



Méthodologie statistique pour étudier l’empreinte parentale dans le cadre des études d’association pan-génomique -Etude de l’association entre le gène ABO et le facteur de Willebrand dans la maladie thromboembolique veineuse

Blandine Gendre

► **To cite this version:**

Blandine Gendre. Méthodologie statistique pour étudier l’empreinte parentale dans le cadre des études d’association pan-génomique -Etude de l’association entre le gène ABO et le facteur de Willebrand dans la maladie thromboembolique veineuse. Santé publique et épidémiologie. 2021. dumas-03377718

HAL Id: dumas-03377718

<https://dumas.ccsd.cnrs.fr/dumas-03377718>

Submitted on 14 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

2^{ème} années du Master Biostatistique

Mention Santé publique

Promotion 2020-2021

Méthodologie statistique pour étudier
l’empreinte parentale dans le cadre des études
d’association pan-génomique - Etude de
l’association entre le gène ABO et le facteur de
Willebrand dans la maladie thromboembolique
veineuse

Blandine GENDRE

Centre de Recherche - Bordeaux Population Health - Inserm U1219
Equipe Vintage

Tuteur de stage :

M. TREGOUET David Alexandre, Directeur de recherche

22/02/2021 - 21/07/2021

Remerciements

Je tiens tout d'abord à remercier mon tuteur de stage : David-Alexandre TREGOUET pour ta bienveillance envers moi durant toutes nos réunions. Merci également d'avoir pris de ton temps pour répondre à mes questions, d'avoir approfondi ma compréhension de certaines notions, et pour nos discussions sur les potentiels projets à venir. Merci pour tes conseils, ta réactivité et ta franchise. Je te remercie de m'avoir inclus dans l'équipe pour mener ce stage et également de compter sur moi pour le projet de thèse que j'espère pouvoir développer ensemble.

Je tiens également à remercier Gaëlle MUNSCH qui n'a pas hésité à me proposer son aide et avec qui j'ai pu discuter.

A cause des mesures prises en raison de la crise sanitaire, je n'ai pas pu rencontrer toute l'équipe, hormis les réunions d'équipe en visioconférence toutes les 2 semaines, mais je tenais également à remercier toute l'équipe VINTAGE que j'espère pouvoir prochainement rencontrer.

Je remercie aussi Pierre-Emmanuel MORANGE, d'avoir pris le temps de discuter sur le jeu de données MARFAST et de trouver des solutions pour mener à bien mon projet.

Je remercie également mon compagnon, Clément, qui me supporte tous les jours, me rassure quand je commence à stresser ou à déprimer, et me pousse vers le haut. Merci d'avoir pris le temps, même le soir, de travailler certaines notions mathématiques, de discuter de mon travail et amener une réflexion et également de relire mon rapport.

Merci à toute ma famille, qui croit en moi et me soutient, j'ai hâte de vous revoir. Une pensée personnelle pour Mélanie, mon acolyte que je n'ai pas quitté depuis le lycée, pour nos déjeuners le midi avant que le stage soit à 100% en télétravail. Merci également de m'avoir aidé sur R, d'avoir pris le temps de discuter sur mon sujet, et de nos journées en co-working qui étaient plus motivantes.

Sommaire

1 Introduction	6
2 Structure du centre de recherche - U1219	8
3 Contexte et Objectifs	9
3.1 Contexte	9
3.1.1 La maladie thromboembolique veineuse	9
3.1.2 Le gène ABO	10
3.1.3 L’empreinte parentale	11
3.2 Objectifs du projet	14
4 Matériel et Méthodes	15
4.1 Méthodologie statistique	15
4.1.1 Empreinte parentale et variance inter groupe	15
4.1.2 Test de Levene	16
4.1.3 Test de Levene tronqué à 25%	17
4.1.4 Test de Brown-Forsythe	17
4.1.5 Stratégie d’analyse : Méta-analyse	20
4.2 Contexte des études	21
4.2.1 Projet MARTHA	21
4.2.2 Etude FARIVE	22
4.2.3 Etude MARFAST	23
5 Résultats	24
5.1 Détection des groupes sanguins potentiellement soumis à empreinte	24
5.2 Validation des résultats dans des données familiales	27
6 Discussion	31
7 Conclusion	33
7.1 Conclusion sur les principaux résultats	33
7.2 Conclusion personnelle	33
Bibliographie	34
I	34

Nomenclature

ADN Acide désoxyribonucléique

ANOVA Analyse de la variance (*ANalysis Of VAriance*)

AVC Accident Vasculaire Cérébral

BPH Bordeaux Population Health

DNMT ADN-méthyltransférase

EP Embolie Pulmonaire

FARIVE FActeurs RISques de la maladie thromboembolique VEineuse

GWAS Etudes d'association pan-génomique (*Genome Wide Association Studies*)

IMC Indice de Masse Corporelle

MARFAST MARseilles FAmily Study on venous Thrombosis

MARTHA MARseille THrombosis Association

MTEV Maladie ThromboEmbolique Veineuse

OMS Organisation Mondiale de la Santé

RC Rapport des Cotes

SNP Polymorphisme d'un Seul Nucléotide (*Single Nucleotide Polymorphism*)

TVP Thrombose Veineuse Profonde

VINTAGE Epidémiologie intégrative et génétique du vieillissement cérébral et vasculaire

vWF Facteur de Willebrand (*von Willebrand Factor*)

Listes des figures

1	Schéma explicatif de la maladie thromboembolique veineuse.	9
2	Triade de Virchow, 1856.	9
3	Représentation d'un polymorphisme d'un seul nucléotide.	10
4	Expériences de transferts nucléaires. McGrath JM, Solter D et Surani Azim. 1984. . .	12
5	Les espèces hybrides : le ligre (en haut) et le tigrion (en bas). <i>Source : Google Images</i>	12
6	Le cycle de l'empreinte génomique [18].	13
7	Flow-chart de la cohorte MARFAST. (n=2865).	28

Listes des tableaux

1	Fréquences alléliques dans les trois cohortes : MARTHA, MARTHA12 et FARIVE. .	24
2	Tableau descriptif de certaines variables recueillies en fonction des trois cohortes : MARTHA, MARTHA12 et FARIVE.	25
3	Mesure d'association ajustée sur le sexe et l'âge, d'après le test de Brown-Forsythe entre le factor von Willebrand et les SNPs chez MARTHA, MARTHA12 et FARIVE. .	26
4	Méta-analyse du test de Brown-Forsythe réalisée dans les trois cohortes : MARTHA, MARTHA12 et FARIVE.	26
5	Mesure du facteur Willebrand en fonction du groupe sanguin A1, chez MARFAST (n=203), moyenne et écart-type.	29

1 Introduction

Selon l'Organisation Mondiale de la Santé (OMS) : *"D'ici 2050, la proportion de la population mondiale de plus de 60 ans doublera pour passer d'environ 11% à 22%. Le nombre de personnes âgées de 60 ans et plus devrait augmenter pour passer de 605 millions à deux milliards au cours de la même période"* [1]. Dans un monde où la population est vieillissante, une réponse globale en terme de santé publique est attendue. L'augmentation de l'âge entraîne de nombreuses maladies et accroît notamment les risques de contracter des maladies vasculaires et neurologiques. Or les maladies cardio-vasculaires sont les premières causes de mortalité dans le monde et les deuxièmes en France (premières pour les femmes). En France, ces maladies sont à l'origine de 140 000 nouveaux cas chaque année, soit une personne toutes les quatre minutes [2]. Il est donc vital d'entreprendre des recherches en amont de cette aggravation car la compréhension de ces mécanismes est une priorité pour prévenir une telle problématique.

L'équipe *Epidémiologie intégrative et génétique du vieillissement cérébral et vasculaire* (VINTAGE) du centre de Recherche U1219 "Bordeaux Population Health" (BPH) y consacre toutes ses recherches. Les trois grandes maladies étudiées sont l'Accident Vasculaire Cérébral (AVC), la démence et la Maladie ThromboEmbolique Veineuse (MTEV) dont David-Alexandre TREGOUET est un des directeurs de recherche de l'équipe qui s'y dédie entièrement. La MTEV représente la 3ème cause de mortalité d'origine vasculaire et affecte environ 1 million de personnes par an en Europe. L'incidence de la MTEV augmente avec l'âge passant de 0,05% avant 40 ans à 0,5% vers 70 ans [3].

La MTEV est la résultante de l'effet de plusieurs facteurs, qu'ils soient "environnementaux", tels que le tabagisme ou la prise de contraception orale, ou génétiques : c'est une maladie multifactorielle. Une trentaine de gènes ont déjà été identifiés comme étant associés à la MTEV grâce aux études d'association pan-génomique, parmi lesquels, le gène ABO est l'un des premiers et l'un des plus importants. Le gène ABO code pour le groupe sanguin et il est connu que ce gène influence les taux plasmatique du facteur de Willebrand (vWF). A ce jour, on sait que le groupe sanguin A1 et B augmentent les taux de Willebrand [14] et que le gène ABO explique 25% de sa variabilité plasmatique. Le facteur de Willebrand est une protéine de la coagulation qui augmente le risque de développer une MTEV si son taux est anormalement élevé.

Les études d'association pan-génomique (GWAS) deviennent des stratégies de recherche très populaires et très efficaces pour identifier des polymorphismes génétiques associés à des facteurs de risques de maladies tels que le facteur de Willebrand pour la MTEV. Les études GWAS font cependant l'hypothèse que les allèles ont les mêmes effets qu'ils soient hérités du père ou de la mère. Or certaines études ont montré que l'origine de l'allèle peut augmenter le risque de développer une maladie s'il a été hérité du père ou de la mère [4]. Ces études ont pour l'instant toujours été réalisées dans des données familiales, où il était possible d'identifier quels allèles étaient hérités des parents.

Récemment, une méthodologie scientifique a été proposée pour identifier la présence d'empreinte parentale en l'absence de données familiales, appliquée à l'Indice de Masse Corporelle (IMC), se basant sur un test de comparaison de variance : le test de Brown-Forsythe. L'avantage de cette étude est qu'il est maintenant possible de détecter des polymorphismes (SNPs) dont les effets sur les variables biologiques étudiées puissent être sujets à empreinte parentale. Au cours de ce stage, cette méthodologie a donc été appliquée pour déterminer si des polymorphismes du gène ABO pouvaient être soumis à empreinte parentale sur le taux de Willebrand. Afin de vérifier ces résultats, les SNPs détec-

tés précédemment ont été analysés dans des données familiales pour tester s'il existait une différence statistiquement significative entre les taux de Willebrand des individus dont l'allèle en question provenait du père ou de la mère.

L'obtention de tels résultats permettrait d'illustrer que potentiellement l'héritage des allèles du groupe sanguin pourrait entraîner un plus grand risque de développer une MTEV.

2 Structure du centre de recherche - U1219

L'Inserm est le seul organisme public de recherche français entièrement dédié à la santé humaine. Il est placé sous la double tutelle du Ministère de la Recherche et du Ministère de la Santé [5].

Le centre de recherche "Bordeaux Population Health ResearchCenter" (BPH) émerge de l'union de grandes équipes de recherches labellisées par l'Inserm et l'université de Bordeaux [6]. Il est dirigé par le professeur Christophe TZOURIO et est composé de 11 équipes de recherche et de 2 équipes en émergence. Chacune des équipes a une thématique de recherche qui lui est propre tout en développant des projets communs.

Parmi les 11 équipes, Stéphanie Debette dirige en co-direction avec Carole Dufouil l'équipe *Epidémiologie intégrative et génétique du vieillissement cérébral et vasculaire* (plus communément appelée VINTAGE). Cette équipe s'intéresse à l'origine des mécanismes de maladies neurovasculaires, liées à l'âge, dont les plus fréquentes : les accidents vasculaires cérébraux, la démence et la maladie thromboembolique veineuse. L'équipe porte un intérêt particulier pour l'identification et l'évaluation de nouveaux biomarqueurs de ces maladies dont le but est d'accélérer le ciblage de nouvelles thérapies, de faciliter la détection d'individus à haut risque de développer ces maladies et d'optimiser les stratégies de prévention. De cette manière, l'équipe contribue à l'élaboration de stratégies pour prévenir ou retarder l'apparition de maladies neurologiques et promeut le bien vieillir cérébral [7].

David-Alexandre Trégouët est Directeur de la Recherche au sein de cette équipe et coordonne un axe sur la maladie thromboembolique veineuse qui est la 3^{ème} maladie cardio-vasculaire la plus fréquente dans les pays développés.

3 Contexte et Objectifs

3.1 Contexte

3.1.1 La maladie thromboembolique veineuse

La maladie thromboembolique veineuse (MTEV) regroupe la thrombose veineuse profonde (TVP) et sa complication immédiate, l'embolie pulmonaire (EP).

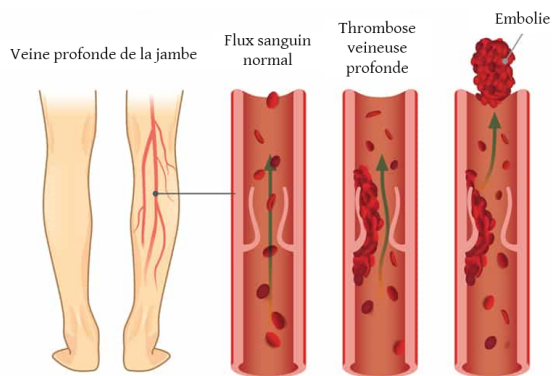


FIGURE 1 – Schéma explicatif de la maladie thromboembolique veineuse.

La thrombose veineuse correspond à la formation d'un thrombus (caillot de sang) dans le système veineux qui peut obstruer la circulation sanguine. Ce caillot de sang peut se créer si le sang stagne (lors d'une immobilisation pendant plusieurs jours), ou bien si, en plus de la stagnation, il y a eu un saignement. Généralement, le thrombus se forme dans les membres inférieurs. Une fois formé, le thrombus peut se détacher (ou se fractionner) puis migrer vers les artères pulmonaires et provoquer une EP (Figure 1).

La MTEV est un enjeu de santé publique car elle correspond à la troisième cause de mortalité d'origine vasculaire après l'infarctus du myocarde et l'accident vasculaire cérébral [8]. C'est une pathologie fréquente, on constate qu'en France, en 2013, l'incidence d'une MTEV était de 1,57 pour 1000 personnes par an, soit plus de 100 000 cas par an [9]. De plus, l'incidence augmente avec l'âge et atteint près d'un cas sur 100 personnes par an chez les sujets âgés de plus de 75 ans. La thrombose veineuse est une pathologie multifactorielle.

Dès 1856, les facteurs à l'origine d'une MTEV sont résumés par la triade de Virchow : la stase veineuse, la lésion de la paroi vasculaire et les altérations de la coagulation (Figure 2). Les altérations de la coagulation représentent une hypercoagulabilité du sang, i.e. sa capacité exagérée à s'épaissir et s'agréger qui peut être causée par une grossesse, une pilule contraceptive ou un cancer. Les lésions de la paroi du vaisseau correspondent à une altération de la paroi veineuse par un traumatisme, une injection ou une chirurgie. Quant à la stase veineuse, elle est caractérisée par un ralentissement du flux sanguin, qui peut être la conséquence d'une immobilisation prolongée ou d'une obésité. Chacun de ces trois facteurs peut être à l'origine d'une thrombose.

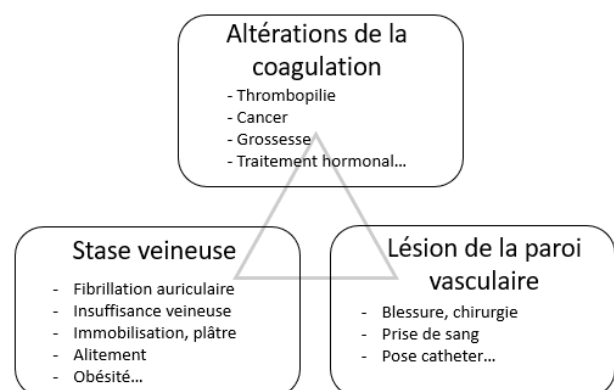


FIGURE 2 – Triade de Virchow, 1856.

De nombreuses études ont montré que le risque de développer une MTEV était multiplié par 2 ou 3 chez les sujets obèses ($IMC > 30 \text{ kg/m}^2$). De même, le cancer et l'immobilisation/chirurgie sont des

facteurs majeurs qui sont retrouvés dans respectivement 20% et 15% des MTEV [10]. Malgré une incidence élevée et une mortalité importante, les mécanismes étiopathologiques de la MTEV restent mal identifiés. En effet, les trois mécanismes décrits par Virchow agissent en interaction et sont souvent très intriqués. Il est donc difficile de mettre en évidence l'implication d'un mécanisme plutôt qu'un autre [11].

Depuis plus de 30 ans, les facteurs génétiques de risques de thrombose sont devenus des facteurs prédisposant fréquents de la MTEV. Parmi ces facteurs biologiques le gène ABO, qui caractérise le groupe sanguin d'un individu, a été le premier identifié, mais d'autres facteurs ont également été retrouvés tels que : le facteur de Willebrand (vWF), le facteur XI ou encore le facteur VIII [12]. Pour la plupart de ces facteurs, leurs niveaux plasmatiques sont influencés par des polymorphismes génétiques.

Le facteur de Willebrand, qui a été étudié lors de ce stage, est une protéine de coagulation qui agit, avec plusieurs autres composantes de ce dernier, pour arrêter les saignements dans le corps. En temps normal, lorsqu'un vaisseau sanguin est endommagé et saigne, le vWF forme une passerelle adhésive qui finit par produire un caillot sanguin stable. Or dans des situations à risque, ce caillot peut grossir et provoquer une TVP. L'effet thrombotique du gène ABO est distingué par des modifications du facteur VIII et du facteur von Willebrand [14]. En effet, certains SNP du gène ABO, notamment les SNP codant pour les groupes sanguins A1 et B, augmenteraient le taux plasmatique de vWF et du facteur VIII.

3.1.2 Le gène ABO

Un gène est une séquence d'ADN sous la forme de successions de nucléotides (bases) : l'Adénine (A), la Cytosine (C), la Guanine (G) et la Thymine (T) ; qui code pour une protéine. Les SNPs (Single Nucleotide Polymorphism) constituent la forme la plus abondante de variations génétiques dans le génome humain.

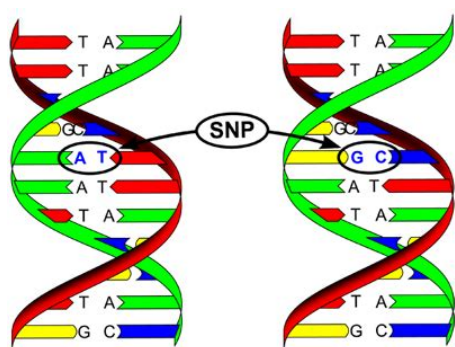


FIGURE 3 – Représentation d'un polymorphisme d'un seul nucléotide.

Il s'agit de la variation (polymorphisme) d'une seule paire de bases du génome (Figure 3). Ils représentent plus de 90% de toutes les différences entre individus. Le gène ABO est le gène responsable du groupe sanguin. Il se situe sur le bras long du chromosome 9 (position 9q34). Les principaux groupes sanguins peuvent être définis par la paire de base présente sur 5 SNPs : la base T sur le SNP rs2519093 codant pour le groupe sanguin A1 (qui est un sous groupe du groupe sanguin A), la base A sur le SNP rs1053878 (A2), la perte de la base G sur le rs8176719 (O1), la base T sur le SNP rs41302905 (O2) et la base T sur rs8176746 (B) [14].

De cette manière, si un individu possède une paire de bases T/T du SNP codant pour le groupe A1 alors cet individu sera identifié comme appartenant au groupe sanguin A1A1. De même si un individu code A/A sur le SNP du groupe sanguin A2 alors il appartiendra au groupe sanguin A2A2. On peut également identifier si un individu est hétérozygote A1, c'est dire qu'il possède qu'une seule base T sur le SNP codant pour le groupe A1 mais par contre il n'est pas possible de déterminer uniquement

avec ce SNP, s'il est : A1A2, A1O1, A1O2 ou A1B. Pour obtenir ces résultats, il faudrait combiner les informations des différents SNPs du gène ABO.

Dans la population française, le groupe sanguin le plus fréquent est le groupe O1 (63%) suivi du groupe A1 (20%).

En 1969, le gène ABO a été l'un des premiers facteurs de risques identifiés de la MTEV et les groupes A1 et B sont ceux qui augmentent le risque de provoquer cette pathologie [15]. Cependant, le gène ABO n'a pas seulement été impliqué dans la MTEV, mais également dans d'autres maladies telles que le paludisme [4], l'hépatite C [16] ou encore l'obésité [17]. Dans ces pathologies, les individus de chaque groupe sanguin ont un risque inégal. *Andrew E. Fry et al* [4] ont confirmé que les individus du groupe sanguin O sont relativement protégés contre le paludisme par rapport aux autres groupes, en particulier des groupes A et AB (respectivement des RC de l'ordre de 1,3 et 1,6). Ces résultats ont été observés dans des données d'individus non apparentés mais étant donné que les auteurs disposaient également d'un jeu de données familiales, il a été possible d'analyser les données généalogiques et ainsi de rechercher des effets sur l'empreinte parentale. Ainsi, *Andrew E. Fry et al* [4] ont remarqué que les individus de groupes sanguins O1 sont associés à un plus grand risque de paludisme si l'allèle a été transmis par la mère (RC=1,38, P=0,0002) plutôt que par le père (RC=1,05, P=0,6) comparés aux individus non porteurs. On peut donc émettre l'hypothèse que certains SNPs du gène ABO pourrait être soumis à une empreinte parentale sur le paludisme. Si de tels résultats ont été observés pour certaines maladies, il semble intéressant de regarder ce phénomène d'empreinte cette fois-ci sur le taux de Willebrand. En effet, peut-être que les taux de Willebrand ne sont pas les mêmes en fonction de l'héritage allélique de nos parents. Par conséquent, le risque de développer une MTEV pourrait être augmenté.

3.1.3 L'empreinte parentale

On dit qu'un gène est soumis à empreinte parentale lorsque, chez l'Homme et certains mammifères, la copie héritée de la mère et la copie héritée du père ne sont pas exprimées de la même manière. Le plus souvent, l'une des deux copies du gène est totalement éteinte alors que l'autre est active. L'expression de ce gène est donc mono-allélique et dépend de l'origine parentale. La découverte de l'empreinte parentale a d'abord commencé chez les animaux et s'est également confirmée chez l'Homme.

Dans les années 1980 des expériences de transfert nucléaire ont été mises en place chez les rongeurs. Des embryons gynogénétiques (portant uniquement un patrimoine génétique d'origine maternelle (cf Figure 4 n°2)) et androgénétiques (uniquement d'origine paternelle (cf Figure 4 n°3)) ont été créés mais sont incapables de se développer normalement à terme malgré leur composition diploïde. Les embryons gynogénétiques se développaient normalement contrairement aux annexes (futur placenta) et vice-versa chez les embryons androgénétiques.

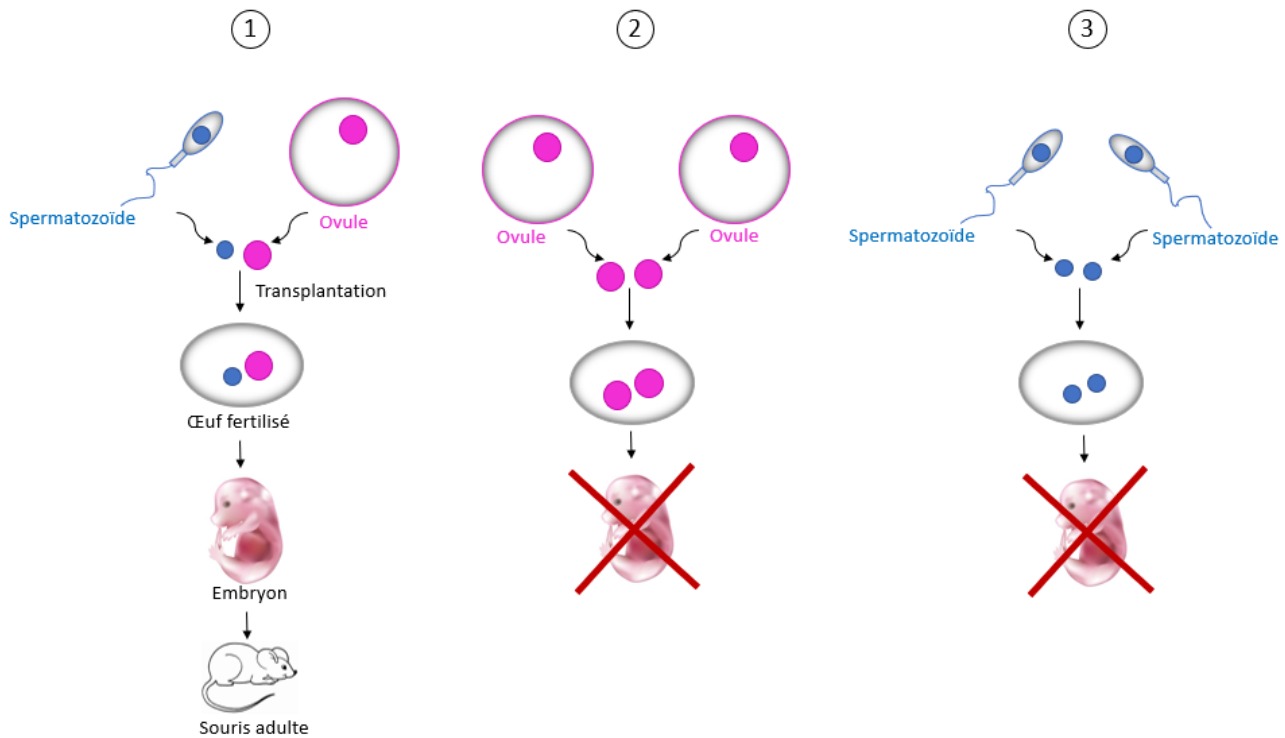


FIGURE 4 – Expériences de transferts nucléaires. McGrath JM, Solter D et Surani Azim. 1984.

D'après ces expériences, les mammifères euthériens (avec développement placentaire) doivent avoir un génome paternel et maternel pour donner un embryon viable.



FIGURE 5 – Les espèces hybrides : le ligre (en haut) et le tigrion (en bas). Source : Google Images

Mais ces expériences ont aussi montré que la contribution de chacun des parents n'est pas la même, il semble bien y avoir une influence parentale pour donner un embryon.

Cette influence parentale peut se retrouver chez les espèces hybrides, le lion et le tigrion en sont un bon exemple (Figure 5). Le ligre est un croisement entre une tigresse et un lion, alors qu'un tigrion est un croisement entre une lionne et un tigre.

Le ligre qui est le plus grand félin du monde peut peser plus de 400 Kg alors que le tigrion a un poids qui reste semblable à celui de sa mère soit environ 150 Kg. De plus on remarque également des différences au niveau de leur pelage. Tous deux ont le même matériel génétique pourtant ce sont deux animaux complètement différents. La seule chose qui les différencie est bien la transmission parentale.

Chez l'Homme, il existe également des différences dues à la transmission parentale, comme par exemple les syndromes de Prader-Willi et d'Angelman. Ces syndromes sont dus à la même mutation, c'est à dire au même défaut génétique mais dépend de la transmission parentale. Si c'est la mère qui transmet la mutation, l'enfant aura le syndrome d'Angelman : retards de développements psychomoteurs et mentaux ; alors que si cette même mutation a été transmise par le père, l'enfant aura le syndrome

de Prader-Willi : problème de satiété qui peut entraîner une obésité.

Cette mutation est en fait une anomalie qui va mettre sous silence la partie paternelle ou maternelle du gène qui devait normalement s'exprimer et qui en fonction du cas va développer l'un des deux syndromes. Ce gène est donc soumis à l'empreinte parentale.

Un des médiateurs les plus connus de ce phénomène est la méthylation de l'ADN. Cette méthylation va aboutir à l'inactivation du gène hérité du père ou de la mère. Pour qu'un individu transmette via ses gamètes une empreinte génomique propre à son sexe (et non à l'empreinte qu'il a reçu de ses parents) une déméthylation totale à lieu dans ces cellules germinales (durant la gamétogénèse). C'est-à-dire qu'en plus d'un effacement de la méthylation du génome complet, les méthylations des gènes soumis à l'empreinte génomique de ses parents sont également effacés. Ce qui permet ensuite un rétablissement de ces empreintes, grâce à des DNMT (ADN-méthyltransférase) de novo, mais spécifique à son sexe (une empreinte maternelle sera établie dans les ovocytes et une empreinte paternelle dans les spermatozoïdes). Cette phase permet d'éviter les problèmes que pourrait rencontrer un embryon si la mère transmettait un chromosome hérité du père sans avoir effacé l'empreinte paternelle. Après la fécondation, les méthylations de l'ADN sont maintenues tout au long des divisions des cellules somatiques (chez l'embryon et l'adulte) grâce à des DNMT de maintenance. Tout ce reprogrammage est important et des erreurs au niveau du processus de l'empreinte (effacement, établissement, maintenance) peuvent avoir des conséquences sur l'embryon et donc par la suite sur l'individu.

Si on reprend le schéma de la Figure 6, les modifications (les bandes bleues et rouges sur les chromosomes) du génome qui constituent l'empreinte sont acquises dans les cellules de la lignée germinale (a), ces cellules qui donnent naissance aux gamètes (ovules et spermatozoïdes) qui fusionnent lors de la fécondation (b).

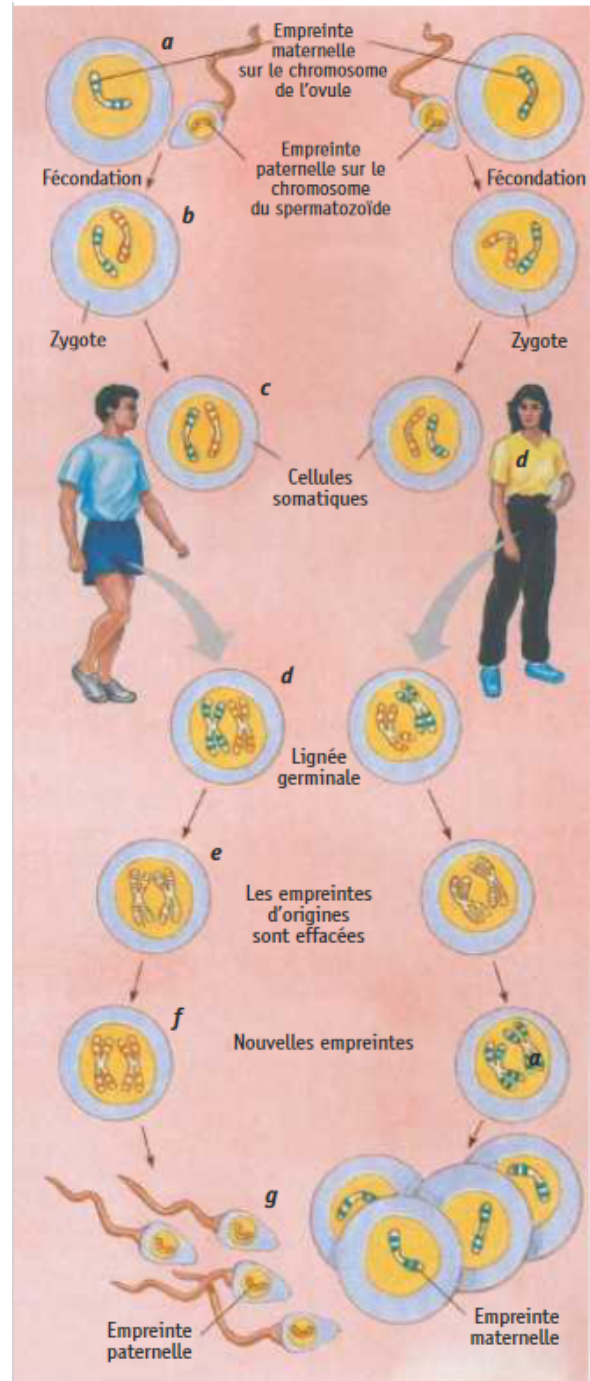


FIGURE 6 – Le cycle de l'empreinte génomique [18].

L'empreinte est maintenue tout au long de la vie de l'organisme, dans les cellules somatiques (c) où elle dirige l'expression des gènes selon leur provenance maternelle ou paternelle. Cependant, dans les cellules de la nouvelle lignée germinale (d), l'empreinte est effacée (e) de façon à ce qu'une nouvelle

soit instaurée (f) dans les gamètes de cette génération (g) [18].

Actuellement, quelques études ont détecté des gènes soumis à empreinte parentale dans certaines maladies (par exemple le gène ABO pour le paludisme [4]). Mais ces détections n'étaient réalisables que dans des données familiales. Or ce type de données est peu commun, coûteux et plus long à mettre en place. En 2014, *Hoggart C et al* [19] ont développé une méthodologie statistique permettant de détecter l'empreinte parentale sur des données non apparentées.

3.2 Objectifs du projet

L'objectif de ce stage est de comprendre la méthodologie statistique issue de l'article de *Hoggart C et al* [19] afin de détecter des gènes soumis à empreinte parentale dans des données génétiques de sujets non apparentés. Cette méthodologie sera appliquée sur des échantillons de données quantitatives (le facteur de Willebrand) de sujets atteints de maladie thromboembolique veineuse pour déterminer si le gène ABO peut-être soumis à empreinte parentale. Dans un premier temps, l'objectif est de déterminer quels SNPs du gène ABO peuvent être sujets à empreinte parentale dans 3 cohortes de patients atteints de la MTEV. La seconde partie avait pour objectif de valider, dans un échantillon de 203 familles nucléaires, les résultats suggérés par la méthodologie proposée par *Hoggart C, et al* [19].

4 Matériel et Méthodes

4.1 Méthodologie statistique

Les études d'association pan-génomique (GWAS en anglais) classiques permettent d'identifier les SNPs impliqués dans la variabilité d'un trait biologique ou d'une variable clinique, mais repose sur l'hypothèse d'absence d'empreinte parentale au niveau des polymorphismes testés. L'étude des phénomènes d'empreinte parentale chez l'Homme s'effectue généralement à partir de données familiales (c'est à dire, quand il est possible de reconstruire les transmissions maternelle et paternelle d'un polymorphisme). Cependant *Hoggart C, et al* [19] ont développé une méthodologie statistique afin d'identifier l'empreinte parentale non pas à partir de données familiales mais à partir simplement de données génétiques de sujets non apparentés. Cette méthode sera détaillée ci-dessous.

4.1.1 Empreinte parentale et variance inter groupe

On dit qu'un individu est homozygote pour un polymorphisme lorsqu'il possède 2 allèles identiques (ex : A/A ou B/B), ou hétérozygote, lorsque les 2 allèles sont différentes (ex : A/B). Généralement, dans les études d'association pan-génomique, pour tester l'effet d'un allèle (par exemple l'allèle A) sur un trait quantitatif, le phénotype d'un individu s'exprime tel que : $\mathbb{E}(y|X) = \mu + \beta X$ avec $X = \{0, 1, 2\}$ en fonction du nombre d'allèle A. Ainsi, la moyenne de la variable dépendante sera pour chaque situation :

Pour un individu homozygote A/A, la moyenne phénotypique est :	Pour un individu hétérozygote A/B, la moyenne phénotypique est :	Pour un individu homozygote B/B, la moyenne phénotypique est :
$\mathbb{E}(y X = 2) = \mu + 2\beta$	$\mathbb{E}(y X = 1) = \mu + \beta$	$\mathbb{E}(y X = 0) = \mu$

Seulement, cette écriture est utilisée quand on ne tient pas compte de l'origine parentale de l'allèle A qui est difficile à déterminer chez les individus hétérozygotes AB. Si on pose maintenant $\beta = \frac{1}{2}(\beta_P + \beta_M)$ avec β_M qui signifie l'effet maternel, par exemple, l'héritage de l'allèle A par la mère et β_P , l'effet paternel, l'héritage de l'allèle A par le père. Alors la moyenne phénotypique pour chaque cas serait :

Pour un individu homozygote A/A, il a hérité des deux allèles A de ses parents, la moyenne phénotypique est :	Pour un individu hétérozygote A/B, il a pu hériter de l'allèle A par son père ou bien par sa mère, il s'agit d'un mélange à 50%-50% de l'héritage des allèles de ces parents, la moyenne phénotypique est donc :	Pour un individu homozygote B/B, la moyenne phénotypique reste la même étant donné que l'individu n'a reçu aucun allèle A :
$\mathbb{E}(y X = 2) = \mu + \beta_M + \beta_P$	$\mathbb{E}(y X = 1) = \mu_{AA} + \pi\beta_M + (1 - \pi)\beta_P$ $\mathbb{E}(y X = 1) = \mu_{AA} + \beta_P + \pi(\beta_M - \beta_P)$ où $\pi \sim \mathcal{B}(\frac{1}{2})$ $\pi = \begin{cases} 1 & \text{si l'allèle A est hérité de la mère} \\ 0 & \text{si l'allèle A est hérité du père} \end{cases}$ $\pi \perp \epsilon$	$\mathbb{E}(y X = 0) = \mu$

La variance phénotypique est donc différente entre les individus homozygotes et les individus hétérozygotes :

Pour un individu homozygote A/A, la variance phénotypique est :

$$\begin{aligned} \text{Var}(y_{AA}) &= \text{Var}(\mu + \beta_M + \beta_P \\ &\quad + \epsilon) \\ \text{Var}(y_{AA}) &= \text{Var}(\epsilon) \\ \text{Var}(y_{AA}) &= \sigma^2 \end{aligned}$$

ϵ représente le terme d'erreur et $\text{Var}(\epsilon) = \sigma^2$

Pour un individu hétérozygote A/B, la variance phénotypique est :

$$\begin{aligned} \text{Var}(y_{AB}) &= \text{Var}[\mu + \beta_P + \pi(\beta_M - \beta_P) + \epsilon] \\ &= \text{Var}[\pi(\beta_M - \beta_P) + \epsilon] \\ \text{car } \text{Var}(\mu) &= \text{Var}(\beta_P) = 0 \\ &= (\beta_M - \beta_P)^2 \text{Var}(\pi) + \text{Var}(\epsilon) \\ &= (\beta_M - \beta_P)^2 [p(1-p)] + \sigma^2 \\ \text{car } \text{Var}(\pi) &= p(1-p), \text{Var}(\epsilon) = \sigma^2 \\ &= (\beta_M - \beta_P)^2 \left[\frac{1}{2} \times \frac{1}{2} \right] + \sigma^2 \\ &= \frac{1}{4} (\beta_M - \beta_P)^2 + \sigma^2 \end{aligned}$$

Sachant que π et ϵ sont des variables aléatoires indépendantes.

Pour un individu homozygote B/B, la variance phénotypique est :

$$\begin{aligned} \text{Var}(y_{BB}) &= \text{Var}(\mu + \epsilon) \\ \text{Var}(y_{BB}) &= \text{Var}(\epsilon) \\ \text{Var}(y_{BB}) &= \sigma^2 \end{aligned}$$

La variance des individus homozygotes est la même ($\text{Var}(y_{BB}) = \text{Var}(y_{AA})$). Cependant, en présence d'empreinte parentale, i.e $\beta_M \neq \beta_P$, la variance des individus hétérozygotes est plus grande que la variance des individus homozygotes ($\text{Var}(y_{AB}) > \text{Var}(y_{AA}) = \text{Var}(y_{BB})$). il est donc possible d'essayer de détecter la présence d'empreinte parentale en réalisant un test de comparaison de variance, en testant l'hypothèse que la variance du phénotype chez les individus hétérozygotes est plus élevée que celle observée chez les individus homozygotes.

4.1.2 Test de Levene

En 1960, Howard Levene publie, suite à ses travaux, une approche paramétrique permettant d'évaluer l'égalité des variances pour deux groupes ou plus : le test de Levene [20]. Ce test est reconnu comme étant plus robuste au non-respect de la normalité des données, et offre donc une alternative intéressante par rapport au test de Bartlett [21]. Le test de Levene équivaut à une analyse de variance à un facteur (ANOVA) où la variable dépendante est la valeur absolue de la différence entre l'observation et la moyenne du groupe d'appartenance de l'observation. Pour détailler, la statistique usuelle du test de Fisher-Snedecor, utilisée lors d'une ANOVA, est

$$F = \frac{N - k}{k - 1} \frac{\sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y}_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2} \sim \mathcal{F}(k - 1)(N - k) \quad (1)$$

où

- Y_{ij} est la $j^{\text{ème}}$ observation du $i^{\text{ème}}$ groupe ;
- $\bar{Y}_i$ est la moyenne dans le $i^{\text{ème}}$ groupe ;
- $\bar{Y}_{..}$ est la moyenne générale ;
- n_i est le nombre d'observation dans le $i^{\text{ème}}$ groupe ;
- K représente le nombre de groupe.

or en posant $d_{ij} = |Y_{ij} - \bar{Y}_{i.}|$, on obtient, la statistique du test de Levene [21] :

$$W = \frac{N - k}{k - 1} \frac{\sum_{i=1}^k n_i (\bar{d}_{i.} - \bar{d}_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (d_{ij} - \bar{d}_{i.})^2} \sim \mathcal{F}(k - 1)(N - k) \quad (2)$$

Ainsi, la statistique du test de Levene suit une loi de Fisher-Snedecor de paramètres $(k - 1, n - k)$ et l'hypothèse H_0 est : "Les variances sont égales ($\sigma_1^2 = \dots = \sigma_k^2$), opposant l'hypothèse alternative H_1 , qu'au moins deux d'entre elles diffèrent.

Seulement pour augmenter la robustesse du test (être moins sensible aux valeurs extrêmes), deux modifications du test de Levene ont été proposées : le test de Levene tronqué à 25% et le test de Brown-Forsythe.

4.1.3 Test de Levene tronqué à 25%

Une des modifications du test de Levene qui semble être plus adaptée pour des échantillons de petite et moyenne tailles (plutôt balancés) est la version tronquée à 25% du test de Levene (25% trimmed en anglais). Il s'agit de remplacer la moyenne par une moyenne "tronquée" qui ignore ainsi les valeurs les plus extrêmes des données qui pourraient être considérées comme aberrantes. Pour calculer cette moyenne tronquée à 25%, il faut enlever 25% des observations les plus hautes et 25% des observations les plus basses et calculer la moyenne sur les observations restantes.

Les résultats obtenus avec cette modification sont très proches du test de Brown-Forsythe. Ce dernier sera uniquement utilisé dans la suite de ce travail.

4.1.4 Test de Brown-Forsythe

En 1974 Morton Brown et Alan Forsythe publient le test de Brown-Forsythe, une généralisation du test de Levene [22]. Le test de Brown-Forsythe utilise non pas la moyenne comme valeur de centrage mais un estimateur de localisation plus robuste : la médiane.

Ainsi la statistique du test de Brown-Forsythe est, à l'instar de celle de Levene :

$$W = \frac{N - k}{k - 1} \frac{\sum_{i=1}^k n_i (\bar{d}_{i.} - \bar{d}_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (d_{ij} - \bar{d}_{i.})^2} \sim \mathcal{F}(k - 1)(N - k) \quad (3)$$

Avec $d_{ij} = |Y_{ij} - \text{médiane}(Y_i)|$. La statistique de test de Brown-Forsythe suit la même loi que celle du test de Levene, et l'hypothèse H_0 est également la même.

Pour revenir à la détection de l'empreinte parentale entre le groupe des homozygotes (AA et BB) et celui des hétérozygotes (AB), le test de Brown-Forsythe sera donc utilisé afin de comparer si les variances sont statistiquement significativement différentes entre les 2 groupes.

Au lieu d'utiliser la variable Y_{ij} directement, *Hoggart C, et al* [19] ont suggéré de réaliser le test de Brown-Forsythe sur les résidus (et leur médiane) obtenus après ajustements sur des facteurs de confusion connus influençant la variable Y d'intérêt. Les résidus de cette régression ont ensuite été récupérés pour être utilisés comme la variable Y_{ij} . Ce qui permet de prendre en compte les médianes résiduelles par groupe qui sont ajustées sur l'âge et le sexe des individus.

Pour éviter l'inflation de la variance dans les effets marginaux du groupe des homozygotes (qui comprend les individus AA et BB), *Hoggart C, et al* [19] ont proposé de centrer la valeur phénotype de chaque génotype par leur valeur phénotypique médiane. De cette manière, les groupes AA et BB, chez les homozygotes, sont conservés. Cette nouvelle variable est notée z :

$$z_i = \begin{cases} y_i - \mu_{AA} & \text{si } g_i = AA \\ y_i - \mu_{AB} & \text{si } g_i = AB \\ y_i - \mu_{BB} & \text{si } g_i = BB \end{cases}$$

Ensuite par analogie entre une ANOVA et la régression linéaire, tester la différence de variance par le test de Brown-Forsythe revient à tester dans une régression linéaire l'effet d'une variable indicatrice définissant le statut hétérozygote d'un individu sur la variable z . Sachant que le test Wald d'une variable revient à une ANOVA à un facteur, le test de Wald de la variable centrée par la médiane permettra de retrouver la p-value du test de Brown-Forsythe. De cette manière, il est possible de passer par un test de Wald dont la statistique de test est égale à : $\hat{\beta}/SE_b \sim \mathcal{N}(0, 1)$. Ainsi, la pente de la régression est :

$$\hat{\beta} = \alpha_{het} - \alpha_{hom} \quad (4)$$

avec $\alpha_{het} = \frac{\sum_{i:g_i=AB} |z_i|}{n_{AB}}$,
et $\alpha_{hom} = \frac{\left(\sum_{i:g_i=AA} |z_i| + \sum_{i:g_i=BB} |z_i| \right)}{(n_{AA} + n_{BB})}$

Et l'erreur standard [1] est, à une constante près, le rapport entre la somme des carrés des résidus et la variance des X_i :

1. Fernandez MA. Modèle de régression linéaire simple, Université de Bordeaux, ISPED, 2016.

$$\begin{aligned}
 SE_{\beta}^2 &= \frac{RSS}{(n-1)S_{XX}^2} \\
 &= \frac{RSS}{(n-1) \sum_{i=1}^n (X_i - \bar{X})^2} \\
 &= \frac{\frac{1}{n} RSS}{(n-1) \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \\
 &= \frac{\frac{1}{n} RSS}{(n-1) \frac{1}{n} \sum_{i=1}^n (X_i^2) - \bar{X}^2} \\
 &= \frac{\frac{1}{n} RSS}{(n-1) \frac{n_{AB}}{n} - (\frac{n_{AB}}{n})^2} \\
 &= \frac{\frac{1}{n} RSS}{(n-1) \frac{n_{AB}}{n} (1 - \frac{n_{AB}}{n})} \\
 &= \frac{RSS}{(n-1) n_{AB} (\frac{n - n_{AB}}{n})} \\
 &= \frac{RSS}{(n-1) n_{AB} (\frac{n_{AA} + n_{BB}}{n})} \\
 &= \frac{RSS}{(n-1)} \cdot \frac{n}{n_{AB} (n_{AA} + n_{BB})}
 \end{aligned}$$

$$\begin{aligned}
 \text{avec } RSS &= \sum (z_i - \hat{z}_i)^2 \\
 &= \sum (z_i - \beta_0 - \beta x_i)^2 \\
 &= \sum (z_i - \bar{z} + \beta \bar{x} - \beta x_i)^2 \text{ car } \beta_0 = \bar{z} + \beta \bar{x} \\
 &= \sum (z_i - \bar{z} - \beta (x_i - \bar{x}))^2 \\
 &= \sum (z_i - \bar{z})^2 - 2\beta \sum (x_i - \bar{x})(z_i - \bar{z}) + \beta^2 \sum (x_i - \bar{x})^2 \\
 &= \sum (z_i - \bar{z})^2 - 2\beta^2 S_{XX} + \beta^2 S_{XX} \text{ car } \beta = S_{XY}/S_{XX} \\
 &= \sum (z_i - \bar{z})^2 - \beta^2 S_{XX} \\
 &= \sum (z_i - \bar{z})^2 - \beta^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\
 &= \sum (z_i - \bar{z})^2 - n\beta^2 (\frac{1}{n} \sum X_i^2 - \bar{X}^2) \\
 &= \sum (z_i - \bar{z})^2 - \beta^2 (\sum X_i^2 - n\bar{X}^2) \\
 &= \sum (z_i - \bar{z})^2 - \beta^2 (n_{AB} - n(\frac{n_{AB}}{n})^2) \\
 &= \sum (z_i - \bar{z})^2 - \beta^2 (n_{AB} - \frac{n_{AB}^2}{n}) \\
 &= \sum (z_i - \bar{z})^2 - \beta^2 (n_{AB} (\frac{n - n_{AB}}{n})) \\
 &= \sum (z_i - \bar{z})^2 - \beta^2 (\frac{n_{AB} (n_{AA} + n_{BB})}{n})
 \end{aligned}$$

où $\bar{z} = \sum_i \frac{|z_i|}{n}$

Pour chaque groupe sanguin, la p-value unilatérale a été calculée. En effet, l'hypothèse H_0 était que la pente soit plus grande que 0 pour que le groupe sanguin soit soumis à empreinte parentale ($\beta > 0$). Ainsi, pour les pentes positives, la p-value unilatérale a été calculée en divisant la p-value bilatérale par 2, tandis que pour les pentes négatives, la pente unilatérale est : $1 - \frac{\text{p-value bilatérale}}{2}$.

A l'issue de ce test, si le β est strictement positif alors le groupe des hétérozygotes aurait une variance plus grande par rapport au groupe des homozygotes. Cela impliquerait donc que le SNP étudié pourrait être sujet à empreinte parentale. A contrario, si le β est négatif ou nul, alors les observations ne seraient pas compatibles avec une hypothèse d'empreinte parentale pour le polymorphisme testé. Afin de conclure sur de tels propos, il est nécessaire de se fier à la p-value du test statistique. Cependant, si cette p-value n'est pas statistiquement significative cela veut dire que la différence ou l'association observée est vraisemblablement le fruit du hasard, ou bien qu'il y a un manque de puis-

sance dans l'étude (une taille de l'échantillon trop petit ou bien une taille d'effet à mettre en évidence trop grande).

4.1.5 Stratégie d'analyse : Méta-analyse

Dans le cadre de mon stage, le test de Brown-Forsythe a été appliqué sur 3 jeux de données pour déterminer si certains polymorphismes du gène ABO pouvaient présenter des effets avec empreinte dans le cadre de leur association avec les taux de vWF. Afin d'obtenir un effet commun des résultats, une méta-analyse pour chaque SNP va être effectuée.

Une méta-analyse est une synthèse, quantitative, des résultats à partir de plusieurs d'études répondant à une même question de recherche. Cela permet d'évaluer l'ensemble des preuves scientifiques disponibles. Plus le nombre de patients inclus dans une étude est important, plus le poids de cette étude au sein de la méta-analyse est grand. Le terme "méta-analyse" a été introduit en 1976 par Glass. Cette technique s'est tout d'abord développée dans le domaine des sciences sociales et de l'éducation avant de faire son apparition en médecine. Par rapport à l'analyse séparée de plusieurs études, la méta-analyse permet :

- d'augmenter la puissance statistique (par exemple, quand les études sont de trop petites tailles) ;
- apprécier la cohérence entre les études et permettre une meilleure estimation de l'effet réel ;
- de réconcilier des résultats apparemment discordants ;
- de synthétiser une somme d'informations parfois très importante ;

Il existe deux modèles de méta-analyse : un modèle à effets fixes et un modèle à effets aléatoires.

Méta-analyse à effets fixes

Dans le modèle à effets fixes, toutes les études estiment la même taille d'effet et une méthode très courante de la méta-analyse à effets fixes est la méthode de Mantel-Haenszel. Elle permet d'estimer l'effet commun en pondérant la moyenne des estimations de chaque étude par l'inverse de leur variance. De cette manière, les études plus importantes, qui comportent des écart-types plus petits, reçoivent plus de poids que les études plus petites qui comportent de écart-types plus importants. Ainsi, l'imprécision de l'estimation de l'effet est minimisée.

Soit $\mathcal{B}_j$ l'effet commun du $j^{\text{ème}}$ SNP :

$$\mathcal{B}_j = \frac{\sum_{i=1}^N \beta_{ij} w_{ij}}{\sum_{i=1}^N w_{ij}} \quad \text{où } w_{ij} = [\text{var}(\beta_{ij})]^{-1} \quad (5)$$

Et la statistique de test est $X_j^2 = \mathcal{B}_j^2 / V_j \sim \chi^2(1)$ où $V_j = [\sum_{i=1}^N w_{ij}]^{-1}$

Test d'hétérogénéité : Test de Cochran et Indicateur I^2

Pour tester la cohérence des effets alléliques entre les études du même SNP, il est possible de réaliser un test d'hétérogénéité : le test de Cochran. La statistique de test est :

$$Q_j = \sum_{i=1}^N w_i (\mathcal{B}_j - \beta_{ij})^2 \sim \chi^2(N_j - 1) \quad (6)$$

Une alternative au test de Cochran est l'indicateur I^2 qui mesure la proportion d'hétérogénéité dans les études qui ne peut pas être expliquée par le hasard. Les valeurs de I^2 égales à 25%, 50% et 75% représentent respectivement une hétérogénéité basse, modérée et forte.

$$I^2 = \frac{[Q_j - (N_j - 1)]}{Q_j} \quad (7)$$

N_j représente le nombre d'études où un effet allélique a été estimé.

Méta-analyse à effets aléatoires

En présence d'hétérogénéité des effets alléliques entre les études, il est préférable d'utiliser un modèle à effet aléatoire. Au lieu de considérer que chaque étude représente une estimation unique du "vrai" effet, le modèle aléatoire va admettre que l'effet réel peut varier d'une étude à l'autre. Par exemple, la taille de l'effet peut être plus élevée si les sujets sont plus âgés, plus instruits ou en meilleure santé ; ou si l'étude a utilisé une variante légèrement plus longue de l'intervention. De cette manière, les études incluses sont supposées être un échantillon aléatoire de la distribution des effets, et l'effet combiné estime l'effet moyen dans cette distribution. Les grandes études peuvent produire des estimations plus précises que les petites études, mais chaque étude estime une taille d'effet différente, et chacune de ces tailles d'effet sert d'échantillon de la population dont nous voulons estimer la moyenne. Par conséquent, par rapport au modèle à effets fixes, les poids attribués sous effets aléatoires sont plus équilibrés. Les grandes études sont moins susceptibles de dominer l'analyse et les petites études d'être banalisées [23].

L'effet allélique combiné de toutes les études au SNP est alors donné par :

$$\mathcal{B}_j^* = \frac{\sum_{i=1}^N \beta_{ij} w_{ij}^*}{\sum_{i=1}^N w_{ij}^*} \quad \text{où} \quad \begin{cases} w_{ij}^* = [\tau_j^2 + \text{var}(\beta_{ij})]^{-1} \\ \tau_j^2 = \max \left(0, \frac{Q_j - (N_j - 1)}{\sum_i w_{ij} - (\sum_i w_{ij}^2 / \sum_i w_{ij})} \right) \end{cases} \quad (8)$$

Avec la statistique $X_j^2 = \mathcal{B}_j^{*2} / V_j^* \sim \chi^2(1)$ où $V_j^* = [\sum_{i=1}^N w_{ij}^*]^{-1}$

4.2 Contexte des études

4.2.1 Projet MARTHA

Etude MARTHA

Le projet MARTHA (MARseille THrombosis Association) a été mis en place par le Pr. Pierre-Emmanuel Morange en 1994 à Marseille et les données ont été recueillies jusqu'en 2010. Chaque

patient venant consulter au centre de Thrombophilie de l'Hôpital de la Timone à Marseille pour une MTEV s'est vu proposé de participer à une étude d'épidémiologie génétique. Ces patients devaient présenter au moins un antécédent documenté de MTEV (TVP et/ou EP) en l'absence de facteur de risques majeur de MTEV. L'objectif de cette étude épidémiologique est de découvrir de nouveaux facteurs de risques génétiques de la MTEV et de ses phénotypes biologiques associés.

Le recueil des données a été réalisé lors de chaque consultation à l'aide d'un questionnaire. Différentes informations cliniques et biologiques ont été recueillies, de plus la détermination du groupe sanguin de chaque patient a été identifiée. Pour chaque participant, des dosages biologiques ont été prélevés (prise de sang, tubes citrates et tubes EDTA). Les tubes EDTA ont permis d'extraire l'acide désoxyribonucléique (ADN), tandis que les tubes citrates ont permis de quantifier grâce à différentes méthodes le taux de Willebrand (VWF).

Etude MARTHA12

L'étude MARTHA12 est une continuité de l'étude MARTHA. Les patients de l'étude MARTHA12 ont été recrutés entre 2010 et 2012 au centre de Thrombophilie de l'Hôpital La Timone, à Marseille, de la même façon que pour l'étude MARTHA. De plus, les mêmes polymorphismes du gène ABO ont été génotypés.

4.2.2 Etude FARIVE

FARIVE pour FActeurs RIques de la maladie thromboembolique VEineuse est une étude cas-témoins multicentrique de la région parisienne incluant des sujets traités en hospitalisation ou en ambulatoire pour un premier épisode de MTEV depuis janvier 2003. Cette étude est conduite par le Pr. Emmerich. Les cas sont des patients hospitalisés, ou venant en consultation externe pour un premier épisode de MTEV, confirmé par veinographie et échographie pour les cas de TVP et par tomographie scanner ventilation/perfusion ou angiographie pour les cas d'EP.

Les patients étaient exclus de l'étude s'ils présentaient au moins un de ces critères :

- mineurs,
- ayant déjà eu une MTEV,
- avec un diagnostic de cancer actif,
- avec des antécédents de malignité de moins de 5 ans,

A l'inclusion, les données de base comprenant des informations sur le premier évènement de MTEV, l'âge, les antécédents médicamenteux, les antécédents personnels et familiaux de maladie thrombotique, les facteurs de risques cardiovasculaires et l'utilisation d'hormones exogènes ont été enregistrées lors d'un entretien en face à face par un médecin ou un assistant de recherche, à l'aide d'un questionnaire standardisé.

A chaque cas inclus, un témoin hospitalisé ou se présentant aux consultations a été apparié sur l'âge, le sexe et l'hôpital. Il ne devait pas présenter de thrombose veineuse ou artérielle, de cancer, d'insuffisance rénale ou hépatique ou des antécédents de MTEV.

Seules les données concernant les cas ont été analysées car dans MARTHA et MARTHA12 les individus sont déjà des cas avec des taux de Willebrand déjà élevés.

4.2.3 Etude MARFAST

Afin de valider les résultats issus de la méta-analyse pour chacun des 5 polymorphismes testés obtenu par la méthode de Brown-Forsythe dans MARTHA, MARTHA12, FARIVE ; ceux-ci ont pu être vérifié dans des données familiales de la cohorte MARFAST.

L'étude MARFAST (MARseilles Family Study on venous Thrombosis) est une cohorte comprenant des familles des alentours de Marseille qui se sont rendues au centre de thrombophilie de l'Hôpital de La Timone entre 1986 et 2013. Une famille était incluse dans la cohorte si elle comprenait au moins deux membres présentant les caractéristiques suivantes : un individu ayant des antécédents personnels de MTEV confirmés ainsi qu'un test de thrombophilie positif ainsi qu'un proche présentant le même défaut. Tous les proches référés au centre de thrombophilie ont ensuite été inclus [24]. Au total, 2865 individus provenant de 656 familles ont été recrutés.

5 Résultats

L'ensemble de cette section est divisée en deux parties. Dans une première partie, l'identification des SNPs du gène ABO potentiellement soumis à l'empreinte parentale par la méthode de Brown-Forsythe est présentée, puis dans une seconde section, la validation de ces SNPs grâce à des données familiales est analysée.

5.1 Détection des groupes sanguins potentiellement soumis à empreinte

La fréquence allélique de 5 polymorphismes dans les trois jeux de données a été calculée. La représentation des différents allèles dans les trois cohortes a ainsi pu être illustrée.

TABLE 1 – Fréquences alléliques dans les trois cohortes : MARTHA, MARTHA12 et FARIVE.

	MARTHA (n=1375)	MARTHA12 (n=732)	FARIVE (n=546)
Allèles			
A1	0,34	0,31	0,31
A2	0,06	0,06	0,05
O1	0,49	0,52	0,53
O2	0,02	×	0,01
B	0,09	0,09	0,1

Le génotypage du groupe O2 n'a pas été obtenu dans la cohorte MARTHA12, mais on peut supposer que ça fait 2%.

Pour chaque SNP codant pour les différents allèles du groupe sanguin, il a été nécessaire de recoder les variables en s'inspirant de l'article de *Goumini et al* [14]. Cela a permis de définir les allèles du groupe sanguin du sujet à partir de la paire de base du SNP. La fréquence d'un allèle a été obtenue par le rapport entre le nombre d'exemplaires de cet allèle dans la population et le nombre total d'allèles dans la population.

Dans le Tableau 1, l'allèle O1 est le plus prépondérant, respectivement 49%, 52% et 53% chez MARTHA, MARTHA12 et FARIVE, suivi ensuite par l'allèle A1 avec une fréquence allélique variant entre 31% et 34% dans les trois cohortes. L'allèle le moins fréquent est l'allèle O2.

Si on compare cette répartition par rapport à la population générale française, on remarque bien qu'il y a plus d'individu de groupe A1 dans une population atteints de MTEV comparée à une population générale.

Les 3 bases de données (MARTHA, MARTHA12 et FARIVE) ont été étudiées et décrites dans le tableau 2 en illustrant seulement les variables nécessaires aux analyses. La cohorte MARTHA est celle qui contient le plus grand nombre d'individu (n=1375) suivi par MARTHA12 (n=732) et enfin FARIVE (n=546).

De manière générale, on remarque qu'il y a majoritairement plus de femmes que d'hommes. En effet, certaines études notent un risque légèrement plus élevé de MTEV chez les femmes [25]. En moyenne, les individus des trois cohortes confondues ont un taux de Willebrand de 165,36 UI/dL qui est caractéristique d'un sujet ayant une MTEV. Les individus malades ont un taux plus élevé et opposent les sujets sains où le taux est plutôt proche de 100 [25]. La moyenne d'âge des individus est relativement proche des 50 ans pour les trois cohortes.

TABLE 2 – Tableau descriptif de certaines variables recueillies en fonction des trois cohortes : MARTHA, MARTHA12 et FARIVE.

		MARTHA (n=1375)			MARTHA12 (n=732)			FARIVE (n=546)		
Variables		n (%)	Moyenne	ET ¹	n (%)	Moyenne	ET ¹	n (%)	Moyenne	ET ¹
Facteur Von Willebrand		1375	152,54	66,56	732	144,62	60,71	546	198,93	86,99
Age à l'inclusion (année)		1375	47,35	15,38	732	51,37	15,79	546	53,48	19,58
Sexe	Femme	907 (66,0)			402 (54,9)			329 (60,3)		
	Homme	468 (34,0)			330 (45,1)			217 (39,7)		
		n (%)	Moyenne ²	ET ³	n (%)	Moyenne ²	ET ³	n (%)	Moyenne ²	ET ³
rs2519093 (A1)	A1A1	131	166,45	62,54	68	172,57	57,25	46	201,07	75,77
	A1X	673	163,60	67,83	348	152,41	61,50	244	205,89	89,85
	XX	571	136,32	62,53	316	130,03	56,93	256	191,90	85,84
rs1053878 (A2)	A2A2	3	111,00	29,31	2	174,00	73,54	0		
	A2X	157	147,46	64,28	79	132,96	52,26	49	185,78	80,97
	XX	1215	153,30	66,88	651	145,94	61,55	497	200,22	87,53
rs8176719 (O1)	O1O1	288	125,13	62,19	180	117,86	50,28	154	191,84	90,16
	O1X	780	156,60	65,42	413	148,68	62,10	275	200,21	87,08
	XX	307	167,96	66,21	139	167,22	56,85	117	205,24	82,51
rs41302905 (O2)	O2O2	1	184,00					0		
	O2X	40	126,45	49,74				15	221,92	112,89
	XX	1334	153,30	66,88				531	198,28	86,19
rs8176749 (B)	BB	14	144,79	64,44	1	117,00		6	190,99	40,35
	BX	230	167,96	65,50	129	165,41	61,57	95	200,64	78,66
	XX	1131	149,50	66,42	602	140,21	59,68	445	198,67	89,21

1 : Ecart-Type.

2 : Moyenne du facteur Von Willebrand.

3 : Ecart-Type du facteur Von Willebrand.

Ce tableau (tableau 2), permet de mettre en évidence quelques hypothèses sur certains SNPs soumis à empreinte parentale. Par exemple, pour le SNP codant pour le groupe A1, on remarque que la variance des hétérozygotes (A1X) est toujours plus élevée que celle des homozygotes (A1A1 et XX), et ce pour les trois cohortes. Or une variance plus élevée chez le groupe des hétérozygotes peut traduire une empreinte parentale. Cette tendance pourra donc être vérifiée avec le test de Brown-Forsythe. Sur le tableau 2, on remarque que les variances du groupe des hétérozygotes pour le SNP A2 sont toujours inférieures aux variances du groupe des homozygotes. On devrait donc s'attendre à ce que le SNP codant pour le groupe A2 n'ait pas de résultat significatif au test de Brown-Forsythe.

Le test de Brown-Forsythe, qui permettait de détecter les SNPs soumis à empreinte parentale, a été appliqué pour les SNPs du gène ABO, sur les trois jeux de données. Les résultats obtenus sont présentés sur le tableau 3 et les méta-analyses issues de ces résultats sont détaillées dans le tableau 4. Quatre méta-analyses des 3 cohortes ont été effectuées en utilisant l'inverse de la variance (une pour chaque groupe sanguin, excepté le groupe O2). La méta-analyse du groupe sanguin O2 n'a pas été effectuée car la cohorte MARTHA12 n'avait pas les données codant pour le groupe O2.

En reprenant le tableau 3, on remarque que les résultats sont cohérents par rapport aux hypothèses émises grâce au tableau 2. Dans les situations où la variance du groupe des hétérozygotes est supérieure à celle du groupe des homozygotes, le test de Brown-Forsythe révèle une pente positive ce

qui implique une variance plus élevée dans le groupe des hétérozygotes. Dans les situations contraires, le test de Brown-Forsythe confirme également les hypothèses évoquées.

Une tendance est également présente chez certains SNPs du gène ABO, notamment pour les SNPs codant pour les groupes A1 et A2. En effet, dans les trois cohortes, le β est assez homogène. On constate, que pour le groupe A1, les β des 3 cohortes sont positifs, tandis que pour le groupe A2, ils sont négatifs. La méta-analyse réalisée pour le groupe A1 montre un effet commun à la limite de la significativité (p-value=0,077 cf tableau 4). Les résultats du tableau 4 montrent une p-value significative, pour le groupe O1, cependant le test d'hétérogénéité indique une p-value à la limite de la significativité du seuil α de 5% et donc pourrait être hétérogène (il y a de trop fortes variations dans les résultats des 3 études), ce qui rend ininterprétable le résultat de cette méta-analyse. En calculant l'indicateur $\mathcal{I}^2$, on remarque une hétérogénéité modérée (55%). Comme on le voit dans le tableau 3, l'association vient principalement de la cohorte MARTHA12, où le coefficient est de 8. Pour cette situation, il est donc préférable de réaliser une méta-analyse à effets aléatoires. Avec la méta-analyse à effets aléatoires, l'effet commun est égale à 2,55, cependant la p-value du test d'hétérogénéité reste encore proche du seuil de significativité α à 5% (p-value Q=0,166) .

TABLE 3 – Mesure d'association ajustée sur le sexe et l'âge, d'après le test de Brown-Forsythe entre le factor von Willebrand et les SNPs chez MARTHA, MARTHA12 et FARIVE.

SNPs	MARTHA (n=1375)			MARTHA12 (n=732)			FARIVE (n=546)		
	β^1	SE_β	p-value ²	β^1	SE_β	p-value ²	β^1	SE_β	p-value ²
rs2519093 (A1)	2,790	2,383	0,121	2,467	2,985	0,205	1,201	4,826	0,402
rs1053878 (A2)	-3,123	3,776	0,796	-4,732	4,774	0,839	-3,887	8,305	0,68
rs8176719 (O1)	1,900	2,402	0,215	8,015	2,961	0,004	-2,839	4,881	0,720
rs41302905 (O2)	-10,785	7,143	0,935				25,792	14,469	0,038
rs8176749 (B)	3,784	3,200	0,119	1,170	3,876	0,382	-13,786	6,249	0,986

1 : pente de régression d'après la formule 4

2 : p-value unilatérale.

TABLE 4 – Méta-analyse du test de Brown-Forsythe réalisée dans les trois cohortes : MARTHA, MARTHA12 et FARIVE.

SNPs	Méta-analyse			Test d'hétérogénéité	
	β combiné	IC à 95%	p-value ¹	Q Cochran	p-value Q
rs2519093 (A1)	2,474	-0,931 ; 5,88	0,077	0,088	0,957
rs1053878 (A2)	-3,758	-9,225 ; 1,709	0,911	0,070	0,965
rs8176719 (O1)	3,414	-0,002 ; 6,829	0,025	4,453	0,108
rs8176749 (B)	0,497	-4,002 ; 4,996	0,414	6,309	0,043

1 : p-value unilatérale.

Il y a tout de même quelques différences entre les résultats obtenus par le test de Brown-Forsythe (BF) (cf tableau 3) et le tableau descriptif (cf tableau 2). Pour le SNP O1, dans la cohorte FARIVE, la variance des hétérozygotes est supérieure à la variance des homozygotes (en faisant une moyenne pondérée des variances du groupe O1O1 et XX, on obtient : $Var(hom) = 86,86 < Var(het) = 87,08$) alors que le test de BF nous indique une pente décroissante. A l'inverse pour le SNP B, dans la cohorte MARTHA, la variance des homozygotes est supérieure à celle des hétérozygotes ($Var(hom) = 66,40 > Var(het) = 65,50$) alors que le test de BF indiquait une variance plus élevée pour le groupe des hétérozygotes ($\beta = 3,784$). Ces différences peuvent être expliquées par le fait que le test de BF

est ajusté sur des covariables. Pour vérifier si cette différence est bien due à l'ajustement, les variances des hétérozygotes et des homozygotes ont été ajustées sur l'âge et le sexe de l'individu. Pour obtenir ces variances ajustées, les résidus de la régression de l'âge du patient et son sexe sur le facteur von Willebrand ont été récupérés. Les variances ont ensuite été calculées sur ces résidus. Ainsi, pour le SNP O1, la variance ajustée des hétérozygotes était bien inférieure à celle des homozygotes et coïncidait avec les résultats du test de BF. De même pour le SNP B, la variance ajustée des homozygotes s'est révélée être inférieure à celle des hétérozygotes.

Pour conclure, s'il devait y avoir un polymorphisme parmi ceux étudiés qui pourrait être sujet à empreinte parentale, cela pourrait être le polymorphisme caractérisant le groupe sanguin A1. Pour chaque cohorte, le test de Brown-Forsythe a démontré une pente positive, même si la p-value du test n'est pas significative. Avec la méta-analyse, la p-value unilatérale semble très proche du seuil de significativité α de 5%. Il semble intéressant de regarder s'il y a des différences entre un allèle A1 hérité du père et un allèle A1 hérité de la mère, pour émettre une hypothèse forte sur l'empreinte parentale de ce SNP.

5.2 Validation des résultats dans des données familiales

Afin de corroborer les résultats obtenus précédemment (comme quoi le SNP codant pour le groupe sanguin A1 serait soumis à empreinte parentale), il semble intéressant de confirmer cette hypothèse dans des données familiales, i.e, dans des études où des individus de la même famille ont été interrogés et qu'il est possible de retrouver l'héritage de leurs allèles. Une comparaison statistique pourra être réalisée pour voir s'il existe bien une différence significative des taux de Willebrand chez les enfants ayant hérité l'allèle A1 de leur mère et ceux l'ayant hérité de leur père.

La base de données de la cohorte MARFAST comporte des données familiales. Pour chaque propositus interrogé, différents membres de la famille ont également intégré l'étude. Ainsi, parmi les 656 familles, 2865 individus font partie de la base de données. Cependant l'héritage des allèles du groupe sanguin n'était pas défini, un travail de data management a donc été nécessaire (voir Figure 7). Parmi les 2865 individus seulement 2143 sujets ont été conservés car le lien entre le propositus et la tierce personne était définie (par exemple : mère, père, fils, soeur, etc..), ce qui veut dire que certaines familles n'étaient pas complètes.

Le but est de constituer une famille nucléaire ou encore un trio (2 parents et 1 enfant) entre génération afin de connaître l'héritage des allèles. Par exemple, si le propositus est un enfant, il faut connaître le groupe sanguin d'au moins un de ces parents. Dans le cas où le propositus est un adulte, il faut avoir les informations d'un de ses enfants et de son/sa conjoint(e).

Ainsi, pour 166 familles, le propositus est un enfant, donc 166 trios, et pour les 490 autres familles il s'agissait d'un propositus parent. Pour ces 490 familles, seulement l'information sur un seul des parents est disponible actuellement et donc ces données n'ont pas été étudiées. Le jeu de données n'a pas été constitué pour faire des recherches sur l'empreinte parentale, il s'agissait plutôt de comprendre la transmission de la MTEV, le conjoint n'avait donc pas été interrogé.

Parmi les 166 trios avec un propositus enfant, 162 trios ont été identifiés en plus, ce sont les frères et soeurs du propositus. Ainsi, au lieu d'avoir uniquement 166 individus qui possèdent des informations sur au moins l'un des deux parents, 328 fratries peuvent être analysées. Sur ces 328 fratries, il a été possible de retrouver la transmission allélique chez uni-

quement 225 fratries. Pour certains enfants, le groupe sanguin était manquant, ou alors il n'était pas possible d'identifier l'allèle hérité quand il y avait le groupe sanguin d'un des deux parents renseignés ou bien quand les deux parents étaient hétérozygotes. A la fin, les données de 203 enfants étaient exploitables car elles possédaient la mesure du facteur Willebrand.

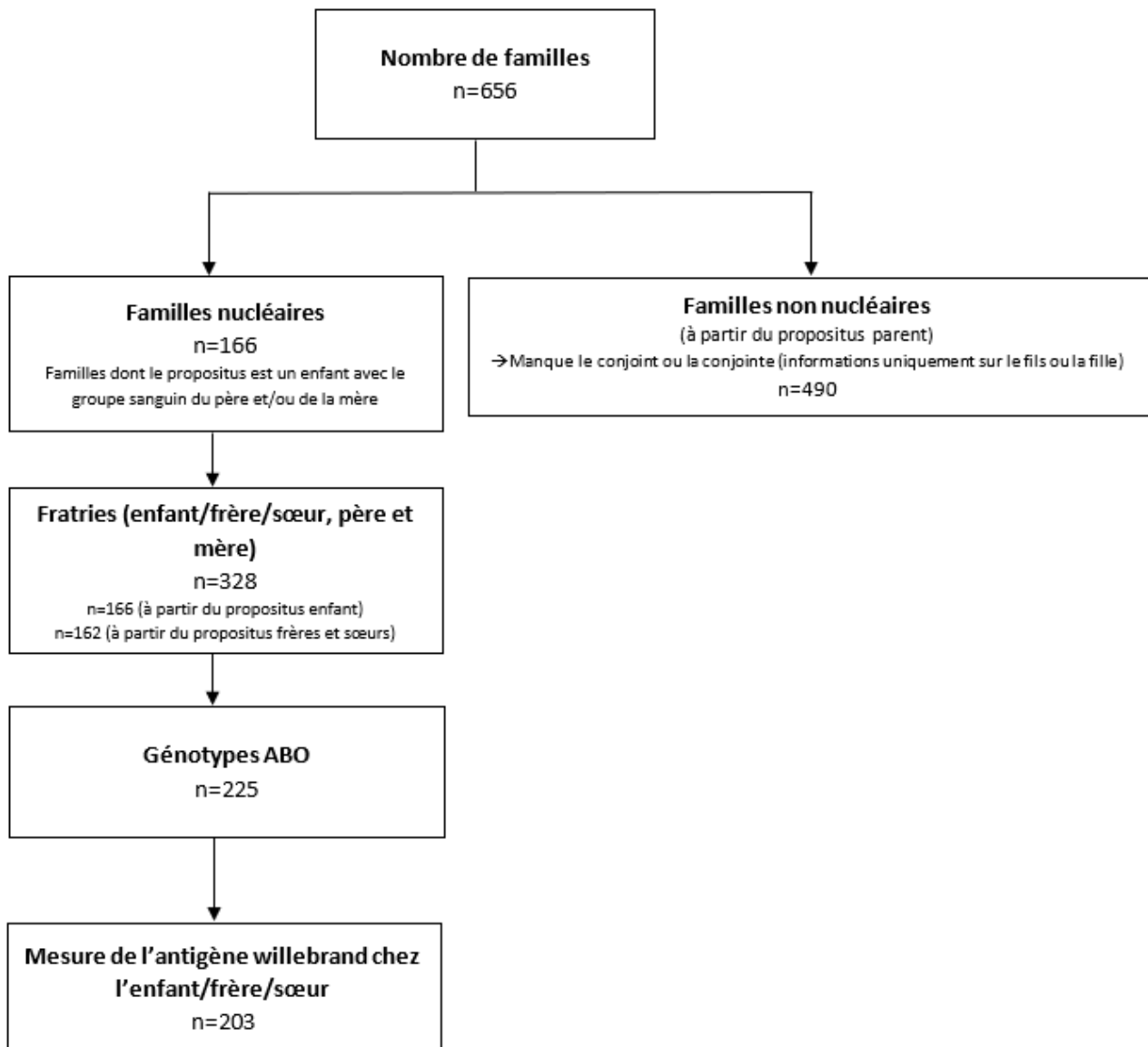


FIGURE 7 – Flow-chart de la cohorte MARFAST. (n=2865).

Pour illustrer la différence entre un allèle A1 hérité du père et un allèle A1 hérité de la mère, les analyses ont été effectuées uniquement sur les enfants hétérozygotes A1 parmi les 203 enfants, ce qui représente 83 enfants.

TABLE 5 – Mesure du facteur Willebrand en fonction du groupe sanguin A1, chez MARFAST (n=203), moyenne et écart-type.

	Non A1		Hétérozygotes A1				Homozygotes A1	
			A1 hérité du père		A1 hérité de la mère			
	n	Moy (ET) ¹	n	Moy (ET) ¹	n	Moy (ET) ¹	n	Moy (ET) ¹
Facteur von Willebrand	104	110,5 (54,28)	48	128,52 (66,76)	35	137,49 (36,21)	16	165 (96,59)

1 : Moyenne (Ecart-Type)

Dans le Tableau 5, on remarque que les homozygotes A1 ont une moyenne plus élevée par rapport aux hétérozygotes A1, qui ont eux-mêmes une moyenne plus élevée par rapport aux individus qui ne sont pas A1. Ce qui est cohérent puisque que le groupe A1 augmente le taux de Willebrand [14]. Si on s'intéresse uniquement à l'origine de l'allèle, on remarque que la moyenne des hétérozygotes qui ont hérité de l'allèle A1 de leur mère est plus élevée que ceux qui l'ont hérité de leur père. On peut donc se demander si cette différence est significative ?

Afin de calculer si cette différence est significative, il peut être envisageable de réaliser un test de comparaison de moyenne, cependant celui-ci ne tiendrait pas compte des facteurs confondants de la MTEV : l'âge et le sexe. Une régression linéaire ajustée sur l'âge et le sexe serait préférable, puis un test de Wald permettrait de tester la différence entre un allèle A1 hérité du père et hérité de la mère.

Dans des modèles linéaires, l'estimation des paramètres se fait sur la base d'observations indépendantes, alors que dans certaines situations, impliquant des données groupées, les réponses peuvent être corrélées. Dans la base de données MARFAST, étant donné que chaque enfant de la famille est associé à ses parents, chaque individu n'est pas indépendant. Pour tenir compte de ces corrélations, et pour corriger les estimations des coefficients de régression ainsi que ceux des variances et covariances, l'approche GEE pour Équations d'Estimation Généralisées sera utilisée [26].

Les équations d'estimations généralisées sont une extension des modèles de régression qui permettent de prendre en compte les corrélations entre les observations. On peut ainsi supposer une structure de corrélation de travail autre que l'indépendance. Pour compenser le fait que cette matrice puisse ne pas être la vraie structure de corrélation, la variance de l'estimateur $\hat{\beta}$ sera estimée par un estimateur de variance robuste.

En l'absence de corrélation, c'est à dire dans un modèle linéaire classique, la matrice est simplement la matrice identité. Une des extensions la plus simple de cette structure est la matrice de corrélation échangeable (aussi appelée commune ou égale) car elle fait l'hypothèse que la corrélation est la même dans toutes les familles.

Une transformation logarithme du facteur Willebrand a été réalisée permettant ainsi de se rapprocher d'une distribution normale. La régression avec l'approche GEE a ensuite été effectuée. Ainsi, à âge et sexe égaux, le taux de Willebrand est en moyenne plus petit quand c'est le père qui a transmis l'allèle A1 par rapport à une transmission maternelle de l'allèle A1, et cette différence est statistiquement significative ($\hat{\beta}=-0,145$; p-value=0,027). Mettre en évidence une différence statistiquement significative entre l'héritage parentale de l'allèle A1 peut permettre de confirmer la forte hypothèse sur une potentielle empreinte parentale sur ce SNP du gène ABO. Cependant, détecter une empreinte parentale reste compliqué. En effet, l'empreinte parentale met sous silence un gène, ce qui veut dire

que celui-ci ne s'exprime pas. En revanche le terme "effet parental" (*parent of origin effect* en anglais) implique juste que l'effet est différent selon que l'allèle a été hérité du père ou de la mère. Ces 2 mécanismes sont très intriqués, il est tout de fois plus simple de détecter un effet parental qu'une empreinte parentale. On ne pourra donc jamais être sûre qu'une différence statistique significative entre l'héritage parentale des allèles puissent être dû à une empreinte parentale, mais l'hypothèse qui peut en découler reste forte. Dans l'idéal, il faudrait que cette différence soit très forte.

Pour conclure, on a vu que le groupe A1 était à la limite de la significativité au test de Brown-Forsythe, ce qui pourrait être la preuve d'une empreinte parentale. De plus, dans une étude familiale, ce groupe a démontré une différence statistiquement significative entre l'allèle A1 hérité du père et l'allèle A1 hérité de la mère. Le père ayant une moyenne plus basse, on peut émettre l'hypothèse que chez les individus de groupe A1, une empreinte parentale paternelle peut être présente et que le risque de développer une MTEV pourrait être plus élevé si l'allèle A1 a été hérité de la mère plutôt que du père.

6 Discussion

L'objectif de mon stage était de détecter si un polymorphisme du gène ABO pouvait être soumis à empreinte parentale sur le Willebrand. Ce qui signifie que l'effet d'un allèle du gène ABO sur le facteur Willebrand pourrait varier selon qu'il a été transmis par le père ou par la mère.

Sur des données de sujets indépendants la méta-analyse des résultats obtenus par le test de Brown-Forsythe a permis d'identifier un effet commun positif sur le groupe A1. Ce qui permet de conclure que le groupe des hétérozygotes A1 aurait une variance plus élevée par rapport au groupe des homozygotes et donc que ce SNP pourrait être sujet à empreinte parentale sur le taux de Willebrand. Les données familiales de la cohorte MARFAST ont permis de confirmer ce résultat puisqu'une différence statistiquement significative des taux de Willebrand a été détectée entre les enfants ayant hérité de l'allèle A1 de leur mère et ceux l'ayant hérité de leur père. De plus il a pu être identifié que la mère serait le parent transmettant l'allèle qui augmenterait le taux de Willebrand. L'ensemble de ces résultats permet de montrer que la transmission de l'allèle A1 provenant de la mère augmenterait ce taux.

Dans les analyses effectuées sur le test de Brown-Forsythe pour détecter une empreinte parentale, il est clairement identifiable que la taille des groupes peut être très hétérogène. Généralement la taille du groupe des hétérozygotes est souvent moins élevée (si ce n'est très petite) comparée à celle des homozygotes. Or, différents auteurs ont remarqué que le test de Levene pouvait ne pas être adapté dès lors que la taille des groupes différait [21]. Il faudrait donc remplacer d_{ij} par $u_{ij} = d_{ij}/\sqrt{1-1/n_i}$ où n_i serait l'effectif du groupe i . Cette modification apportée au test de Levene pourrait également s'appliquer pour le test de Brown-Forsythe qui est très homologue à ce dernier.

Il faut préciser que le test de Brown-Forsythe utilisé pour détecter une empreinte parentale offre une alternative non négligeable pour détecter l'empreinte parentale sur des données non apparentées. Celle-ci reste quand même moins précise que l'identification d'une empreinte parentale dans les données familiales. En effet dans le test développé par *Hoggart C, et al* [19] il est possible de distinguer s'il y a un effet parental par la différence de variances entre le groupe des homozygotes et celui des hétérozygotes, mais il n'est pas possible de préciser si l'effet parental est maternel ou paternel (si c'est l'allèle hérité de la mère qui ne s'exprime pas ou si c'est celui du père).

Différentes perspectives sont à envisager et pourrait se réaliser par la suite : comme la possibilité d'étendre la méthodologie du test de Brown-Forsythe à l'analyse d'un trait binaire. Ce qui permettrait d'étudier, par exemple, l'impact de l'empreinte parentale sur le risque de la maladie. Pour cela, il serait envisageable d'appliquer les tests de Brown-Forsythe aux résidus obtenus à partir d'une régression logistique. Cette variable binaire serait régressée sur des covariables (tels que l'âge ou bien le sexe...) et les premières principales composantes dérivées de l'ensemble des polymorphismes génotypés comme cela est généralement fait dans les études d'association pan-génomique classiques.

Durant ce stage, seulement 5 polymorphismes ont été étudiés qui définissent le type de groupe sanguin, comme A1, A2 ou bien B. On peut également se poser la question s'il n'y a pas d'autres polymorphismes de gène ABO qui ne traduisent pas le type de groupe sanguin mais qui régulent l'activité de la protéine qui va déterminer le groupe sanguin et qui eux pourraient être sujets à empreinte parentale.

On remarque également que si l'effet paternel et l'effet maternel sont très différents, voir de signes opposés (par exemple : $\beta_M = -\beta_P$) alors le coefficient β sera très proche de 0. Ce qui voudrait dire, qu'en temps normal, dans les études pan-génomiques, l'identification d'un gène associé à une maladie pourrait ne pas être déterminée du fait que les signes des effets soient très différents. Il est donc probable, que des gènes qui n'ont pas été identifiés comme étant associés statistiquement significativement puissent être des gènes soumis à empreinte. Dans notre cas, les études pan-génomiques avaient identifié que le gène ABO était significativement associé à la MTEV. Les résultats obtenus dans ce projet montrent une susceptibilité d'empreinte parentale sur le SNP A1 mais qui est à la limite de la significativité. Il est probable que des gènes qui n'ont pas encore été identifiés comme significativement associés à la MTEV puissent montrer des résultats plus concluants. Il serait donc intéressant de se focaliser sur des polymorphismes connus dans le mécanisme de l'empreinte parentale : la méthylation de l'ADN.

Comme il a pu être dit précédemment, l'effet thrombotique du gène ABO se distingue par des modifications du facteur VIII et du facteur de Willebrand qui sont tous deux des protéines de coagulation qui peuvent augmenter le risque de développer une MTEV. Il serait donc intéressant de regarder si les résultats obtenus entre le gène ABO et le facteur de Willebrand se retrouve également sur le facteur VIII.

7 Conclusion

7.1 Conclusion sur les principaux résultats

Pour conclure, les études d'association pan-génomique sont des outils très puissants pour identifier des polymorphismes génétiques associés à des traits qualitatifs et quantitatifs. L'origine parentale des allèles transmis à un individu d'un polymorphisme n'est cependant pas pris en compte. Le test de comparaison de variance de Brown-Forsythe a été utilisé afin de détecter si de potentiel polymorphisme pouvait être sujet à une empreinte parentale dans des données non apparentées. La compréhension de cette méthodologie a été nécessaire pour pouvoir l'utiliser sur des données de personnes atteintes de maladies thromboemboliques veineuses. Le groupe sanguin A1 serait susceptible d'être soumis à une empreinte parentale sur le taux plasmatique de Willebrand. Ce qui signifie que l'effet de l'allèle A1 sur les taux de Willebrand pourrait être différent selon qu'un individu ait reçu l'allèle A1 de son père ou de sa mère.

Afin de vérifier cette hypothèse, l'empreinte parentale, qui pouvait déjà s'observer dans des données familiales, a pu être vérifiée dans les données issues de la cohorte MARFAST. Après quelques remaniements, et une prise en compte de la corrélation entre les sujets, une différence significative a été testée statistiquement. Celle-ci a permis de montrer que le taux de Willebrand était statistiquement significativement plus élevé quand l'allèle A1 a été hérité de la mère par rapport à un héritage paternel.

7.2 Conclusion personnelle

Ce mémoire synthétise ce que j'ai réalisé durant ce stage. Les dates de soutenances orale et écrite ont été anticipées étant donné que je souhaite continuer ce projet en thèse avec mon tuteur David-Alexandre TREGOUET. Ce stage a été très enrichissant (et il l'est toujours), j'ai appris énormément de notions notamment en génétique, où mes connaissances remontaient à celles acquises au lycée. J'ai été impressionnée par ce sujet qui m'était inconnu mais auquel David m'a initié. Je n'imaginais pas que la provenance parentale de nos allèles puissent avoir un impact sur notre santé. Mis à part l'intérêt que je porte à ce sujet que j'espère approfondir ces prochaines années, l'une des principales difficultés que j'ai rencontrée a été au niveau de "l'organisation" de mon travail. Il aurait été préférable pour moi de poser "à plat" les notions que je découvrais au fur et à mesure pour pouvoir prendre du recul sur les résultats, permettre une meilleure interprétation et ainsi de présenter de nouvelles suggestions.

Bibliographie

- [1] OMS. (page consultée le 31/05/2021). Faits marquants sur le vieillissement [en ligne].
<http://www.who.int/ageing/about/facts/fr/>
- [2] Ministère des Solidarités et de la Santé. (page consultée le 31/05/2021). Maladies cardiovasculaires - Maladies cardio-neurovasculaires. [en ligne]. 2021.
<https://solidarites-sante.gouv.fr/soins-et-maladies/maladies/maladies-cardiovasculaires/article/maladies-cardiovasculaires>
- [3] Arshad N, Isaksen T, Hansen J-B, Brækkan SK. Time trends in incidence rates of venous thromboembolism in a large cohort recruited from the general population. *Eur J Epidemiol.* 2017;32(4) :299-305.
- [4] Fry AE, Griffiths MJ, Auburn S, Diakite M, Forton JT, Green A, et al. Common variation in the ABO glycosyltransferase is associated with susceptibility to severe *Plasmodium falciparum* malaria. *Hum Mol Genet.* 2008;17(4) :567-76.
- [5] Inserm. (page consultée le 09/03/2021). La science pour la santé. [en ligne].
<https://www.inserm.fr/>
- [6] Bordeaux Population Health. (page consultée le 09/03/2021). Livret d'accueil Centre de recherche Inserm U1219. 2017. [en ligne].
https://issuu.com/valeriegarcia1/docs/livret_accueil_u1219-oct2017-5
- [7] VINTAGE | Bordeaux population Health. (page consultée le 09/03/2021). Epidémiologie intégrative et génétique du vieillissement cérébral et vasculaire. [en ligne].
<https://www.bordeaux-population-health.center/les-equipes/epidemiologie-integrative-et-genetique-du-vieillissement-cerebral-et-vasculaire-vintage/>
- [8] Toledano A. (page consultée le 15/03/21). Guide pratique de la maladie veineuse thromboembolique. [en ligne]. Paris. Éditions Med-Line.
<https://www.med-line.fr/catalogue/guide-pratique-de-la-maladie-veineuse-thromboembolique/>
- [9] Masson E. Recommandations de bonne pratique pour la prise en charge de la maladie veineuse thromboembolique chez l'adulte. Version courte [en ligne]. EM-Consulte. (page consultée le 17/03/2021). <https://www.em-consulte.com/rmr/article/1277640>
- [10] Morange P-E, Emmerich J, Tregouet D-A. Update on the genetic risk factors for venous thrombosis. *Sang Thrombose Vaisseaux.* 2010;22(8) :421-7.
- [11] Olié V. La maladie veineuse thromboembolique : étude des facteurs de risque de récurrence [Thèse]. Paris XI; 2012 :189. <https://tel.archives-ouvertes.fr/tel-00719318/document>
- [12] Bertina RM. Elevated clotting factor levels and venous thrombosis. *Pathophysiol Haemost Thromb.* 2003;33(5-6) :395-400.
- [13] Cushman M. Epidemiology and Risk Factors for Venous Thrombosis. *Semin Hematol.* 2007;44(2) :62-9.
- [14] Goumidi L, Thibord F, Wiggins KL, Li-Gao R, Brown MR, van Hylckama Vlieg A, et al. Association of ABO haplotypes with the risk of venous thrombosis : impact on disease risks estimation. *Blood.* 2021; 137(17) :2394-2402

- [15] Jick H, Westerholm B, Vessey MartinP, Lewis GeorgeP, Slone D, Inman WilliamHW, et al. Venous thromboembolic disease and ABO blood type : A Cooperative Study. *Lancet*. 1969 ;293(7594) :539-42.
- [16] Poujol-Robert A, Boëlle P-Y, Wendum D, Poupon R, Robert A. Association between ABO blood group and fibrosis severity in chronic hepatitis C infection. *Dig Dis Sci*. sept 2006 ;51(9) :1633-6.
- [17] Kabrhel C, Varraso R, Goldhaber SZ, Rimm EB, Camargo CA. Prospective Study of BMI and the Risk of Pulmonary Embolism in Women. *Obesity*. 2009 ;17(11) :2040-6.
- [18] Génome humain et médecine. (page consultée le 17 mai 2021). Dossier Pour la Science [en ligne]. 2005 ; (46)
<https://www.pourlascience.fr/sd/genetique/dossier-pour-la-science-46-563.php>
- [19] Hoggart C, Venturini G, Mangino M, Gomez F, Ascari G, Zhao JH, et al. Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index. *PLoS Genet*. 2014 ;10.
- [20] Levene H. Robust tests for equality of variances. In Ingram Olkin ; Harold Hotelling ; et al. (eds.). *Contributions to Probability and Statistics : Essays in Honor of Harold Hotelling*. Stanford University Press.Palo Alto. pp. 1960, 278–292.
- [21] Gastwirth J, Gel Y, Miao W. The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Stat Sci*. 2 oct 2010 ;24.
- [22] Brown, M.B. and Forsythe, A.B. (1974) Robust Tests for Equality of Variances. *J Am Stat Assoc*, 69, 364-367.
- [23] Borenstein M, Hedges L, Rothstein H. (page consultée le 11 mai 2021). Meta-Analysis Fixed effect vs. random effects. :162.
https://www.meta-analysis.com/downloads/M-a_f_e_v_r_e_sv.pdf
- [24] Suchon P, Resseguier N, Ibrahim M, Robin A, Venton G, Barthet M-C, et al. Common Risk Factors Add to Inherited Thrombophilia to Predict Venous Thromboembolism Risk in Families. *TH Open*. 2019 ;3(1) :e28-35.
- [25] Antoni G. Identification de facteurs génétiques modulant deux phénotypes intermédiaires de la maladie thrombo-embolique veineuse : les taux de facteurs VIII et von Willebrand : Intérêt de l'utilisation de différentes approches de recherche pangénomique [Thèse]. Sciences agricoles : Université Paris XI ; 2012. <https://www.theses.fr/2012PA11T019>
- [26] Seck B. Estimation pour les modèles linéaires généralisés : Approche marginale, approche conditionnelle et application. [Thèse]. Faculté des Sciences et de génie : Université Laval, Québec ; 2006 :63. https://archimede.mat.ulaval.ca/theses/B-Seck_06.pdf

Résumé

La maladie thromboembolique veineuse (MTEV) est la 3^{ème} maladie cardiovasculaire la plus fréquente et est responsable d'environ 10 000 décès chaque année en France. Elle résulte de la formation d'un thrombus (caillot de sang) dans le système veineux qui peut obstruer la circulation sanguine. Les dernières études d'association pan-génomiques ont identifié une trentaine de gènes associés à cette pathologie, dont le gène ABO qui est responsable du groupe sanguin. Les études d'association pan-génomiques permettent d'identifier les polymorphismes génétiques impliqués dans la variabilité d'un trait quantitatif ou qualitatifs, mais repose sur l'hypothèse d'absence d'empreinte parentale au niveau des polymorphismes testés. Cependant quelques travaux ont montré que l'origine de l'allèle peut augmenter le risque de développer une maladie s'il a été hérité du père ou de la mère. Ainsi, dans ce rapport, l'objectif était d'étudier l'empreinte parentale du gène ABO sur le taux plasmatique de Willebrand, une variable biologique qui peut augmenter le risque de développer une MTEV. Ainsi, grâce à une nouvelle méthodologie basée sur la méthode de Brown-Forsythe appliquée sur 2643 individus, le groupe A1 s'est révélé être potentiellement soumis à empreinte parentale sur le taux de Willebrand. Cette hypothèse a pu être validée dans une étude de données familiales où le taux de Willebrand était significativement plus bas quand un enfant a hérité l'allèle A1 de son père plutôt que de sa mère ; ce qui pourrait montrer que cette empreinte serait une empreinte parentale paternelle.

Mots-clés : Empreinte parentale, effet de l'origine parentale, polymorphismes du gène ABO, facteur von Willebrand, test de Brown-Forsythe, équations d'estimations généralisées, méta-analyse.

Abstract

The Venous ThromboEmbolism (VTE) is the 3rd most frequent vascular disease in Europe and is responsible of around 10 000 deaths per year in France. It results from blood clot forms (thrombus) which can obstruct blood circulation. Last genome wide association studies identified a thirties genes associated with this disease, including ABO gene which determines blood group. Genome wide association studies permit to identify genetics polymorphisms involved in quantitative or qualitative traits variances but is based on the assumption of parental origin effect absence. However some works have shown parental origin of the allele can increase thrombos disease risk depending if the allele is herited from the father or the mother. Thus, in this research intership, using a new methodology based on Brown-Forsythe test applied on 2 643 individuals, A1 group has been revealed potentially prone to parental imprinting on the von Willebrand factor (vWF). This hypothese is confirmed on a family data where vWF was significantly lower when a child has herited A1 allele from his father rather than his mother ; which can show that this parental origin effect is potentially paternal.

Keywords : Parental imprinting, parental origin effect, ABO gene polymorphisms, von Willebrand factor, Brown-Forsythe test, generalized estimation equations, meta-analysis.