 (18)
Onde P	103.51 (20)	104.5 (17)	103 (16)	106 (17)
Intervalle PQ	181.2 (33)	173.6 (29)	171 (27)	180.2 (27)
Intervalle RR	876.3 (171)	844 (177)	875 (145)	934 (166)
QT de Bazett	411 (38)	408 (37)	401 (35)	393 (26)

Tableau 12 : Etapes de l'Etude Multi-centrique Brugada
Caractéristiques démogrpahiques et cliniques de 4 populations du Projet 2 (GWAs, étude de réplication Européenne, étude de réplication Japonaise). Il est à noter que l'étude de réplication Européenne a été organisée en deux vagues mais l'analyse a été effectuée sur tout l'échantillon.

G.III.4.b) *Population témoin*

D.E.S.I.R. est une étude sur les facteurs de risque du syndrome de l'insulino-résistance et du diabète de type II; 5212 hommes et femmes, âgés de 30 à 65 ans ont été recrutés sur la période 1994-1996, à partir des consultants des Centres d'Examens de Santé de la Sécurité Sociale, dans le centre ouest de la France.
Les consultants ont participé dans les examens cliniques et biologiques tous les trois ans, pendant neuf années.

Dans mon étude, j'ai utilisé deux sous-ensembles de cette population. Le premier sous-ensemble (appelé ici DESIR) consiste en 890 individus génotypés sur la puce Axiom Affymetrix, comme les patients. Ces individus ont été utilisés comme témoins pour l'analyse d'association génome entier proprement dite.
Par ailleurs, nous avons également eu accès au sous-ensemble de 697 individus de la population D.E.S.I.R. génotypés en Illumina 300 k HumanHap et initialement utilisés pour une analyse d'association pour le diabète de type II (groupe appelé ici DESIR-REP). Les génotypes de ces individus ont été utilisés dans la deuxième phase de réplication Européenne. Nous avons dû imputer les génotypes des SNPs d'intérêt car ils n'étaient pas présents sur la puce HumanHap.

Comme individus témoins, nous avons, dans un premier temps, utilisé la population DESIR et retenu les patients qui s'appariaient parfaitement (**Analyse I**).
Toutefois, le caractère multicentrique de l'étude avait pour conséquence une perte importante d'un grand nombre de cas issus des centres non français (Figure 58).

J'ai alors proposé une stratégie complémentaire se basant sur les données disponibles du projet 1000 génomes. Dans ce projet, décrit en D.V.3., plus de 1000 individus issus de différentes populations mondiales ont été séquencés (et pour certains génotypés) sur tout le génome. Nous avons ainsi les génotypes pour tous ces individus et il nous est possible, par analyse de stratification classique de tester la possibilité d'appariement démographique entre les individus issus du projet 1000 génomes et nos patients issus de centres non français. Cette analyse est expliquée en (G.III.5.a.2).

Nous avons donc ensuite effectué une méta-analyse (**appelée analyse II**) incluant tous les individus atteints issus des différents centres Européens. Ces cas ont été comparés aux contrôles DESIR mais également aux individus issus Européens issus des 1000 génomes (voir étape de stratification précédente). Il s'agissait d'une analyse de confirmation et d'une tentative pour mieux séparer les SNPs ayant une forte chance de réplication immédiate dans un échantillon de taille moyenne.

G.III.5. Résultats

G.III.5.a) *Stratification*

Dans ces analyses de stratification, nous avons appliqué une analyse en composantes principales comme présenté dans D.VI.6. Les individus sont présentés suivant leur position sur deux Axes (en général les axes principaux 1 et 2).

G.III.5.a.1) Détection des individus non-Européens :

Dans cette analyse, l'IBD est estimée en utilisant tous les individus des 1000 génomes (Européens, Africains et Asiatiques).
Nous voyons sur la Figure 57 que la plupart de nos patients atteints du syndrome de Brugada se placent aux mêmes positions que les individus d'origine Européenne du panel de référence des 1000 génomes. Les individus qui s'en éloignent ne seront pas retenus dans l'analyse.

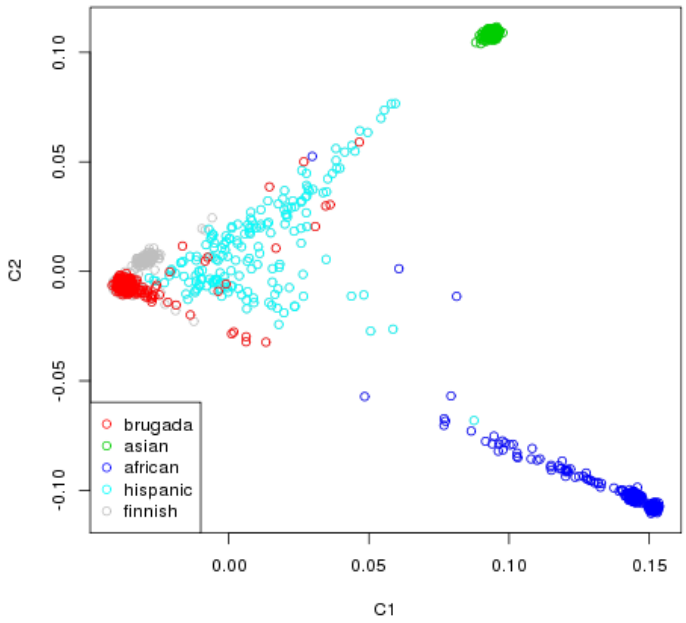


Figure 57 : l'Analyse en Composantes Principales – localisation des individus.
Les individus sont présentés selon leur position sur le plan des axes des deux premières composantes. Les individus atteints de SBr sont en rouge. Les individus issus des 1000 génomes sont colorés en fonction de leur population d'origine.

G.III.5.a.2) Stratification à l'échelle de l'Europe :

L'analyse est effectuée cette fois-ci à l'échelle Européenne. Les IBDs sont calculée0s en incluant les individus des 1000 Génomes Européens à l'exclusion des Finlandais qui sont assez différents dans leur profil génétique.

Nous voyons sur la Figure 58 que les coordonnées des patients Brugada dans les composantes 1 et 2 recouvrent les différentes populations Européennes présentes dans les 1000 génomes. Et, sans surprise, un sous ensemble, seulement, s'apparie avec les témoins. Ce résultat est attendu car les témoins DESIR sont recrutés en France (plus précisément dans l'Ouest).

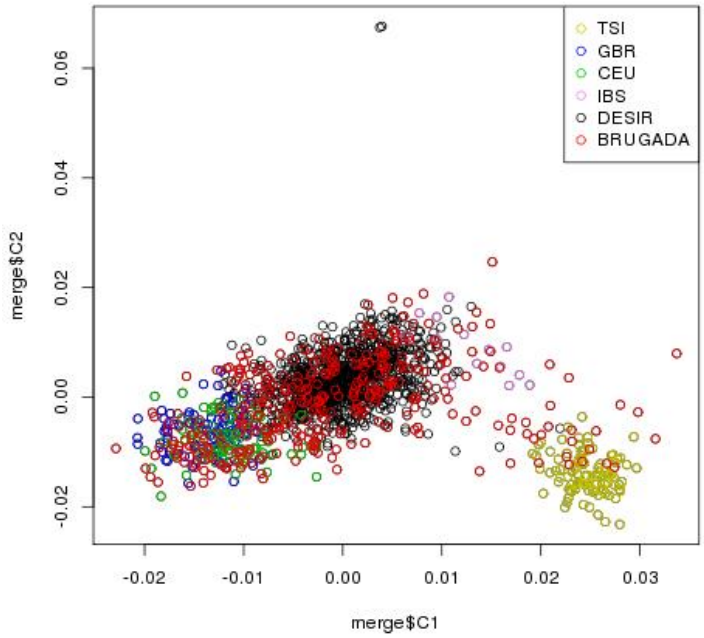


Figure 58 : Positionnement sur les composantes principales en Europe.
Les individus sont présentés selon leur position sur le plan des axes des deux premières composantes. Nous avons représentés les témoins et les patients – tous centres confondus – et les individus issus des 1000 Génomes. Ce graphe permet de retrouver des populations « témoins » qui s'apparient avec les cas de notre étude qui ne peuvent être intégrés dans l'analyse stricte (Analyse I).

Pour vérifier qu'il y a bien un appariement démographique cohérent, nous avons ré-analysé ces résultats en identifiant chaque centre recruteur (Figure 59).

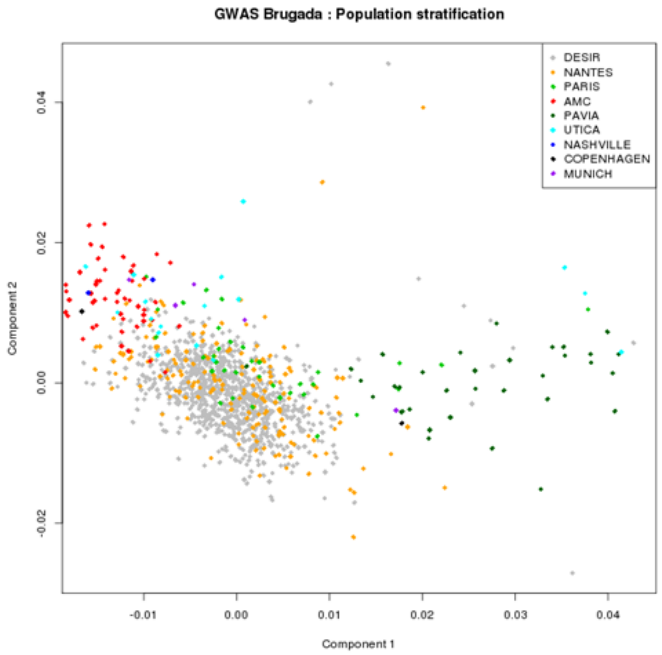


Figure 59 : position des individus de chaque centre sur les composantes 1 et 2
Position des individus selon les axes des composantes 1 et 2. Les individus sont colorés en fonction de leur centre de recrutement (pour les patients) ou de leur statut témoin. Les individus sont relativement bien regroupés suivant le centre d'où ils sont issus.

Nous voyons que, comme c'est le cas pour la plupart des études impliquant des échantillons provenant de plusieurs centres Européens, les positions sur les deux premières composantes reflètent la position géographique de l'origine des individus. Ici, les échantillons français sont au milieu, entre les échantillons « nordiques » (Amsterdam, Copenhague) et les échantillons provenant de Pavie.
Cette analyse va nous permettre de créer une étude cas contrôles démographiquement homogène.
Nous utilisons pour cela une méthode en deux étapes avec une sélection empirique basée sur les positions suivie d'une analyse de recherche de valeurs aberrantes.

G.III.6. Résultats

G.III.6.a) *Analyse sur les cas et témoins français :*

L'analyse est d'abord effectuée en utilisant les témoins français et les patients qui s'apparient graphiquement au niveau de leur origine génétique. Il s'agit, par la force des choses, d'une analyse d'association dans une population française. Les résultats sont d'abord montrés en fonction de la position des SNPs sur le génome (Figure 60).

Ces résultats sont très encourageants et font ressortir un locus très associé sur le chromosome 3 ainsi qu'une forte probabilité d'association (non significative à l'échelle du génome toutefois) sur le chromosome 6.

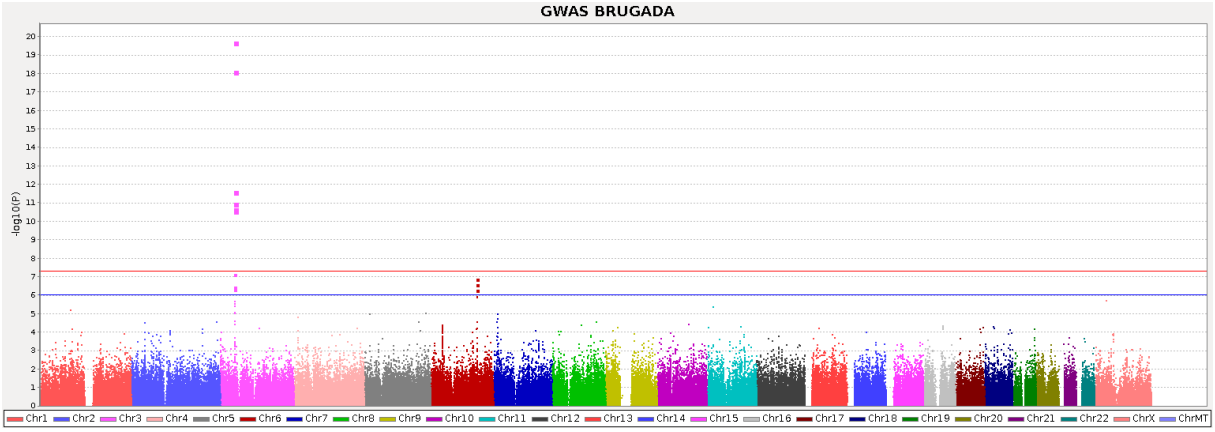


Figure 60 : Résultats d'analyse d'association Brugada (analyse I) – Manhattan Plot
Cette figure appelée Manhattan Plot (D.VI.1) montre les résultats, sous forme de p-valeur, en fonction de la position des SNPs sur le génome. La ligne rouge correspond à une p-valeur de 5.10^{-8} et la ligne bleue à une p-valeur de 10^{-6}. Sur le chromosome 3, en rose foncé, il y a un groupe de SNPs fortement significatifs, même après correction pour le nombre de tests. Sur le chromosome 6, un autre groupe de SNPs s'approche du seuil de 5.10^{-8} (en rouge).

Il convient ensuite de vérifier si ces résultats ne sont pas dus à une augmentation systématique de la statistique, à une stratification ou à une différence dans la qualité de génotypage, grâce à une QQ plot (Figure 61).

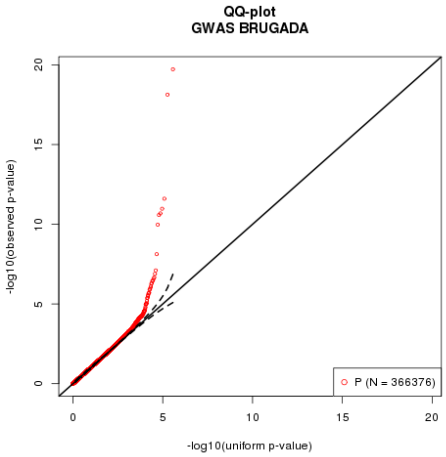


Figure 61 : Résultats d'analyse d'association Brugada (analyse I) - QQplot
Les p-values observées (rouge - ici transformées en log de base 10) sont comparées à leur distribution sous l'hypothèse nulle. On observe ici une identité entre les deux courbes jusqu'à une valeur de 4.8 (p-valeur=1.6x10^{-5}) et une séparation juste après. On en conclut qu'il n'y a pas d'inflation générale de la statistique et qu'à partir de 4.8, il y a un enrichissement en vrais positifs.

Par contre, il y a une proportion non négligeable des cas qui ne sont pas utilisés en raison de leur origine géographique qui est éloignée de celle des témoins disponibles français (~30%). Or ces cas pourraient être utiles pour discriminer entre les SNPs potentiellement réplicables des autres. C'est le but de l'analyse II : en effet, la phase de réplication doit être réduite à moins de 10 SNPs pour des raisons financières et également de rapidité de publication. Pour cela, nous voulons sélectionner des SNPs ayant les plus grandes chances de réplication étant donnée la taille de la population utilisée en phase 2.

G.III.6.b) *Extension de l'analyse aux populations Européennes :*

Dans un deuxième temps nous allons utiliser les populations issues des 1000 génomes comme contrôles pour nos patients issus des centres Européens. Il apparaît que cette approche a aussi été utilisée par [Génin et al.,**2011**] sur une pathologie rare avec recrutement multicentrique.

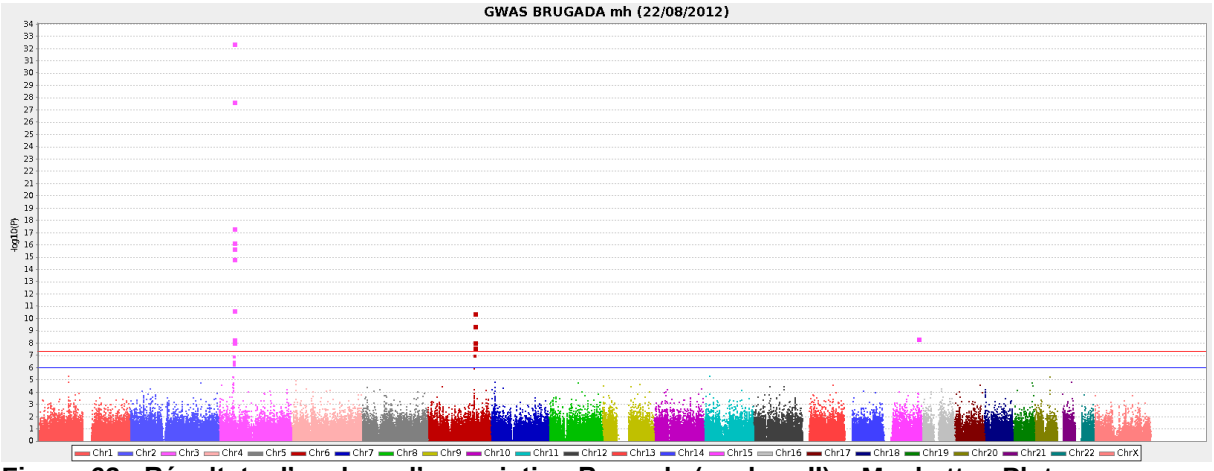


Figure 62 : Résultats d'analyse d'association Brugada (analyse II) – Manhattan Plot
Cette figure montre les résultats, sous forme de p-valeur, en fonction de la position des SNPs sur le génome. La ligne rouge correspond à une p-valeur de 5.10^{-8} et la ligne bleue à une p-valeur de 10^{-6}. Sur le chromosome 3, en rose foncé, le résultat des SNPs fortement significatifs est fortement augmenté. Sur le chromosome 6, le groupe de SNPs initialement proches du seuil rouge 5.10^{-8} le dépassent largement.

Les résultats (Figure 62) permettent de discriminer parfaitement entre les loci réplicables dans un échantillon de taille modeste et ceux qui sont soit de faux positifs, soit nécessiteraient un échantillon beaucoup plus grand pour donner une p-valeur significative de $5*10^{-08}$.

G.III.6.c) *Contrôles a-posteriori*

La première vérification que nous avons entreprise est l'inspection des résultats des SNPs dans chaque région. Il s'agit de tester si des SNPs corrélés (en déséquilibre de liaison) montrent la même association.

G.III.6.c.1) Locus du chromosome 3

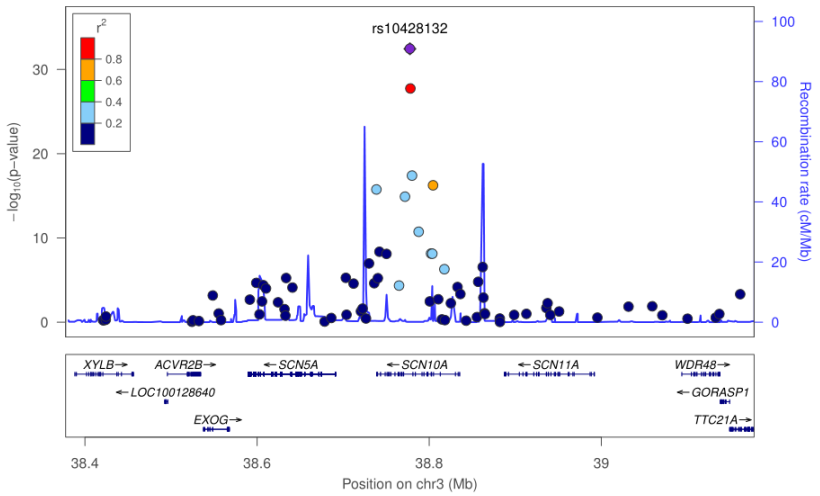


Figure 63 : Association au niveau du locus chromosome 3

Représentation de la force de l'association (log 10 (p-valeur)) au niveau local pour la plus forte association du chromosome 3. Le SNP le plus significatif est indiqué en violet. Pour les autres SNPs, le déséquilibre de liaison avec le SNP le plus significatif est indiqué selon un code de couleurs (rouge pour les SNPs en fort DL). En bleu (légende à droite), est montré le taux de recombinaison local. Sur cette figure, les SNPs corrélés au SNP principal montrent également un fort signal d'association. Enfin, les gènes et leurs positions dans cette région sont montrés dans la partie basse de la figure.

Les SNPs qui sont en déséquilibre de liaison montrent également un signal d'association fort (Figure 63). L'intensité de ce signal (-log10(p-valeur)) est proportionnel au déséquilibre de liaison. Nous pouvons sélectionner ce locus pour une réplication.
Nous notons que le locus du chromosome 3 est très proche du gène SCN5A, dont le rôle dans la survenue du syndrome de Brugada a été largement démontré.

G.III.6.c.2) Locus du chromosome 6

Nous observons, pour le chromosome 6, la même corrélation entre intensité de l'association. Cette association se limite au bloc de recombinaison qui entoure les gènes *HEY2* et *NCOA7* (Figure 64).

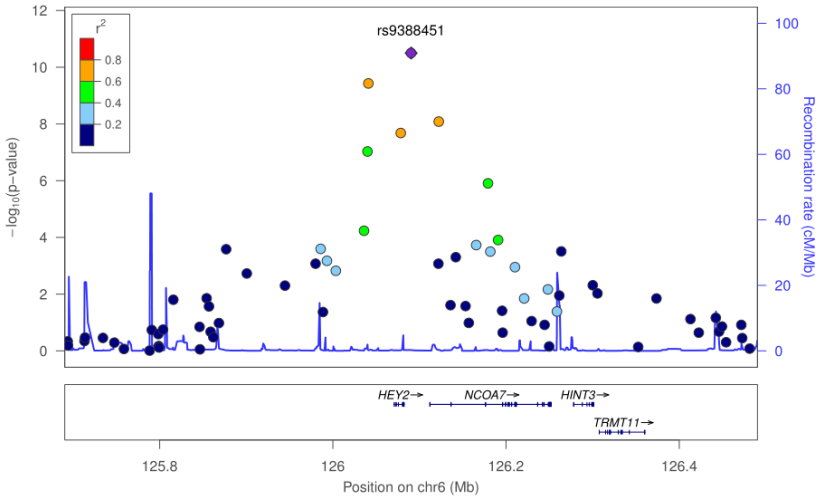


Figure 64: Association au niveau du locus chromosome 6

Représentation de la force de l'association (log 10 (p-valeur)) au niveau local pour la plus forte association du chromosome 6. Le SNP le plus significatif est indiqué en violet. Pour les autres SNPs, le déséquilibre de liaison avec le SNP le plus significatif est indiqué selon un code de couleurs (rouge pour les SNPs en fort DL). En bleu (légende à droite), est montré le taux de recombinaison local. Sur cette figure, les SNPs corrélés au SNP principal (en orange) montrent également un fort signal d'association.

Enfin, les gènes et leurs positions dans cette région sont montrés dans la partie basse de la figure.

G.III.6.c.3) Locus du chromosome 15

L'analyse fine de l'association sur le locus du chromosome 15 fait apparaître un seul SNP fortement associé (Figure 65), de façon isolée. Les SNPs se trouvant fortement corrélés avec le SNP rs12592167 ne montrent pas de signal d'association (p-value > 0.02), et en tous cas pas du même ordre que le signal principal (p-valeur < 10^{-8}). Il est très probable que ce SNP soit un faux positif technique, même s'il a passé les différents filtres.

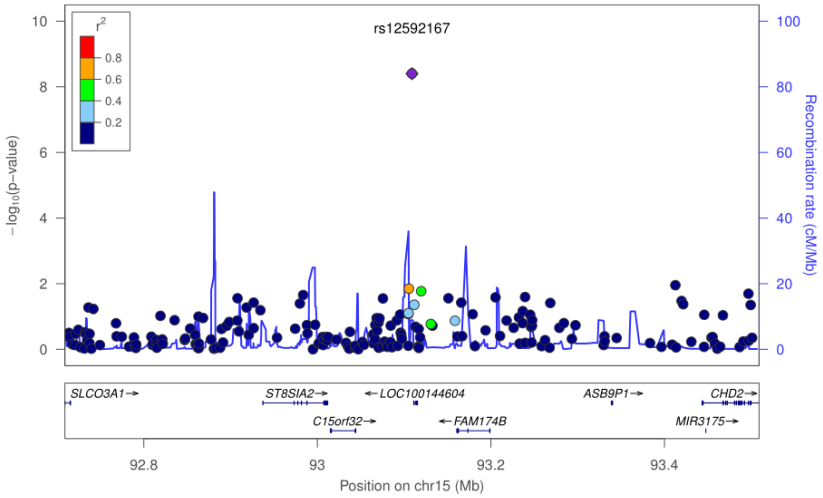


Figure 65 : Association au niveau du locus chromosome 15

Représentation de la force de l'association (log 10 (p-valeur)) au niveau local pour la plus forte association du chromosome 15. Le SNP le plus significatif est indiqué en violet. Pour les autres SNPs, le déséquilibre de liaison avec le SNP le plus significatif est indiqué selon un code de couleurs (rouge pour les SNPs en fort DL). En bleu (légende à droite), est montré le taux de recombinaison local.
Enfin, les gènes et leurs positions dans cette région sont montrés dans la partie basse de la figure.

Une vérification du graphe de clusterisation (Figure 66) permet de réaliser que l'association sur le chromosome 15 peut-être due un génotypage imparfait qui induirait un faux positif.

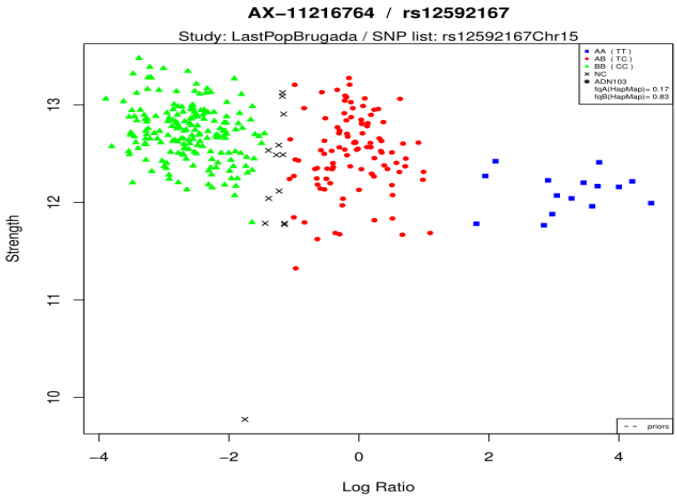


Figure 66 : Cluster Graphe chez les patients

Les génotypes sont mal différenciés. Les homozygotes AA (en vert) sont ainsi très mal séparés des hétérozygotes(en rouge). Il y a un excès d'individus mis de façon forcée au statut inconnu pour les individus qui sont soit AA, soit AB. Il en résulte une baisse artificielle de la fréquence de l'allèle A chez les patients.

Nous avons toutefois vérifié s'il n'y avait pas d'association sous un modèle récessif. En effet, le groupe homozygote rare est (relativement) bien identifié (Figure 66). Par ailleurs, sous un modèle récessif, la corrélation entre génotypes (et non plus entre allèles) réduit la puissance d'un test pour un SNP en déséquilibre de liaison. Cette possibilité a été exclue (p= 0.3).

G.III.6.d) *Sélection et réplication*

Au vu de ces résultats nous avons donc mis en place un essai de réplication dans deux autre échantillons. Il est à noter que dans notre cas, nous avons déjà une significativité énorme et que la réplication est nécessaire plutôt pour s'assurer du caractère général de cette association – en utilisant une autre technique sur d'autres patients par exemple – que

pour augmenter la significativité statistique. Par ailleurs, nous avons identifié un SNP localisé dans l'intron du gène *SCN5A*, sur le bras court du chromosome 3, qui montrait une association modeste mais indépendante de l'association principale (Tableau 13). Ce SNP étant également associé à l'intervalle PR dans les études d'association génome entier publiées, nous l'avons inclus dans la phase de réplication.

gène	CHR	SNP	BP	A1	F_A	F_U	P	OR
SCN5A	3	rs11708996	38633923	C	0.2203	0.1542	5.807e-06	1.549
SCN10A	3	rs10428132	38777554	T	0.6911	0.4164	3.563e-33	3.136
HEY2	6	rs3799708	126078530	G	0.2935	0.3929	2.085e-08	0.641
HEY2	6	rs9388451	126090377	T	0.3829	0.49	3.165e-11	0.645

Tableau 13 : Fréquence des allèles dans les *loci* sélectionnés dans l'analyse I

Les gènes correspondent aux gènes les plus proches du SNP le plus associé. Ce ne sont pas automatiquement les gènes causant cette association. Pour le locus du gène HEY2, nous avons choisi deux SNPs. Les fréquences allélique sont calculées dans la population française (**Analyse I**) mais le degré de significativité montré est celui de la méta-analyse Européenne (**Analyse II**).

G.III.6.e) *Distribution des fréquences en Europe*

Du fait de l'hétérogénéité de population des patients, il est important d'évaluer la différence des fréquences alléliques, entre centres, pour ces différents SNPs (Figure 67). Si les fréquences SNPs au niveau de gènes SCN10A et SNC5A semblent assez homogènes à travers l'Europe, cela semble être moins le cas pour les deux SNPs proches du gène *HEY2*.

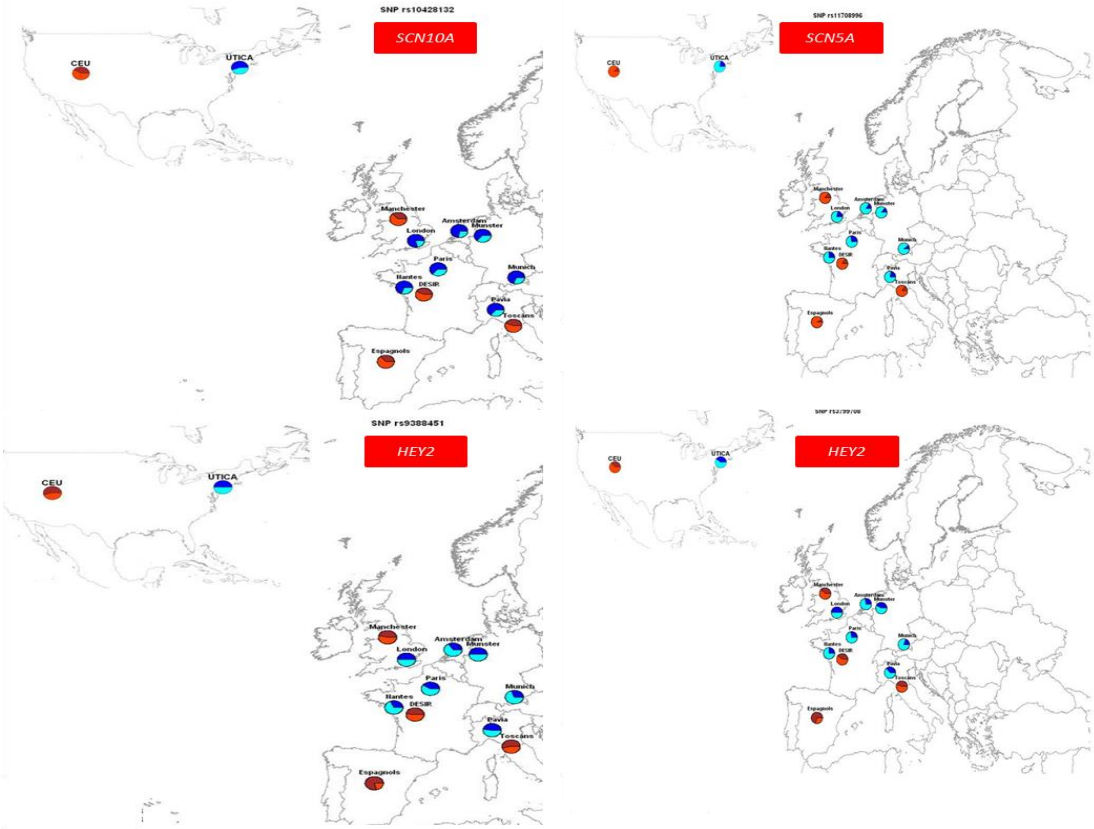


Figure 67 : Distribution Géographique des fréquences alléliques

Les fréquences sont présentées sous forme de disques avec un dégradé différent suivant l'allèle. Les deux surfaces de chaque camembert sont proportionnelles à la fréquence allélique. Les disques bleus représentent les fréquences alléliques des patients et sont placés géographiquement sur leur origine. Les disques rouges représentent la population témoin DESIR ou une population issue de l'étude des 1000 Génomes. Les patients centres d'origine des patients sont représentés par un disque bleu et ceux des témoins par un disque orange.

G.III.7. Réplication

G.III.7.a) *Réplication Européenne*

L'échantillon Européen pose des problèmes à plusieurs égards. Il s'agit, comme pour l'analyse initiale, d'une étude multicentrique. Il peut donc y avoir des fréquences allélique différentes par dérive génétique. Par ailleurs, nous n'avons pas la possibilité de vérifier la stratification puisque nous n'allons pas effectuer de génotypage génome entier. Nous avons donc réuni un échantillon de 599 patients Européens.

gène	chr	SNP	BP	A1	F_A	F_U	P	OR
SCN5A	3	rs11708996	38633923	C	0.23489	0.15257	3.64e-08	1.7
SCN10A	3	rs10428132	38777554	T	0.64845	0.41874	4.13e-32	2.37
HEY2	6	rs3799708	126078530	G	0.30717	0.40299	1.91e-07	0.66
HEY2	6	rs9388451	126090377	T	0.41385	0.49507	1.62e-05	0.72

Tableau 14 : Association des SNPs de SCN5A, SCN10A et HEY2 dans une étude Européenne de 599 cas et 780 témoins.

Nous observons une association très significative pour tous les SNPs. Ces associations restent très significatives lorsqu'on analyse la population française seulement.
Les associations observées en analyse d'association génome entier sont donc répliquées et nous pouvons en conclure qu'il s'agit de vrais positifs.

G.III.7.b) *Réplication Japonaise*

La population de réplication comprenait plusieurs échantillons.
Les patients sont constitués de deux groupes. Il y a d'abord un groupe de 205 patients ayant un diagnostic de Brugada décrits dans [Ohno et al.,**2011**]. Les patients viennent de trois centres (Nagasaki, Osaka et Chiba). En parallèle, 1000 témoins japonais ont été génotypés, venant d'un centre à Nagasaki.

gène	CHR	SNP	BP	A1	F_A	F_U	P	OR
SCN5A	3	rs11708996	38633923	C	0.08904	0.03998	1.521e-05	2.347
SCN10A	3	rs10428132	38777554	T	0.4406	0.23080	2.418e-19	2.626
HEY2	6	rs3799708	126078530	G	0.21	0.31600	1.075e-05	0.575
HEY2	6	rs9388451	126090377	T	0.2785	0.39360	6.233e-06	0.594

Tableau 15 : Association des SNPs de SCN5A, SCN10A et HEY2 dans une étude Japonaise de 205 cas et 1000 témoins.

Il y a clairement réplication de tous les SNPs. Nous pouvons donc en conclure non seulement que ces associations sont des vrais positifs, ce qui était déjà .

G.III.8. Effets en population

Pour estimer les effets de notre modèle polygénique en population, nous avons décidé d'utiliser les effets observés dans la réplication japonaise,

G.III.8.a) *Risque de récurrence familial λ_S :*

Le risque de récurrence pour germain (frère/sœur) d'un proposant atteint λ_S (Équation 11) et (D.III.9) est de 1.41, quand on le calcule avec les paramètres suivants : fréquence allélique et OR de la population japonaise. Ce λ_S est relativement haut dans un contexte d'effets de variants fréquents. Il reflète évidemment l'effet fort (toujours dans le cadre des variants fréquents) observé pour chaque variant individuellement ainsi que leur fréquence élevée dans la population. Pour comparaison, et comme montré au premier paragraphe des résultats, le λ_S pour le diabète de type 2 est de 1.10.

G.III.8.b) *Héritabilité :*

L'héritabilité estimée (Équation 12) à partir d'une prévalence de 5.10^{-3} est de 8% dans la population japonaise. L'héritabilité générale du syndrome de Brugada n'a jamais été estimée et il est donc difficile de réaliser la proportion de l'héritabilité totale expliquée par ces trois loci. Il n'est pas fréquent, pour des pathologies communes, que les trois marqueurs les plus associés expliquent à eux seuls 8% de l'héritabilité. Il est difficile d'en tirer des conclusions en termes de proportion de l'héritabilité attendue car cette héritabilité n'a pas été calculée en population.

G.III.9. Architecture génétique

Mon objectif était d'évaluer comment le Syndrome de Brugada s'inscrit dans le paysage des pathologies complexes et quelle est sa proximité avec d'autres entités phénotypiques à travers les résultats de la génétique. Il est à noter que l'effet le plus fort est le fait d'un SNP (rs10428132) fortement associé à l'intervalle PR – qui donne une information principalement sur le courant auriculaire. Il est également, mais beaucoup plus modestement, associé au complexe QRS, qui est un marqueur de la conduction cardiaque. Des troubles de conduction étant identifiés dans le syndrome de Brugada, cette association semblerait plus logique et interprétable du point de vue de la physiologie. Toutefois, l'importance d'une modification du courant auriculaire dans la survenue du syndrome de Brugada ne saurait être négligée.

G.III.9.a) *Effets des SNPs associés aux phénotypes ECG*

Nous avons cherché l'association des SNPs de notre étude avec des traits de l'ECG. Notre idée était d'améliorer notre connaissance des bases du Syndrome de Brugada par comparaison d'association.

SNP	chr	BP	Cas fa	Tém fa	p-valeur	Position	Gène(s)	maf	Trait
rs10865879	3	38577362	0.28048	0.22643	5.7271e-03	3p22_2	*SCN5A, SCN10A*	0.26	QRS
rs11129795	3	38589163	0.28210	0.22710	5.5191e-03	3p22_2	*SCN5A, SCN10A*	0.25	QRS
rs11708996	3	38633923	0.24016	0.15344	1.4787e-04	3p22_2	*SCN5A, SCN10A*	0.17	PR/QRS
rs12053903	3	38593393	0.38994	0.32000	3.6771e-03	3p22_2	*SCN5A*	0.29	Conduction
rs12053903	3	38593393	0.38994	0.32000	3.6771e-03	3p22_2	*SCN5A*	0.29	QTc
rs1362212	7	35305306	0.23044	0.18670	7.7290e-03	7p14_3	*TBX20*	0.13	QRS
rs17020136	2	37248015	0.22835	0.17731	4.3898e-03	2p22_2	*HEATR5B, STRN*	0.23	QRS
rs3807989	7	116186241	0.48228	0.39965	2.5891e-03	7q31_1	*CAV1/CAV2*	0.43	PR
rs6599222	3	38648062	0.28795	0.22498	5.3864e-03	3p22_2	*SCN5A, SCN10A*	0.23	PR
rs6795970	3	38766675	0.66223	0.40992	4.7195e-22	3p22_2	*SCN5A, SCN10A*	0.42	PR/QRS
rs6798015	3	38798836	0.59965	0.36804	2.7742e-20	3p22_2	*SCN5A, SCN10A*	0.38	PR
rs9851724	3	38719935	0.21860	0.30936	3.3496e-04	3p22_2	*SCN5A, SCN10A*	0.28	QRS

Tableau 16 : Association pour les SNPs identifiés comme déjà associés avec des traits ECG [Kolder et al.,2012**].**

Les résultats sont extraits de **l'analyse I** sur les populations françaises seulement. Ce Tableau ne montre que les SNPs qui ont une p-valeur inférieure à 0.01. Les résultats statistiques ont été obtenus par imputation. Les SNPs testés sont inclus car ils ont montré une association significative à l'échelle du génome (p-valeur < $5x10^{-8}$) pour au moins un trait ECG.
BP = position en paires de bases
Cas fa = fréquence allélique chez les cas de notre population
Tém fa = fréquence allélique chez les témoins de notre population
p-valeur = p-valeur dans l'analyse cas-témoins Brugada
Position = bande chromosomique
Gène = nom du ou des gènes les plus proches
Maf = fréquence allélique dans les études publiées.
Trait = phénotype avec lequel l'association a été trouvée.

Les résultats des associations sont tirés de [Kolder et al.,**2012**], qui est une revue de toutes les associations publiées jusqu'à 2012 avec des traits ECG, et présentées dans le Tableau 16. Nous observons que les SNPs décrits comme associés aux traits ECG dans la littérature ne sont pas, majoritairement associés avec le statut SBr. Le résultat le plus important est le SNP rs6795970 (et le SNP en fort DL rs6798015), sur le chromosome 3. Il s'agit donc de SNPs associés avec l'intervalle PR et le QRS.

Pour les SNPs se trouvant hors de cette région 3p22.2, il s'agit également de SNPs qui sont associés avec l'intervalle PR et le complexe QRS.

G.III.9.b) *Enrichissement en gènes*

Nous avons utilisé le test d'enrichissement de gènes connectés par la littérature pour rechercher d'éventuels voies, ou groupes de gènes importants dans le risque au Syndrome de Brugada.

Les mots clés reliant les gènes les plus connectés étaient :
'kinesin', 'cardiac', 'guidance', 'homeodomain', 'heart', 'developing', 'homeobox', 'development', 'neuronal', 'axons', 'embryos', 'embryonic', 'neurons', 'left', 'bhlh', 'axon', 'connexin', 'morphogenesis', 'ventricular', 'differentiation'.

Ces mots clés semblent indiquer que les bases génétiques du Syndrome de Brugada se situent également au niveau du développement cardiaque, et pas seulement au niveau des protéines canaux directement (Figure 68).

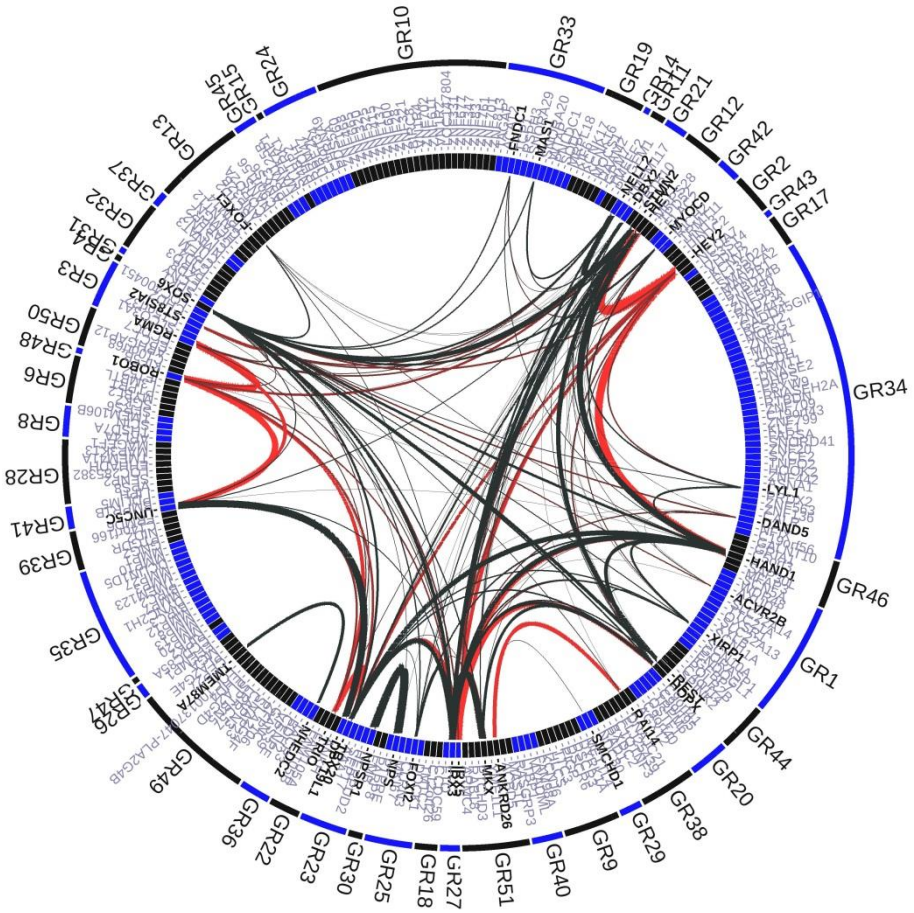


Figure 68 : Graphe de connexion des régions associées à SBr

Les groupes d'association sont placés autour du cercle. Chaque groupe comprend les gènes à une distance de 500 kb (en 3' et en 5'). Le numéro du groupe donne son rang dans le résultat statistique (GR1 est la région autour du SNP le plus significatif). Dans la deuxième couche sont notés les gènes appartenant à chaque groupe. Les connexions entre deux gènes sont notées par une ligne les reliant. Les lignes rouges témoignent d'une p-valeur < 0.01 et la force du lien est proportionnelle à l'épaisseur du trait. Une grande partie des gènes liés par des lignes rouges sont impliqués dans le développement, qu'il soit ou non cardiaque.

Nous retrouvons le gène *TBX20* dans ce groupe hautement corrélé, qui est également associé au QRS (G.III.9.a) dans les analyses d'association génome entier.

G.III.10. Synthèse des résultats Brugada

Nous avons mis en évidence trois variants génétiques très fortement associés à la survenue du Syndrome de Brugada. Ces variants sont fréquents et, qui plus est, dans le cas des *HEY2*, c'est l'allèle (haplotype) fréquent qui augmente le risque de survenue de cette pathologie.

Le locus *SCN5A-SCN10A* était déjà identifié comme modulant très fortement l'onde PR. Les mêmes SNPs sont donc également fortement associés à la survenue du Syndrome de Brugada. Le gène *SCN5A*, qui code pour le canal Nav1.5 était également le principal gène candidat dans la survenue du syndrome de Brugada.

En parallèle, un autre variant, localisé près du gène *HEY2* fortement impliqué dans le développement embryonnaire par la voie de signalisation Delta-Notch a été mis en évidence par notre analyse. L'invalidation de ce gène chez la souris entraîne des défauts cardiaques graves [Sakata et al.,**2002**][Gessler et al.,**2002**][Donovan et al.,**2002**][Sakata et al.,**2006**]. Aucun polymorphisme dans cette région n'avait été mis en évidence dans les études d'association génome entier pour quelque pathologie que ce soit. Cette découverte met en évidence le possible rôle du développement cardiaque dans le syndrome de Brugada, ce qui n'était pas forcément attendu.

Le locus au niveau de *SCN10A* avait déjà été associé avec l'intervalle PR et l'intervalle QRS, qui sont deux mesures de la conduction cardiaque, dans la population générale [Chambers et al.,**2010**][Holm et al.,**2010**][Pfeufer et al.,**2010**]. *SCN10A* est localisé juste en amont du gène codant pour la sous unité α du canal sodium, *SCN5A*. Nous avons vu la part importante jouée par ce gène dans le risque de SBr. Récemment, il a été démontré que ce SNP modifie un nucléotide hautement conservé qui se trouve à l'intérieur d'un site de fixation de type T-box. Le polymorphisme atteint l'activité du segment *enhancer*, réduisant ainsi l'activité de ce segment [van den Boogaard et al.,**2012**]. Dans cette étude, Van den Boogaard et ses collègues font donc l'hypothèse que ce polymorphisme touche l'expression de *SCN5A*/*SCN10A* régulée par *TBX5*/*TBX3 in vivo.* Étant donné la contribution majeure de *SCN5A* au courant Sodium et à la conductivité cardiaque, il nous semble très probable que ce soit une modification de l'expression de *SCN5A* (plutôt que de *SCN10A*) qui soit responsable de l'effet de ce SNP sur le système de conduction et sur l'apparition du phénotype SBr de Type 1. Nous pensons que la région mise en évidence par ces SNPs sur le chromosome 3p a une action sur la survenue du syndrome de Brugada à travers la régulation de *SCN5A* et, possiblement de *SCN10A*.

G.IV. Papier n° 3 : Dina et al.

G.V. Papier n° 4 : Bezzina et al.

H. Projet 3 : Pathologie Cardiaques - Le système valvulaire

- Recherche des bases génétique du Prolapsus Valvulaire Mitral

H.I. Les valves dans le cycle cardiaque

Dans le système cardiaque qui nécessite un passage régulier du sang, les valves jouent un rôle majeur.

Les quatre valves, **tricuspide** (oreillette droite – ventricule droit), **pulmonaire** (ventricule droit – artère pulmonaire), **mitrale** (oreillette gauche – ventricule gauche) et **aortique** (ventricule gauche – aorte), assurent le passage unidirectionnel du sang d'une cavité à l'autre en empêchant le reflux (Figure 69). Les valves sont constituées de trois feuillets à l'exception de la valve mitrale qui est constituée de deux feuillets. Les valves auriculo-ventriculaires (mitrale et tricuspide) sont ancrées au ventricule par des cordages tendineux.

Le fonctionnement normal de ces valves est appréhendé par l'auscultation à l'aide d'un stéthoscope. Le premier bruit du cœur est grave, assourdi, et correspond à la fermeture des valves auriculo-ventriculaires, marquant le début de la systole. Le second bruit du cœur est plus fort, et coïncide à la fermeture des valves aortiques et pulmonaires, marquant le début de la diastole.

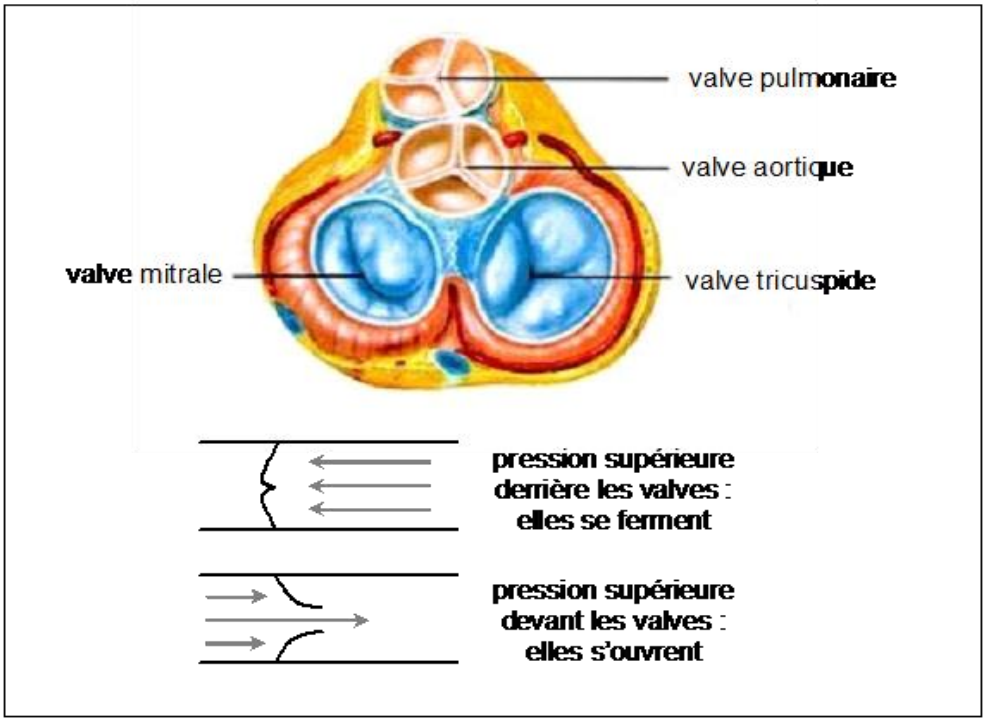


Figure 69 : Les valves dans l'anatomie du cœur

Vue des valves à partir de l'atrium (après ablation). Adapté d'après [Netter et al.,**2011**]

Les valves cardiaques sont des structures élastiques, non contractiles, empêchant le reflux du sang d'une cavité cardiaque vers une autre. Dans certains cas, elles peuvent présenter un dysfonctionnement, qu'il soit inné ou qu'il survienne au cours de la vie. Une valve peut dysfonctionner de deux manières. Soit elle ne s'ouvre pas correctement : on parle de rétrécissement, soit elle ne se ferme pas convenablement : on parle d'insuffisance.

Les valves de la partie gauche du cœur sont atteintes plus précocement que les droites. Cela est expliqué par le stress physique du à une pression plus élevée dans la partie gauche. Les valves droites peuvent être touchées ultérieurement mais l'on diagnostiquera la pathologie par les manifestations symptomatiques des valves gauches.

Selon les types de valvulopathies, un individu peut rester asymptomatique ou présenter des symptômes non-spécifiques (palpitations, malaise, etc.) liés aux complications.

On parle de valvulopathies aortiques (souvent un Rétrécissement Aortique Calcifié ou Osseux) et valvulopathies mitrales (comme le Prolapsus Valvulaire Mitral).
Selon sa gravité, une valvulopathie peut nécessiter une prise en charge allant de la simple surveillance cardiologique régulière à une intervention de correction ou de remplacement. Les modalités d'intervention dépendent du patient (âge en particulier), de la valve atteinte, de sa défaillance (fuite ou rétrécissement), de l'origine de la maladie, de son évolution et de la présence ou non de complications.

En général, en cas de valvulopathie, on remplace les valves cardiaques défectueuses par des valves prothétiques.
Les prothèses ne sont pas toutefois pas sans inconvénients (thrombose, traitement anticoagulant, dégénérescence, désinsertion). Les plasties valvulaires, qui permettent de réparer la valve malade sans la changer sont encore en phase de développement.
Actuellement, pour la chirurgie de la valve aortique, les valvuloplasties n'ont pas obtenu de bons résultats en raison d'un taux de faisabilité faible et d'un manque de standardisation des techniques.

En revanche, en ce qui concerne la valve mitrale, une réparation est plus souvent réalisable (2/3 des cas). Cependant, son résultat demeure dépendant du mécanisme de la fuite, de l'extension des lésions, ainsi que de la technique de réparation utilisée.

L'absence de solution chirurgicale optimale et les complications induites par la présence des prothèses motivent surtout la recherche de bio-marqueurs permettant de prévenir au mieux cette pathologie.

H.II. Prolapsus Valvulaire Mitral

H.II.1. Description

H.II.1.a) *La Valve Mitrale*

La valve mitrale est située entre l'oreillette gauche et le ventricule gauche. Elle s'ouvre en diastole, permettant le remplissage ventriculaire et se ferme durant systole (Figure 70).

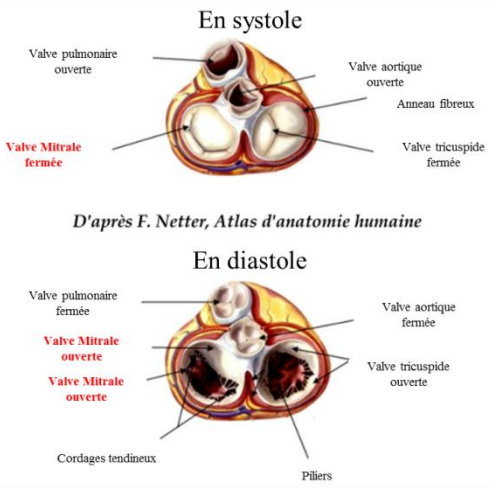


Figure 70 : dynamique des valves
Vue de la base (après ablation des oreillettes). La valve mitrale est fermée en systole, afin d'empêcher le sang du ventricule de remonter dans l'auricule. Adapté de [Netter et al.,**2011**]

L'ensemble de l'appareil mitral, qui comprend l'annulus, la voile, le cordage et les muscles papillaires a un mouvement propre dans lequel interviennent les différents constituants – particulièrement les muscles papillaires et les cordages (Figure 71).

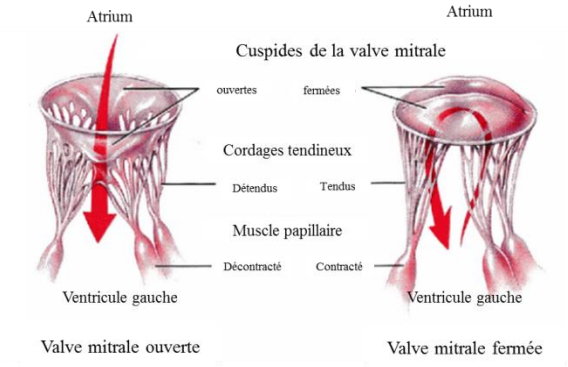


Figure 71 : Mouvements et structure de l'appareil mitral
L'appareil mitral, constitué de la valve, des cordages tendineux et du muscle papillaire est un système coordonné qui évolue de façon active grâce au muscle papillaire. Adapté de [Netter et al.,**2011**]

H.II.1.a.1) Histologie

Sur le plan histologique, la valve mitrale est un tissu constitué de quatre couches distinctes, qui sont, de la surface auriculaire à la région ventriculaire :
l'auricularis : c'est une fine couche de cellules endothéliales reposant sur du tissu conjonctif et une fine couche de cellules musculaires lisses.
la spongiosa : elle est formée de fibres élastiques, de fibroblastes, de fibres de collagène organisées de manière lâche et situées dans une matrice de muco-polysaccharides. Elle est riche en protéoglycanes et assure la protection de la valve face aux pressions exercées sur les feuillets valvulaires. Elle atténue les frottements entre l'auricularis e la fibrosa.
la fibrosa : c'est une couche dense et compacte de fibres de collagène en continuité avec l'anneau fibreux valvulaire et la partie centrale des cordages tendineux. Elle assure la rigidité et la résistance des feuillets de la valve mitrale.
la ventricularis : fine couche de cellules endothéliales reposant sur du tissu conjonctif.

C'est donc une structure lamellaire où le tissu conjonctif est majoritaire. Le maintien de cette structure et donc les liens entre les cellules valvulaire interstitielles et la matrice extracellulaire jouent sans doute un rôle important.

H.II.1.a.2) Cellules de la valve mitrale

La majorité des cellules constituant le myocarde sont des fibroblastes [Eghbali et al.,**1991**]. Ces cellules sont primordiales dans la constitution et le maintien du tissu conjonctif par le renouvellement du collagène et des protéines des fibres. Elles assurent aussi sa destruction par l'action des métallo-protéases (collagénases et protéases). La présence de ces cellules est encore plus importante dans le tissu conjonctif valvulaire mitral.

H.II.1.b) *Le Prolapsus Valvulaire Mitral*

Le Prolapsus Valvulaire Mitral est une pathologie où un (ou même deux) feuillets rentrent dans l'atrium gauche, n'empêchant plus le sang de refluer depuis le ventricule dans cette chambre atriale gauche (Figure 72). Ce reflux du sang est la régurgitation mitrale (RM) qui, selon sa gravité, peut déboucher sur une insuffisance ventriculaire.

En général, on observe une dégénérescence myxoïde du tissu valvulaire mitral dont la couche spongieuse (*spongiosa*) vient envahir et fragiliser la couche fibreuse (*fibrosa*) par dépôt de protéoglycans. Il y a une désorganisation marquée de certains types de collagènes (I et II) ainsi que de l'élastine et d'autres protéines du tissu conjonctif.

La valve en devient remarquablement épaissie, avec des cordages souvent également épaissis, fusionnés ou même calcifiés, et des ruptures de ces cordages. Si la valve en dégénérescence myxoïde prend la forme d'un parachute, et bombe au-delà de la jonction atrio-ventriculaire dans l'oreillette gauche, nous parlons de **prolapsus mitral**, dans lequel une partie d'une valve, une valve entière ou même les deux valves peuvent être atteintes.
Les dystrophies valvulaires par dégénérescence myxoïde sont des pathologies cardiaques fréquentes et potentiellement graves. Nous nous intéressons à la forme la plus répandue, le prolapsus valvulaire mitral (PVM), encore appelé maladie de Barlow.

Nous notons que dans le Prolapsus Valvulaire Mitral, on inclut également la dégénérescence conjonctive avec déficience en collagène, en élastine et en protéoglycans, résultant en un amincissement du tissu valvulaire (dégénérescence fibro-élastique - DFE). Bien que ces deux formes soient histologiquement différentes (et même opposées avec un amincissement du tissu conjonctif dans le DFE), leurs conséquences semblent proches et elles sont classées en tant que PVM.

On pense, en général, que ce processus de changement de la population cellulaire et de la matrice intercellulaire est une réponse à un stress trop important. Il s'agit alors d'un cercle vicieux où l'affaiblissement de la valve mitrale entraîne un effet encore plus grave de la pression exercée.

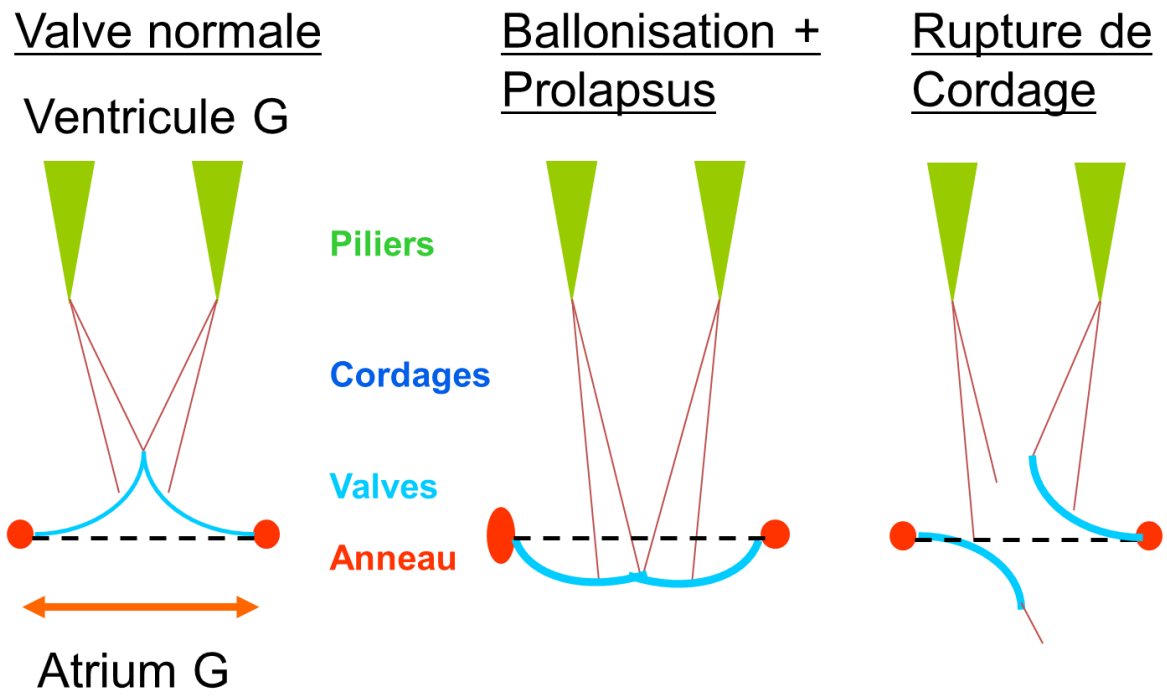


Figure 72 : Schéma comparatif d'une valve mitrale normale et atteinte de PVM
À gauche : valve en condition physiologique normale, la valve est jointe aux piliers du cœur par l'intermédiaire des cordages et fixée également au tissu cardiaque au niveau des anneaux.
Au centre : valve atteinte et touchée par le prolapsus, avec un aspect de ballonisation, orientation anormale vers l'oreillette gauche (nommée Atrium G sur le schéma).
À droite : rupture de cordage ne permettant plus d'assurer la liaison de la valve à l'un des piliers. Source : Pr. Thierry LE TOURNEAU, CHU Nantes, communication personnelle.

Une valvulopathie importante va fatiguer progressivement le muscle cardiaque et aboutir à une insuffisance cardiaque.
Elle peut également entraîner un trouble du rythme de l'oreillette (particulièrement dans le cas de la valvulopathie mitrale). De plus, une valve malade est fragilisée, ce qui la rend beaucoup plus sensible aux infections, notamment aux endocardites.
Il n'existe actuellement pas de traitement spécifique médical et la seule solution thérapeutique reste le remplacement valvulaire lorsque les symptômes apparaissent.

H.II.2. Épidémiologie

Deux à 3 % de la population adulte est porteuse d'une valvulopathie, la plus fréquente étant l'insuffisance mitrale [Freed et al.,**1999**][Freed et al.,**2002**].
Cette prévalence s'accroît avec l'âge pour atteindre entre 10 et 15% chez les patients de plus de 75 ans. On s'attend à ce que cette prévalence continue à croître avec le vieillissement de la population.
Il existe plusieurs étiologies de valvulopathies. Dans le laboratoire, et dans le cadre de mon travail de thèse, nous nous intéressons plus particulièrement aux formes dégénératives non-syndromiques (dégénérescence fibroblastique, maladie de Barlow). Nous éviterons les formes syndromiques (syndrome de Marfan, syndrome d'Ehlers-Danlos, syndrome de Loeys-Dietz) ainsi que les formes d'origine ischémique, infectieuse, inflammatoire (par exemple à la suite d'un rhumatisme articulaire aigu) ou traumatique.

Toutefois, les formes syndromiques, souvent des maladies génétiques du tissu conjonctif, et incluant un Prolapsus Valvulaire Mitral permettent de générer des hypothèses de gènes candidats. Il s'agit principalement de gènes impliqués dans le maintien du tissu conjonctif (collagène III, FBN-I …) et le maintien ou la différentiation des fibroblastes (*TGFβ1* par exemple).

On distingue deux types de prolapsus suivant que l'épaisseur des feuillets mitraux est égale ou supérieure (forme dite classique) ou inférieure (forme non classique) à 5 mm. La forme classique (1,3 % des cas) est répartie à peu près également entre les deux sexes. La forme non classique (1,1 % des cas) est à nette prédominance féminine et d'évolution généralement bénigne. Dans notre étude, une partie des patients que nous allons étudier ont souffert de cette forme dite « non classique » (dégénérescence fibro-élastique - DFE).

H.II.3. Bases Génétiques

Contrairement au Diabète de Type 2, il n'y a pas eu d'étude permettant d'estimer une héritabilité du Prolapsus Valvulaire Mitral dans la population. Toutefois, dans la définition même du Syndrome de Barlow, qui inclut le PVM, il y avait, intrinsèquement, l'idée de transmission génétique [Barlow, Pocock,**1979**]. Par ailleurs, des études sur la survenue de la pathologie chez des apparentés du premier degré de proposants atteints concluaient à une concentration familiale, sans doute d'origine génétique [Strahan et al.,**1983**][Weiss et al.,**1975**][Devereux et al.,**1982**]. Le modèle de transmission le plus probable était autosomique dominant avec une expressivité variable en fonction de l'âge et du sexe [Strahan et al.,**1983**].
Par ailleurs, des formes familiales plus marquées étaient décrites régulièrement dans la littérature. Par exemple Monteleone et Fagan décrivaient une dystrophie valvulaire familiale avec une transmission liée au chromosome X, forme plus rare de cette pathologie [Monteleone, Fagan,**1969**].
L'équipe dirigée par Xavier Jeunemaître a décrit le premier locus de prolapsus valvulaire mitral à transmission autosomique dominante en 1999 [Disse et al.,**1999**] situé en 16p11-13 dans deux familles. Deux autres loci ont été récemment identifiés par une équipe américaine. Ils sont situés sur le chromosome 11p15.4 [Freed et al.,**2003**] et sur le chromosome 13q31-32 [Nesta et al.,**2005**]. À ce jour les gènes sous-tendant ces liaisons ne sont pas encore identifiés. La connaissance des bases moléculaires à l'origine des valvulopathies myxoïdes est donc encore très limitée.
Notre équipe a identifié et décrit le premier locus de prolapsus valvulaire mitral avec une transmission liée au chromosome X en 1998 [Kyndt et al.,**1998**]. Ce travail a été réalisé à partir d'une grande famille de 318 membres dans laquelle l'atteinte valvulaire, était associée à une hémophilie A mineure. Par ailleurs, de façon importante, le séquençage de la région codante du gène *FLNA* codant pour la filamine A a permis d'identifier une mutation faux sens c.1910C>A dans l'exon 13 entraînant la substitution p.Pro637Gln [Kyndt et al.,**2007**]. Il s'agissait donc de la première analyse de liaison débouchant sur l'identification d'un gène du Prolapsus Valvulaire Mitral.
Une recherche systématique de mutations dans ce gène dans un ensemble de familles où la pathologie ségrégait suivant un modèle de liaison au chromosome X a permis d'identifier de nouvelles mutations : deux mutations faux-sens, p.Gly288Arg et p.Val711Asp ainsi qu'une délétion de 1944 paires de base ont renforcé l'hypothèse que ce gène portait des mutation augmentant le risque de PVM [Kyndt et al.,**2007**].

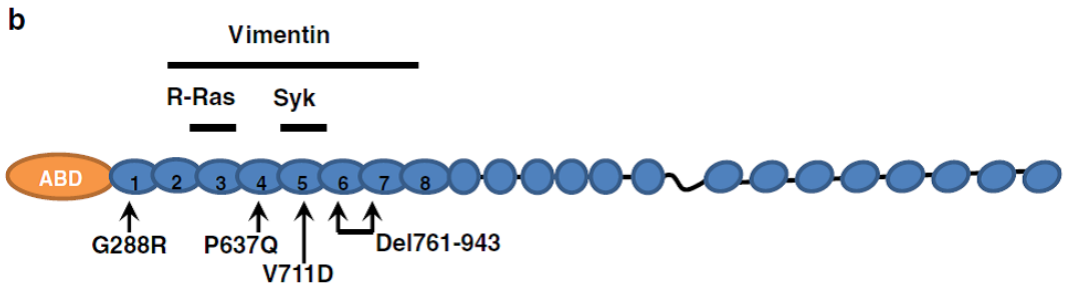


Figure 73 : Schéma de la filamine A.
Emplacement des différentes mutations *FLNA* et domaines d'interaction avec différentes protéines partenaires.). Nous montrons ici les protéines interagissant seulement avec le domaine où se trouvent les mutations identifiées. Les protéines partenaires de la filamine A sont beaucoup plus nombreuses.
ABD : Actin Binding Domain
Extrait de [Lardeux et al.,**2011**]

Des études récentes ont identifié de nouveaux partenaires liés à ces domaines en N-terminal de la protéine, où sont localisées les mutations trouvées jusqu'à maintenant, pouvant avoir des conséquences physiopathologiques en cas de dysfonctionnement : elles incluent la petite GTPase R-Ras, la tyrosine-kinase splénique (Syk) et une protéine de filament intermédiaire, la Vimentine (interaction illustrées dans la figure représentant la filamine A Figure 73). De façon générale, l'identification de la Filamine A permet de générer des hypothèses de gènes candidats à partir des protéines qui interagissent avec cette protéine (comme la voie du TGF-B interagissant avec ses médiateurs cellulaires, les « smads »).

L'ensemble de ces résultats renforce l'hypothèse d'une composante héréditaire. Il arrive que certaines revues présentent le PVM comme étant majoritairement héritable [Guy, Hill,**2012**]. Contrairement à d'autres pathologies, comme le Diabète de Type 2, nous n'avons pas d'estimation de l'héritabilité qui puisse nous permettre de nous rendre compte, dans une phase ultérieure, quelle proportion de l'héritabilité nous pourrons expliquer avec les variants fréquents. Il semble malgré tout raisonnable d'initier une recherche de variants fréquents modulant le risque de Prolapsus Valvulaire Mitral par une étude d'association génome entier.

H.II.4. Description de l'étude

Cette étude, que nous appellerons MITRAL se base également sur un Consortium, MITRAL, financé par la Fondation Leducq et comprenant des équipes des deux côtés de l'Atlantique (Figure 74).

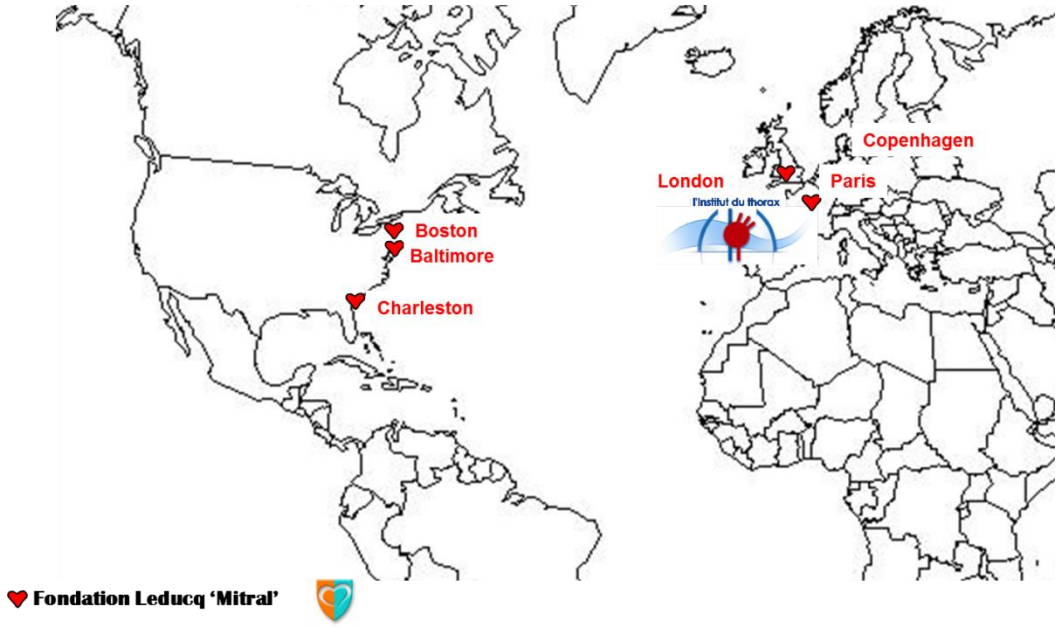


Figure 74 : Laboratoires du Réseau Leducq
Boston : Massachussets General Hospital, Brigham and Women's Hospital, Children's Hospital Boston and Harvard, **Baltimore** : Johns Hopkins University - School of Medicine, Chaleston : Medical University of South Carolina, **Londres** : Imperial College, **Paris** : Institut National de la Santé et de la Recherche Médicale U970 (PARCC) , Hôpital Européen Georges Pompidou, **Nantes** : Institut du thorax/Université de Nantes.

L'objectif principal de ce réseau est d'étudier les facteurs étiologiques de la pathologie de la valve mitrale.
Dans ce cadre, mon objectif est la recherche de polymorphismes génétiques fréquents associés à la pathologie « Prolapsus Valvulaire Mitral ».

Pour réaliser cet objectif, deux analyses cas-populations à l'échelle du génome ont été réalisées. Dans la première, nous avons combiné 522 cas recrutés à Nantes et Angers avec 780 témoins issus de la population DESIR. Ces individus ont été génotypés avec le système Axiom Affymetrix.
En parallèle, 980 patients ont été génotypés sur une puce Illumina 650k (étude PVM France décrite ci-dessous (H.II.4.e.1..i). Ils ont été comparés à la population SU.VI.MAX, comprenant 1673 contrôles et génotypés sur puce Illumina 300k. Pour réaliser cette comparaison, nous n'avons retenu que les SNPs communs aux deux puces.

H.II.4.a) *PVM-France*

PVM France est une étude multicentrique nationale est en cours en France, ayant pour objectif de rechercher des gènes de susceptibilité du prolapsus valvulaire mitral idiopathique par une étude prospective sur l'ensemble du génome. Son but est également de mieux appréhender les différences existant entre DFE et maladie de Barlow.
Cette collection prospective nationale, placée sous l'égide de la Société française de cardiologie, est basée sur une échographie diagnostique (pour les patients non opérés) et un prélèvement sanguin pour l'étude génétique (seul examen réalisé chez les patients déjà opérés).
Cette étude a inclus, en France, 1 000 patients opérés ou non d'un prolapsus mitral.

H.II.4.b) *PVM-Nantes*

PVM-Nantes est une étude basée au CHU de Nantes. Il s'agit du recrutement systématique de patients opérés d'un prolapsus valvulaire mitral au service de cardiologie. La plus grande

partie des patients souffrent d'un syndrome de Barlow. Le nombre de patients ainsi inclus dans l'étude a dépassé 550.

H.II.4.c) *Caractéristiques cliniques des patients*

Les deux études avaient à l'origine une différence de stratégie de recrutement et il n'est pas étonnant d'observer en final une différence dans la structure finale des cohortes (Tableau 17).

		PVM-France	PVM-Nantes
Sexe (H/F)		798/319	330/155
Type PVM			
	Barlow	710	435
	DFE	191	36
	Non déterminé	216	100
Opération			
	Oui	814	565
	Non	303	7
Combinaisons	*Barlow Opéré*	495	430
	Barlow non Opéré	215	5
	DFE Opéré	129	36
	DFE Non Opérés	62	0
	Non déterminé Opéré	190	98
	Non déterminé Non Opéré	88	2
Année de Naissance	*Moyenne (écart type)*	1945 (14.5)	1936 (12.3)

Tableau 17 : Patients dans les deux études PVM

Nous observons une grande disproportion dans la constitution des cohortes. Dans la cohorte Nantaise, il y a une écrasante majorité de patients opérés. Cela peut être dû au fait que ces patients sont plus âgés. Toutefois, lorsqu'on inspecte les cas de PVM-France, on observe que la moyenne des années de naissance des patients opérés est 1948 (1945 pour les non opérés). Il semble donc qu'il ne s'agit pas juste d'une question d'aggravation du phénotype avec l'âge.

H.II.4.d) *Population générale SU.VI.MAX*

SU.VI.MAX (SUpplémentation en VItamines et Minéraux Anti-oXydants) est une étude lancée le 11 octobre 1994 en vue de constituer une source d'informations sur la consommation alimentaire des français et leur état de santé. En fin d'étude 12,735 sujets ont été inclus [Hercberg et al.,**1998**]. Le groupe de témoins de notre étude comprend 1673 individus, un sous-groupe d'individus de l'étude SU.VI.MAX génotypés en Illumina 317k – Illumina HumanHap 300- (utilisés par exemple dans [Wain et al.,**2011**]). Comme pour la population D.E.S.I.R., ce groupe d'individu génotypés sera analysé indépendamment de ses phénotypes (qui ne sont de toute façon pas accessibles). Il s'agit donc d'une étude cas-population générale.

H.II.4.e) *Stratification :*

H.II.4.e.1) Stratification Intercontinentale :

Nous avons appliqué d'abord l'analyse classique permettant de placer nos échantillons par rapport à des individus ayant un fonds génétique très différent. La matrice d'Identité par Etat

(Identity By State) entre les individus de notre étude (cas et témoins) et tous les individus issus de l'étude « 1000 Génomes » (D.V.3) a été réalisée en utilisant le logiciel PLINK[Purcell et al.,**2007**]. A partir de cette matrice de distance génétique entre individus, le logiciel PLINK permet de réaliser par MDS (Multi-Dimension Scaling) une réduction de dimension. Chaque individu est ainsi représenté par des coordonnées correspondant à ses valeurs propres (comme expliqué dans D.VI.6).

H.II.4.e.1..i) PVM France

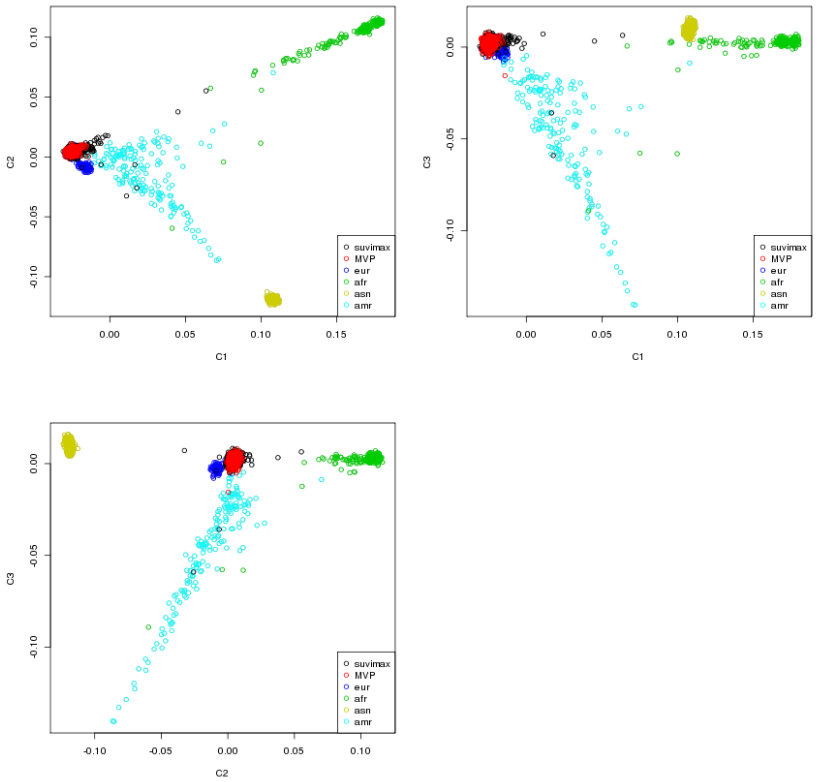


Figure 75 : Localisation sur les composantes pour une analyse intercontinentale
L'analyse de stratification intercontinentale permet de mettre en évidence une séparation claire des individus Africains (vert), Asiatiques (jaune) Européens (bleu foncé).

Nos cas et nos témoins se placent bien dans la population Européenne des 1000 génomes. Les quelques individus de notre étude se trouvant placés en dehors du regroupement de points représentant les Européens seront exclus.

H.II.4.e.2) Stratification Européenne

Nous observons des différences entre populations Européennes. Nos populations de patients et de témoins (population générale) sont placées entre les deux les populations Européennes nordiques (GBR et CEU) et les deux populations du Sud (TSI et IBS) - Figure 76.

H.II.4.e.2..i) PVM-France

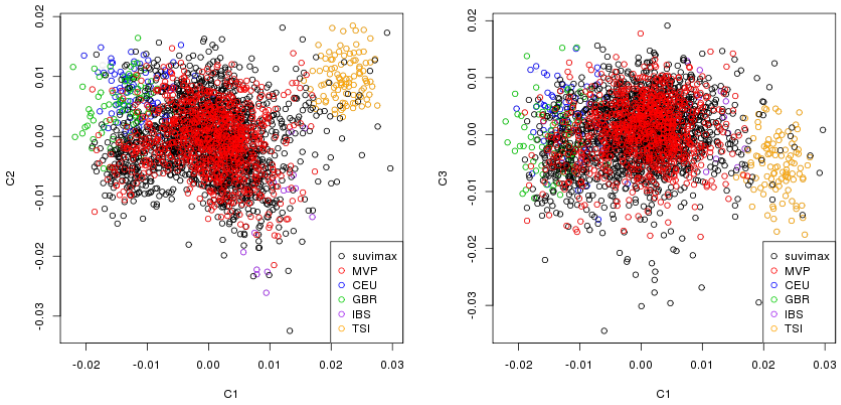


Figure 76 : Localisation des individus sur les composantes principales.
Gris : tous les individus. Verts : britanniques. Bleu : Européens de l'Utah. Jaune : Toscans. Violet : ibériques (en fait Espagnols). Noir : témoins. Rouge : patients. Les cas (rouge) et les témoins (noir) se recouvrent de façon assez satisfaisante.

Une deuxième vue plus détaillée (Figure 77) permet de mieux identifier les différences entre populations et, surtout, la superposition des cas et des témoins permet de conclusre que nous pouvons réaliser une étude d'association comparant ces deux groupes.

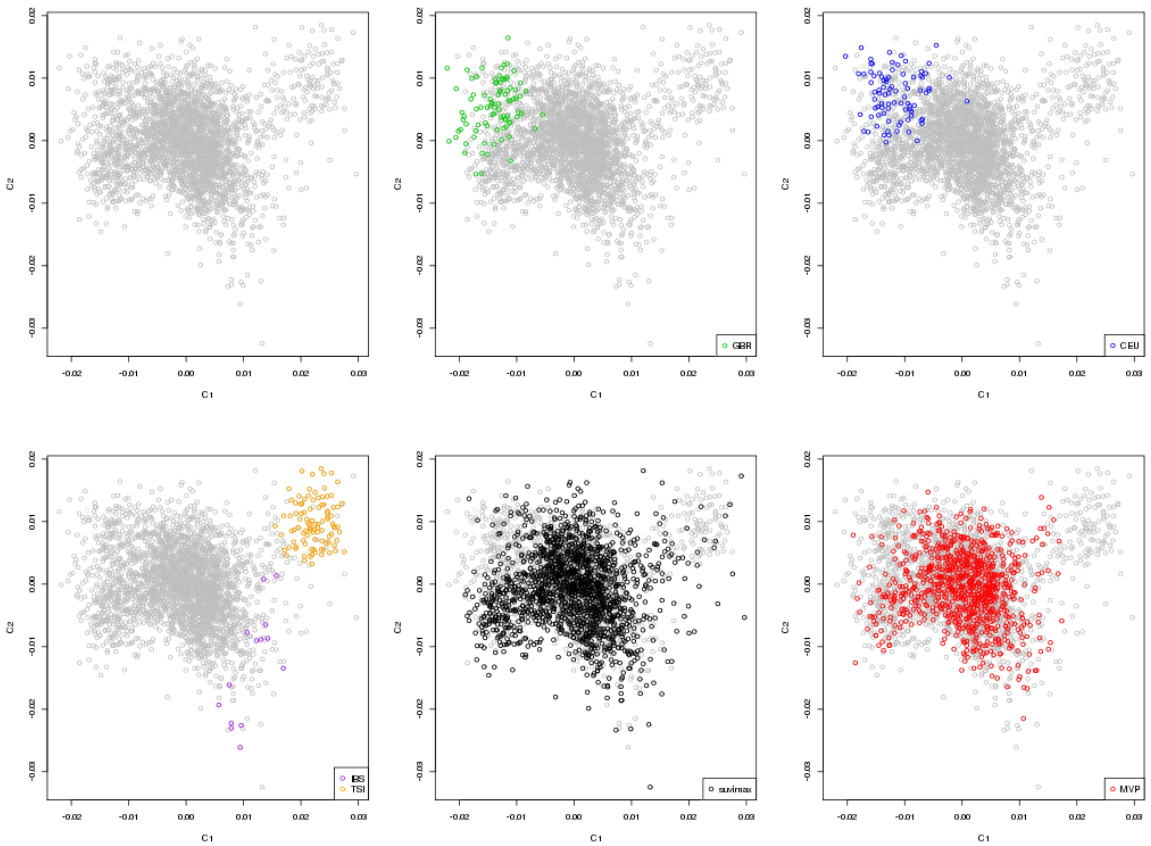


Figure 77 : stratification continentale et nationale - détails.
Gris : tous les individus. Verts : britanniques. Bleu : Européens de l'Utah. Jaune : Toscans. Violet : ibériques (en fait Espagnols). Noir : témoins. Rouge : patients. Dans la distribution générale, les cas et témoins français se distribuent entre les différentes populations Européennes et se recouvrent de façon satisfaisante.

Les populations françaises se retrouvent bien de façon systématique entre les individus 1000 génomes « nordiques » (CEU et GBR) et les individus « méridionaux » (TSI et IBS – Toscans et Espagnols).
Quelques individus de notre étude recouvrent les populations Espagnoles, Italiennes et Nordiques, ce qui est très vraisemblable étant donné l'histoire démographique de notre pays.

H.II.4.e.2..ii) PVM-Nantes

Nous faisons une observation similaire dans l'étude Nantaise. Les cas recrutés à Nantes ainsi que les témoins D.E.S.I.R se retrouvent placés sensiblement de la même façon que PVM-France/SU.VI.MAX. : entre les populations Européennes d'origine Nordique et les Toscans et Espagnols.
Le recouvrement entre cas et témoins est satisfaisant.

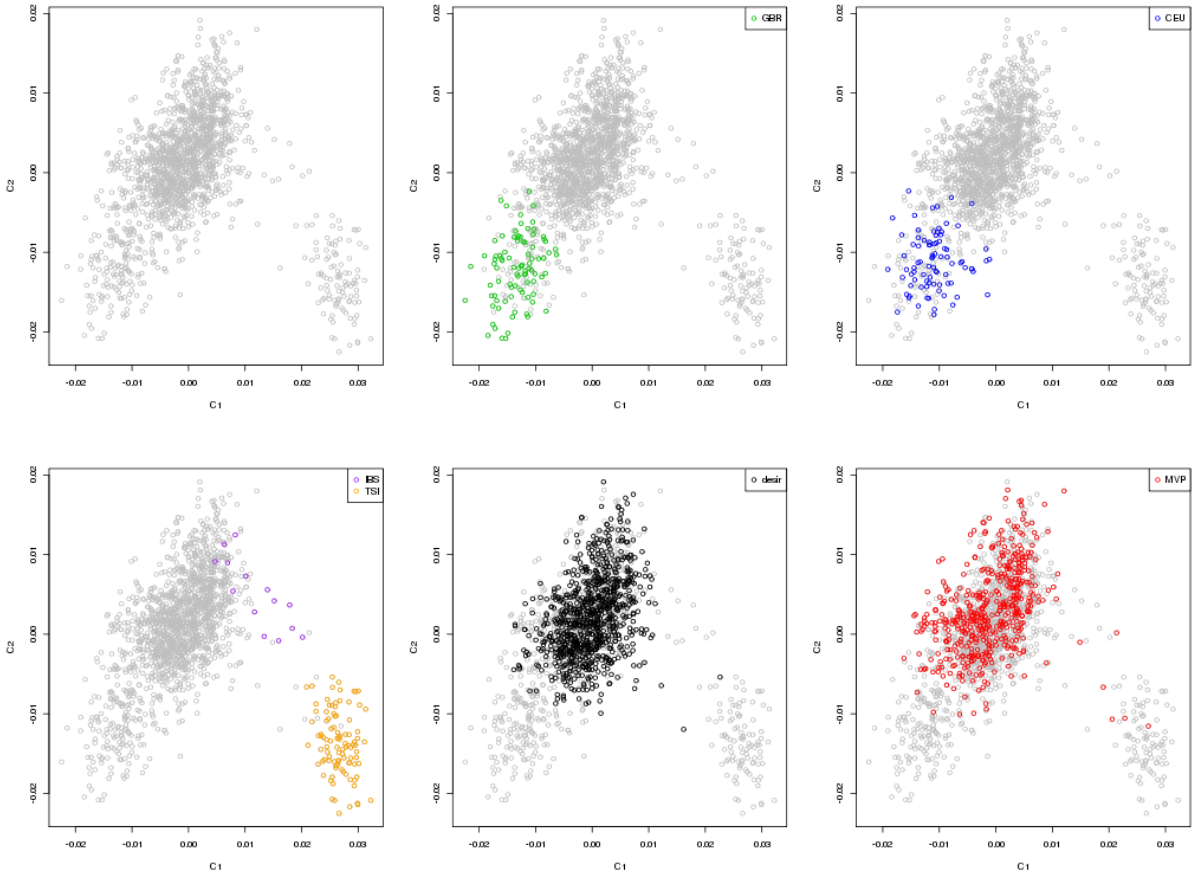


Figure 78 : stratification continentale et nationale – détails
Gris : tous les individus. Verts : britanniques. Bleu : Européens de l'Utah. Jaune : Toscans. Violet : ibériques (en fait Espagnols). Noir : témoins. Rouge : patients. Dans la distribution générale, les cas et témoins français se distribuent entre les différentes populations Européennes et se recouvrent de façon satisfaisante.

H.II.4.f) *Méthode d'analyse*

Les données génotypiques ont été préparées pour l'imputation selon la méthode décrite dans (D.VI.2.a) . Nous avons pré-phasé les données afin de pouvoir facilement imputer les données à partir de nouvelles versions des 1000 génomes dans le futur. Le pré-phasage a été effectué en utilisant le programme SHAPEIT [Delaneau et al.,**2012**]. Nous avons ensuite

utilisé le programme IMPUTE afin d'imputer les génotypes manquants. J'ai choisi l'option qui permettre d'utiliser les SNPs déjà génotypés.
Trois phénotypes ont été testés : PVM, Barlow, Opérés.
J'ai ensuite utilisé la méthode de l'inverse de la normale pour combiner les résultats des analyses des deux études (D.VI.2.b).
J'ai choisi de corriger sur les 5 premières composantes. La stratégie était d'inclure les deux premières composantes qui étaient très associées aux coordonnées géographiques du lieu de naissance.

H.II.4.g) *Résultats de la meta-analyse :*

H.II.4.g.1) Prolapsus Valvulaire Mitral

Nous observons une inflation (Figure 79) de la p-valeur (1.07) avec toutefois, plusieurs SNPs qui sont associés avec une p-valeur inférieure à 5.10^{-08} (Figure 80).

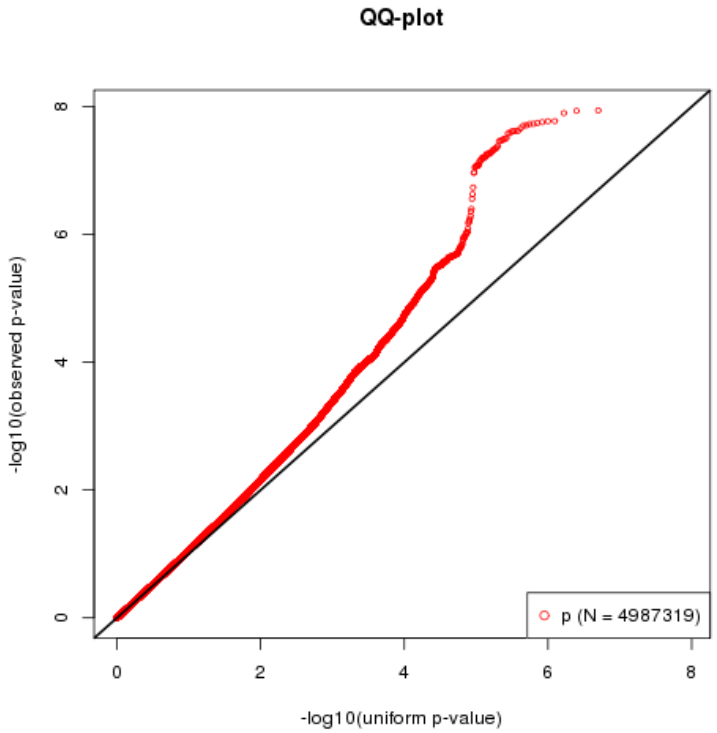


Figure 79 : QQ-Plot pour le Prolapsus Valvulaire Mitral
Les p-valeurs observées (rouge - ici transformées en log de base 10) sont comparés à leur distribution sous l'hypothèse nulle. On observe ici une séparation entre les deux courbes à un niveau assez bas, qui témoigne d'une légère inflation de la statistique – qui peut être dû à des différences démographiques résiduelles mais également à une différence de qualité de l'ADN. Il y a un décollement plus important à partir de 1.10^{-6}.

Sur cette Figure 80, il apparaît que plusieurs régions sont intéressantes et devront être testées dans des cohortes additionnelles. Il s'agit des régions situées sur le chromosome 2p, 17p et 21p.

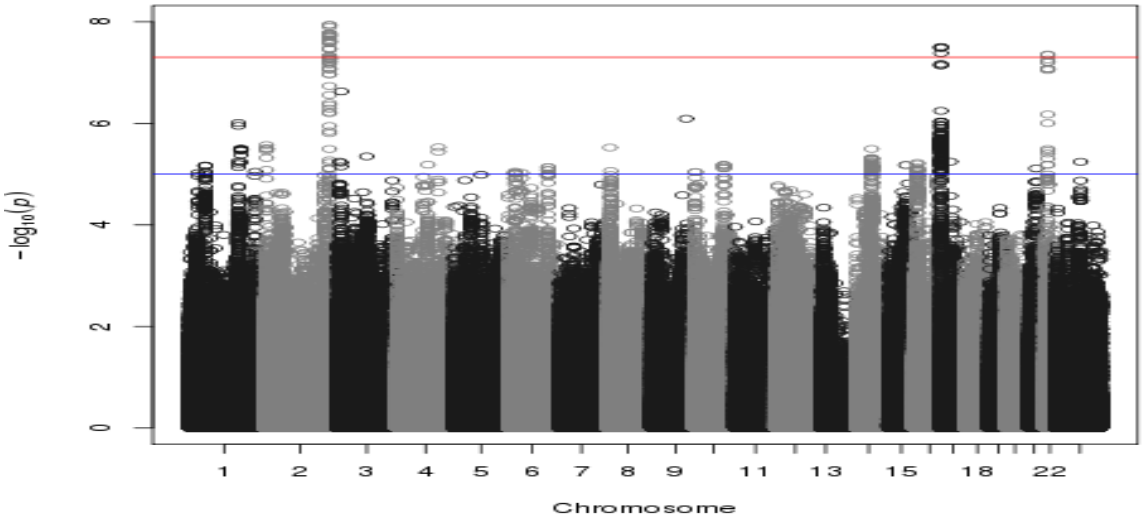


Figure 80 : Manhattan Plot pour le Prolapsus Valvulaire Mitral

Ce Manhattan Plot (D.VI.1) montre les résultats, sous forme de p-valeur, en fonction de la position des SNPs sur le génome. La ligne rouge correspond à une p-valeur de 5.10^{-8} et la ligne bleue à une p-valeur de 10^{-6}. Sur le chromosome 2, le chromosome 17 et le chromosome 22, on observe une association qui dépasse le seuil rouge de 5.10^{-8}.

H.II.4.g.2) PVM opérés

Les résultats sont assez semblables pour une analyse restreinte aux patients opérés. Il y a toujours une inflation de la statistique (λ_{GC}=1.05 cette fois) - Figure 81.

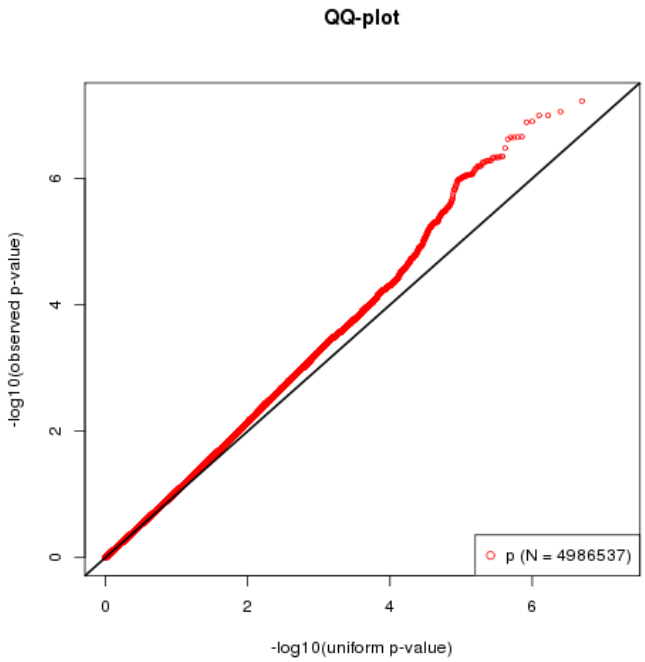


Figure 81 : QQ-Plot pour le Prolapsus Valvulaire Mitral - Opéré

Les p-values observées (rouge - ici transformées en log de base 10) sont comparés à leur distribution sous l'hypothèse nulle. On observe ici une séparation entre les deux courbes à un niveau assez bas, qui témoigne d'une légère inflation de la statistique – qui peut être dû à des différences démographiques résiduelles mais également à une différence de qualité de l'ADN. Il y a un décollement plus important à partir de 1.10^{-6}.

Les loci du génome apparaissant comme intéressants (Figure 82) recouvrent largement ceux observés pour l'analyse PVM totale. Toutefois, nous voyons l'apparition d'une nouvelle région au niveau du chromosome 16.

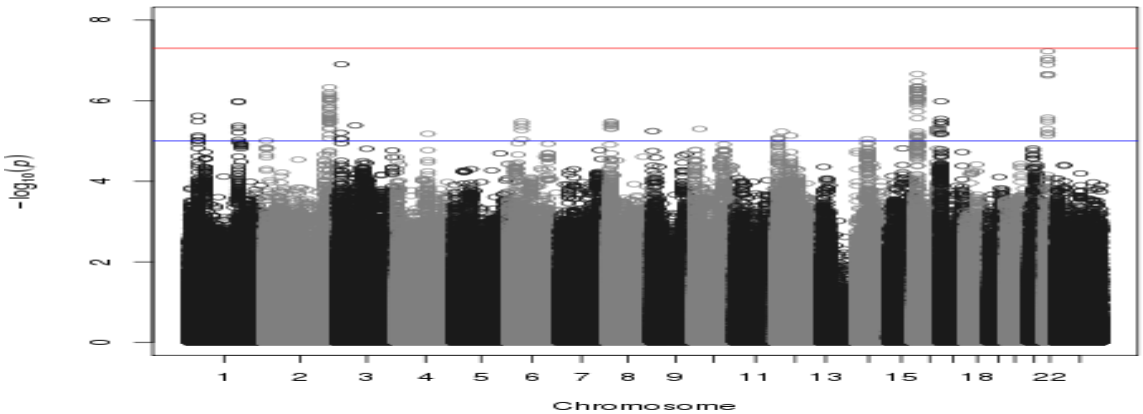


Figure 82 : Manhattan Plot pour le Prolapsus Valvulaire Mitral - Opéré

Ce Manhattan Plot (D.VI.1) montre les résultats, sous forme de p-valeur, en fonction de la position des SNPs sur le génome. La ligne rouge correspond à une p-valeur de 5.10^{-8} et la ligne bleue à une p-valeur de 10^{-6}. Sur le chromosome 2, le chromosome 17 et le chromosome 22, on observe une association qui dépasse le seuil rouge de 5.10^{-8}.

H.II.4.g.3) PVM Barlow

De façon assez surprenante, les résultats de l'analyse menée sur les patients souffrant de la forme de Barlow ne montrent pas de résultats potentiellement différents (Figure 83, Figure 84), alors que nous aurions pu nous attendre à ce que l'exclusion des formes de dégénérescence fibro-élastiques permette, grâce à un phénotype plus homogène, de faire ressortir des signaux spécifiques.

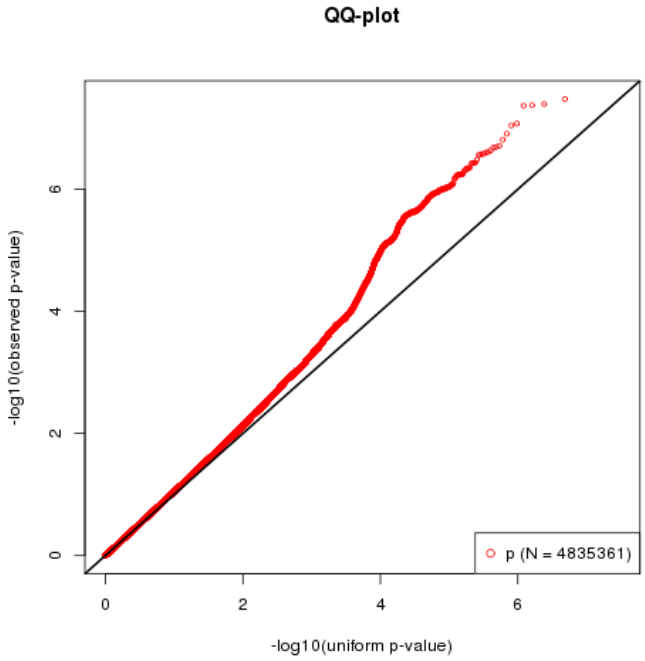


Figure 83 : QQ-Plot pour le Prolapsus Valvulaire Mitral - Barlow
Les p-values observées (rouge - ici transformées en log de base 10) sont comparés à leur distribution sous l'hypothèse nulle. On observe ici une séparation entre les deux courbes à un niveau assez bas, qui témoigne d'une légère inflation de la statistique – qui peut être dû à des différences démographiques résiduelles mais également à une différence de qualité de l'ADN. On n'observe pas ici de décollement.

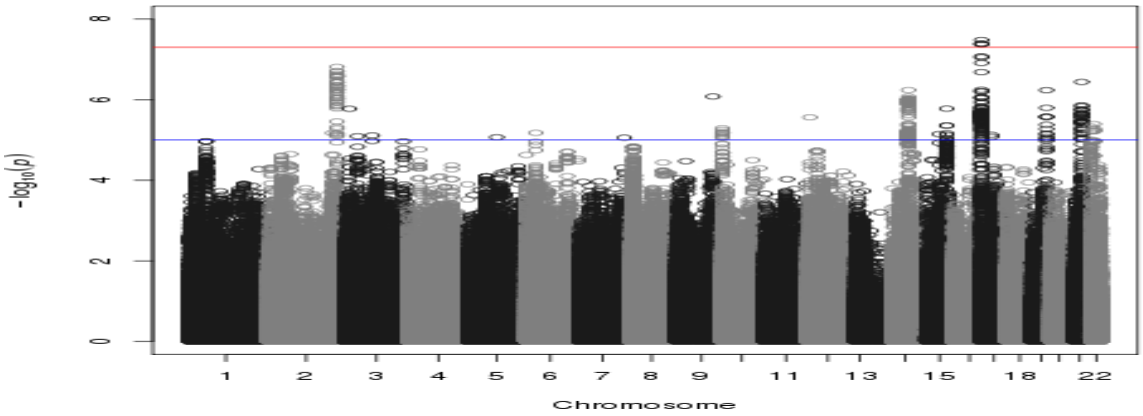


Figure 84 : Manhattan Plot pour le Prolapsus Valvulaire Mitral - Barlow
Ce Manhattan Plot (D.VI.1) montre les résultats, sous forme de p-valeur, en fonction de la position des SNPs sur le génome. La ligne rouge correspond à une p-valeur de 5.10^{-8} et la ligne bleue à une p-valeur de 10^{-6}. Sur le chromosome 17 on observe une association qui dépasse le seuil rouge de 5.10^{-8}. C'est le même signal que celui observé dans l'analyse PVM globale (Figure 80).

H.II.4.g.4) Gènes impliqués dans le PVM

Les résultats ne sont pas forcément très informatifs en termes de gènes. L'association la plus forte pointe sur une zone où se trouve le gène *TNS1* (Tableau 18). La protéine codée par ce gène est localisée au niveau des adhésions focales, à l'endroit où la membrane de la cellule s'attache à la matrice extracellulaire. C'est un résultat très intéressant car il est tout à fait possible que le Prolapsus Valvulaire Mitral résulte du défaut du développement puis du remodelage de la matrice extracellulaire par les cellulaires interstitielles des valves.

chr	start	end	p	Gène	Gènes à moins de 1MB
2	217394403	218394403	1.14e-08	NA	DIRC3 IGFBP2 IGFBP5 TNP1 TNS1
17	1692839	2692839	3.13e-08	SMG6	SMG6 SMYD4 TSR1 SRR METTL16 RPA1 MIR212 HIC1 RTN4RL1 DPH1 MIR132 OVCA2 LOC284009 SNORD91B SNORD91A SGSM2 MNT PAFAH1B1 KIAA0664 MIR1253
22	27706912	28706912	4.40e-08	NA	PITPNB MN1 TTC28 TTC28-AS1 MIR3199-2 MIR3199-1
3	13806607	14806607	2.35e-07	NA	GRIP2 WNT7A CHCHD4 TPRXL C3orf20 LOC100132526 LSM3 XPC TMEM43 SLC6A6 C3orf19
1	163794495	164794495	9.84e-07	NA	PBX1 LOC100505795
4	138673327	139673327	2.96e-06	NA	SLC7A11-AS1 LINC00499 LINC00616 SLC7A11
8	7675136	8675136	2.99e-06	NA	FAM66E MFHAS1 DEFB105B DEFB105A DEFB106A DEFB104A SPAG11A SPAG11B DEFB106B DEFB104B DEFB103A DEFB4A USP17L8 LOC100132396 DEFB103B USP17L3 DEFB109P1B MIR548I3 FLJ10661 SGK223 CLDN23
2	8036278	9036278	3.12e-06	NA	LINC00299 KIDINS220 MBOAT2 LOC339788 ID2
1	172291512	173291512	3.16e-06	NA	DNM3 TNFSF18 LOC100506023 PIGC C1orf105 SUCO FASLG TNFSF4
14	71252344	72252344	3.18e-06	NA	SIPA1L1 PCNX MAP3K9 SNORD56B LOC145474
3	8108920	9108920	5.70e-06	LMCD1	LOC100288428 SRGAP3 CAV3 LMCD1 LINC00312 C3orf32 OXTR RAD18
23	85736938	86736938	5.72e-06	NA	DACH2
16	13513666	14513666	6.00e-06	NA	MKL2 ERCC4 MIR365A MIR193B
22	26257994	27257994	6.59e-06	SEZ6L	MYO18B SEZ6L HPS4 CRYBB1 ASPHD2 MIR548J SRRD TPST2 TFIP11 MIAT CRYBA4
15	77330306	78330306	6.63e-06	NA	PEAK1 HMG20A TSPAN3 LINC00597 LINGO1 LOC253044 LOC645752 TBC1D2B LOC91450
6	131897902	132897902	7.26e-06	NA	ENPP3 MOXD1 ARG1 MED23 CTAGE9 OR2A4 ENPP1 CTGF LOC100507254 TAAR6 TAAR8 TAAR9 STX7
21	36960271	37960271	7.75e-06	LOC100133286	MORC3 DOPEY2 LOC100133286 CLDN14 MIR802 SETD4 CBR1 CBR3 CBR3-AS1 CHAF1B
6	22101548	23101548	8.93e-06	NA	LINC00340 LOC729177 PRL HDGFL1
1	219764913	220764913	8.99e-06	RNU5F-1	EPRS RNU5F-1 RAB3GAP2 MARK1 SLC30A10 IARS2 MIR194-1 MIR215 BPNT1 MIR664 SNORA36B AURKAPS1
6	43994470	44994470	9.45e-06	NA	SUPT3H TMEM63B CAPN11 MRPL14 SLC35B2 SLC29A1 HSP90AB1 TMEM151B AARS2 MIR4647 NFKBIE TCTE1 SPATS1 MIR4642 CDC5L
1	29618576	30618576	9.51e-06	A	

Tableau 18 : Liste des plus fortes associations à l'échelle du génome

Le choix des SNPs à répliquer est plus compliqué, comparativement à mon étude du projet 2 (Syndrome de Brugada). En effet, les analyses d'imputation peuvent donner des résultats peu conservateurs (optimistes) pour des SNPs qui ne sont pas suffisamment bien imputés.
Nous avons donc mis en place une procédure par laquelle nous vérifions que l'association repose sur au moins un SNP génotypé dans une des deux études, qui montre une association au minimum égale à 10^{-3}x p (p étant la p-valeur du SNP le plus associé dans une région).

H.II.4.h) *Réplication (1ere phase)*

H.II.4.h.1) Description de la cohorte Framingham :

L'étude de Framingham a été mise en place en 1948, permettant de suivre la santé cardiaque de la population d'une petite ville américaine et de corréler différents paramètres

avec le risque de survenue d'une maladie cardio-vasculaire. Cette cohorte mise en place afin d'identifier des facteurs de risque épidémiologiques a également été utilisée très rapidement dans des analyses génétiques [Govindaraju et al.,**2008**].

H.II.4.h.2) Résultats :

Nous n'observons pas de résultat significatif si on corrige par le nombre de tests. Toutefois plusieurs choses sont extrêmement intéressantes et encourageantes. D'abord, j'observe que, sur 30 tests statistiques indépendants, 22 ont un effet dans la même direction entre GWAs et Framingham. La probabilité d'observer 22 effets ou plus en sachant que l'on en attend 15 est de 0.008 (loi binomiale). J'effectue ici le même type de calcul que ce qui a été précédemment fait dans l'analyse du Diabète de Type 2 [Morris et al.,**2012**]. Dans le même ordre d'idées, nous observons une corrélation entre les effets (sous la forme de coefficient de régression beta dans la régression logistique). Cette corrélation est significative (p=0.01).
Il apparait ainsi que la distribution est significativement différente de ce qui est attendu sous l'hypothèse nulle, avec un excès de petites p-valeurs.

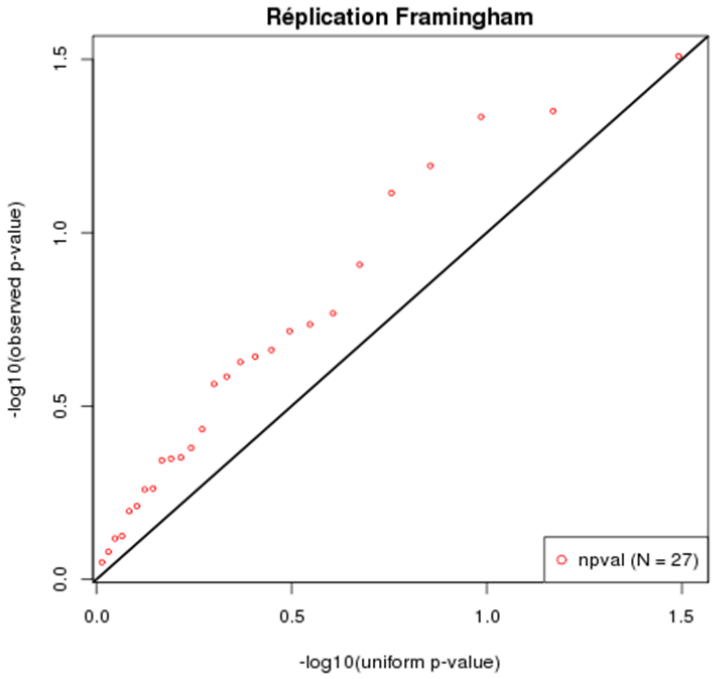


Figure 85 : QQ-plot pour les SNPs testés dans Framingham
QQ-plot pour un sous ensemble de SNPs indépendants (un par locus) dans la population de Framingham. Les valeurs observées s'éloignent de la distribution attendue sous l'hypothèse nulle.

J'en conclus que nous avons une bonne possibilité de répliquer une partie des associations observées – surtout celles qui apparaissent significatives à l'échelle du génome en phase 1 - dans les mois qui viennent.
Plus particulièrement, les trois associations les plus fortes pour le phénotype PVM semblent très prometteuses car elles sont parmi les plus fortes observées dans cette phase modeste de réplication (Tableau 19).

chr	P-valeur initiale	P-valeur Framingham
2	1.687301e-08	0.009031498
2	1.145784e-08	0.044567255
2	6.333854e-07	0.006275249
17	3.132746e-08	0.046333000
17	3.172415e-08	0.048043148
22	3.203484e-06	0.307619501
22	4.400178e-08	0.170792403
3	4.034121e-04	0.065132561
3	2.355571e-07	0.076830472

Tableau 19 : Résultats de l'analyse Framingham pour les 3 régions très associées en phase 1
Les p-valeurs sont estimées dans l'analyse cas-témoins sur la population de Framingham par un test d'association unilatéral. Une petite p-valeur signifie donc qu'il y a une (modeste) association dans la même direction que l'association initiale.

H.II.5. Synthèse des résultats PVM

Cette étude a été plus compliquée à mener à bien des égards. D'abord, la nécessité de contrôles de qualités beaucoup plus poussés s'est fait sentir. En effet, des marqueurs de mauvaise qualité risquent d'entraîner des mauvaises imputations et des faux positifs.
Ensuite, il apparaît clairement que les bases génétiques du Prolapsus Valvulaire Mitral sont proches de celles d'autres maladies complexes, sans polymorphisme génétique montrant de Risques Relatifs supérieurs à 1.7. Il ne semble pas qu'il y ait même d'équivalent de marqueurs tels que ceux de *FTO* dans l'obésité [Dina et al.,**2007**][Frayling et al.,**2007**], ou *TCF7L2* dans le diabète de Type 2 [Grant et al.,**2006**].

Toutefois, il semble que nous soyons en mesure d'identifier des associations significatives qui permettront d'avancer dans la connaissance de la physiologie de cette maladie.

I. Discussion et conclusion

I.I. Rappel de l'état des connaissances et sur les hypothèses avant cette thèse

Les analyses d'association « génome entier » récentes ont permis de mettre en évidence un très grand nombre d'associations entre des variants génétiques (souvent fréquents) et des traits phénotypiques (incluant des pathologies). Dans ce sens, nous pouvons considérer qu'il s'agit d'un succès. Toutefois, la plupart de ces marqueurs génétiques ne se situent pas dans des phases ouvertes de lecture, sont parfois introniques et la plupart du temps intergéniques. Un exemple frappant est l'association entre le SNP rs17782313 et *MC4R* [Loos et al.,**2008**]. Il en résulte une difficulté à identifier le gène d'intérêt et donc la voie biosynthétique impliquée dans la survenue de la pathologie.
Les effets individuels de ces marqueurs sont par ailleurs assez faibles et, par conséquent, la proportion de l'héritabilité expliquée par les variants fréquents (issus des GWAs), pour ces pathologies, était très petite. Comme l'effet des variants qu'on pourra trouver en augmentant la taille des échantillons est de plus en plus petit, il était légitime de faire l'hypothèse que l'héritabilité globale expliquée par ces variants atteindrait un seuil qui serait assez bas [Dina,**2008**].
Ces observations ont amené beaucoup de chercheurs à remettre en question l'importance des résultats des analyses d'association génome-entier dans la compréhension des pathologies et de l'architecture génétique de ces pathologies [McClellan, King,**2010**][Zuk et al.,**2012**][Manolio et al.,**2009**].
Un temps, l'importance potentielle des CNVs a également été mise en avant. Il semblerait toutefois, que les CNVs fréquents n'expliquent également qu'une faible proportion [Craddock et al.,**2010**] des pathologies humaines. Cette observation était assez prévisible dans la mesure où il semble bien que les CNVs fréquents étaient, comme tout autre marqueur, en déséquilibre de liaison avec des SNPs fréquents. Ainsi, ils ne présentaient pas une information complètement indépendante des SNPs déjà interrogés dans les GWAs.
L'hypothèse d'un impact très fort des variants rares dans certaines pathologies complexes s'est alors imposée avec force [McClellan, King,**2010**][Gibson,**2011**].
Quoi qu'il en soit, l'hypothèse d'héritabilité manquante (*missing heritability* et/ou *dark matter*) était très prégnante.

Les travaux de thèse auxquels j'ai participé ont permis d'évoluer dans notre compréhension des bases moléculaires de ces pathologies (et sans doute des pathologies en général). D'abord nous avons mis en évidence une nouvelle notion, celle de l'effet de variants fréquents dans des pathologies rares à travers l'étude du Syndrome de Brugada. Par ailleurs, en approfondissant la recherche des bases génétiques du diabète de type 2, nous avons abouti à la conclusion que la théorie du variant commun / maladie commune, même si elle n'était pas vérifiée dans son intégralité, correspondait à un modèle réel.
En effet, il semble qu'il existe un grand nombre de variants fréquents ayant un effet, quelque petit qu'il soit sur la survenue du diabète de Type 2. Par contre, il apparaît maintenant certain que la plupart de ces variants génétiques n'auront pas d'effet individuel fort. L'enjeu principal, en dehors de la mise en évidence des bases physiologiques à travers l'identification de ces variants reste également d'évaluer la proportion de l'héritabilité restant à expliquer et donc la place qu'il reste dans ce modèle pour les variants plus rares (MAF < 5%). Cela est important dans la mesure où les projets actuels, basés sur l'explosion des méthodes de séquençage à haut débit nouvelle génération, s'orientent vers la recherche de ce type de variants dans l'étiologie des maladies complexes [McClellan, King,**2010**].

I.II. Pathologie Rare, Variant Fréquent :

Le résultat initial permet de mettre en évidence l'effet de trois variants fréquents dans la survenue d'une pathologie considérée comme rare : le syndrome de Brugada. Par ailleurs, nous avons démontré qu'il y a une forte proportion de l'héritabilité qui est expliquée par ces

trois variants. Les effets sont toutefois insuffisants pour expliquer la concentration familiale qui semble être observée dans le syndrome de Brugada.
La proportion du risque due à des facteurs génétiques fréquents est inattendue, même si elle a pu déjà être observée dans certains cas. Le risque d'être atteint est 39 fois plus grand pour les individus (1 individu sur 1000) portant les 6 allèles (2 allèle à risque pour chaque SNP) délétères comparativement à ceux n'en portant aucun (5 individus sur 100), dans une population Européenne.
Il est étonnant qu'un allèle dont l'effet est assez grave pour la santé humaine puisse atteindre une fréquence aussi haute dans la population humaine. Le Syndrome de Brugada est étudié du fait de l'augmentation de risque de Mort Subite cardiaque inexpliquée qu'il semble induire – ou dont il semble témoigner. Or ce phénotype n'est pas un phénotype à survenue tardive – il touche tous les âges y compris les âges de reproduction humaine.
Une analyse *in silico* des allèles rares portés par l'haplotype à risque ne semble pas démontrer d'accumulation spécifique de variants rares dans certains gènes. Cette voie doit être cependant approfondie – toujours par analyse *in-silico*.
A l'inverse, on peut se demander si les variants qui augmentent le risque de survenue du Syndrome de Brugada (et donc de la mort subite) n'ont pas un effet très favorable à la survie par ailleurs. On peut imaginer une configuration où les caractéristiques cardiaques induites par la combinaison génétique à risque pour le Syndrome de Brugada confèrent un avantage – par exemple dans les performances physiques utiles à des activités telles que la chasse. Le « prix » à payer pour un avantage moyen dans la population serait un risque de Mort Subite augmenté au niveau individuel. Alternativement, il est possible que le variant augmentant le risque de SBr soit en très grand déséquilibre de liaison avec un autre variant ayant un avantage sélectif suffisante pour en contrebalancer l'effet délétère. La recherche de telles signatures, soit d'une sélection équilibrée ou d'un scenario évolutif de *hitch-hiking*, c'est-à-dire présence sur le même haplotype d'un autre allèle ayant un avantage sélectif, sera une des prochaines étapes de nos études.
Quoi qu'il en soit, la notion de variant fréquent pour une pathologie rare est relativement nouvelle et n'a pas encore, à ma connaissance, été rapportée dans la littérature.
Alternativement, il est tout à fait possible que cette découverte mette en évidence l'importance d'un phénotype homogène. De façon intéressante, l'article qui lançait la notion d'héritabilité manquante [Maher,**2008**] a également proposé l'hypothèse que les pathologies fréquentes seraient plutôt l'addition d'une constellation de maladies rares présentant des symptômes proches et abusivement classées sous la même étiquette. Est-ce que notre analyse sur le Syndrome de Brugada n'est pas une manifestation de cette architecture génétique et phénotypique, le SBr n'étant qu'un phénotype homogène classé dans un ensemble plus vaste que sont l'arythmie et la Mort Subite. Dans ce cadre, il serait intéressant de mener des études d'association génome entier sur d'autres pathologies aussi homogènes comme le Syndrome du QT long.

I.III. Composante polygénique du Diabète de Type 2 :

Lors de mon parcours, j'ai participé à la première étude d'analyse génome entier pour le diabète de Type 2, puis à une des méta-analyses les plus importantes (présentée dans ce travail comme papier issu de la thèse) et enfin, de façon moins impliquée, à l'analyse de type Metabo-chip. L'idée que nous avions de l'architecture génétique du Diabète de Type 2 a peu à peu évolué au cours de ces trois travaux. Dans la première analyse nous avons trouvé un nombre d'associations significatives peut-être moins important que ce qui était prévu. L'article de méta-analyse (dit DIAGRAM+) a augmenté le nombre d'associations, les nouveaux variants génétiques ayant chacun un effet de moins en moins fort. Cette analyse a aidé à la réalisation de la puce Metabo-Chip – sélection des SNPs à répliquer. À partir de cette puce nous avons pu mettre en évidence, dans le papier suivant [Morris et al.,**2012**] l'importance d'une composante polygénique non négligeable.
Nous nous orientons de plus en plus vers une architecture où un grand nombre de variants communs à effet très faible expliquent une proportion non négligeable de l'héritabilité. Cette

hypothèse n'est pas complètement nouvelle, elle a déjà été mise en avant par Shaun Purcell et son équipe dans l'analyse d'association de la schizophrénie [Purcell et al.,**2009**].
Les conséquences de cette observation pour la stratégie de la recherche génétique dans les maladies complexes sont diverses. D'abord, on peut se demander si, dans la pratique médicale quotidienne, il est intéressant de connaître les gènes dont les variations ont des effets si petits. Cette composante polygénique est difficilement réductible à un facteur simple de prédiction et encore moins de soins. On peut imaginer des tests multi-SNPs sur cette composante polygénique tel que ceux proposés dans [Wei et al.,**2009**], qui se basent sur des techniques de fouille de données, par exemple, pour créer la combinaison des marqueurs optimale permettant une bonne prédictibilité.
Une connaissance plus précise des voies biologiques responsables, non seulement du maintien de l'homéostasie du glucose (d'autres voies peuvent participer à la survenue de cette pathologie) mais de la différence entre les individus dans leur susceptibilité au diabète de Type 2, peut aider à orienter les stratégies de soins et, éventuellement, d'élaboration de médicaments.

I.IV. Qu'est-ce qu'un gène ?

Lorsque des équipes de génétique veulent communiquer de façon spectaculaire (et c'est important à notre époque), il n'est pas rare de voir des titres comme celui-ci : « Des chercheurs de l'Université de Nantes ont trouvé le gène de la Mort Subite ». Même si les chercheurs sont conscients du degré de simplification de cette affirmation, il n'en est pas moins qu'ils lancent souvent des études afin de trouver les gènes responsables d'une pathologie donnée.
Toutefois, des études comme celle menées sur le Diabète de Type 2 mais également sur des phénotypes quantitatifs comme la taille [Lango Allen et al.,**2010**] concluent à l'influence d'un grand nombre de variants (690 pour l'étude sur la taille). On peut bien sûr faire l'hypothèse que tous ces variants, répartis sur tout le génome, n'interviennent que dans la régulation d'un seul gène (ou d'une poignée de gènes). Cela semble peu probable même si la régulation *trans* semble plus importante que ce qui était précédemment supposé [Elbein et al.,**2012**].
De cela, on peut en conclure que le nombre de gènes ayant une influence réelle sur une pathologie (ou un phénotype) est très grand. Si une telle observation s'applique à d'autres pathologies, cela signifie, fatalement qu'un nombre non négligeable de gènes sont impliqués dans plusieurs pathologies.
Cette conception n'est en rien nouvelle puisqu'on a déjà formulé l'hypothèse d'effet pléïotrope depuis longtemps. Récemment, l'idée des phenomeScans [Jones et al.,**2005**] est apparue, qui propose d'étudier pour un SNP tous les phénotypes possibles – de façon simultanée ou non.
De telles études, selon leurs résultats, pourraient faire évoluer notre perception de l'effet des gènes pour passer des gènes responsable d'une maladie (d'un phénotype) aux gènes responsables des différences interindividuelles prises dans un sens large. Les résultats que nous avons obtenus dans l'article [Voight et al.,**2010**] et que d'autres ont obtenu semblent aller dans le même sens.

Inversement, l'étude réalisée dans ce travail de thèse sur le Syndrome de Brugada revigore l'idée de gène majeur. En effet, il apparaît que le gène *SCN5A*, mis en évidence par des analyses de type mono-génique, avec un effet très fort de variants rares, soit également impliqué par sa régulation. L'association de l'haplotype fréquent du SNP qui est localisé dans le gène *SCN10A* semble pointer sur un locus ayant un effet sur l'expression de *SCN5A*. A cela s'ajoute le polymorphisme localisé dans le gène lui-même, qui est indépendant de la précédente. Les trois facteurs les plus forts de la composante génétique du Syndrome de Brugada sont attribuables à ce gène canal sodique.

Par contre, il semble que la deuxième partie de la composante génétique soit plutôt le fait de gènes impliqués dans le développement cardiaque, à commencer par *HEY2*, mais également un ensemble d'autres gènes qui semblent interagir entre eux (G.III.9.b).

I.V. Le futur des génotypes

Le projet de recherche de variants fréquents pour expliquer la variabilité phénotypique humaine et le grand nombre d'études d'association génome entier qui en sont la conséquence nous met dans une situation où des dizaines de milliers d'individus ont été génotypés, à un prix relativement important.
La prise en compte de l'environnement est, peut-être, indispensable mais elle s'avère très compliquée. Si le test systématique de tous les éléments du génome semble faisable, à un certain prix, la détermination tout aussi systématique des variables environnementales semble hors de portée, tant la notion d'environnement est multiforme et infinie. S'il faut (pour l'instant) renoncer à réaliser des EWAS (« Environment Wide Association Studies »), il existe d'autres approches qui peuvent permettre d'essayer de mieux appréhender cette dimension dans l'analyse génétique. Je pense principalement à l'analyse de la différence des effets génétiques dans le temps et l'espace. Dans ce contexte, la récente étude démontrant la variation d'héritabilité chez des paires de jumeaux en fonction de la localisation géographique de ces données est très intéressante [Davis et al.,**2012**]. Les auteurs concluent à un effet important de l'environnement dans la modulation des effets génétiques. Nous pouvons imaginer nous intéresser, dans le cadre des Consortiums d'études génétiques de maladies fréquentes (comme DIAGRAM+), à la variation d'association entre études pour un SNP donné – et non plus à la significativité globale. A l'intérieur des études, il sera également possible de tester l'interaction entre chaque SNP et un marqueur de la géolocalisation d'un individu (latitude et longitude, 1^{ere} et 2eme composantes, localisation géographique discrète).
Il est également possible d'explorer la dimension du Temps (après celle de l'Espace). A ce titre, les comparaisons des résultats d'association pour des cohortes recrutées à différentes époques peut-être riche d'enseignement. Des études comme celle de Framingham, qui s'étendent sur plusieurs générations peuvent donner matière à une recherche des interactions gène-environnement très fructueuses. Par exemple, il est possible et même très probable que certains variants génétiques n'aient d'effet que dans un environnement précis. Or les paramètres de l'environnement sont multiples et très difficiles à contrôler. Par contre, différentes générations connaissent un environnement qui a tendance à être globalement différent (le cas de l'accès à la nourriture est l'exemple le plus parlant).

I.VI. Conclusion

Est-ce que les chercheurs de l'époque Victorienne étaient plus doués que nous pour prédire un phénotype en fonction de données génétiques ? L'équipe de Yurii Aulchenko soulignait avec raison que la capacité prédictive du modèle de Galton, en 1886, se basant simplement sur la taille des parents, dépassait largement celle de notre modèle actuel basé sur 54 SNPs et leurs effets, pourtant issus des toutes dernières découvertes [Aulchenko et al.,**2009**]. La technologie, aussi bien matérielle, qui fait appel à des appareils de plus en plus sophistiqués, qu'intellectuelle, qui fait appel à des modèles de plus en plus avancés, a-t-elle été réellement mise au service d'une avancée de notre connaissance des bases génétiques et biologiques des pathologies humaines ?
Les différents projets de ma thèse concouraient à cette tentative de mettre à jour ce qu'on appelle l'architecture génétique de traits complexes afin d'essayer d'en mieux comprendre les mécanismes. Il s'agissait également de proposer des modèles génétiques permettant de mieux prédire le risque de survenue de pathologies graves. Au cours de ce travail, j'ai été amené à étudier un spectre assez large de situations. Les études génétiques sur le Diabète

de Type 2 sont anciennes et les données assemblées pour ces analyses sont impressionnantes. En ce qui concerne le Prolapsus Valvulaire Mitral, nous ne sommes qu'au début de ce programme de recherche. Enfin, pour le Syndrome de Brugada (et la Mort Subite cardiaque), si la recherche de variants génétiques fréquents en était également à ses début, les connaissances sur les processus d'action de variants rares intra-géniques était assez avancées. C'est dans ce contexte qu'on a pu dire que l'objet de ma thèse était de trouver des polymorphismes modificateurs des mutations présentes dans le gène *SCN5A*.
De ces trois études, il m'est apparu que les études d'association génome entier ont apporté une information nouvelle et importante pour la mise en évidence des mécanismes physiopathologiques, avec une prépondérance des défauts de production de l'Insuline par les cellules Bêta pour le DT2 et l'importance – confirmée – du canal sodique mais également des facteurs de développement embryonnaire pour le SBr.
Dans ce dernier cas, nous avons été assez surpris de mettre à jour une nouvelle architecture, qui est celle de polymorphismes génétiques fréquents pour des maladies rares.
Au vu de ces résultats et des projets en cours dans le domaine de la génétique des pathologies complexes, il me semble que nous pourrons améliorer nos capacités prédictives et surtout de compréhension de ces pathologies. Cela sera possible aussi bien en utilisant de nouvelles technologies (séquençage de nouvelle génération) que grâce au matériel déjà accumulé en termes de génotypes. Je pense donc qu'un généticien ou un mathématicien de l'époque Victorienne miraculeusement transporté à notre époque aurait l'impression que notre connaissance a progressé de façon significative depuis l'époque qu'il a quitté.

Bibliographie

F Galton (**1886**) Regression towards mediocrity in hereditary stature 246‑263.

Jean Bouyer, Denis Hémon, Sylvaine Cordier et al. Livre - Epidemiologie principes et methodes quantitatives -

WS Sutton (**1902**) On the Morphology of the Chromoso Group in Brachystola Magna 1,24‑39.

OT Avery et al. (**1944**) 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 2,137‑158.

NM Stevens (**1905**) Studies in spermatogenesis

EB Wilson (**1905**) THE CHROMOSOMES IN RELATION TO THE DETERMINATION OF SEX IN INSECTS 564,500‑502.

JD WATSON, FH CRICK (**1953**) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid 4356,737‑738.

DN Cooper, M Krawczak (**1993**) Human Gene Mutation

MW Nachman, SL Crowell (**2000**) Estimate of the Mutation Rate per Nucleotide in Humans 1,297‑304.

A Kong et al. (**2012**) Rate of de novo mutations and the importance of father’s age to disease risk 7412,471‑475.

1000 Genomes Project Consortium et al. (**2010**) A map of human genome variation from population-scale sequencing 7319,1061‑1073.

ES Lander et al. (**2001**) Initial sequencing and analysis of the human genome 6822,860‑921.

JC Venter et al. (**2001**) The sequence of the human genome 5507,1304‑1351.

M Przeworski et al. (**2000**) Adjusting the focus on human variation 7,296‑302.

GH Hardy (**1908**) Mendelian Proportions in a Mixed Population 706,49‑50.

W Weinberg (**1908**) Über den Nachweis der Vererbung beim Menschen 368‑382.

W-C Lee (**2003**) Searching for Disease-Susceptibility Loci by Testing for Hardy-Weinberg Disequilibrium in a Gene Bank of Affected Individuals 5,397‑400.

JK Wittke-Thompson et al. (**2005**) Rational Inferences about Departures from Hardy-Weinberg Equilibrium 6,967‑986.

K Song, RC Elston (**2006**) A powerful method of combining measures of association and Hardy–Weinberg disequilibrium for fine-mapping in case-control studies 1,105–126.

S Lukić, J Hey (**2012**) Demographic Inference Using Spectral Methods on SNP Data, With an Analysis of the Human out-of-Africa Expansion

AJ Jeffreys et al. (**2001**) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex 2,217-222.

N Arnheim et al. (**2003**) Hot and Cold Spots of Recombination in the Human Genome: the Reason We Should Find Them and How This Can Be Achieved 1,5-16.

KW Broman et al. (**1998**) Comprehensive human genetic maps: individual and sex-specific variation in recombination. 3,861-869.

VG Cheung et al. (**2007**) Polymorphic variation in human meiotic recombination 3,526-530.

G Coop et al. (**2008**) High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans 5868,1395-1398.

A Kong et al. (**2008**) Sequence Variants in the RNF212 Gene Associate with Genome-Wide Recombination Rate 5868,1398-1401.

R Chowdhury et al. (**2009**) Genetic Analysis of Variation in Human Meiotic Recombination 9,e1000648.

RC Lewontin (**1964**) The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models 1,49-67.

MJ Daly et al. (**2001**) High-resolution haplotype structure in the human genome 2,229-232.

JW James (**1971**) Frequency in relatives for an all-or-none trait 1,47–49.

N Risch (**1990**) Linkage strategies for genetically complex traits. I. Multilocus models. 2,222.

N Risch, K Merikangas (**1996**) The future of genetic studies of complex human diseases 5281,1516-1517.

BA Rybicki, RC Elston (**2000**) The Relationship between the Sibling Recurrence-Risk Ratio and Genotype Relative Risk 2,593-604.

NR Wray et al. (**2010**) The genetic interpretation of area under the ROC curve in genomic profiling 2,e1000864.

T Reich et al. (**1972**) The use of multiple thresholds in determining the mode of transmission of semi-continuous traits 2,163-184.

F Bernstein (**1931**) Zur Grundlegung der Chromosomentheorie der Vererbung beim Menschen 1,113-138.

JBS Haldane (**1936**) A SEARCH FOR INCOMPLETE SEX-LINKAGE IN MAN 1,28–57.

NE MORTON (**1958**) Segregation analysis in human genetics 3289,79-80.

F Clerget-Darpoux et al. (**1986**) Effects of misspecifying genetic parameters in lod score analysis 2,393‑399.

N Risch (**1984**) Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. 2,363.

LS Penrose (**1935**) THE DETECTION OF AUTOSOMAL LINKAGE IN DATA WHICH CONSIST OF PAIRS OF BROTHERS AND SISTERS OF UNSPECIFIED PARENTAGE 2,133–138.

JK Haseman, RC Elston (**1972**) The investigation of linkage between a quantitative trait and a marker locus 1,3‑19.

P Holmans (**1993**) Asymptotic properties of affected-sib-pair linkage analysis 2,362‑374.

PM Fishman et al. (**1978**) A robust method for the detection of linkage in familial disease. 3,308.

HHH Göring, JD Terwilliger (**2000**) Linkage Analysis in the Presence of Errors IV: Joint Pseudomarker Analysis of Linkage and/or Linkage Disequilibrium on a Mixture of Pedigrees and Singletons When the Mode of Inheritance Cannot Be Accurately Specified 4,1310‑1327.

BC Martin et al. (**1992**) Familial clustering of insulin sensitivity 7,850‑854.

DE Weeks, K Lange (**1988**) The affected-pedigree-member method of linkage analysis. 2,315.

S Purcell et al. (**2007**) PLINK: a tool set for whole-genome association and population-based linkage analyses 3,559‑575.

A Gusev et al. (**2011**) DASH: A Method for Identical-by-Descent Haplotype Mapping Uncovers Association with Recent Variation 6,706‑717.

SR Browning, EA Thompson (**2012**) Detecting rare variant associations by identity-by-descent mapping in case-control studies 4,1521‑1531.

DJ Balding (**2006**) A tutorial on statistical methods for population association studies 10,781‑791.

KG Ardlie et al. (**2002**) Patterns of linkage disequilibrium in the human genome 299‑309.

KA Frazer et al. (**2007**) A second generation human haplotype map of over 3.1 million SNPs 7164,851‑861.

The International HapMap Consortium (**2005**) A haplotype map of the human genome 7063,1299‑1320.

GR Abecasis et al. (**2001**) Extent and Distribution of Linkage Disequilibrium in Three Genomic Regions 1,191‑197.

SA Tishkoff, BC Verrelli (**2003**) Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping 6,569‑575.

A Chakravarti et al. (**1984**) Nonuniform recombination within the human beta-globin gene cluster 6,1239‑1258.

LR Cardon, GR Abecasis (**2003**) Using haplotype blocks to map human complex trait loci 3,135‑140.

International Human Genome Sequencing Consortium (**2004**) Finishing the euchromatic sequence of the human genome 7011,931‑945.

The International HapMap Consortium (**2005**) A haplotype map of the human genome 7063,1299‑1320.

International HapMap Consortium et al. (**2007**) A second generation human haplotype map of over 3.1 million SNPs 7164,851‑861.

International HapMap 3 Consortium et al. (**2010**) Integrating common and rare genetic variation in diverse human populations 7311,52‑58.

Y Li et al. (**2009**) Genotype Imputation 1,387‑406.

JC Mueller et al. (**2005**) Linkage disequilibrium patterns and tagSNP transferability among European populations 3,387‑398.

PIW de Bakker et al. (**2006**) Transferability of tag SNPs in genetic association studies in multiple populations 11,1298‑1303.

JD Terwilliger, T Hiekkalinna (**2006**) An utter refutation of the 'Fundamental Theorem of the HapMap' 4,426‑437.

B Howie et al. (**2011**) Genotype imputation with thousands of genomes 6,457‑470.

J Marchini et al. (**2007**) A new multipoint method for genome-wide association studies by imputation of genotypes 7,906‑913.

SR Browning, BL Browning (**2007**) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering 5,1084‑1097.

Y Li et al. (**2010**) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes 8,816‑834.

B Howie et al. (**2012**) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing 8,955‑959.

O Delaneau et al. (**2012**) A linear complexity phasing method for thousands of genomes 2,179‑181.

SA Stouffer et al. (**1949**) The American soldier: adjustment during army life. (Studies in social psychology in World War II, Vol. 1.)

B Han, E Eskin (**2011**) Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies 5,586‑598.

Affymetrix Analysis Guide Axiom^{TM} Genotyping Solution Data Analysis Guide

F Dudbridge, A Gusnanto (**2008**) Estimation of significance thresholds for genomewide association scans 3,227‐234.

I Pe'er et al. (**2008**) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants 4,381‐385.

R Sladek et al. (**2007**) A genome-wide association study identifies novel risk loci for type 2 diabetes 7130,881‐885.

F Dudbridge (**2006**) A note on permutation tests in multistage association scans 6,1094‐1095; author reply 1096.

Y Benjamini, Y Hochberg (**1995**) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing 289‐300.

Y Benjamini (**2001**) The control of the false discovery rate in multiple testing under dependency 4,1165‐1188.

Y Benjamini (**2001**) The control of the false discovery rate in multiple testing under dependency 4,1165‐1188.

JT Leek, JD Storey (**2008**) A general framework for multiple testing dependence 48,18718‐18723.

C Friguet, D Causeur (**2011**) Estimation of the proportion of true null hypotheses in high-dimensional data under dependence 9,2665‐2676.

Y Blum et al. (**2010**) A factor model to analyze heterogeneity in gene expression 368.

P Menozzi et al. (**1978**) Synthetic maps of human gene frequencies in Europeans 4358,786‐792.

AE Mourant (**1954**) The Distribution of the Human Blood Groups

AL Price et al. (**2006**) Principal components analysis corrects for stratification in genome-wide association studies 8,904‐909.

HM Kang et al. (**2010**) Variance component model to account for sample structure in genome-wide association studies 4,348‐354.

JF Gusella et al. (**1983**) A polymorphic DNA marker genetically linked to Huntington's disease 5940,234‐238.

M Macdonald (**1993**) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes 6,971‐983.

JR Riordan et al. (**1989**) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA 4922,1066‐1073.

JM Rommens et al. (**1989**) Identification of the cystic fibrosis gene: chromosome walking and jumping 4922,1059‐1065.

AP Monaco et al. (**1986**) Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene 6089,646‐650.

J Amberger et al. (**2009**) McKusick's Online Mendelian Inheritance in Man (OMIM) Database issue,D793-796.

ES Lander (**1996**) The New Genomics: Global Views of Biology 5287,536-539.

M KIMURA, JF CROW (**1964**) THE NUMBER OF ALLELES THAT CAN BE MAINTAINED IN A FINITE POPULATION 725-738.

M Kimura (**1969**) The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations 4,893-903.

A Chakravarti (**1999**) Population genetics—making sense out of sequence 56-60.

RJ Klein et al. (**2005**) Complement Factor H Polymorphism in Age-Related Macular Degeneration 5720,385-389.

AM Saunders et al. (**1993**) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease 8,1467-1472.

J Poirier et al. (**1993**) Apolipoprotein E polymorphism and Alzheimer's disease 8873,697-699.

EH Corder et al. (**1993**) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families 5123,921-923.

RJ Klein et al. (**2005**) Complement factor H polymorphism in age-related macular degeneration 5720,385-389.

JL Haines et al. (**2005**) Complement factor H variant increases the risk of age-related macular degeneration 5720,419-421.

AO Edwards et al. (**2005**) Complement factor H polymorphism and age-related macular degeneration 5720,421-424.

M Perola et al. (**2007**) Combined Genome Scans for Body Stature in 6,602 European Twins: Evidence for Common Caucasian Loci 6,e97.

PM Visscher et al. (**2006**) Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings 3,e41.

DF Gudbjartsson et al. (**2008**) Many sequence variants affecting diversity of adult human height 5,609-615.

G Lettre et al. (**2008**) Identification of ten loci associated with height highlights new biological pathways in human growth 5,584-591.

MN Weedon et al. (**2008**) Genome-wide association analysis identifies 20 loci that influence adult height 5,575-583.

B Maher (**2008**) Personal genomes: The case of the missing heritability 7218,18-21.

TA Manolio et al. (**2009**) Finding the missing heritability of complex diseases 7265,747-753.

SP Dickson et al. (**2010**) Rare Variants Create Synthetic Genome-Wide Associations 1,e1000294.

A Grimaldi (**2005**) Guide pratique du diabète

Langerhans, Paul (**1869**) Über den feineren Bau der Bauchspeicheldrüse

P Marchetti et al. (**2004**) Pancreatic Islets from Type 2 Diabetic Patients Have Functional Defects and Increased Apoptosis That Are Ameliorated by Metformin 11,5535-5541.

AE Butler et al. (**2003**) β-Cell Deficit and Increased β-Cell Apoptosis in Humans With Type 2 Diabetes 1,102-110.

SD Guerra et al. (**2005**) Functional and Molecular Defects of Pancreatic Islets in Human Type 2 Diabetes 3,727-735.

M Stumvoll et al. (**2008**) Type 2 diabetes: pathogenesis and treatment 9631,2153-2156.

M Stumvoll et al. (**2007**) Pathogenesis of type 2 diabetes 1-2,19-37.

M Stumvoll et al. (**2005**) Type 2 diabetes: principles of pathogenesis and therapy 9467,1333-1346.

O Kusnik-Joinville et al. (**2008**) Prevalence and treatment of diabetes in France: trends between 2000 and 2005 3,266-272.

C Bonaldi et al. (**2011**) A first national prevalence estimate of diagnosed and undiagnosed diabetes in France in 18- to 74-year-old individuals: the French Nutrition and Health Survey 2006/2007 5,583–589.

H Beck-Nielsen et al. (**2003**) Metabolic and genetic influence on glucose metabolism in type 2 diabetic subjects--experiences from relatives and twin studies 3,445-467.

JC Florez et al. (**2003**) The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits 257-291.

P Poulsen et al. (**1999**) Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study 2,139-145.

RL Hanson et al. (**2001**) Family and genetic studies of indices of insulin sensitivity and insulin secretion in Pima Indians 4,296-303.

GW Mills et al. (**2004**) Heritability estimates for beta cell function and features of the insulin resistance syndrome in UK families with an increased susceptibility to type 2 diabetes 4,732-738.

MA Permutt et al. (**2005**) Genetic epidemiology of diabetes 6,1431-1439.

SS Deeb et al. (**1998**) A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity 3,284-287.

D Altshuler et al. (**2000**) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes 1,76-80.

AL Gloyn et al. (**2003**) Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes 2,568‑572.

O Laukkanen et al. (**2004**) Polymorphisms of the SUR1 (ABCC8) and Kir6.2 (KCNJ11) genes predict the conversion from impaired glucose tolerance to type 2 diabetes. The Finnish Diabetes Prevention Study 12,6286‑6290.

E Zeggini et al. (**2007**) Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes 5829,1336‑1341.

V Steinthorsdottir et al. (**2007**) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes 6,770‑775.

LJ Scott et al. (**2007**) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants 5829,1341‑1345.

E Zeggini et al. (**2008**) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes 5,638‑645.

JRB Perry, TM Frayling (**2008**) New gene variants alter type 2 diabetes risk predominantly through reduced beta-cell function 4,371‑377.

BF Voight et al. (**2010**) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis 7,579‑589.

BF Voight et al. (**2012**) The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits 8,e1002793.

AP Morris et al. (**2012**) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes

DR Matthews et al. (**1985**) Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man 7,412‑419.

TM Wallace et al. (**2004**) Use and abuse of HOMA modeling 6,1487‑1495.

R Saxena et al. (**2010**) Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge 2,142‑148.

S Raychaudhuri (**2011**) Mapping Rare and Common Causal Alleles for Complex Human Diseases 1,57‑69.

PD Thomas et al. (**2003**) PANTHER: A Library of Protein Families and Subfamilies Indexed by Function 9,2129‑2141.

H Mi et al. (**2005**) The PANTHER database of protein families, subfamilies, functions and pathways Database issue,D284‑D288.

K Wang et al. (**2009**) Diverse Genome-wide Association Studies Associate the IL12/IL23 Pathway with Crohn Disease 3,399‑405.

LA Hindorff et al. (**2009**) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits 23,9362‑9367.

MR Boyett et al. (**1996**) A list of vertebrate cardiac ionic currents Nomenclature, properties, function and cloned equivalents 3,455‐481.

Jaakko Malmivuo, Robert Plonsey Bioelectromagnetism

X Jouven, D Escande (**2006**) Sudden cardiac death: toward the identification of susceptibity genes 806‐812.

DP Zipes, HJ Wellens (**1998**) Sudden cardiac death 2334‐2351.

HV Huikuri et al. (**2001**) Sudden death due to cardiac arrhythmias 1473‐1482.

C Napolitano, SG Priori (**2002**) Genetics of ventricular tachycardia 3,222‐228.

A Sarkozy, P Brugada (**2005**) Sudden cardiac death and inherited arrhythmia syndromes 8‐20.

HL Tan et al. (**2005**) Sudden unexplained death: heritability and diagnostic yield of cardiological and genetic examination in surviving relatives 207‐213.

R Roberts, R Brugada (**2003**) Genetics and arrhythmias 257‐267.

M Cerrone, SG Priori (**2011**) Genetics of sudden death: focus on inherited channelopathies 17,2109‐2118.

AL Pond, JM Nerbonne (**2001**) ERG Proteins and Functional Cardiac IKr Channels in Rat, Mouse, and Human Heart 7,286‐294.

C Napolitano et al. (**2012**) Sudden Cardiac Death and Genetic Ion Channelopathies Long QT, Brugada, Short QT, Catecholaminergic Polymorphic Ventricular Tachycardia, and Idiopathic Ventricular Fibrillation 16,2027‐2034.

SG Priori (**2010**) The Fifteen Years of Discoveries That Shaped Molecular Electrophysiology Time for Appraisal 4,451‐456.

DE Arking et al. (**2011**) Identification of a Sudden Cardiac Death Susceptibility Locus at 2q24.2 through Genome-Wide Association in European Ancestry Individuals 6,e1002158.

CR Bezzina et al. (**2010**) Genome-wide association study identifies a susceptibility locus at 21q21 for ventricular fibrillation in acute myocardial infarction 8,688‐691.

DF Gudbjartsson et al. (**2007**) Variants conferring risk of atrial fibrillation on chromosome 4q25 7151,353‐357.

DE Arking et al. (**2006**) A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization 6,644‐651.

A Pfeufer et al. (**2009**) Common variants at ten loci modulate the QT interval duration in the QTSCD Study 4,407‐414.

C Newton-Cheh et al. (**2009**) Common variants at ten loci influence QT interval duration in the QTGEN Study 4,399‐406.

H Holm et al. (**2010**) Several common variants modulate heart rate, PR interval and QRS duration 2,117-122.

JC Chambers et al. (**2010**) Genetic variation in SCN10A influences cardiac conduction 2,149-152.

N Sotoodehnia et al. (**2010**) Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction 12,1068-1076.

A Pfeufer et al. (**2010**) Genome-wide association study of PR interval 2,153-159.

YS Cho et al. (**2009**) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits 5,527-534.

M Eijgelsheim et al. (**2010**) Genome-wide association analysis identifies multiple loci related to resting heart rate 19,3885-3894.

CM Albert et al. (**2010**) Common variants in cardiac ion channel genes are associated with sudden cardiac death 3,222-229.

WHL Kao et al. (**2009**) Genetic Variations in Nitric Oxide Synthase 1 Adaptor Protein Are Associated With Sudden Cardiac Death in US White Community-Based Populations 7,940-951.

C Dina (**2011**) Of 508 mice and 40,000 humans 3,377-379.

P Brugada, J Brugada (**1992**) Right bundle branch block, persistent ST segment elevation and sudden cardiac death: a distinct clinical and electrocardiographic syndrome. A multicenter report 6,1391-1396.

R Brugada et al. (**2006**) Electrocardiogram interpretation and class I blocker challenge in Brugada syndrome 4 Suppl,S115-118.

C Antzelevitch et al. (**2005**) Brugada syndrome: from cell to bedside 1,9-54.

MM Gallagher et al. (**2008**) Prevalence and significance of Brugada-type ECG in 12,012 apparently healthy European subjects 1,44-48.

K Hong et al. (**2004**) Value of electrocardiographic parameters and ajmaline test in the diagnosis of Brugada syndrome caused by SCN5A mutations 19,3023-3027.

L Eckardt et al. (**2005**) Long-term prognosis of individuals with right precordial ST-segment-elevation Brugada syndrome 3,257-263.

B Benito et al. (**2008**) Gender differences in clinical manifestations of Brugada syndrome 19,1567-1573.

C Antzelevitch et al. (**2005**) Brugada syndrome: report of the second consensus conference: endorsed by the Heart Rhythm Society and the European Heart Rhythm Association 5,659-670.

RJ Havlik et al. (**1980**) Variability of heart rate, P-R, QRS and Q-T durations in twins 1,45-48.

J Li et al. (**2009**) Familial aggregation and heritability of electrocardiographic intervals and heart rate in a rural Chinese population 2,147‐152.

S Mutikainen et al. (**2009**) Genetic influences on resting electrocardiographic variables in older women: a twin study 1,57‐64.

MW Russell et al. (**1998**) Heritability of ECG measurements in adult male twins 64‐68.

Q Chen et al. (**1998**) Genetic basis and molecular mechanism for idiopathic ventricular fibrillation 6673,293‐296.

JD Kapplinger et al. (**2010**) An international compendium of mutations in the SCN5A-encoded cardiac sodium channel in patients referred for Brugada syndrome genetic testing 1,33‐46.

B London et al. (**2007**) Mutation in Glycerol-3-Phosphate Dehydrogenase 1-Like Gene (GPD1-L) Decreases Cardiac Na+ Current and Causes Inherited Arrhythmias 20,2260‐2268.

C Antzelevitch et al. (**2007**) Loss-of-function mutations in the cardiac calcium channel underlie a new clinical entity characterized by ST-segment elevation, short QT intervals, and sudden cardiac death 4,442‐449.

H Watanabe et al. (**2008**) Sodium channel β1 subunit mutations associated with Brugada syndrome and cardiac conduction disease in humans

E Delpon et al. (**2008**) Functional Effects of KCNE3 Mutation and Its Role in the Development of Brugada Syndrome 3,209‐218.

D Hu et al. (**2009**) A Mutation in the 3 Subunit of the Cardiac Sodium Channel Associated With Brugada ECG Phenotype 3,270‐278.

D Kattygnarath et al. (**2011**) MOG1: A New Susceptibility Gene for Brugada Syndrome 3,261‐268.

S Ohno et al. (**2011**) KCNE5 (KCNE1L) variants are novel modulators of Brugada syndrome and idiopathic ventricular fibrillation 3,352‐361.

JR Giudicessi et al. (**2011**) Transient outward current (Ito) gain-of-function mutations in the KCND3-encoded Kv4.3 potassium channel and Brugada syndrome 7,1024‐1032.

P Berne, J Brugada (**2012**) Brugada syndrome 2012 7,1563‐1571.

V Probst et al. (**2009**) SCN5A mutations and the role of genetic background in the pathophysiology of Brugada syndrome 6,552‐557.

CR Bezzina et al. (**2006**) Common sodium channel promoter haplotype in asian subjects underlies variability in cardiac conduction 3,338‐344.

L Crotti et al. (**2012**) Spectrum and Prevalence of Mutations Involving BrS1- Through BrS12-Susceptibility Genes in a Cohort of Unrelated Patients Referred for Brugada Syndrome Genetic Testing: Implications for Genetic Testing

E Génin et al. (**2011**) Genome-wide association study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe 1,52.

ICRM Kolder et al. (**2012**) Common genetic variation modulating cardiac ECG parameters and susceptibility to sudden cardiac death 3,620‑629.

Y Sakata et al. (**2002**) Ventricular septal defect and cardiomyopathy in mice lacking the transcription factor CHF1/Hey2 25,16197‑16202.

M Gessler et al. (**2002**) Mouse gridlock: no aortic coarctation or deficiency, but fatal cardiac defects in Hey2 -/- mice 18,1601‑1604.

J Donovan et al. (**2002**) Tetralogy of fallot and other congenital heart defects in Hey2 mutant mice 18,1605‑1610.

Y Sakata et al. (**2006**) The spectrum of cardiovascular anomalies in CHF1/Hey2 deficient mice reveals roles in endocardial cushion, myocardial and vascular maturation 2,267‑273.

M van den Boogaard et al. (**2012**) Genetic variation in T-box binding element functionally affects SCN5A/SCN10A enhancer 7,2519‑2530.

FH Netter et al. (**2011**) Atlas d'anatomie humaine

M Eghbali et al. (**1991**) Cardiac fibroblasts are predisposed to convert into myocyte phenotype: specific effect of transforming growth factor beta 3,795‑799.

LA Freed et al. (**1999**) Prevalence and clinical outcome of mitral-valve prolapse 1,1‑7.

LA Freed et al. (**2002**) Mitral valve prolapse in the general population: the benign nature of echocardiographic features in the Framingham Heart Study 7,1298‑1304.

JB Barlow, WA Pocock (**1979**) Mitral valve prolapse, the specific billowing mitral leaflet syndrome, or an insignificant non-ejection systolic click 3,277‑285.

NV Strahan et al. (**1983**) Inheritance of the mitral valve prolapse syndrome. Discussion of a three-dimensional penetrance model 6,967‑972.

AN Weiss et al. (**1975**) Echocardiographic detection of mitral valve prolapse. Exclusion of false positive diagnosis and determination of inheritance 6,1091‑1096.

RB Devereux et al. (**1982**) Inheritance of mitral valve prolapse: effect of age and sex on gene expression 6,826‑832.

PL Monteleone, LF Fagan (**1969**) Possible X-linked congenital heart disease 5,611‑614.

S Disse et al. (**1999**) Mapping of a first locus for autosomal dominant myxomatous mitral-valve prolapse to chromosome 16p11.2-p12.1 5,1242‑1251.

LA Freed et al. (**2003**) A locus for autosomal dominant mitral valve prolapse on chromosome 11p15.4 6,1551‑1559.

F Nesta et al. (**2005**) New locus for autosomal dominant mitral valve prolapse on chromosome 13: clinical insights from genetic studies 13,2022‑2030.

F Kyndt et al. (**1998**) Mapping of X-linked myxomatous valvular dystrophy to chromosome Xq28 3,627‑632.

F Kyndt et al. (**2007**) Mutations in the gene encoding filamin A as a cause for familial cardiac valvular dystrophy 1,40-49.

A Lardeux et al. (**2011**) Filamin-a-related myxomatous mitral valve dystrophy: genetic, echocardiographic and functional aspects 6,748-756.

TS Guy, AC Hill (**2012**) Mitral Valve Prolapse 1,277-292.

S Hercberg et al. (**1998**) Background and rationale behind the SU.VI.MAX Study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. SUpplementation en VItamines et Minéraux AntioXydants Study 1,3-20.

LV Wain et al. (**2011**) Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure 10,1005-1011.

S Purcell et al. (**2007**) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses 3,559-575.

DR Govindaraju et al. (**2008**) Genetics of the Framingham Heart Study population 33-65.

C Dina et al. (**2007**) Variation in FTO contributes to childhood obesity and severe adult obesity 6,724-726.

TM Frayling et al. (**2007**) A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity 5826,889-894.

SFA Grant et al. (**2006**) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes 3,320-323.

RJF Loos et al. (**2008**) Common variants near MC4R are associated with fat mass, weight and risk of obesity 6,768-775.

C Dina (**2008**) New insights into the genetics of body weight 4,378-384.

J McClellan, M-C King (**2010**) Genetic heterogeneity in human disease 2,210-217.

O Zuk et al. (**2012**) The mystery of missing heritability: Genetic interactions create phantom heritability 4,1193-1198.

N Craddock et al. (**2010**) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls 7289,713-720.

G Gibson (**2011**) Rare and common variants: twenty arguments 2,135-145.

J McClellan, M-C King (**2010**) Genetic Heterogeneity in Human Disease 2,210-217.

SM Purcell et al. (**2009**) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder 7256,748-752.

Z Wei et al. (**2009**) From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes 10,e1000678.

H Lango Allen et al. (**2010**) Hundreds of variants clustered in genomic loci and biological pathways affect human height 7317,832‑838.

SC Elbein et al. (**2012**) Genetic risk factors for type 2 diabetes: a trans-regulatory genetic architecture? 3,466‑477.

R Jones et al. (**2005**) The search for genenotype/phenotype associations and the phenome scan 4,264–275.

OSP Davis et al. (**2012**) Visual analysis of geocoded twin data puts nature and nurture on the map 9,867‑874.

YS Aulchenko et al. (**2009**) Predicting human height by Victorian and genomic methods 8,1070‑1075.

ANALYSE D'ASSOCIATION GENOME ENTIER DE 3 PATHOLOGIES - LE DIABETE DE TYPE 2, LE SYNDROME DE BRUGADA ET LE PROLAPSUS VALVULAIRE MITRAL : OBSERVATIONS SUR L'ARCHITECTURE GENETIQUE DE TRAITS COMPLEXES

Les maladies cardio-vasculaires et métaboliques représentent une proportion de plus en plus importante des facteurs réduisant la qualité et l'espérance de vie et entrainant une hausse vertigineuse des dépenses publiques dans le domaine de la santé. Cette augmentation est une des conséquences du vieillissement de la population dans nos sociétés occidentales ainsi que d'un changement marqué du mode de vie. Ce même phénomène touche également de plus en plus les pays en voie de développement.

Devant cette urgence à remédier aux conséquences d'une telle explosion, la recherche des bases moléculaires de ces pathologies ainsi que de bio-marqueurs permettant d'augmenter leur prédictibilité représente une priorité. L'approche génétique, dans ce contexte, est une des voies les plus prometteuses. Cette approche présente de nombreuses variantes. Une des plus populaires dans la dernière décennie est l'approche des études d'association génome entier. La stratégie développée repose sur l'hypothèse d'un rôle important joué par variants génétiques fréquents pour des pathologies fréquentes. Ce paradigme a été nommé l'hypothèse de « Maladie fréquente, polymorphisme génétique fréquent » (« Common Variant / Common Disease »).

Dans le cadre de ma thèse, j'explore l'effet de variants fréquents sur trois pathologies, le Diabète de Type 2, la Prolapsus Valvulaire Mitral, pathologies fréquentes et le Syndrome de Brugada, qui est rare dans la population. Ces trois pathologies présentent un poids important dans l'explosion des besoins sanitaires des populations, soit par la gravité des complications pour le Diabète de Type 2, soit par les besoins d'intervention chirurgicale lourde pour le Prolapsus Valvulaire Mitral ou enfin par le risque de Mort Subite inexpliquée pour le Syndrome de Brugada.

J'ai appliqué différentes techniques génétiques que sont l'imputation, la méta-analyse et la correction de stratification afin de contribuer à mettre en évidence leurs bases génétiques.

Dans le diabète de Type 2, la mise en évidence de l'architecture génétique était déjà bien avancée et j'ai participé à l'approfondissement de ces connaissances. Ces travaux ont permis de mettre en évidence jusqu'à 40 gènes associés après correction pour le nombre de tests. De façon important, nous avons montré qu'il existe une composante polygénique importante et que la plupart des gènes impliqués pointent vers un dysfonctionnement des cellules bêta.

Les études sur le Prolapsus Valvulaire Mitral sont à un stade moins avancé. J'ai sélectionné des variants génétiques montrant une association possible et ces variants sont en cours de réplication. Les résultats préliminaires sur l'étude Framingham montrent la possible implication de gènes de la matrice extracellulaire.

Enfin, pour le Syndrome de Brugada, j'ai clairement identifié trois loci qui montrent une association très significative et ne laissant pas de place au doute. Ces loci ont été répliqués aussi bien dans une population Européenne que dans une population Japonaise. Si l'implication de gènes codant pour des canaux ionique est confirmée, une autre voie liée au développement cardiaque est également mise en évidence.

Enfin, au cours de ma thèse, j'ai également contribué à faire apparaître la notion de variant fréquent pour pathologie rare, dans l'étude d'association portant sur le Syndrome de Brugada.

Mots-clés : 1/ Génétique 2/ Epidémiologie 3/ Association 4/ Valve 5/ Arythmie 6/ Diabète 7/ Polymorphisme 8/ Héritabilité

GENOME SCAN ASSOCIATION STUDIES OF 3 PATHOLOGIES - TYPE 2 DIABETES, BRUGADA SYNDROME AND MITRAL VALVULAR PROLAPSE : THE GENETIC ARCHITECTURE OF COMPLEX TRAITS

The escalating prevalence of cardio-vascular and metabolic disorders, and the limitations of currently available preventive and therapeutic options are increasingly important factors reducing the quality and life expectancy resulting in a dramatic increase in public spending in the health field. This emergency highlights the need for a more complete understanding of the pathogenesis of these diseases as well as the need for bio-markers to increase their predictability is a priority. The genetic approach, in this context, is among the most promising strategies. This approach has many variants. One of the most popular in the last decade is the approach of genome-wide association studies. The strategy is based on the assumption of an important role played by common genetic variants for common diseases. This paradigm has been called the assumption of "common variant, common disease".

As part of my thesis, I explored the effect of common variants in three diseases, Diabetes Type 2, Mitral Valvular Prolapse, both being common pathologies and the Brugada syndrome, which is rare in the population. These three diseases strongly contribute to the explosion of population health needs, either by the severity of complications for Type 2 Diabetes, through the need of major surgery for Mitral Valvular Prolapse and through the increased risk of Sudden Death for Brugada Syndrome. I applied various techniques such as genetic imputation, meta-analysis and correction of stratification to help highlight their genetic bases. In Type 2 diabetes, highlighting of the genetic architecture was already well advanced and I participated in the deepening of knowledge. This work helped identify up to 40 genes. We have also shown that there is a substantial polygenic component underlying the genetic architecture of this disease and that most of the identified genes point to a dysfunction of beta cells.

Studies on Mitral Valvular Prolapse are less advanced. I selected genetic variants showing a possible association and these variants are being replicated. Preliminary results on the Framingham study showed the possible involvement of genes of the extracellular matrix.

Finally, for Brugada Syndrome, I clearly identified three loci that show a highly significant association with the disease. These loci were replicated as well in a European population in Japanese population. If the involvement of genes coding for ion channel proteins (SCN5A and SCN10A) seems to be confirmed, strengthening the definition of Brugada Syndrome as a channelopathy, another pathway possibly related to cardiac development was also identified (through the gene HEY2). Finally, during my PhD, I also contributed to create the concept of common variant for rare disease (CV/CR).

Keywords : 1/ Genetics 2/ Epidemiology 3/ Association 4/ Valve 5/ Arythmia 6/ Diabetes 7/ Polymorphism 8/ Heritability