UNIVERSITE PARIS XI

ECOLE DOCTORALE DE SANTE PUBLIQUE - ED420

Champ disciplinaire : Statistique et Santé

Année : 2012 N° attribué par la bibliothèque

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE PARIS XI
Spécialité : Santé Publique, option Génétique Statistique

Présentée et soutenue publiquement par
Guillemette Antoni
Le 25 avril 2012

Titre

IDENTIFICATION DE FACTEURS GENETIQUES MODULANT DEUX PHENOTYPES INTERMEDIAIRES DE LA MALADIE THROMBOEMBOLIQUE VEINEUSE : LES TAUX DE FACTEUR VIII ET DE VON WILLEBRAND

Sous-titre

Intérêt de l'utilisation de différentes approches de recherche pangénomique

Dirigée par David-Alexandre Trégouët et co-dirigée par France Gagnon

JURY

Mme Sophie Tezenas du Montcel	Rapporteur
M Grégoire Le Gal	Rapporteur
M Jean-Charles Lambert	Examinateur
M Pierre-Yves Scarabin	Examinateur
Mme France Gagnon	Co-Directrice
M David-Alexandre Trégouët	Directeur

Remerciements

Je remercie monsieur François Cambien et madame Laurence Tiret de m'avoir accueillie au sein de leur laboratoire INSERM UMRS_937 « Génomique Cardiovasculaire », ainsi que tous les membres de ce laboratoire dont les tempéraments s'accordent pour rendre l'U937 un lieu de travail agréable et paisible.

Je remercie les rapporteurs de ma thèse, madame Sophie Tézenas du Montcel et monsieur Grégoire Le Gal, pour leur relecture approfondie et minutieuse, autant critique que bienveillante. Je remercie les examinateurs, monsieur Pierre-Yves Scarabin et monsieur Jean-Charles Lambert, d'avoir accepté de consacrer de leur temps pour prendre part à mon jury.

Je remercie David Trégouët, mon directeur de thèse, pour avoir su établir une stratégie cohérente, intégrant les multiples bases de données qu'il a mises à ma disposition pour la réalisation de mon travail de thèse. Je te remercie, David, car malgré nos personnalités peut-être un peu divergentes, tu as tenu bon dans tes exigences tout en continuant à m'accorder ta confiance, lorsque je n'aspirais qu'à travailler à ma guise… et dans une direction qui ne convenait pas toujours à ta recherche d'efficacité et de productivité ! Un grand merci enfin pour ta réactivité notamment concernant les corrections de mon manuscrit, tes conseils avisés et ta parfaite gestion du temps.

Je remercie France Gagnon, ma co-directrice de thèse, qui a fait naître en moi le désir de réaliser dans le cadre d'une thèse un travail de recherche portant sur l'épidémiologie génétique de la maladie thrombo-embolique veineuse. Je lui suis très reconnaissante pour nos discussions fructueuses qui m'ont permis de mieux appréhender la question de la liaison génétique et m'ont donné les premières clés pour faire « rouler Loki ». France, mes remerciements viennent du fond du cœur pour ton enthousiasme, tes qualités humaines, ton écoute, tes encouragements, et tes attentions quasi maternelles lors de mes séjours à Toronto. Mes stages dans ton laboratoire, où tu sais si bien rendre tes « *trainees* » heureux de travailler, comptent parmi les périodes de ma thèse les plus épanouissantes.

Je remercie le professeur Pierre-Emmanuel Morange, professeur d'hématologie au CHU La Timone à Marseille et directeur de l'UMR_S 626 à Marseille. Sa soif de connaissance et sa grande curiosité intellectuelle sont à l'origine du projet MARTHA. Ses connaissances de la biologie de l'hémostase et son désir de les partager enrichissent substantiellement la réflexion issue d'un travail d'analyse statistique.

Je remercie le professeur Joseph Emmerich, directeur de l'UMR 765 à l'Hôpital Européen Georges Pompidou et responsable de l'enquête FARIVE dont les données ont été mises à ma disposition. Alors que je n'ai eu qu'une occasion de le croiser lors d'un congrès, je n'oublierai pas ses quelques mots d'encouragements, emplis d'humanité et de simplicité, qui m'ont été d'un grand réconfort.

Je remercie Noémie Saut pour la qualité du génotypage des échantillons STANISLAS, MARTHA05 et FARIVE qu'elle réalise avec diligence au sein de l'UMR_S 626 à Marseille.

Je remercie Marine Germain, Ingénieur de Recherche à l'UMR_S 937, pour sa parfaite gestion des bases de la GWAS *in silico* et des bases de MARTHA08 et MARTHA10. Son sens de l'organisation, sa rigueur et sa disponibilité m'ont permis de facilement retrouver mes petits lors de diverses opérations de sélections de SNPs.

Je remercie vivement Apostolos Dimitromanolakis et Simon Luo, bio-informaticiens dans le laboratoire de France Gagnon, qui m'ont aidée, avec une extrême gentillesse, à écrire les scripts nécessaires à l'utilisation des logiciels Loki et Solar.

Je remercie Tiphaine Oudot-Mellakh, qui a réalisé durant son stage post-doctoral à l'UMR-S 937 plusieurs analyses GWAS de MARTHA08 et MARTHA10. Sa bonne humeur à toute épreuve et sa simplicité dans ses rapports aux autres ont fait naître de notre travail d'équipe sur les études d'associations génome-entier une grande confiance réciproque, et ont été pour moi une immense source de plaisir.

Je remercie mon relecteur, correcteur d'orthographe et rectificateur de phrase alambiquée, Roman Pétrouchine, qui a accepté de laisser de côté sa plume littéraire, pour s'adapter à la sécheresse du style scientifique et son horrible passé composé. Cher Roman, tes grands éclats de rire, à la lecture des méthodes et résultats de cette thèse, bien que m'apparaissant parfaitement mystérieux, me permirent de retrouver un sens à ce travail au moment où je l'en croyais dénué.

Je remercie également ma relectrice scientifique, Sophie Garnier, MCU qui effectue son travail de recherche à l'UMR_S 937, pour sa relecture attentive et ses suggestions pertinentes.

Si j'ai pu surmonter les quelques moments difficiles de ces années de thèse, c'est en grande partie grâce à l'équipe des doctorants (ou ex-doctorants, maintenant) de l'UMR_S 937. Ce fut pour moi une grande chance de côtoyer quotidiennement Raphaële Castagné, Maxime Rotival et Nicolas Greliche. Merci à Maxime, qui non seulement est mon dernier ami à ne pas posséder de téléphone portable, mais a eu la patience et le mérite de parvenir à me faire comprendre les

principes des statistiques bayésiennes, depuis Londres, via la messagerie instantanée de gmail. Merci également à Nicolas, dont l'humour absolument unique et totalement absurde relève souvent du génie, et dont la générosité s'est exprimée à maintes reprises, qu'il s'agisse de prendre le temps de régler mes diverses tracasseries informatiques ou encore d'entraîner physiquement une équipe de Charlies-Patte-Folle pour l'accompagner jusqu'au bout d'une course de 10 km . Merci enfin à Raphaële pour sa joie de vivre si communicative !

Je remercie le professeur Laurence Meyer, directrice du service d'Epidémiologie et de Santé Publique de l'hôpital de Bicêtre, chez qui je travaille actuellement, et qui m'a prodigué de nombreux encouragements pour achever ma thèse dans des délais raisonnables. Je la remercie pour sa compréhension et pour tout le temps qu'elle m'a laissé libre afin de terminer dans de bonnes conditions ce travail.

Je remercie enfin ma famille et mes amis pour leur fidèle soutien et pour avoir adouci ou égayé ces dernières années. Dans l'ordre alphabétique (et non selon une quelconque importance), je remercie : les amateurs d'os à moelle et de fonction exponentielle, les as du kouglof, les asthmatiques allergiques au plastique des tentes Décathlon, les bébés nés à la Saint Parfait et leur maman, les buveurs de piscola, les cachotiers du vendredi midi, les champions en herbe de Hockey sur glace et leur grand frère artiste dessinateur, les colleurs d'affiches, les duettistes qu'ils soient flûtistes, violonistes, violoncellistes ou encore chanteurs, les éleveurs de chèvres, les flûtistes de l'opéra de Paris, les « graines de poète » semeurs de Haïkus, les grimpeurs de Roc 14, les hallucinées et leur compagnon en habit de cosmonaute, les hôtes de grande sérénité dont la convivialité (et les ponchos) compensent largement les insuffisances de la chaudière, les inconditionnelles de DSLZ, les internes et leurs gentilles attentions (merci pour le muffin anti-déprime !), les judokas 5ème dan, les piliers du Bar du Marché, les professeurs de harpe, les tamponneuses de cookies, les zélateurs de la décroissance… Ils se reconnaîtront ! Je souhaiterais fermer cette page de remerciements en exprimant toute mon affection pour mes petits parents, à la fois tellement aimants et tellement discrets… Votre soutien m'est toujours aussi précieux…

Résumé

La Maladie Thrombo-Embolique Veineuse (MTEV) est une maladie dont les facteurs de risque sont à la fois environnementaux et génétiques. Les facteurs de risque génétiques bien établis sont les déficits en anti-thrombine, en protéine S, en protéine C, la mutation du Facteur V de Leiden (FVL), la mutation du Facteur (F) II G20210A, ainsi que le gène *ABO* dont les allèles A1 et B augmentent le risque de MTEV par rapport aux allèles A2 et O. Alors qu'une part importante de l'héritabilité de la MTEV reste inexpliquée, les études contemporaines se heurtent à un manque de puissance pour découvrir de nouveaux facteurs génétiques dont les effets sont de plus en plus faibles.En vue d'augmenter la puissance de détection de nouveaux gènes de susceptibilité à la MTEV, j'ai recherché les déterminismes génétiques de deux de ses phénotypes intermédiaires : les taux d'activité plasmatique du FVIII et les taux d'antigénémie de sa protéine de transport, le Facteur de von Willebrand (vWF).

Dans un premier temps, j'ai réalisé une analyse de liaison des taux de FVIII et de vWF à partir d'un échantillon de cinq grandes familles franco-canadiennes (totalisant 255 personnes) recrutées *via* un cas de MTEV avec mutation FVL. Quatre régions liées aux taux de FVIII et/ou vWF ont été identifiées. L'une de ces régions correspondait au locus du gène *ABO* déjà connu pour influencer les taux de FVIII et vWF. La recherche de gènes candidats au sein des autres signaux de liaison s'est effectuée par l'étude *in silico* d'une analyse d'association pangénomique de la MTEV incluant 419 cas et 1228 témoins. Deux gènes candidats ont été identifiés : *STAB2* et *BAI3*. J'ai ensuite réalisé des études d'associations de cinq polymorphismes de *BAI3*. L'un d'entre eux était d'une part associé à une élévation des taux de vWF (résultat obtenu dans un échantillon de 108 familles nucléaires en bonne santé et reproduit dans un échantillon de 916 patients non apparentés atteints de MTEV), et d'autre part associé au risque de survenue de MTEV parmi les sujets non porteurs de mutations FVL et FII de deux échantillons cas-témoins (respectivement 916 cas et 801 témoins, et 250 cas et 607 témoins). Quant à *STAB2*, durant le courant de ma thèse, deux de ces polymorphismes ont été décrits comme associés aux taux de FVIII et vWF au cours d'une vaste étude d'association pangénomique (GWAS) menée par le consortium CHARGE rassemblant 23 600 personnes.

Dans un second temps, j'ai réalisé une méta-analyse de trois GWAS des taux de FVIII et vWF. Ces analyses avaient été conduites avec l'échantillon des cinq grandes familles franco-canadiennes et deux échantillons de 972 et 570 patients atteints de MTEV. Elles étaient ajustées sur les polymorphismes du gène ABO permettant de distinguer les allèles A1, A2, B et O, dans l'optique d'augmenter la puissance des analyses en diminuant la variance résiduelle des phénotypes. Aucun polymorphisme n'était associé ni aux taux de vWF ni à ceux de FVIII après prise en compte de la correction de Bonferroni pour tests multiples ($p<10^{-7}$). Cependant, parmi les onze gènes qui présentaient des polymorphismes associés aux taux de vWF ou de FVIII avec une significativité $p<10^{-5}$, de manière intéressante se trouvait *STAB2*. Cette étude a de plus permis de confirmer les associations nouvellement découvertes de polymorphismes situés dans les gènes *VWF*, *STXBP5* et *STX2*.

Mots-Clés : Maladie thrombo-embolique veineuse, Facteur VIII, Facteur de von Willebrand, analyse de liaison génétique, analyse d'association pangénomique (GWAS)

PRODUCTION SCIENTIFIQUE

Articles Publiés

Antoni G*, Oudot-Mellakh T*, Dimitromanolakis A, Germain M, Cohen W, Wells P, Lathrop M, Gagnon F, Morange PE, Tregouet DA. Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. *BMC Med. Genet*. 2011;12:102.
* Co-premiers auteurs

Antoni G, Morange PE, Luo Y, Saut N, Burgos G, Heath S, Germain M, Biron-Andreani C, Schved JF, Pernod G, Galan P, Zelenika D, Alessi MC, Drouet L, Visvikis-Siest S, Wells PS, Lathrop M, Emmerich J, Tregouet DA, Gagnon F. A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism. *J. Thromb. Haemost*. 2010 déc;8(12):2671-2679.

Contribution des travaux de la thèse à d'autres articles

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.

Morange P-E, Saut N, **Antoni G**, Emmerich J, Trégouët D-A. Impact on venous thrombosis risk of newly discovered gene variants associated with FVIII and VWF plasma levels. **J**. *Thromb. Haemost.* **2011** janv;9(1):229-231.

Communication orale

G.**Antoni** : A multi-stage strategy identifies a new QTL for von Willebrand Factor on chromosome 6: a possible link with Venous Thromboembolism (VTE)? International Society of Thrombosis and Haemostasis, 2009

G.**Antoni** : Identification of genetics factors influencing plasmatic levels of FVIII and von Willebrand Factor, two intermediary phenotypes of Venous Thrombosis Diseases, Journée de l'IFR 2011

Communication affichée

G.**Antoni**, T.Oudot, A.Dimitromanolakis, M.Germain, W.Cohen, P.Wells, M.Lathrop, F.Gagnon, P-E.Morange, D-A.Tregouet : New candidate loci modulating vWF and FVIII plasma levels: results from a meta-analysis of three GWAS in selected samples, International Society of Thrombosis and Haemostasis, 2011

G.**Antoni**, N.Saut, Y.Luo, P.S. Well, J.Emmerich, D.A.TrégouëtP.E.Morange, F.Gagnon, : A multi-stage strategy identifies a new QTL for von Willebrand Factor on chromosome 6: a possible link with Venous Thromboembolism (VTE)? American Society of Human Genetic 2009 et International Genetic Epidemiology Society 2009

LISTE DES PRINCIPALES ABREVIATIONS

BF : Facteur Bayésien

EE Equation d'Estimation

FVIII : Facteur VIII

FDR : False Discovery Rate out aux de faux positifs

FVL : Facteur V Leiden

GWAS : GenomeWide Association Study ou étude d'association pangénomique

IBD : identique par descendance

MAF : fréquence de l'allèle rare

MCMC : Méthode de Monte-Carlo par Chaîne de Markov

MTEV : Maladie Thrombo-Embolique Veineuse

QTL : Locus de Trait (ou caractère) Quantitatif

TQ : Trait (ou caractère) Quantitatif

VC : Décomposition de la Variance (Variance Component)

vWF : Facteur de von Willebrand

TABLE DES MATIERES

INTRODUCTION

I. La Maladie Thromboembolique Veineuse : définitions, épidémiologie, physiopathologie

I.1. Définitions

La Maladie ThromboEmbolique Veineuse (MTEV) recouvre deux entités nosographiques : la thrombose veineuse profonde (communément appelée phlébite), et sa principale complication, l'embolie pulmonaire. La thrombose veineuse profonde est la conséquence de la formation d'un caillot sanguin dans une veine profonde, le plus souvent située dans un membre inférieur, parfois dans un membre supérieur, voire dans l'abdomen, ou encore, de façon exceptionnelle, dans le cou ou la tête. Tout ou partie du caillot sanguin est susceptible de se détacher de la paroi du vaisseau, constituant ainsi un embole. Celui-ci, entraîné par la circulation sanguine, migre dans le cœur droit, puis dans l'artère pulmonaire. Il risque alors de boucher cette dernière, ou l'une de ses branches dont les calibres vont en s'amenuisant, constituant ainsi une embolie pulmonaire.

I.2. Epidémiologie générale

En raison de sa relative fréquence et de sa gravité en terme de mortalité et de complications à long terme, la MTEV a de lourdes conséquences sur la santé publique des pays dits "développés". D'après plusieurs travaux portant principalement sur des populations d'origine caucasienne [1][2][3] [4], (rev. dans [5]), l'incidence de la MTEV est estimée entre 1 et 2‰ par personne et par an. Environ un tiers des patients avec une MTEV symptomatique ont développé une embolie pulmonaire, tandis que les deux tiers restants n'ont présenté que des signes de thrombose veineuse [2][4][6]. Cependant, la proportion d'embolie pulmonaire est plus élevée dans les études incluant des données issues d'autopsies [1][3].

Le risque de récidive de MTEV est élevé. Selon les études, on observe de 6% à 10% de récidives à 6 mois[6][7][8], 7% à 13% à un an[9][4][8], et jusqu'à 30% à dix ans[7][8]. Le pronostic à court terme est sévère : d'après l'étude LITE (Longitudinal Investigation of Thromboembolism Etiology), la mortalité s'élève à 15% au cours du premier mois suivant une embolie pulmonaire [4]. La MTEV peut par ailleurs se compliquer d'un syndrome post-thrombotique dans 20% à 50% des cas (rev. dans [10]). Pathologie chronique probablement sous-diagnostiquée, le syndrome post-thrombotique cause une sensation de « jambes lourdes », et des douleurs chroniques souvent très importantes et invalidantes. En raison de l'inflammation et des phénomènes de revascularisation consécutifs à la MTEV, les valves

veineuses assurant le retour du sang vers le cœur et les poumons sont altérées. L'hypertension veineuse qui en découle provoque un œdème, une hypoxie des tissus sous-jacents, et dans les cas les plus graves, un ulcère cutané. Les conséquences sont lourdes sur la qualité de vie des patients et sur les dépenses de santé. Le bilan de la MTEV, en terme de santé publique, s'alourdit encore si l'on considère les risques iatrogènes que font courir les traitements anticoagulants : 1 à 3% des patients sous anticoagulants sont victimes d'hémorragie grave par an, dont plus de 10% sont fatales (revue dans [11]).

I.3. Physiologie de l'hémostase

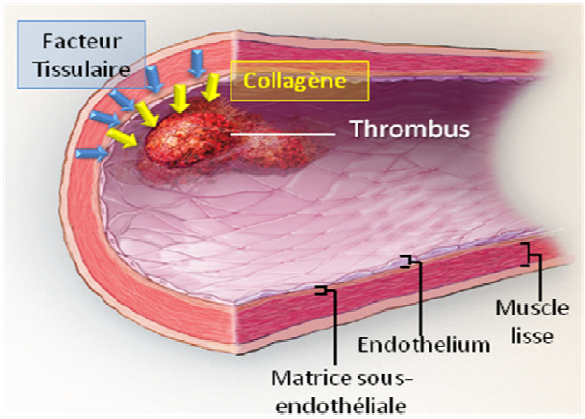
Cette section a été rédigée en grande partie à partir de l'article de revue « The role of procoagulants and anticoagulants in the development of venous thromboembolism » [12]. *Certains points de détails ont été éclaircis par l'article de revue "Mechanisms of Thrombus Formation"* [13]*, dans lequel les lecteurs intéressés par les études de fonctionnalité des multiples acteurs de l'hémostase trouveront de nombreuses références.*

L'hémostase est un phénomène physiologique qui interrompt les saignements.

I.3.1. L'hémostase primaire

Elle intervient dans les secondes suivant une brèche vasculaire. La première réponse est une vasoconstriction. Elle diminue le flux sanguin et entraîne de microturbulences favorisant les réactions hémostatiques à venir. Par ailleurs, le collagène, situé sous l'endothélium de la paroi vasculaire, se trouve alors en contact avec le sang. Il induit une activation des plaquettes qui s'agrègent alors sur leurs ligants, le facteur de Willebrand (*von Willebrand Factor* ou vWF) et le fibrinogène, pour former le clou plaquettaire ou thrombus (**figure 1**).

Figure 1 : Formation d'un thrombus en réponse à une brèche vasculaire (d'après [13])



Le collagène et le Facteur Tissulaire sont des constituants de la paroi vasculaire. Le Collagène (flèches jaunes), situé dans la matrice sous-endothéliale, et le Facteur Tissulaire (flèches bleues), localisé dans la média (muscle lisse) et dans l'adventice, ne sont pas au contact du sang dans des conditions normales, mais seulement après effraction de la paroi vasculaire. Le collagène est la première ligne de défense anti-hémorragique par activation de l'hémostase primaire, tandis que le facteur tissulaire est la deuxième ligne, par initialisation de l'hémostase secondaire.

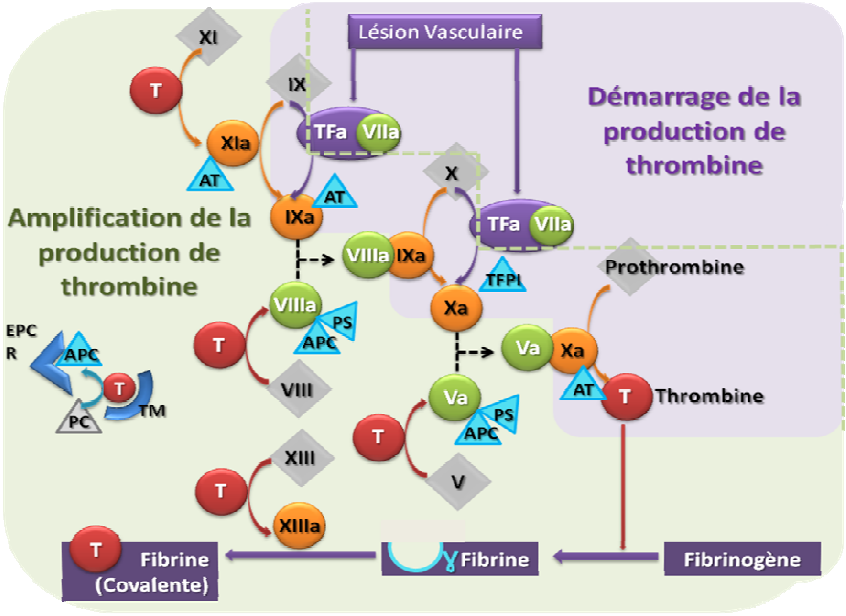
Le facteur de Willebrand est une glycoprotéine dont le gène, situé en 12p13, est exprimé par les mégacaryocytes et les cellules endothéliales. Il circule sous forme de multimères dont la taille est régulée par protéolyse par la protéase ADAMTS13. Il interagit avec les plaquettes *via*

le récepteur plaquettaire GP1bα. L'agrégation plaquettaire est d'autant plus importante que les multimères de vWF sont grands. Le clou plaquettaire formé au cours de l'hémostase primaire est fragile et doit être consolidé par un réseau de fibrine, qui est constitué au cours de l'hémostase secondaire.

I.3.2. L'hémostase secondaire

Dans un deuxième temps, l'hémostase secondaire permet de consolider le clou plaquettaire grâce à la formation d'un réseau de fibrine. Elle fait intervenir et interagir de nombreux facteurs procoagulants et anticoagulants. Cette réaction en chaîne est connue sous le nom de « cascade de la coagulation ». La thrombine (ou Facteur II activé, FIIa) en est la protéine centrale. C'est elle qui transforme le fibrinogène en fibrine. Par ailleurs, comme nous allons le voir plus en détail, elle est omniprésente dans la cascade de la coagulation, en tant qu'activatrice de nombreux facteurs et co-facteurs. Il existe trois phases dans le processus de production de la thrombine : une phase de démarrage (appelée parfois voie extrinsèque), une phase d'amplification (appelée parfois voie intrinsèque), et une phase d'arrêt. La figure 2 illustre les prochains paragraphes. Elle existe également en annexe sur une page dépliable afin d'être visible tout au long de la lecture.

Figure 2 : Cascade de la coagulation



La cascade de la coagulation démarre suite au contact du Facteur Tissulaire (TF) avec le sang, induit par une brèche vasculaire. Dans un premier temps (phase de démarrage ou voie extrinsèque, sur fond mauve), le complexe FVII-TF active FIX et FX. La coagulation est maintenue dans un second temps par des réactions initiées par le facteur IXa (phase d'amplification ou voie intrinsèque, sur fond vert). Les deux voies convergent vers une même voie dans laquelle la Prothrombine est convertie en thrombine. Cette dernière modifie le fibrinogène en fibrine. Celle-ci est est stabilidée par FXIIIa

I.3.2.1. Démarrage de la production de thrombine

Suite à un traumatisme de la paroi vasculaire, le sang entre en contact avec le Facteur Tissulaire (TF), qui forme un complexe activé avec le facteur FVII (TF-VIIa). Le complexe TF-VIIa active à son tour le FX. Le FXa active la prothrombine (FII) en thrombine (FIIa). Cette

voie de production de la thrombine est directe et très rapide. On la compare à un « starter ». Le complexe TF-VIIa active également, de façon plus lente, mais finalement prépondérante, le FIX. Ce dernier participe également à l'activation de FX, permettant ainsi d'amplifier la formation de thrombine.

Les FIXa et FXa agissent de concert avec leur cofacteur respectif, les FVIII et FV. Sans activité enzymatique propre, les deux co-facteurs FVIII et FV (sous forme inactive dans un premier temps) forment un complexe avec respectivement FIXa et FVa. Quand ils sont activés, FVIIIa et FVa renforcent considérablement (jusqu'à 100 fois) l'activité des FIXa et FXa. Cependant, leur activation est provoquée par la thrombine, absente au début du processus hémostatique. Ainsi, leur présence ne devient efficace qu'à la fin de la phase d'initiation, et leur action caractérise essentiellement la phase d'amplification. Le FVIII est stabilisé par sa protéine de transport, le vWF, sans laquelle il est rapidement clivé par des sérines protéases.

Cette phase de démarrage est contrôlée, en premier lieu, par les réserves limitées en TF. De plus, l'Inhibiteur de la voie (Pathway) du Facteur Tissulaire (TFPI) bloque l'action du FXa par la formation d'un complexe Xa-TFPI. Ce dernier bloque à son tour l'action du complexe TF-VIIa par la formation d'un complexe TF-VIIa-Xa-TFPI. La propension de TFPI à se complexer à FXa est potentialisée par la Protéine S. Enfin, l'activité protéasique de l'antithrombine (AT ou parfois AT III) désactive la thrombine, les FXa et FIXa.

I.3.2.2 Amplification de la production de thrombine :

Cette phase est plus lente à se mettre en place que la phase de démarrage, car elle nécessite qu'il y ait déjà eu production de thrombine. Au cours de cette phase, la thrombine active de plus en plus les co-facteurs FV et FVIII. La thrombine joue ainsi un rôle de rétro-contrôle positif permettant de maintenir et d'amplifier sa propre formation. Par ailleurs, FXIa, en activant FIX, se substitue au complexe TF-VIIa, qui n'intervient plus à ce stade. L'activation du FXI serait provoquée par la thrombine. La formation de thrombine devient autonome : c'est la thrombine elle-même qui agit en amont de la cascade. Sa formation devient indépendante du TF.

Finalement, la thrombine active les plaquettes et favorise leur agrégation. En outre, la thrombine transforme le fibrinogène (FI) soluble en fibrine (FIa) insoluble. Celle-ci se polymérise de manière à former un véritable réseau, emprisonnant le clou plaquettaire et les hématies. Après activation par la thrombine, FXIIIa stabilise le polymère de fibrine en créant des lésions covalentes entre chaque monomère.

La phase d'amplification de la formation de la thrombine est contrôlée par la fibrine elle-même, dont certains monomères possèdent un site de capture de la thrombine. Ce site est la conséquence d'un épissage alternatif du gène codant pour l'une des trois chaînes du fibrinogène, le fibrinogène γ. L'AT modère également l'ampleur de la production de thrombine en inhibant cette dernière et en diminuant les concentrations de facteurs FXIa, FXa et FIXa.

Plusieurs questions concernant la formation de la thrombine ne sont que partiellement résolues (rév dans [13]):

- Parmi elles, l'activation du FXI par la thrombine a été remise en cause. Si elle est bien observée *in vitro*, un doute subsiste quant à sa réalité *in vivo*. Ainsi, le maintien de la formation de la thrombine après amenuisement, puis disparition du TF, resterait sans explication.
- Une autre question en suspens concerne le FXII. Il est souvent présenté tout en amont de la cascade, activant le FXI ; son action ne pourrait être qu'un artefact *in vitro*, provoqué par le contact du sang avec la verrerie de laboratoire. En effet, si un déficit en facteur XII entraîne un allongement très important du temps de coagulation *in vitro*, il n'entraîne en revanche aucun risque hémorragique chez les patients.
- Les mécanismes de formation des complexes IXa-VIIIa et Xa-Va sont toujours inconnus. Ils nécessitent le support d'une membrane cellulaire. On a longtemps cru qu'il s'agissait de la membrane de plaquettes activées, jusqu'à ce que des expériences chez la souris invalident cette hypothèse.

I.3.2.3 Arrêt de la production de thrombine

L'antithrombine, nous l'avons vu, freine l'ensemble de la cascade, en inhibant l'action de la thrombine, de FXa, FIXa et FXIa. Cependant, l'arrêt complet du processus revient à la Protéine C Activée (APC). Elle inactive par protéolyse les deux cofacteurs FVa et FVIIIa. L'APC est issue du clivage de la protéine C, sous l'action de la thrombine. La protéine C est d'autant plus activée sous forme d'APC qu'elle est portée par le Récepteur Endothélial de la Protéine C (EPCR), et présentée à la thrombine, elle-même portée par un récepteur endothélial, la ThromboModuline (TM). L'action de l'APC est accrue par son cofacteur, la protéine S.

I.3.3. Fibrinolyse

La fibrinolyse (**figure 3**) permet la dissolution du caillot et dégage le vaisseau. Elle intervient quelques jours après la formation du réseau de fibrine, grâce à l'action protéolytique de la plasmine. Le plasminogène, précurseur inactif de la plasmine, présente une forte affinité

pour la fibrine. Il est incorporé au caillot sanguin dès sa formation. Ce n'est que quelques jours plus tard qu'il est transformé en plasmine grâce à l'Activateur tissulaire du Plasminogène(t-PA) et à l'urokinase. La plasmine fragmente la fibrine dont les résidus se dissolvent et sont éliminés par voie rénale ou hépatique.

Le contrôle de la fibrinolyse a plusieurs origines. L'Inhibiteur de Fibrinolyse Activable par la Thrombine (TAFI) élimine les sites de la fibrine qui lui permettent d'être reconnue par le plasminogène ou la plasmine. Par ailleurs, la plasmine est inactivée par l'α2-antiplasmine. Enfin, l'urokinase et le t-PA sont inhibés par les Inhibiteurs de l'Activateur du Plasminogène (PAI-1 et PAI-2).

Figure 3 : Schéma de la fibrinolyse

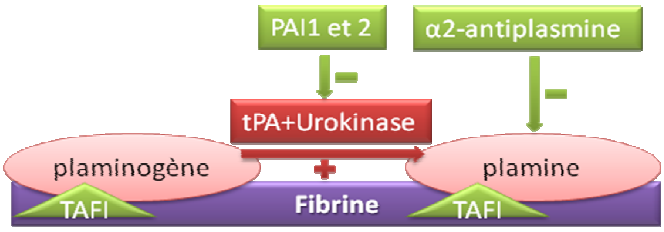


Tableau 1 : Récapitulatif des facteurs procoagulants, anticoagulants, profibrinolytiques et antifibrinolytiques.

Procoagulant	(1)	**Anti-coagulant**
Facteur Tissulaire (TF)		
FVIIa		Inhibiteur de la voie (Pathway) du Facteur Tissulaire (TFPI)
FXa		Anti-Thrombine (AT)
Thrombine		Fibrinogène γ'
FIXa		
FXIa		Protéine C activée (APC)
FVIIa		Protéine S (PS)
FVa		Récepteur endothélial à la protéine C (EPCR)
FVIIIa		ThromboModuline (TM)
FXIa		

Anti-fibrinolytique	**Pro-fibrinolytique**
Inhibiteur de la Fibrinolyse activable par la Thrombine (TAFI)	Plasmine
α2-antiplasmine	Activateur tissulaire du Plasminogène (t-PA)
Inhibiteurs de l'activateur du plasminogène (PAI-1 et PAI-2)	Urokinase

(1) Les flèches représentent la ou les protéines cibles des facteurs anti-procoagulants et des facteurs anti-fibrinolytiques

I.4. Physiopathologie de la MTEV

La physiopathologie des thromboses (veineuses ou artérielles), décrite dès 1856 par Virchow, repose sur trois éléments connus sous le nom de « triade de Virchow ». La formation d'une thrombose est favorisée par (1) une modification du flux sanguin, pouvant être soit une stase, soit une turbulence locale ; (2) un traumatisme de la paroi vasculaire ; (3) une modification de la composition sanguine perturbant l'équilibre entre la coagulation, l'anticoagulation et la fibrinolyse, à l'origine d'une hypercoagulabilité, encore appelée thrombophilie.

La formation du complexe TF-VIIa – à l'origine de la cascade de la coagulation - ne fait pas seulement suite à une brèche vasculaire, comme nous venons de le décrire au **§I.3**. Au contraire, le plus souvent, le TF provient d'une activation de l'endothélium, conséquence d'une inflammation. L'endothélium libère des granules contenant du vWF et de la P-selectin. Ces protéines ont la capacité de recruter des leucocytes, dont certains (les monocytes, particulièrement) sécrètent du TF. La stase veineuse induit également une activation de l'endothélium, en favorisant une hypoxie par désaturation des globules rouges en hémoglobine. De plus, elle favorise l'accumulation de facteurs prothrombotiques. En effet, la baisse de flux sanguin diminue l'acheminement de la thrombine vers son principal lieu de désactivation, le lit capillaire pulmonaire, tapissé de thrombomoduline. Enfin, elle ne favorise pas le déploiement de la molécule du vWF, et rend ainsi peu accessible les sites de clivages par la protéase ADAMTS13.

La MTEV survient lors de situations favorisant un ou plusieurs éléments de la triade de Virchow. La section suivante en présente les principaux facteurs de risque.

II. Etiologie de la MTEV

La MTEV est une maladie multifactorielle (ou "complexe"): ses facteurs de risque sont à la fois génétique et environnementaux (au sens large du terme, soit tout ce qui n'est pas génétique). Différentes études, réalisées à partir d'échantillons composés soit de paires de jumeaux [14], soit de familles plus élargies [15][16], ont estimé l'héritabilité de la MTEV (*i.e* la part du rôle des facteurs génétiques dans la survenue d'une MTEV, voir **annexe** « **les différents types d'études épidémio-génétiques », pA3**). Selon ces études, l'héritabilité varie entre 55% et 62%. De plus, une étude de ségrégation (voir **annexe pA3**) conduite sur l'un de

ces échantillons [16], conclut que le mode de transmission de la MTEV est concordant avec un modèle multigénique impliquant de nombreux gènes aux effets modestes.

Dans ce paragraphe consacré à l'étiologie de la MTEV, nous présenterons d'abord les facteurs de risque de MTEV dits « démographique » (*ie* le sexe, l'âge et l'origine ethnique). Nous verrons ensuite les facteurs de risque acquis, c'est-à-dire toute circonstance dans la vie d'un patient ayant pu favoriser la survenue d'une thrombose. Nous terminerons enfin par un exposé des facteurs de risque génétiques connus actuellement.

II.1. Facteurs démographiques : sexe, âge, ethnie

L'incidence de la MTEV est peu différente entre les hommes et les femmes d'après la plupart des études (revues dans [5]). Si certaines études notent un risque légèrement plus élevé de MTEV chez les femmes jeunes par rapport aux hommes du même âge (vraisemblablement en raison de grossesses ou de la prise de contraception orale) [3], cette tendance s'inverse après 40 ans [3][4][17]. En revanche, l'incidence de la MTEV augmente avec l'âge de manière exponentielle, particulièrement après 40 ans. Elle est estimée à 0.3‰ par an et par personne entre 25 et 35 ans, et s'élève entre 3 et 5‰ par an et par personne entre 70 et 79 ans[2][3][1][17]. Cependant, bien que de nombreuses études corroborent ce résultat, aucune ne fournit d'explication physiopathologique permettant d'en expliquer l'ampleur. Plusieurs facteurs participent probablement à ce phénomène : une baisse de la mobilité des personnes âgées, une augmentation de la prévalence de pathologies induisant de forts risques de MTEV, une dégradation des valves veineuses affectant ainsi le retour du sang vers le cœur.

L'incidence de la MTEV varie de façon importante suivant l'ethnie. D'après une étude californienne incluant des sujets de différentes origines, les personnes d'origine hispanique et asiatique ont une incidence de MTEV respectivement 1.67 et 3.70 fois moindre que les personnes d'origine caucasienne, tandis qu'elle est 1.27 fois plus élevées chez les personnes d'origine africaine [18]. Si des raisons socio-économiques ont pu être avancées, elles n'ont pas permis d'expliquer cette diversité. Il semblerait en revanche qu'elle soit la conséquence de facteurs génétiques, dont la fréquence dépend de l'origine ethnique. Par exemple, la mutation du facteur V Leiden, responsable d'une augmentation du risque de survenue de MTEV (voir **§II.3.2 p13**), n'est présente que chez 0,5% de la population asiatique, contre près de 5% de la population caucasienne [19][20]. Cependant, la fréquence de la mutation Facteur V Leiden observée chez les américains d'origine africaine est du même ordre de grandeur que celle observée chez leurs concitoyens d'origine asiatique. D'autres facteurs génétiques sont donc

vraisemblablement à l'origine des inégalités observées entre les prévalences de la MTEV de ces deux populations [19][21].

II.2. Facteurs de risque acquis

Contrairement à la thrombose artérielle qui fait suite à une longue évolution de maladie inflammatoire des artères (l'artériosclérose), la MTEV survient de façon brutale sur des veines saines. Dans environ la moitié des cas, elle se déclare à la suite d'un événement déclenchant facilement identifiable:

- Les traumatismes et les opérations chirurgicales induisent un fort risque de survenue de MTEV. En particulier, en absence de prophylaxie, la chirurgie orthopédique peut provoquer dans 30% à 50% des cas une MTEV [22][23]. Les poly-traumatisés ont 50 à 60% de risque de développer une MTEV [24]. Les chirurgies abdominales et pelviennes ne sont pas non plus dénuées de risque (environ 30%) [25][26][27]. La plupart de ces événements est heureusement désormais évitée par un traitement anticoagulant préventif. Cependant, malgré la diffusion en pratique courante hospitalière de la prophylaxie antithrombotique, on observe encore 1% à 3% de MTEV dans les suites d'une pose de prothèse de hanche ou de genou. Le risque de MTEV est multiplié environ par 4 en cas d'opération chirurgicale majeure (revu dans [28]).

- Toute situation engendrant une immobilisation, en favorisant la stase veineuse, provoque une augmentation du risque de survenue de MTEV : l'immobilisation d'un membre dans un plâtre, le confinement au lit ou au fauteuil, ou encore devant un écran d'ordinateur à rédiger son mémoire de thèse (phénomène récent, connu sous le nom de « eThrombosis » [29]), par exemple… A ce titre, les longs trajets en avion sont à haut risque de MTEV [30][31]. De surcroît, l'hypoxie hypobare induite par l'altitude pourrait occasionner une activation de la coagulation, quoique ce phénomène soit encore débattu. Il ne pourrait concerner que les personnes présentant d'autres facteurs de thrombophilie, en particulier constitutionnelle, ou induite par la prise de contraception orale [32].

- Les cancers sont associés à une augmentation du risque de MTEV, d'après une étude cas-témoins récente incluant 3220 MTEV et 2131 témoins. Le risque est augmenté d'environ 7 fois, tout cancer confondu. Les néoplasies les plus à risques sont les hémopathies, suivies des cancers du poumon et de l'estomac [33]. Une étude longitudinale de 537 patients atteints d'un cancer du poumon a montré que le risque de MTEV était environ 20 fois supérieur à celui de la population générale. Parmi ces patients, ceux qui présentaient un adénocarcinome avaient 3 fois plus de risque que ceux qui présentaient un cancer à petites cellules [34]. Les mécanismes

incriminés sont multiples : si la baisse de la mobilité consécutive à la maladie ou encore une compression veineuse par la tumeur elle-même sont des raisons évidentes, on invoque également une augmentation de la thrombophilie due à la libération de facteurs pro-coagulant, tel le TF, par la néoplasie [35]. La iatrogénie est également mise en cause. Plus de 10% des patients développent une MTEV au niveau d'un cathéter veineux central [36]. Par ailleurs, certaines chimiothérapies ont des effets thrombogéniques [37].

- Les hormones féminines, et plus particulièrement les dérivés oestrogèniques utilisés pour la contraception orale, augmentent l'activité coagulante du sang. En effet, métabolisées dans le foie, elles interagissent avec la plupart des facteurs de la coagulation synthétisés dans le foie. Ainsi, la contraception orale, qui associe des dérivés de l'œstrogène et de la progestérone, augmente d'environ quatre fois le risque de MTEV (revu dans [28]). D'un côté, l'incidence de la MTEV étant très faible chez les femmes jeunes, le risque absolu de MTEV reste également faible chez les utilisatrices de contraception orale. Mais de l'autre côté, la contraception orale étant largement prescrite chez les femmes jeunes, l'impact en termes de santé publique n'est pas négligeable. Le traitement hormonal substitutif est quant à lui associé à une augmentation de 2 à 4 fois du risque de MTEV (revu dans [28]).

- Les facteurs hormonaux sont également incriminés dans les MTEV survenant au cours de la grossesse ou du postpartum. Une augmentation des taux de PAI-1 et 2, à laquelle s'ajoutent le repos au lit, une modification de la circulation sanguine, ou encore l'apparition d'une pré-éclampsie, pourrait peut-être contribuer à l'augmentation du risque de MTEV au cours - ou au décours - de la grossesse, jusqu'à plus de 10 fois supérieur à celui de la population des femmes du même âge (rev. dans [28]).

- La contraception orale interagit avec d'autres facteurs. Parmi eux, deux facteurs fréquemment rencontrés chez les jeunes femmes sous pilule oestro-progestative sont le tabagisme et l'obésité. On observe une synergie entre le tabagisme et la contraception orale. L'effet du tabac sur le risque de MTEV est difficile à mettre en évidence, et souvent controversé. Une étude prospective de 18 954 personnes issues de la Copenhagen City Heart Study, considérant quatre classes de tabagisme, allant des non fumeurs aux fumeurs de plus de 25g de tabac par jour, observe un Hazard Ratio (HR) de 1,5 entre les deux classes extrêmes, [38]. Dans une autre étude, les femmes fumeuses et sous contraception orale avaient un risque de MTEV 9 fois supérieur aux non-fumeuses sans contraception orale [39].

- La surcharge pondérale, et *a fortiori* l'obésité, augmentent en elles-mêmes modérément le risque de survenue de MTEV, principalement par stase veineuse et augmentation des taux de PAI-1. Les risques relatifs de MTEV des personnes dont l'indice de masse corporelle (IMC) est supérieur à 30 kg/m^2, et des personnes dont l'IMC est situé entre 25 et 30 kg/m^2, sont estimés respectivement à 2 et 1,5 , par rapport aux personnes dont l'IMC est inférieur à 25 kg/m^2 [40]. Cependant, d'après une étude cas-témoins, les femmes en surcharge pondérale (IMC>25 kg/m^2) ont dix fois plus de risque de présenter une MTEV lorsqu'elles sont sous contraception orale (comparativement aux femmes en surcharge pondérale sans contraception orale). A titre comparatif, l'odds ratio (OR) de la contraception orale est de 4,6 dans le groupe des femmes dont l'IMC est inférieur à 25 [41].

- Le syndrome métabolique associe l'obésité à une hypertension artérielle, une insulinorésistance, et une dyslipidémie. Ce syndrome ne définit pas tant une maladie qu'un ensemble de symptômes favorisés par un même habitus de sédentarité et d'alimentation riche en calories. Plusieurs études cas-témoins (rev dans [42]) ont établi une augmentation du risque de MTEV soit avec le syndrome métabolique pris comme entité nosographique, soit avec une ou plusieurs de ses composantes. Les causes envisagées de l'état prothrombotique sont multiples : perturbation des diverses fonctions de l'endothelium, hyperactivité plaquettaire, perturbation du métabolisme des facteurs coagulants et surtout antifibrinolytiques (augmentation du taux de PAI-1 essentiellement). Cependant, si les études cas-témoins réalisées établissent une synergie entre les différentes composantes du syndrome métabolique, une étude longitudinale récente n'impute le risque de MTEV qu'à la seule obésité [43].

- La présence d'anticorps anti-phospholipides dans le plasma peut engendrer une MTEV. Il s'agit d'anticorps, dirigés contre des constituants de la membrane cellulaire. Certains d'entre eux se lient également à la β2-microglobunine et induisent, par son intermédiaire, une agrégation plaquettaire à l'origine d'une MTEV. On parle alors de syndrome des anticorps anti-phospholipides. L'origine de ces anticorps est inconnue. Ils peuvent survenir dans le cadre d'une maladie auto-immune (tel le lupus érythématheux disséminé), d'un syndrome lymphoprolifératif (leucémies lymphoïdes, lymphomes), ou encore d'une infection grave. Dans la moitié des cas cependant, la présence d'anticorps antiphospholipides est isolée. D'après une étude cas-témoin incluant près de 500 cas et 500 témoins, les anticorps antiphospholipides étaient détectables dans le plasma de 0,9% des contrôles, et 3,1% des patients atteints de MTEV, soit un OR de 3,6, s'élevant à 10 en cas de présence d'anticorps anti-β2-microglobunine [44].

II.3. Facteurs de risque génétiques bien établis

II.3.1. Les déficits en inhibiteurs de la coagulation : antithrombine, protéine C, protéine S

Dès la première moitié du vingtième siècle, l'observation de familles dont plusieurs membres souffraient de MTEV permit d'avancer l'hypothèse que des facteurs génétiques pouvaient être responsables de cette pathologie. Le terme de thrombophilie fut employé pour la première fois par Jordan and Nandorff, en 1956. Mais ce fut seulement en 1965 qu'Egeberg *et al.* caractérisèrent le premier cas de MTEV familial (ou héréditaire). La découverte dans le plasma de leurs patients d'un déficit en antithrombine (AT), qui ségrégeait conformément aux lois de Mendel, les amena à la conclusion que ce déficit était dû à une mutation génétique, seule responsable de la MTEV [45]. Durant les deux décennies suivantes, les propriétés coagulantes du sang issu de différentes familles thrombophiles furent attentivement étudiées. C'est ainsi que furent découverts, au début des années 80, deux autres déficits congénitaux en protéines inhibitrices de la coagulation : le déficit en protéine C (PC), et le déficit en protéine S (PS) [46][47].

Le séquençage des gènes de ces trois protéines (AT, PC et PS), réalisé ultérieurement, a mis en évidence des centaines de mutations différentes. Chacune de ces mutations est extrêmement rare, et semble spécifique à chaque famille thrombophile. Au total, les déficits en AT ne concernent qu'une personne pour 5000, tandis que ceux en PC et PS sont présents dans moins de 1% de la population [48][49] (également revu dans [50]). Ils multiplient par dix environ le risque de survenue de MTEV, les effets les plus forts étant observés pour les déficits en AT [51][52][53]. Les déficits en AT, PC et PS ne représentent qu'une minorité de patients (1 à 3%) [51][54]. Les mutations retrouvées chez ces patients sont généralement à l'état hétérozygote, l'homozygotie étant exceptionnelles et vraisemblablement létales dans la plupart des cas.

II.3.2. Les mutations de facteurs de la coagulation : mutation du Facteur V de Leiden (FVL), et mutation du Facteur II G20210A (mutation du gène de la prothrombine)

Le premier cas familial de résistance à la protéine C activée (APCR) a été découvert en 1994 : une mutation du Facteur V (connue sous le nom de Leiden) altère son affinité avec la protéine C, et par là sa désactivation par cette dernière [55]. La présence de cette mutation à l'état hétérozygote augmente le risque de MTEV de trois à cinq fois, et jusqu'à 50 à 80 fois, à l'état homozygote [56][57]. Cette mutation n'est pas rare dans la population caucasienne : sa

fréquence est aux alentours de 5%. Dix à vingt pour cent des patients atteints de MTEV présentent cette mutation, tandis qu'environ 90% des porteurs hétérozygotes de cette mutation ne développeront jamais, au cours de leur vie, de MTEV[58]. Ainsi, la valeur prédictive de la mutation FVL est limitée.

La découverte de cette mutation, fréquente et à faible pénétrance, permit de considérer la MTEV comme une maladie complexe, et d'envisager d'autres stratégies de découverte de nouveaux gènes de susceptibilité. C'est ainsi qu'en 1996, une étude cas-témoins mit en évidence, grâce au séquençage de l'ensemble du gène du Facteur II, une mutation dans la région régulatrice 3' de ce dernier (mutation du Facteur II G20210A). Il s'agit d'une mutation « gain-de-fonction », induisant une augmentation de la reconnaissance du site de clivage, à l'origine d'une augmentation de la production de mRNA et par conséquent de synthèse de prothrombine. Cette mutation, présente dans 2 à 3% de la population générale, induit une augmentation des taux plasmatiques de Facteur II, et un risque de MTEV environ trois fois supérieur à celui de la population générale [59].

II.3.3. Le groupe sanguin ABO

Le groupe sanguin ABO est un phénotype déterminé par la présence de résidus sucrés à la surface des grobules rouges. La variabilité d'un individu à l'autre de ces résidus est induite par la présence de polymorphismes dans le gène *ABO*. Dès la fin des années 60, Jick *et al* observèrent, dans le cadre d'un programme nord-américain de surveillance pharmacologique, une sous-représentation de patients du groupe O parmi ceux traités pour une MTEV par un anticoagulant [60]. Selon une méta-analyse récente incluant 21 études prospectives ou rétropectives totalisant 6 720 patients [61], l'OR de MTEV des personnes des groupes A, B ou AB, par rapport aux personnes du groupe O, était de 1,79 (avec un intervalle de confiance à 95%, IC95, compris entre 1,56 et 2,05). De plus, trois de ces études étudiaient les génotypes du gène *ABO* et distinguaient les allèles A1 et A2. Prenant comme référence les génotypes $OO/OA_2/A_2A_2$, l'OR de MTEV des génotypes $A_1O/BO/A_2B$ était 2,11 (IC95 [1,66-2,68]). Quant à celui des génotypes A1A1/A1B/BB, il était de 2,44 (IC95 [1,79-3,33]. Les risques de MTEV associés aux génotypes $A_1O/BO/A_2B$ et A1A1/A1B/BB n'étaient pas significativement différents. Les personnes du groupe O présentent, dans de nombreuses études revues dans [62], des taux de vWF plus faibles d'environ 25% que les personnes des groupes A, B ou AB. C'est certainement la principale raison pour laquelle les personnes du groupe O sont moins exposées au risque de MTEV. Nous reviendrons au **§IV.3.1.1 p24** sur les mécanismes biologiques reliant ABO, vWF et FVIII.

Les mutations génétiques engendrant une augmentation du risque de survenue de MTEV (déficits en inhibiteurs de la coagulation et mutations des gènes du FV et FII) ne sont retrouvées que dans ~30% des cas de MTEV (rev dans [63]). Parmi ces mutations, la mutation FVL présente le plus grand risque attribuable (14%) en raison de sa relative fréquence dans la population. Les groupes sanguins non-O ont, quant à eux, un risque attribuable de près de 30%. Le risque attribuable conjoint des mutations des gènes des FII et FV et des polymorphismes d'*ABO* s'élève à 40% d'après une étude récente [64]. Il est ainsi couramment admis qu'une part importante des facteurs de prédisposition génétique reste à découvrir.

III. Recherche contemporaine de nouveaux facteurs de risque génétique de la MTEV

III.1. Généralités

Alors que les premiers facteurs de risque génétiques de MTEV découverts étaient à la fois extrêmement rares et à haut risque de MTEV, la recherche de nouveaux facteurs de risque génétiques s'oriente vers la découverte de polymorphismes plus fréquents et conférant des risques de MTEV de plus en plus faibles. On distingue deux approches stratégiques pour la découverte de nouveaux gènes de susceptibilité. La première, l'approche « gènes candidats » consiste à tester des hypothèses *a priori* issues de connaissances antérieures concernant les fonctions des gènes. Si elle peut se révéler efficace dans la découverte de nouveaux polymorphismes associés à la maladie, elle ne permet pas de mettre en lumière de nouvelles voies physiopathologiques. Au contraire, la deuxième approche, dite « genome-wide » ou pangénomique, balaie indistinctement l'ensemble du génome sans considération biologique.

Il existe en épidémiologie génétique deux méthodes d'analyse de données issues du génotypage de marqueurs : les analyses de liaison et d'association (voir **annexe pA2 et A3**).

Les analyses de liaison cherchent à localiser sur le génome le(s) gène(s) influençant le phénotype à l'étude. Pour cela, elles étudient la co-transmission d'un parent à son enfant du gène - dont on cherche la position - et des marqueurs génétiques - de position connue. Si un marqueur et le(s) polymorphisme(s) de susceptibilité sont proches, ils sont transmis ensemble, plus souvent que ne le voudrait le hasard. Ainsi, des frères dont les phénotypes sont identiques partagent, pour tout marqueur proche du gène, une plus grande proportion d'allèles identiques que ne le voudraient les lois de transmission mendéliennes. On parle d'allèles identiques par

descendance (par opposition à des allèles identiques par état), car ce qui importe ici n'est pas tant que les frères présentent les mêmes formes alléliques pour le marqueur, mais que le parent qui leur a transmis l'allèle responsable de leur phénotype commun leur ait également transmis à l'un et l'autre le même allèle marqueur ancestral.

De façon succincte, les analyses de liaison, fondées sur l'observation des transmissions alléliques, nécessitent un échantillon familial. On observe un phénomène de liaison lorsque des apparentés de phénotype similaire présentent des allèles identiques *par descendance* en proportion plus grande que ne le voudrait le hasard des transmissions. D'une famille à l'autre, la forme allélique liée aux phénotypes peut tout à fait différer (mais pas au sein d'une même famille). Le marqueur lié au phénotype n'est pas responsable de la variabilité phénotypique observée. Il est simplement localisé à proximité du gène (ou de l'un des gènes) impliqué(s). Les études de liaison mettent en évidence des régions chromosomiques d'intérêt, longues en général de quelques megabases, ou dizaines de megabases, et qui incluent plusieurs dizaines, ou quelques centaines de gènes.

Les analyses d'association, quant à elles, cherchent à identifier le ou les polymorphismes dont les différentes formes alléliques participent à la variabilité du phénotype. En cas d'association avec un marqueur, la probabilité de présenter un phénotype particulier dépend des allèles présentés par ce marqueur. Les études d'association se réalisent aisément en population générale, les études cas-témoins étant particulièrement adaptées aux études d'association d'un phénotype binaire. Il est cependant possible de réaliser des analyses d'association au sein d'échantillons familiaux au moyen d'une méthodologie idoine. Les études d'association sont plus précises que les analyses de liaison. Pour autant, l'association entre un marqueur et le phénotype ne signe pas nécessairement une relation de causalité (on parle alors de polymorphisme « fonctionnel »), mais seulement un déséquilibre de liaison (voir annexe **pA5**) entre le marqueur et le polymorphisme fonctionnel.

Les analyses de liaison et d'association sont en théorie toutes deux applicables tant aux approches de génération d'hypothèses (pangénomiques), qu'à celles testant des hypothèses sur des gènes-candidats. Cependant, jusqu'à la fin du XX^{ème} siècle, les limites techniques ne permettaient pas d'effectuer des études d'association pangénomiques. Les études pangénomiques étaient des analyses de liaison, réalisées au moyen de quelques centaines de microsatellites criblant l'intégralité du génome. Les régions candidates ainsi mises en évidence faisaient ensuite l'objet d'une cartographie plus fine, grâce à un un maillage génotypique plus

serré au moyen de polymorphismes biallélіques (Single Nucleotide Polymorphisms, SNPs) en vue de réaliser des analyses d'association.

Les avancées technologiques du début du XXIème siècle permirent de réaliser, au moyen de biopuces, un génotypage à très haut débit de plusieurs centaines de milliers de SNPs couvrant l'intégralité du génome. La conception de ces biopuces s'appuya sur les connaissances fondamentales apportées par le projet HapMap. Celui-ci entreprit de cataloguer les quelques 10 millions de SNPs identifiés par le séquençage du génome, et de mesurer le déséquilibre de liaison entre ces SNPs. Cette connaissance permet un génotypage optimal, qui couvre plus de 90% de la variabilité et évite toute redondance d'information. Il est désormais possible, grâce à ces puces, de balayer l'ensemble du génome à la recherche de signaux d'association.

III.2. Approche gène candidat

La recherche de facteurs de risque génétiques de MTEV par une approche gène candidat est à l'origine d'une littérature abondante. Les résultats les plus robustes, confirmés par des études ultérieures et des arguments fonctionnels, ont été obtenus par l'étude de polymorphismes du gène du fibrinogène γ [65]. Le SNP rs2066865 influence le risque de MTEV, vraisemblablement en modifiant l'efficacité de l'épissage alternatif à l'origine de la formation du fibrinogène γ'. Nous avons vu **§I.3.2.2 p6** et **tableau 1 p7** que cet épissage crée un site de capture de la thrombine, crucial pour le contrôle de l'amplification de la formation de celle-ci.

Les résultats des autres études de gènes candidats ne semblent pas aussi robustes. L'étude la plus complète, au moment où je commençais ce travail, était certainement celle menée par Smith *et al*.[66] qui étudiaient les polymorphismes de 24 gènes codant pour des protéines intervenant dans la coagulation, l'anticoagulation, la fibrinolyse et l'antifibrinolyse. D'autres études sont parues depuis, examinant un nombre toujours plus grands de gènes. Chaque étude observe de nombreuses et nouvelles associations sans nécessairement reproduire les résultats précédents [67][64]. Cependant, plusieurs gènes présentant des polymorphismes associés à la MTEV sont communs aux trois études, bien que ces polymorphismes soient différents. Il s'agit des gènes du FII, du FV, du FXI, du fibrinogène α et de la PC.

III.3. Approche pangénomique

III.3.1. Analyse de liaison suivie d'association au sein de régions candidates

Une seule étude de liaison pangénomique sur la MTEV a été publiée. Elle a été réalisée chez une unique famille franco-canadienne de 289 membres, sur quatre générations, dont 28 avaient présenté une MTEV. Un déficit en protéine C ségrégeait dans cette famille, mais n'expliquait que partiellement les cas de MTEV. L'analyse de liaison avait donc pour but de mettre en évidence la présence de gène(s) expliquant la pénétrance incomplète et la phénocopie de la mutation dans cette famille. Outre la région de la Protéine C, l'analyse a décelé trois régions liées à la MTEV, en 10p12, 11q23 et 18p11 [68]. Après avoir séquencé 109 gènes localisés au sein de ces régions et décelé près de 500 polymorphismes, les auteurs ont mis en évidence une association entre la MTEV et des SNPs de *CADM1* ($p<10^{-5}$), gène à l'origine d'une protéine intervenant dans la migration des cellules endothéliales[69]. A notre connaissance, ce résultat n'a pas été reproduit.

III.3.2. Les études d'association génome-entier (*GWAS* en anglais)

Un premier pas vers les GWAS fut réalisé par une étude qui analysait environ 20 000 SNPs situés dans 10 000 gènes [70]. Les SNPs étaient sélectionnés sur un critère de fonctionnalité : seuls avaient été génotypés les SNPs non synonymes -qui induisent une modification d'acide aminé dans la protéine- ou situés dans le promoteur du gène et susceptibles de provoquer une modification dans la quantité du gène exprimé. Les auteurs ont adopté une stratégie en trois étapes, afin de se prémunir des fausses découvertes liées aux tests multiples. Les fréquences alléliques de l'ensemble des SNPs ont d'abord été comparées chez 443 cas et 453 témoins. Dans une seconde étape, 1206 SNPs présentant des associations significatives ($p<0,05$) avec la MTEV ont été étudiés dans un échantillon indépendant de 1398 cas et 1757 témoins. Enfin, parmi les dix-huit associations répliquées, neuf ont fait l'objet d'une troisième analyse chez 1314 cas et 2877 témoins. Au terme de ces trois étapes, trois SNPs étaient associés au risque de MTEV. Deux d'entre eux étaient situés dans des gènes dont la fonction est directement liée à l'hémostase. Ils étaient situés dans les gènes *SERPINC1* (gène de l'anti-thrombine), et *GP6* (gène codant pour une glycoprotéine qui est présente à la surface des plaquettes et joue un rôle dans l'activation et l'agrégation des plaquettes). Quant au troisième, il était situé dans le gène *CYP4V2*, lui-même proche du gène *F11* (codant pour le facteur XI). Les OR de ces SNPs étaient estimés à 1,24 IC95 [1,11-1,37] pour

CYP4V2-rs13146272, 1,29 IC95 [1.10-1.49] pour *SERPINC1*-rs2227589 et 1,15 IC95 [1,01-1,30] pour *GP6*-rs1613662.

La première véritable GWAS sur la MTEV a été réalisée par l'équipe de David Trégouët en collaboration avec le groupe du Professeur Pierre-Emmanuel Morange (Hopital de la Timone, Marseille). Elle incluait 419 patients de moins 50 ans et 1228 témoins chez lesquels un panel d'environ 300 000 SNPs a été génotypé [71]. Les seules associations dont la significativité était inférieure à $p=1.7\ 10^{-7}$ (correspondant à la correction de Bonferroni pour test multiples, voir **annexe pA7**) étaient situées dans les gènes déjà connus *F5* et *ABO*. L'étude avait une puissance proche de 100% de déceler toute association du même ordre que celle obtenue avec les polymorphismes d'ABO (OR=1,9 et fréquence de l'allèle mineur (MAF) proche de 0,5), à condition bien sûr que le polymorphisme fonctionnel fût en déséquilibre de liaison suffisamment fort avec l'un des SNPs du panel génotypé (voir annexe «**déséquilibre de liaison**» **pA5**). La puissance valait un peu plus de 60% pour déceler des associations du même ordre que celle obtenue pour les SNPs de *F5* (OR=2,3 ; MAF<0,10). Elle était ainsi insuffisante pour détecter de faibles effets, à l'instar de ceux habituellement trouvés dans les études GWAS. En comparaison, une GWAS portant sur la maladie coronaire avait inclus près de 90 000 individus, auxquels s'ajoutaient 50 000 individus pour l'étude de réplication, afin de pouvoir déceler des OR d'environ 1.12 [72].

NB : *Les résultats de cette étude étaient connus et en voie de publication au moment où j'entreprenais ce travail. A plusieurs reprises, j'ai consulté ces résultats à la recherche d'associations observées dans certaines régions du génome, ou encore pour quelques SNPs particuliers. Ces recherches seront désignées dans ce document par le terme «* in silico *».*

Ainsi, à l'instar de toute maladie complexe, la MTEV possède de multiples facteurs de risques. Considérés individuellement, ceux-ci modifient de manière minime la susceptibilité à la maladie. Les analyses de liaison sont peu adaptées à leur découverte. En effet, l'hétérogénéité génétique réduit la puissance de découverte en cas d'échantillon multifamilial. D'un autre côté, la découverte d'un gène de susceptibilité dans un unique et grand pedigree est excessivement difficile à reproduire. Les analyses d'association (GWAS), quant à elles, paraissent plus adaptées à la recherche de facteurs multiples. Cependant, les effets qu'elles visent à découvrir sont généralement faibles et nécessitent des échantillons de très grande taille pour pouvoir les détecter et ensuite les valider dans des cohortes indépendantes.

III.4. Etude de phénotypes intermédiaires de la MTEV

III.4.1. Généralités

La susceptibilité à la MTEV peut être vue comme un phénomène quantitatif si l'on considère que la MTEV est la conséquence de modifications quantitatives de divers facteurs physiologiques. Il peut s'agir soit d'une modification pathologique, comme dans le cas par exemple d'un déficit en protéine C, soit de légères modifications qui restent dans des fourchettes de variation normales mais dont l'accumulation est à l'origine d'une augmentation de risque de MTEV. Plusieurs protéines plasmatiques ont été ainsi trouvées associées au risque de survenue de MTEV (voir **§III.4.2** ci-après) ; les taux plasmatiques de ces protéines sont alors communément appelés des phénotypes intermédiaires de la MTEV.

La recherche de gènes intervenant dans la variabilité de phénotypes intermédiaires permet d'augmenter la puissance de détection de nouveaux gènes de susceptibilité à la MTEV. En individualisant les différents phénotypes intermédiaires, on réduit considérablement l'hétérogénéité génétique. En effet, on s'attend à ce que le nombre de gènes qui contrôlent chaque phénotype intermédiaire soit très inférieur à ceux qui contrôlent la susceptibilité à la MTEV. Par ailleurs, les phénotypes intermédiaires sont plus proches de l'action du gène que ne l'est la maladie elle-même. L'effet du gène sera donc moins atténué induisant ainsi une puissance statistique plus importante pour le détecter. Un autre avantage de l'étude des phénotypes intermédiaires est qu'il peut être étudié avant l'apparition de la maladie. Enfin, elle permet de mieux individualiser les différents mécanismes physiopathologiques sous-jacents.

La première étape vers la détection de nouveaux gènes de susceptibilité à la MTEV consiste à identifier de bons phénotypes intermédiaires, en recherchant les facteurs biologiques dont les taux plasmatiques sont associés au risque de survenue de MTEV, puis en étudiant leur héritabilité. Cette recherche est à l'origine de nombreuses études dont nous proposons un aperçu.

III.4.2. Taux plasmatiques de facteurs biologiques associés à la MTEV

<u>Les protéines intervenant dans la cascade de la coagulation</u>

Les taux plasmatiques des protéines intervenant dans la cascade de la coagulation et de la fibrinolyse sont tout naturellement candidats pour faire partie de la liste des phénotypes intermédiaires de la MTEV. De nombreuses études ont étudié l'association entre ces taux plasmatiques et le risque de MTEV. De manière succincte, nous présentons ici les principaux

facteurs dont l'association avec le risque de MTEV est désormais bien établie. Nous reviendrons plus en détail, au **§IV.2 p23**, sur l'association entre la MTEV et les deux facteurs qui font l'objet de mon travail de thèse, le FVIII, et sa protéine de transport, le vWF.

Les **facteurs procoagulants** dont les associations avec la MTEV sont les plus convaincantes sont le FVIII (OR du dernier quartile versus le permier quartile = 4,8), le FIX (OR du 90ème percentile = 2,5) et le FXI (OR du 90ème percentile = 2,2), d'après l'étude LETS qui incluaient entre 300 et 500 témoins et autant de cas suivant l'avancée du recrutement [73][74][75]. Les effets des autres facteurs, s'ils existent, sont plus difficiles à mettre en lumière (rev. dans [76]). Concernant les **facteurs anticoagulants**, les taux d'Inhibiteur de la Voie du Facteur Tissulaire (TFPI) inférieurs au dixième percentile sont associés à une augmentation du risque de MTEV de 1,7 fois d'après le projet LETS [77]. Enfin, pour les **facteurs intervenant dans la fibrinolyse**, nous retiendrons une augmentation du risque de MTEV (OR=1,7) chez les personnes présentant des taux d'Inhibiteur de la Fibrinolyse Activable par la Thrombine (TAFI) au-delà du 90ème percentile [78]. Nous avons déjà évoqué, lors de la présentation du syndrome métabolique, le rôle potentiellement thrombogène d'une augmentation du PAI-1, quoique celui-ci soit moins clairement objectivé que ceux des facteurs précédents (voir **§II.2 p12**).

L'homocystéine

Plusieurs études ont montré que la MTEV était associée à la présence de taux plasmatiques d'homocystéine modérément élevés (revues dans [28]). L'homocystéine est un acide aminé, dérivé du métabolisme de la méthionine, elle-même issue de la digestion des protéines. Le métabolisme de la méthionine nécessite l'intervention de la vitamine B, dont la carence alimentaire constitue la cause la plus fréquente d'hyperhomocystéinémie. Un polymorphisme génétique fréquent du gène de la Méthylen Tetrahydrofolate Reductase (MTHFR C677T) est associé à une élévation des taux d'homocystéine [79].

Cependant, alors qu'on s'attendrait à observer une augmentation du risque de MTEV induite par MTHFR C677T, celle-ci – si elle existe – semble difficile à mettre en évidence. Alors qu'une méta-analyse, totalisant 8 400 patients, suggère un effet faible de *MTHFR* C677T sur le risque de MTEV (OR=1,20)[80], une étude récente incluant 4375 patients et 4856 témoins ne montre aucun effet significatif (OR=0,94) [81]. Une absence d'association entre MTHFR C677T et la MTEV ne remettrait en question ni l'association entre MTHFR C677T et l'hyperhocystéinémie, ni entre l'hyperhocystéinémie et la MTEV. Elle pourrait en effet

s'expliquer par une absence de cause à effet de l'association statistique entre l'hyperhocystéinémie et la MTEV. Par ailleurs, une faible augmentation de l'homocystéine par l'allèle MTHFR C677T expliquerait que l'impact sur la MTEV de celui-ci soit minime voire nul.

III.4.3. Héritabilité des phénotypes intermédiaires de la MTEV

Les traits quantitatifs (TQs) associés à la MTEV ont un déterminisme à la fois environnemental et génétique, dans des proportions variables. Seuls ceux dont le déterminisme génétique est important sont de bons phénotypes intermédiaires de MTEV. Au moment où je commençais ma thèse, une nouvelle ère démarrait, au cours de laquelle la recherche de nouveaux facteurs génétiques de MTEV *via* l'étude de ses phénotypes intermédiaires se fit plus systématique. Elle fut préparée par la GAIT Study qui, en estimant l'héritabilité de 27 traits quantitatifs associés, ou susceptibles d'être associés, à la MTEV permit de choisir les phénotypes intermédiaires les plus prometteurs [82]. L'héritabilité est la part de la variabilité expliquée par des facteurs génétiques. La GAIT Study est une étude familiale, incluant 21 grandes familles espagnoles dont 12 étaient recrutées *via* un cas de thrombophilie idiopathique. Les résultats sont rapportés dans le **tableau 2**.

Les estimations d'héritabilité, plus encore que toute estimation statistique, n'ont qu'une valeur indicatrice. Elles peuvent être surestimées en cas de recrutement particulier, ou encore en présence d'un environnement partagé par l'ensemble de la famille et mimant une ségrégation mendelienne. Elles seraient à l'inverse sous-estimées en cas de mauvaise modélisation des effets génétiques (voir **§VI.2.1**). Les héritabilités des FVIII et vWF, par exemple, ont été estimées à 40% et 32% dans la GAIT, alors qu'ils s'élevaient à 61% et 75% dans une grande étude de jumeaux, incluant 149 paires de monozygotes et 352 paires de dizygotes[83].

Tableau 2: Héritabilité de traits quantitatifs associés à la MTEV, estimée dans la GAIT

APTT[1]	**APCR[2]**	**FXII**	**FVII**	**TFPI**	**PT[3]**	**PC**	**FII**	**AT**
0.83 ± 0.07	0.71 ± 0.08	0.67 ± 0.09	0.52 ± 0.09	0.5 ± 0.09	0.50 ± 0.09	0.50 ± 0.09	0.49 ± 0.09	0.49 ± 0.09
PS	**FXI**	**FV**	**FX**	**FVIII**	**FIX**	**FG[4]**	**vWF**	**PAI-1**
0.46 ± 0.09	0.45 ± 0.10	0.44 ± 0.09	0.43 ± 0.13	0.40 ± 0.09	0.39 ± 0.09	0.34 ± 0.10	0.32 ± 0.11	0.30 ± 0.08
tPA	**HCY[5]**	**PG[6]**	**TF**	**DD[7]**	**tPA**			
0.27 ± 0.07	0.24 ± 0.08	0.24 ± 0.10	0.17 ± 0.08	0.11 ± 0.09	0.27 ± 0.07			

1. APTT : Activated Partial Thromboplastin Time ou TCA (Temps de Céphaline Activée), exploration globale de la voie « intrinsèque » ; **2. APCR** : Résistance à la Protéine C Activée ; **3. PT** : Temps de Prothrombine : exploration globale de la voie « extrinsèque ; **4. FG**: Fibrinogène ; **5. HCY**: Homocystéine ; **6. PG**: Plasminogène ; **7. DD** : D-Dimères, produits de dégradation de la fibrine

IV. Objectif de la thèse

IV.1. Enoncé de l'objectif

L'objectif de ma thèse fut de contribuer à la recherche de nouveaux facteurs de risque génétique de la MTEV en étudiant deux de ses phénotypes intermédiaires : le FVIII, et sa protéine de transport, le vWF. La stratégie générale adoptée reposait sur des analyses de liaison et d'association pangénomiques des taux plasmatiques de ces deux facteurs. Des études cas-témoins ont ensuite permis de tester si les régions chromosomiques et les polymorphismes identifiés étaient associés au risque de survenue de MTEV.

IV.2. Justification de l'objectif

Le FVIII est un phénotype intermédiaire particulièrement intéressant de la MTEV. En effet, son importante héritabilité se conjugue à une forte association avec le risque de MTEV. Les personnes présentant des taux de FVIII au-delà de 150 UI/dL ont cinq fois plus de risque de développer une MTEV que les personnes dont les taux sont inférieurs à 100 UI/dL [73]. Les taux de FVIII sont très corrélés à ceux vWF, sa protéine de transport vWF. Cette dernière explique près de 50% de la variabilité des taux de FVIII [84]. Cela est dû au rôle protecteur de vWF : à l'état libre, le FVIII a une demi-vie d'une heure contre neuf lorsqu'il est porté par vWF [85]. Le vWF est lui-même associé à la MTEV. Alors qu'une étude cas-témoin (300 cas et 300 témoins) estime que l'effet du vWF est intégralement expliqué par son rôle de protecteur du FVIII [73], une étude longitudinale (LITE) de 19 237 personnes, dont 159 ont présenté une MTEV pendant un suivi médian de huit ans, décèle un effet du vWF indépendamment de celui du FVIII [86]. En effet, les HR de MTEV des valeurs hautes (dernier *versus* premier quartile) de FVIII et vWF y sont respectivement de 2,6 [1,6-4,3] et 3,8 [2,0-7,2] dans un modèle multivarié incluant simultanément les taux de FVIII et vWF.

Comme mentionné ci-dessus, les taux de FVIII et vWF présentent une forte héritabilité quoique leurs estimations soient sujettes à quelque imprécision. En dépit de cela, seul le gène *ABO* était, au moment où j'ai débuté ce travail, connu pour influencer sans ambigüité et de manière constante à travers les études les taux vWF et de FVIII. Une étude de jumeaux a estimé que le gène *ABO* expliquait environ 30% de l'héritabilité du FVIII [87]. Une étude familiale a corroboré cette estimation en montrant que la corrélation entre les taux de FVIII de deux germains était de l'ordre de 0,37 alors que celle observée entre deux époux (qui sont génétiquement indépendants) était de l'ordre de 0,08 [84][84]. Ces observations sont en faveur de la prédominance des facteurs génétiques sur les facteurs environnementaux dans le

déterminisme des taux de FVIIII. De plus, la corrélation entre germains est peu affectée par un ajustement sur le groupe ABO (0,30), ainsi que par un ajustement supplémentaire sur les taux de vWF (0,24). Cela traduit la présence de facteurs génétiques autre qu'*ABO* modulant les taux de FVIII, indépendamment de leurs effets sur vWF.

Notons qu'en dehors des facteurs génétiques, la variabilité des taux de FVIII et vWF est expliquée essentiellement par l'âge : une augmentation des taux de FVIII et de vWF avec l'âge est systématiquement observée dans les études. On retrouve, mais de façon inconstante selon les études, des taux de FVIII et vWF plus élevés chez les femmes que chez les hommes [88] [89], des taux de FVIII moins élevés chez les fumeurs [88][90], et plus élevés chez les utilisatrices de la contraception orale [91].

IV.3. Etat des connaissances sur le sujet de cette thèse

Au moment où j'entreprenais ce travail, plusieurs études avaient déjà permis d'identifier divers facteurs génétiques influençant les taux de FVIII et/ou vWF, par des approches « gènes candidats » ou pangénomiques. Hormis les associations avec le gène *ABO*, les résultats n'étaient pas toujours reproductibles. Peu d'études avaient étudié la répercussion de ces facteurs génétiques sur le risque de survenue de MTEV.

IV.3.1. Approche gènes candidats

IV.3.1.1. Association entre le gène *ABO* (chromosome 9q34) et les taux de FVIII et vWF

Le gène *ABO* code pour une galactosyltransferase. Cette protéine participe à la maturation post-traductionnelle en ajoutant des sucres complexes sur des protéines. On retrouve ces complexes sur la membrane des globules rouges, des plaquettes, de cellules épithéliales et endothéliales. On connaît trois protéines plasmatiques concernées par cette maturation post-traductionnelle : le vWF, le FVIII et l'α2-macroglobuline. Les polymorphismes du gène *ABO*, à l'origine du sérogroupe ABO, détermine le type d'oligosaccharides (appelés A, B ou H) ajoutés aux protéines.

Les allèles A1 et B du gène ABO induisent une augmentation des taux de vWF et FVIII d'environ 20 à 30% par rapport aux allèles O et A2. (rev dans [62];[84]). Le mécanisme par lequel le groupe ABO influence les taux de vWF et FVIII est partiellement élucidé. La présence de l'oligosaccharide H (groupe O) favorise le clivage de vWF par ADAMTS13, et, par ailleurs, une plus forte clearance hépatique de vWF (rev dans [62]). L'association entre ABO et FVIII est au moins en partie expliquée par le rôle protecteur de vWF vis-à-vis de FVIII. Cependant,

cette association reste significative après ajustement sur vWF ([84]). Une action du groupe ABO sur FVIII indépendamment de vWF est confortée par la présence de résidus A, B ou H sur FVIII. Cependant, contrairement à ce qui a pu être observé pour vWF, il semblerait que le type de résidu porté par FVIII n'ait aucune inflence sur la clairance de ce dernier. Ainsi, le seul mécanisme physiologique identifié pour expliquer la variation des taux de FVIII en fonction du groupe ABO est une action indirecte de celui-ci *via* les taux de vWF (rev. dans [92]).

IV.3.1.2. Gènes structuraux du vWF et du FVIII

Le gène *VWF* (chromosome 12p13)

Le gène structurel de vWF, situé en 12p13, mesure 180 kb et contient 52 exons. Quatre SNPs de la région régulatrice 5' [-3268 C>G (rs7965425), -2709 T>C (rs7964777), -2661 G>A (rs7954855), -2527 A>G (rs7965413)] en fort déséquilibre de liaison se trouvaient significativement associés aux taux de vWF dans un échantillon de 261 personnes en bonne santé du groupe O [93][94]. A partir de la séquence d'ADN de la région, les auteurs de ce travail ont déduit les facteurs transcriptionnels susceptibles de réguler la transcription de VWF. Ils ont montré que l'un de ces facteurs, NFκB, présentait une meilleure affinité avec -3268C et -2709T qu'avec -3268G et -2709C. De plus, un autre polymorphisme de la région régulatrice en 5' (une répétition en tandem GT) semblait intervenir dans l'augmentation de transcription de vWF dans des conditions de stress des cellules endothéliales (« shear stress ») [95]. Cependant, les associations entre ces divers polymorphismes et les taux de vWF n'ont pas été reproduites dans une étude épidémiologique incluant 400 personnes en bonne santé, même après stratification sur le groupe ABO [96].

Un autre polymorphisme, rs1800386 A>G, rare (<0.5%), situé dans l'exon 28 de *VWF*, responsable d'un changement d'acide aminé Tyr / Cys en position 1584 a été trouvé enrichi dans un échantillon de patients atteints de maladie de von Willebrand (maladie hémorragique liée à un défaut quantitatif ou qualitatif du vWF) [97]. L'acide aminé Cys confère à la protéine vWF une plus grande susceptibilité à l'activité protéolytique d'ADAMTS13. Enfin, les taux de vWF étaient plus bas parmi les porteurs de l'allèle rs1800386-G dans un échantillon de 5 000 donneurs de sang en bonne santé, probablement en raison d'une augmentation de la clearance de vWF [98][99].

Enfin, une étude récente a sélectionné 27 SNPs au sein de *VWF*, capturant entre 70 et 95% de la variabilité génétique de celui-ci, dont la totalité des polymorphismes communs (MAF>5%). Trois d'entre eux (rs7306706, rs216318 et rs4764478) étaient associés aux taux de

vWF, mesurés dans une étude cas-témoins qui incluaient 421 jeunes patients ayant présenté une thrombose artérielle [100]. De plus, deux autres SNPs (rs1063857, rs216293) étaient associés au risque de survenue de thrombose artérielle. Il est intéressant de noter que rs1063857 était également associé au taux de vWF dans une étude pangénomique (décrite au **§IV.3.2.2 p28**) [101].

Le gène F8 (chromosome Xq28)

Le gène *F8* contient 26 exons et s'étend sur 186 kb. Tandis qu'environ mille mutations non-sens sont responsables de déficit en FVIII à l'origine de thrombophilie, un seul polymorphisme (et les polymorphismes en déséquilibre de liaison avec lui) a été décrit à ce jour comme étant peut-être associé au taux de FVIII. En effet, Viel et al [102] ont séquencé toutes les régions fonctionnelles du gène *F8* chez 147 personnes issues de 7 populations différentes (espagnols caucasiens, afro-américains, chinois, japonais, mexicains indiens, americains andins, asie du sud-est) et identifié 48 polymorphismes. Ils ont ensuite testé, dans l'échantillon familial de la GAIT Study (voir §**III.4.3 p22)**, les associations entre les 12 polymorphismes présents chez les caucasiens et l'activité du FVIII. Une seule était significative ($p<0,007$). Elle était observée pour un SNP non synonyme, rs1800291C>G, dont l'allèle C était associé à une élévation des taux de FVIII ($p=0,02$).

La même association était observée chez près de 300 femmes dont la moitié présentait une MTEV [103], et dans une étude combinant trois échantillons cas-témoins, totalisant 2 500 individus dont le tiers présentait une MTEV [104]. Par ailleurs, ces deux études cas-témoins, ainsi qu'une étude prospective de 5 000 personnes, dont 184 ont développé une MTEV durant un suivi médian de 9 ans [67], observaient une tendance à une association entre rs1800291-C et une augmentation du risque de survenue de la MTEV, bien qu'aucune n'atteignît la significativité.

IV.3.1.3. Low-density Lipoprotein Receptor-related Protein (gène *LRP-1*), en 12q13 et *LDLR* en 19p13

La protéine LRP (gène *LRP-1*) est un récepteur d'endocytose ubiquitaire multifonctionnel, internalisant divers ligands extracellulaires. Des expériences *in vitro* ont montré que le FVIII était capable de se fixer à la LRP [105][106]. Par ailleurs, il a été montré chez la souris que la désactivation de la LRP entraînait une élévation du FVIII et de vWF [107]. Plusieurs équipes ont étudié les associations entre les taux de FVIII et quelques polymorphismes rares (MAF<5%) du gène *LRP-1*, parmi lesquelles trois étaient significatives : D2080N (rs34577247 G>A), -25C>G (rs35282763) et 663 C>T (rs138358068),

[84][108][109]. En particulier, le polymorphisme 663 C>T était de plus associé au risque de MTEV dans une étude incluant 150 cas et 200 (OR=3,3 IC95 [1,3-8,5]) [109]. Par ailleurs, les domaines de fixation à la LRP du FVIII ont fait l'objet d'une recherche de polymorphismes par Morange *et al.* Aucun polymorphisme n'a été retrouvé dans un échantillon de 20 individus non apparentés et présentant des valeurs très contrastées de FVIII [84]. *LRP-1* appartient à la famille des gènes LDLR. Il coopère avec *LDLR*, un autre gène de cette famille, qui participe également au catabolisme du FVIII. Plus récemment, une étude cas-témoins de la maladie coronaire a mis en évidence l'association entre deux SNPs (rs688 et rs222867) et le taux d'activité plasmatique de FVIII, ainsi que l'association entre rs688 et la maladie coronaire [110].

IV.3.2. Approche pangénomique

IV.3.2.1. Analyse de liaison suivie d'association au sein de régions candidates

Jusqu'au milieu de la dernière décennie, les stratégies de recherche pangénomique reposaient sur les analyses de liaison génétique. C'est dans cet esprit que la « GAIT study », constituée de 21 familles dont 12 recrutées *via* un cas de thrombose veineuse ou artérielle, fut mise en place pour identifier de nouveaux déterminants génétiques d'un grand nombre de phénotypes intermédiaires de la MTEV. A ce jour, seule l'étude du FXII a abouti à la découverte d'un polymorphisme situé au sein d'un signal de liaison, dans le gène *F12*, et associé à la fois au taux de FXII et au risque de survenue de thrombose [111][112]. Cette étude confirmait le bien-fondé de la stratégie de recherche adoptée dans cette thèse. Notons toutefois que l'association de ce polymorphisme avec le risque de thrombose était retrouvée de manière très inconstante dans d'autres études. Une méta-analyse combinant les informations de 16 études n'a pas retrouvé d'effet de ce polymorphisme sur le risque de thrombose, malgré une puissance supérieure à 80% de détecter un OR de 1,2 [113]. Ce résultat pourrait être dû à une hétérogénéité entre les études et plus particulièrement une hétérogénéité génétique.

Les résultats obtenus par l'analyse des autres phénotypes étaient cependant moins aboutis. En particulier, une analyse de liaison de l'APCR, ajustée sur la mutation FVL, a mis en évidence un signal sur le chromosome **18p11** [114]. Ce signal était plus important dans une analyse bivariée avec l'APCR et les taux de FVIII, suggérant que le QTL (*ie* l'emplacement sur le chromosome du facteur modulant le trait quantitatif, *Quantitative Trait Locus* en anglais) 18p11 pourrait avoir un effet sur les deux phénotypes (effets pléïotropique). La région identifiée était longue d'environ 40 cM mais ne comportait aucun gène candidat évident.

Concernant l'analyse de liaison des taux de vWF, le signal le plus fort était observé en **9q34**, correspondant au gène *ABO* [115]. D'autres signaux, en **1p36, 2q23, 5q31, 6p22, 22q11**, bien que nettement plus faibles, étaient évocateurs de liaison, mais aucune de ces régions n'a pu être confirmée par la suite. Notons qu'aucun signal de liaison n'a été observé dans la région du gène structural de vWF (2p13).

Une autre étude, incluant 13 familles *via* un cas de MTEV avec des taux élevés de FVIII, a permis une analyse de liaison du taux de FVIII corrigé (« FIII-R ») afin de s'affranchir des effets de l'âge, du sexe, des taux de vWF et du groupe ABO. Les coefficients de correction avaient préalablement été estimés au moyen d'une régression multiple dans un échantillon de donneurs de sang. Un locus lié à FVIII-R a été détecté sur le chromosome **8p21-22** [116]. Trois gènes candidats ont été identifiés dans cette région : les gènes de la Lipoprotein Lipase (LPL) et du t-PA, qui sont des ligants du récepteur LPR, ainsi que ADAMDEC1 (ADAM-like, decysin 1), qui code pour une protéase, dont l'un des sites d'action pourrait être l'ectodomaine de LRP. Les associations entre la MTEV et une quarantaine de polymorphismes situés dans ces trois gènes ont été testées chez 164 patients ayant des antécédents de MTEV avec FVIII élevé (au-delà du 98ème percentile) *versus* 214 donneurs de sang en bonne santé. Aucune des associations testées n'était significative [117].

IV.3.2.2. Analyse d'association pangénomique (GWAS)

La *Framingham Heart Study*

La première étude « GWAS » des taux plasmatiques de vWF a été conduite à partir de familles nucléaires incluant au total 883 individus issus de la cohorte Framingham et génotypés pour environ 100 000 SNPs. Aucune association n'était significative pour un risque de première espèce de 5%, après correction pour les tests multiples [118].

Le consortium CHARGE

Au cours de ma thèse, le consortium CHARGE a publié les résultats d'une grande étude d'association pangénomique des taux plasmatiques des FVII, FVIII et vWF, portant sur 23 600 personnes [101]. Ce consortium regroupe cinq cohortes, parmi lesquelles quatre disposaient des phénotypes étudiés: Arterosclerosis Risk in Communities Study, Framingham Heart Study, Rotterdam Study, et Cardiovascular Heart Study. De plus, cette étude incluait des personnes issues de la cohorte British 1958 Birth. Toute association significative après correction pour les tests multiples (voir **annexe pA7**) a fait l'objet d'une étude de réplication chez 7 600 personnes. Outre les gènes *ABO* et *VWF* (*via* le rs1063857), les SNPs associés

significativement aux taux de FVIII et vWF étaient situés dans les gènes : *STXBP5* (6q24), *SCARA5* (8p21), et *STAB2* (12q23). Les SNPs qui n'étaient associés qu'au taux de vWF étaient situés dans les gènes *STX2* (12q24), *TC2N* (14q32), *CLEC4M* (19p13). Pris ensemble, ces SNPs expliquaient près de 13% de la variabilité des traits quantitatifs. Toutes les associations, sauf celle du SNP *STX2*, ont été reproduites dans l'échantillon indépendant au seuil de significativité de 5%. Le **tableau 3 p30** explicite les associations identifiées et propose une synthèse du **§IV.3**.

En conclusion, la puissance de détection de nouveaux facteurs de risque génétiques de la MTEV pourrait être augmentée par l'étude de ses phénotypes intermédiaires. D'après plusieurs études, les taux de FVIII et de sa protéine de transport, le vWF, semblent être de bons phénotypes intermédiaires de MTEV (association au risque de MTEV et héritabilité importante). Au moment où je commençai ma thèse, l'étude de quelques gènes candidats avait mis en évidence une influence de polymorphismes des gènes *ABO* et *VWF* sur les taux de vWF, et des gènes *ABO*, *FVIII*, *LRP-1* et *LDLR* sur les taux de FVIII. Les recherches par une approche pangénomique étaient encore infructueuses, qu'elles aient reposé sur des analyses de liaison ou d'association. L'approche adoptée dans mon travail de thèse est une approche pangénomique reposant sur des analyses de liaison et d'association. De plus, j'ai évalué l'influence des SNP associés au taux de FVIII et/ou vWF sur le risque de MTEV. Les échantillons et les méthodes d'analyse utilisées pour cela sont décrits dans la partie suivante.

Tableau 3. Facteurs génétiques influençant les taux plasmatiques de vWF et/ou de FVIII d'après une recherche bibliographique

Réf	locus	Gène	Polymorphisme*	Phénotype	Approche stratégique	Robus-tesse	Répercussion clinique
[62] [84]	9q34	*ABO*	allèle **A1 et B** *versus* A2 et O	vWF FVIII	Association Gène candidat	Oui	MTEV
[93] [94] [95]	12p13	*VWF*	rs7965425 C>**G** rs7964777 T>**C** rs7954855 G>**A** rs257965413 A>**G** (GT)n court/**long**	vWF	Association Gène candidat	Oui	-
[99][98]	12p13	*VWF*	rs1800386 **A**>G (très rare)	vWF	Association Gène candidat		syndrome hémorragique
[100]	12p13	*VWF*	rs7306706 **A**>G rs216318 **A**>C rs4764478 T>**A**	vWF	Association Gène candidat	-	-
[102] [103] [104]	Xq28	*FVIII*	rs1800291**C**>G	FVIII	Association Gène candidat	Oui	tendance MTEV [103][104][67]
[84] [108] [109]		*LRP-1*	rs34577247**G**>A rs35282763**C**>G rs138358068C/**T** **(rare)**	FVIII	Association Gène candidat	Oui	MTEV [109]
[110]		*LDLR*	rs688 C>**T** rs222867 C>**T**	FVIII	Association Gène candidat	-	maladie coronaire [110]
[114]	18p11	-	-	FVIII	Liaison Pangénomique	-	-
[115]	1p36 2q23 5q31 6p22 22q11	-	-	vWF	Liaison Pangénomique	-	-
[116]	8p21-22	*LPL1*§ *t-PA*§ *ADAMDEC*§	-	FVIII	Liaison Pangénomique	-	-
[101]	6q24	*STXBP5*	rs9390459 **G**>A	vWF et FVIII	Association Pangénomique	Oui Sauf STX2	thrombose artérielle pour rs1063857 [100]
	8p21	*SCARA5*	rs2726953X>**X**	vWF			
			rs9644133 **C**>T	FVIII			
	12p13	*VWF*	rs1063857 T>**C**	vWF et FVIII			
	12q23	*STAB2*	r4981022 **T**>C	vWF			
			rs12229292 G/**T**	FVIII			
	12q24	*STX2*	rs7978987 G/**A**	vWF			
	14q32	*TC2N*	rs10133762 G/**T**	vWF			
	19p13	*CLEC4M*	rs868875 **A**/G	vWF			

* L'allèle en gras est associé à une élévation des taux plasmatiques.

§ Gènes candidats situés au sein d'un signal de liaison, non confirmés par des études d'association [117]

SUJETS, MATERIEL ET METHODES

V. DONNEES DISPONIBLES POUR LA REALISATION DE CE TRAVAIL

De 2005 à 2006, France Gagnon a colligé un échantillon de cinq grandes familles franco-canadiennes avec pour objectif d'identifier de nouveaux facteurs de risque génétique de la MTEV par l'intermédiaire de l'étude de ses phénotypes intermédiaires. Les taux plasmatiques de nombreux facteurs impliqués dans le risque thrombotique ont ainsi été mesurés chez 255 apparentés appartenant à ces familles. La stratégie initialement adoptée pour identifier de nouveaux gènes fut celle des analyses de liaison grâce à une méthode d'analyse bayésienne, méthode sur laquelle je reviendrai dans la partie **VI.1 p44**.

Au moment où le recueil des données se terminait, une collaboration a vu le jour entre le groupe de France Gagnon et les groupes de David-Alexandre Trégouët (INSERM UMR_S 937) et de Pierre-Emmanuel Morange (INSERM UMR_S 626). Ces derniers développaient en effet depuis quelques années un important axe de recherche sur l'épidémiologie génétique de la MTVE à partir d'échantillon cas-témoins et de familles nucléaires saines. L'objectif de cette collaboration était de confronter les résultats obtenus dans des échantillons de nature différente pour identifier de manière robuste de nouveaux gènes de susceptibilité à la MTEV.

C'est dans ce contexte que s'est déroulé mon travail de thèse qui consistait dans un premier temps à identifier de nouveaux polymorphismes impliqués dans la variabilité plasmatique des taux de vWF et FVIII, et dans un second temps à étudier l'effet des polymorphismes identifiés sur le risque de MTEV. Pour cela, j'ai commencé par réaliser des études de liaison de ces deux phénotypes à partir de l'échantillon de grandes familles. Ensuite, j'ai recherché des polymorphismes situés dans les signaux de liaison et associés aux deux phénotypes à partir d'analyses réalisées chez les familles nucléaires saines et chez les cas d'une étude cas-témoins. J'ai également exploré l'intégralité du génome au moyen d'analyses d'association pangénomique (GWAS) réalisées chez les grandes familles et chez des sujets non apparentés ayant un antécédent de MTEV. Enfin, les effets des polymorphismes identifiés sur le risque de MTEV ont été testés dans des études cas-témoins.

La section suivante détaille la nature des échantillons ainsi que celle des données génotypiques et phénotypiques dont je disposais pour réaliser mon projet.

V.1. Les sujets étudiés

V.1.1. Echantillons dans lesquels les taux de vWF et l'activité plasmatique de FVIII ont été mesurés

- **L'échantillon Familles-FVL**

France Gagnon a étudié les arbres généalogiques de 61 patients du Dr Wells qui avaient consulté avant 2005 pour une MTEV à la clinique de thrombophilie d'Ottawa. Elle a sélectionné les cas idiopathiques (*i.e.* absence de facteur acquis tels qu'un cancer ou syndrome myéloprolifératif, une grossesse ou un post-partum, une immobilisation prolongée, un traumatisme, une chirurgie, un syndrome des anticorps antiphospholipides), sans mutation induisant un déficit en AT, en PS ou en PC, mais porteurs de la mutation FVL. France Gagnon a choisi pour son étude les cinq plus grandes familles franco-canadiennes (**voir tableau 4**). Elles s'étendaient sur quatre à cinq générations. Les fratries pouvaient inclure jusqu'à dix-sept individus. Les arbres généalogiques de ces familles sont présentés en annexe **pA8-A9**. Deux cents cinquante cinq membres, issus de ces cinq familles, ont répondu positivement à l'invitation à participer à l'étude. Parmi eux, 62 individus étaient porteurs de la mutation FVL, dont 12 avaient un antécédent de MTEV. Trois personnes avaient un antécédent de MTEV sans présenter la mutation FVL.

Ce recrutement, *via* un cas de MTEV avec mutation FVL, avait pour objectif de diminuer l'hétérogénéité génétique entre les familles et, ainsi, d'augmenter la puissance des analyses de liaison. De plus, ce recrutement pouvait peut-être permettre d'étudier l'hypothèse de l'existence de gènes modificateurs de la mutation FVL agissant sur les taux de traits quantitatifs liés à la cascade de la coagulation. Le travail présenté dans ce document s'intègre entièrement à ce projet puisqu'il recherche des facteurs génétiques modulant les taux de FVIII et vWF. Cependant, à aucun moment je n'ai étudié spécifiquement l'hypothèse que ces facteurs génétiques soient modificateurs de la mutation FVL.

Tableau 4 : Effectifs des cinq familles ayant permis les études de liaison et d'association génome entier

Famille	Nombre d'individus*	Taille des fratries	Nombre de générations
1	27	1-4	4
2	25	2-8	4
3	40	1-16	4
4	10	1-3	4
5	151	1-17	5
Total	253	1-17	4-5

*Génotypés et pour lesquels on dispose des taux de FVIII et de vWF

- **La cohorte Stanislas.**

En 1993, le Dr Visvikis-Siest et son équipe INSERM CIC 9501 (Nancy) commencèrent le recrutement d'une cohorte de familles nucléaires *(i.e* deux parents et leurs enfants) en bonne santé, dans la région des Vosges et de Meurthe et Moselle. Le but de ce projet est d'étudier spécifiquement les déterminants génétiques et environnementaux des facteurs de risques des maladies cardio-vasculaires (tels l'hypertension artérielle, l'obésité, etc…, et les taux plasmatiques de divers facteurs de la coagulation.). Après avoir répondu à l'appel de la Caisse National d'Assurance et s'être rendues au Centre de Médecine Préventive de Vandoeuvre-lès-Nancy, 1006 familles constituées des deux parents et d'au moins deux enfants biologiques âgés de 6 à 28 ans, tous en bonne santé, ont été incluses dans la cohorte entre 1993 et 1995.

Une banque de données cliniques a été constituée, regroupant des informations issues d'un questionnaire standardisé et d'un examen clinique : poids, taille, âge, sexe, pression artérielle, habitude de vie (dont le tabagisme, la nutrition, activité physique, …), échographie cardiaque… A cette banque clinique s'ajoute une banque d'ADN ainsi qu'une banque de données biologiques constituée d'un très grand nombre de mesures plasmatiques en rapport avec la fonction cardio-vasculaire (bilan lipidique, glucidique, …). Depuis leur inclusion, ces familles ont été invitées à 3 reprises à se rendre à une visite de suivi (1998-2000, 2003-2005, 2011-2012) au cours de laquelle les différentes banques de données ont pu être mises à jour et complétées.

Les données que j'ai utilisées dans mon travail de thèse sont celles qui avaient été obtenues au sein d'un échantillon de 108 familles (pour un total de 451 individus) tirées au sort parmi celles qui s'étaient rendues à la première visite de suivi.

- **Le Projet MARTHA**

Le projet MARTHA pour « MARseille THrombosis Association » a été mis en place en 1994 par le Pr Morange. Depuis 1994, ce dernier propose aux patients venant le consulter au centre de Thrombophilie de l'Hôpital La Timone, à Marseille, de participer à une étude d'épidémiologie génétique de la maladie thrombo-embolique veineuse. Pour être inclus dans cette étude, les patients doivent avoir un antécédent de phlébite ou d'embolie pulmonaire, et ne présenter aucun des facteurs de risque constitutionnels conférant un fort risque de MTEV (déficit en AntiThrombine, en Proteine C, en Proteine S, homozygotie pour le FV Leiden, homozygotie pour la mutation FII G20210A). La thrombose veineuse doit être documentée par

une veinographie, une échographie-doppler, une angiographie pulmonaire, et/ou un scanner ventilation/perfusion. Le recrutement est toujours en cours.

L'échantillon « MARTHA05 » est un échantillon cas-témoins composé de 1150 patients d'origine caucasienne inclus dans la période 1994 à 2005 et de 801 témoins. Les sujets témoins sont issus de la cohorte FITENAT, mise en place par le GEHT (Groupe d'Etude sur l'Hémostase et la Thrombose) [119] dans 11 villes françaises par l'intermédiaire des Caisses Primaires d'Assurance Maladie. Ils sont tous en bonne santé et sans antécédent de maladie cardiovasculaire. Leurs parents sont nés dans la région d'inclusion et leurs grands-parents en Europe. Ces sujets ont fait l'objet d'un génotypage des mutations du FVL et du FII G20210A. Les témoins de MARTHA05 ont été sélectionnés comme suit : 475 habitaient la région de Marseille et ne présentaient pas de mutation FVL ni du FII G20210A ; 326 étaient porteurs sains des mutations FVL ou du FII G20210A à l'état hétérozygote et habitaient l'une des 11 villes françaises. Le choix particulier de ce groupe de témoins trouve sa justification dans l'objectif initial poursuivi par les investigateurs de MARTHA, qui était d'étudier les interactions entre de nouveaux facteurs de risque de MTEV et les mutations FVL ou FII G20210A.

Les échantillons « MARTHA08 » et « MARTHA10 » n'incluent quant à eux que des sujets atteints de MTEV. Ces échantillons ont été constitués ultérieurement avec pour objectif de réaliser des études d'association génome-entier de la MTEV. Le groupe témoin n'était pas encore disponible durant ma thèse. MARTHA08 comporte 1006 patients inclus entre 1994 et 2008 parmi lesquels 793 ont été inclus entre 1994 et 2005 et appartiennent à l'échantillon MARTHA05. Les 213 patients restants ont été inclus entre 2005 et 2008. Quant à l'échantillon MARTHA10, il est constitué de 576 autres patients inclus entre 2008 et 2010.

Les mesures des taux plasmatiques de FVIII et vWF ont été effectuées sur l'ensemble des cas de MARTHA05 (n = 1150), sur les 213 patients de MARTHA08 non inclus dans MARTHA05 et sur l'ensemble des patients de MARTHA10 (n = 576).

V.1.2. Echantillons cas-témoins sur la MTEV

Afin de pouvoir étudier l'impact sur le risque de MTEV des polymorphismes que j'aurai trouvés au cours de mon travail de thèse associés aux taux de FVIII et vWF, j'avais à ma disposition trois échantillons cas-témoins pour lesquels une banque d'ADN étaient disponibles. Outre l'étude MARTHA05 mentionnée ci-dessus, j'avais accès à :

L'étude FARIVE

FARIVE (Facteurs de Risques et de Récidives de la Maladie Thromboembolique Veineuse) est une enquête cas-témoin hospitalière multicentrique de la région parisienne, conduite par le Professeur Emmerich (INSERM UMR_S 765). Les cas sont des patients hospitalisés, ou venant en consultation externe pour un premier épisode de MTEV, confirmée par veinographie et échographie pour les cas de thrombose veineuse, et par tomographie, scanner ventilation/perfusion ou angiographie pour les cas d'embolie pulmonaire. Les patients doivent avoir plus de 18 ans, ne présenter aucun antécédent de cancer dans les 5 dernières années. A chaque cas inclus, un témoin hospitalisé ou se présentant aux consultations a été apparié sur l'âge, le sexe, et l'hôpital. Il devait ne présenter aucun antécédent de MTEV, de thrombose artérielle, de cancer, d'insuffisance hépatique ou rénale. Le recrutement est toujours en cours. Au moment des analyses, cette étude avait inclus 607 cas et 607 témoins.

Une étude GWAS sur la MTEV (données *in silico)*

Au moment où je débutais ma thèse au sein de l'UMR_S 937, une analyse génome entier d'une étude cas-témoin incluant 419 cas de MTEV et 1228 témoins venait d'être terminée par l'équipe de David-Alexandre Trégouët [71]. Je me suis appuyée sur les résultats de cette GWAS pour élaborer une stratégie de sélection de gènes candidats.

Les patients, d'origine européenne, proviennent de quatre centres hospitaliers français (Grenoble, Marseille, Montpellier et Paris). Ils ont consulté entre 1999 et 2006 pour une MTEV, confirmée par veinographie, échographie, angiographie et/ou scanner ventilation/perfusion. La particularité de ces patients est l'âge précoce de survenue de la MTEV. En effet, ils devaient avoir moins de 50 ans au moment du diagnostic. En outre, ces patients ne devaient présenter aucun des facteurs de risque constitutionnels de MTEV (déficit en AntiThrombine, en Proteine C, en Proteine S, homozygotie pour le FV Leiden, homozygotie pour la mutation FII 20210 A), ni facteurs de risque acquis (chirurgie, hospitalisation, grossesse et post partum, contraception orale, cancer, maladie auto-immune). Les témoins, non appariés, proviennent de la cohorte Suvimax [120]. Ils devaient être d'origine européenne et exempts de toute pathologie chronique ou traitement médical.

V.2. Les données génétiques

V.2.1. L'échantillon FVL

Microsatellites : Un criblage du génome de l'ensemble de l'échantillon a d'abord été obtenu au moyen d'une puce incluant plus de 1000 microsatellites, conçue par la société deCODE. Ces marqueurs, particulièrement adaptés aux études de liaison génétique, étaient répartis en moyenne tous les 4 cM réalisant une couverture du génome particulièrement dense. Ces données étaient disponibles dès le début de ce travail.

La totalité de l'information génotypique obtenue grâce aux microsatellites a été intégrée afin de vérifier que les liens de parentés rapportés dans la base de données étaient compatibles avec les génotypes. J'ai utilisé à cet effet le logiciel RelPair [121]. Pour chacune des paires d'individus réalisables à partir de l'échantillon, le lien de parenté le plus probable est estimé par le calcul de la vraisemblance des génotypes observés sous diverses hypothèses de lien de parenté. La quasi-totalité des liens de parentés rapportés était confirmée par les génotypes. Le programme a toutefois permis de corriger le statut monozygote chez des jumeaux en réalité dizygotes. Il a de plus identifié des fratries présentant des demi-frères ou sœurs. Dans ces cas, j'ai créé artificiellement dans la base de données un individu fondateur supplémentaire afin que les informations de transmission génétique soient transmises de manière exacte au logiciel d'analyse de liaison.

SNPs : Pendant le déroulement de ma thèse, les individus de l'échantillon FVL ont été génotypés au moyen de la puce Illumina Human 660W-Quad contenant 547 886 SNPs. L'objectif était de compléter les études de liaison par des analyses d'association génome-entier. Deux filtres successifs ont été appliqués à l'ensemble de ces SNPs (voir **tableau 5 p38**). Le premier a éliminé des analyses les 543 SNPs dont le génotypage avait été obtenu pour moins de 90% de l'échantillon. Le deuxième a éliminé les 57 260 SNPs dont l'allèle le moins fréquent était observé moins de 20 fois dans l'ensemble de l'échantillon. Les analyses ont finalement été réalisées sur 490 083 SNPs. L'équilibre d'Hardy-Weinberg (voir encadré **pA6**) a été testé chez les membres fondateurs pour tous les SNPs présentant des associations intéressantes. Le non respect de cet équilibre n'a cependant pas été retenu comme critère de filtrage.

Grâce à un calcul de pourcentage d'allèles Identiques par Descendance (IBD), implémenté dans le logiciel PREST [122], conduit sur l'ensemble des SNPs, Apostolos Dimitromanolakis, statisticien dans le laboratoire de France Gagnon, a confirmé l'absence d'erreur concernant les liens de parentés rapportés dans la base nettoyée préalablement grâce

aux microsatellites. De plus, le taux d'erreurs mendeliennes était très faible : seules 14 949 erreurs mendeliennes ont été rapportées, soit environ 0.01% de tous les génotypes. Les génotypes considérés comme erronés par PREST ont été retirés des analyses.

V.2.2. MARTHA08 et MARTHA10

Les patients de MARTHA08 ont été génotypés à l'aide de la puce Illumina Human 610-Quad alors que ceux de MARTHA10 l'ont été à l'aide de la puce Illumina 660W-Quad. Le nombre de SNPs génotypés étaient de 567 589 dans MARTHA08 et de 543 391 dans MARTHA10. Trois filtres successifs ont été appliqués. Le premier requérait un taux de génotypage par SNP supérieur à 99% (522 041 SNPs dans MARTHA08 et 523 194 SNPs dans MARTHA10) ; le deuxième, une MAF (fréquence de l'allèle le plus rare) supérieure à 5% (494 958 dans MARTHA08 et 501 773 dans MARTHA10) ; et le troisième, une absence de déséquilibre de Hardy-Weinberg au seuil de significativité $p<10^{-5}$. Les analyses d'association ont donc portées sur respectivement 494 722 et 501 693 SNPs dans MARTHA08 et MARTHA10.

Une sélection plus stricte des individus de MARTHA08 et MARTHA10 a été réalisée par Marine Germain, ingénieure de recherche à l'UMR_S 937, à partir des données génotypiques obtenues. En effet, les individus présentant des liens de parentés ont pu être identifiées grâce au logiciel Plink utilisant le calcul d'allèles identiques par état (*IBS* en anglais pour « Identical By State ») et le « Multi-Dimensional Scaling » [123]. De plus, le programme Eigenstrat, qui étudie la stratification des échantillons par analyse en composante principale, a permis d'identifier les personnes dont le fond génétique diffère de celui de la population européenne [124]. Enfin, les individus dont le taux de génotypage était inférieur à 95% ont également été retirées de l'échantillon. Finalement, 972 individus de MARTHA08 et 570 de MARTHA10 ont été gardés pour les analyses d'association.

Tableau 5: nombre de SNPs génotypés, filtrés et analysés dans les échantillons familles-FVL , MARTHA08 et MARTHA10

	FVL	MARTHA08	MARTHA10
Génotypés	547 886	567 589	543 391
Filtrés en raison du/de la :			
Taux de détermination	543	45 548	20 197
MAF	57 260	27 083	21 421
Test de HW	0	236	80
Analysés	490 083	494 722	501 693
En commun dans les trois études	442 728		

Le génotypage de trois SNPs du gène *ABO,* extrait des données des puces Illumina Human 610 et 660Quad, a permis de déterminer les allèles A1, A2, B ou O de ce dernier [125], dans les échantillons FVL, MARTHA-08 et MARTHA-10, comme indiqué dans le **tableau 6**.

Tableau 6. Correspondance entre trois SNPs et les allèles A1, A2, B et O du gène *ABO*

	rs8176704	**rs8176746**	**rs505922**
A1	G	C	G
A2	**A**	C	G
B	G	**A**	G
O	G	C	**A**

V.2.3. L'étude GWAS sur la MTEV (données *in silico)*

Tous les sujets de cette étude ont été génotypés au moyen de la puce Illumina Sentrix HumanHap300 contenant 317 139 SNPs. Trois filtres ont été appliqués: - taux de génotypage par SNP inférieur à 97%, - MAF inférieure à 5% chez les patients et à 1% chez les témoins, - déviation à l'équilibre d'Hardy-Weinberg significative au seuil $p<10^{-5}$. Les analyses d'association ont été réalisées avec les 291 872 SNPs qui passaient ces filtres. Les personnes dont le taux de génotypage était inférieur à 95% ont été exclues des analyses. Comme indiqué précédemment, l'information génotypique sur un très grand nombre de SNPs a permis d'identifier et d'exclure les personnes d'origine autre qu'européenne, de même que les personnes apparentées. Finalement les analyses ont porté sur 419 patients et 1228 témoins. Le nettoyage de la base de données, et les analyses ont été effectuées par Simon Heath au Centre National de Génotypage. Les résultats ont été transmis à David Trégouët et son équipe.

V.2.4. Stanislas, MARTHA05 et FARIVE

Les échantillons Stanislas, MARTHA05 et FARIVE ont servi à reproduire les résultats obtenus dans les échantillons FVL, MARTHA08, MARTHA10 et l'étude "GWAS" *in silico*. Les banques d'ADN de ces études étaient gardées dans le laboratoire du Pr Morange. Le génotypage de SNPs candidats y a été réalisé par la méthode TaqMan (Applied Biosystems). L'équilibre d'Hardy-Weinberg des SNPs génotypés a été testé chez les parents de l'échantillon Stanislas et chez les témoins de MARTHA05 et FARIVE (voir annexe **pA6**). De plus, d'éventuelles erreurs de génotypage ont été recherchées dans l'échantillon Stanislas en vérifiant la cohérence des génotypes entre les apparentés d'une même famille. Les quelques incohérences retrouvées, moins de 2% de l'ensemble des génotypes, ont parfois pu être corrigées lorsqu'il n'y avait aucune ambigüité sur l'origine de l'erreur. Dans le cas contraire,

les génotypes de tous les membres de la famille concernée ont été considérés comme manquants.

V.3. Mesure des traits quantitatifs

Les taux d'activité plasmatique du FVIII et les taux d'antigénémie plasmatique du vWF ont été obtenus à partir des mêmes techniques de mesure dans les échantillons Familles-FVL, Stanislas, MARTHA05 (seulement les cas), MARTHA08 et MARTHA10. L'activité du FVIII a été mesurée grâce à un coagulateur automatique (Behring Coagulation System, de Siemens, pour l'échantillon Familles-FVL ; Star-R, de Diagnostica Stago pour les quatre autres). L'antigénémie du vWF a été obtenue au moyen d'un test ELISA (kit Asserachrom vWF, de Diagnostica Stago pour les cinq échantillons). Par simplification d'écriture, les taux d'activité plasmatique du FVIII et les taux d'antigénémie plasmatique du vWF seront par la suite appelés taux de FVIII et taux de vWF.

Ils ont été mesurés après une nuit de jeûne, y compris tabagique. Pour les cas index de l'échantillon Familles-FVL et les patients du projet MARTHA, ils ont été mesurés à distance de l'événement thrombothique. Ces mesures ont été obtenues chez 253 sujets de Familles-FVL (sur 255 sujets), et chez la totalité des 451 individus de Stanislas. Dans l'échantillon MARTHA05, les taux de FVIII et de vWF étaient connus pour respectivement 578 et 1 048 patients (sur 1 150 cas). Ces chiffres étaient de 541 et 834 parmi les 972 patients de MARTH08, et de 548 et 537 parmi les 570 patients de MARTHA10. Les distributions de ces traits sont présentées **figures 4 et 5**.

Figure 4: taux plasmatiques de vWF et de FVIII dans les échantillons Familles-FVL, Stanislas et MARTHA

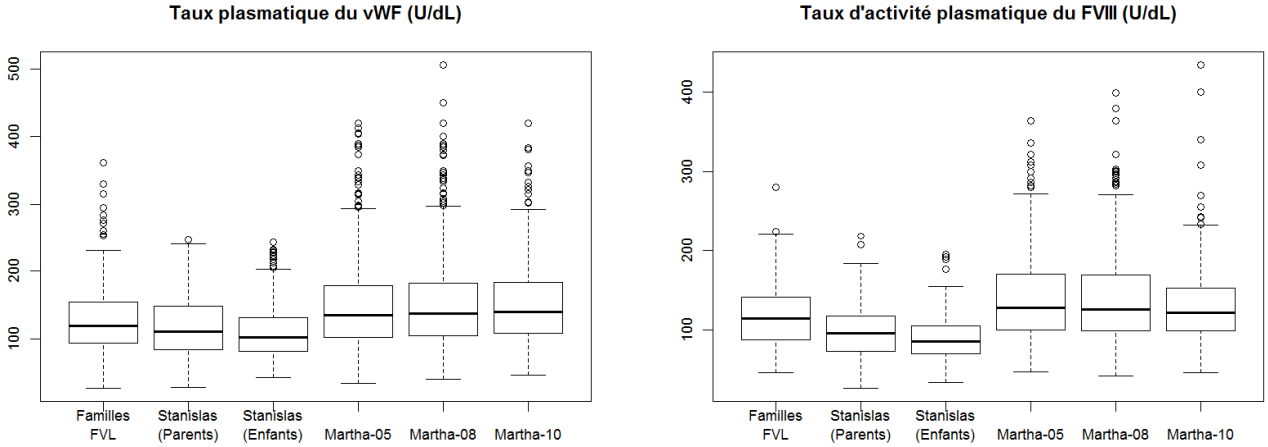
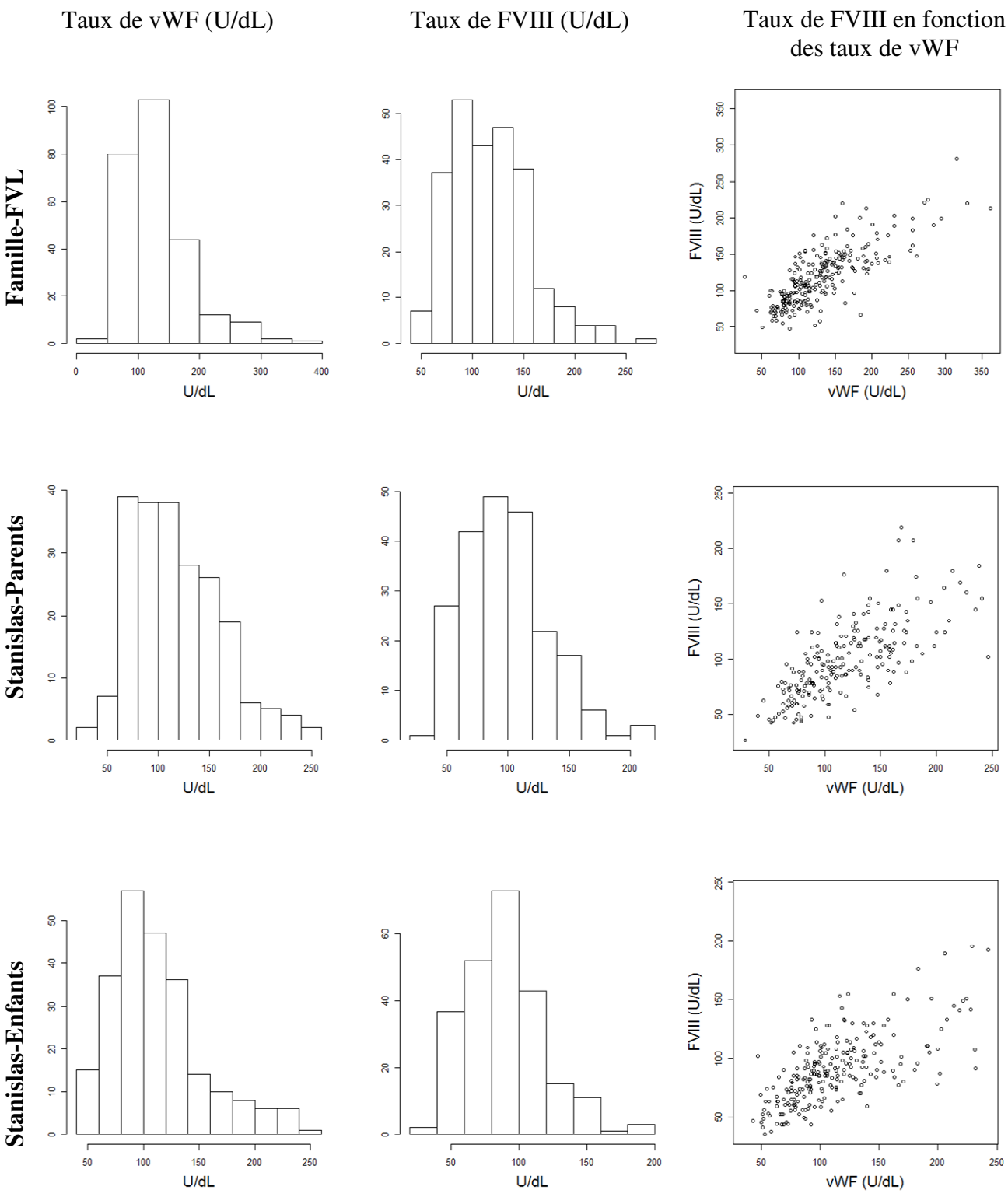
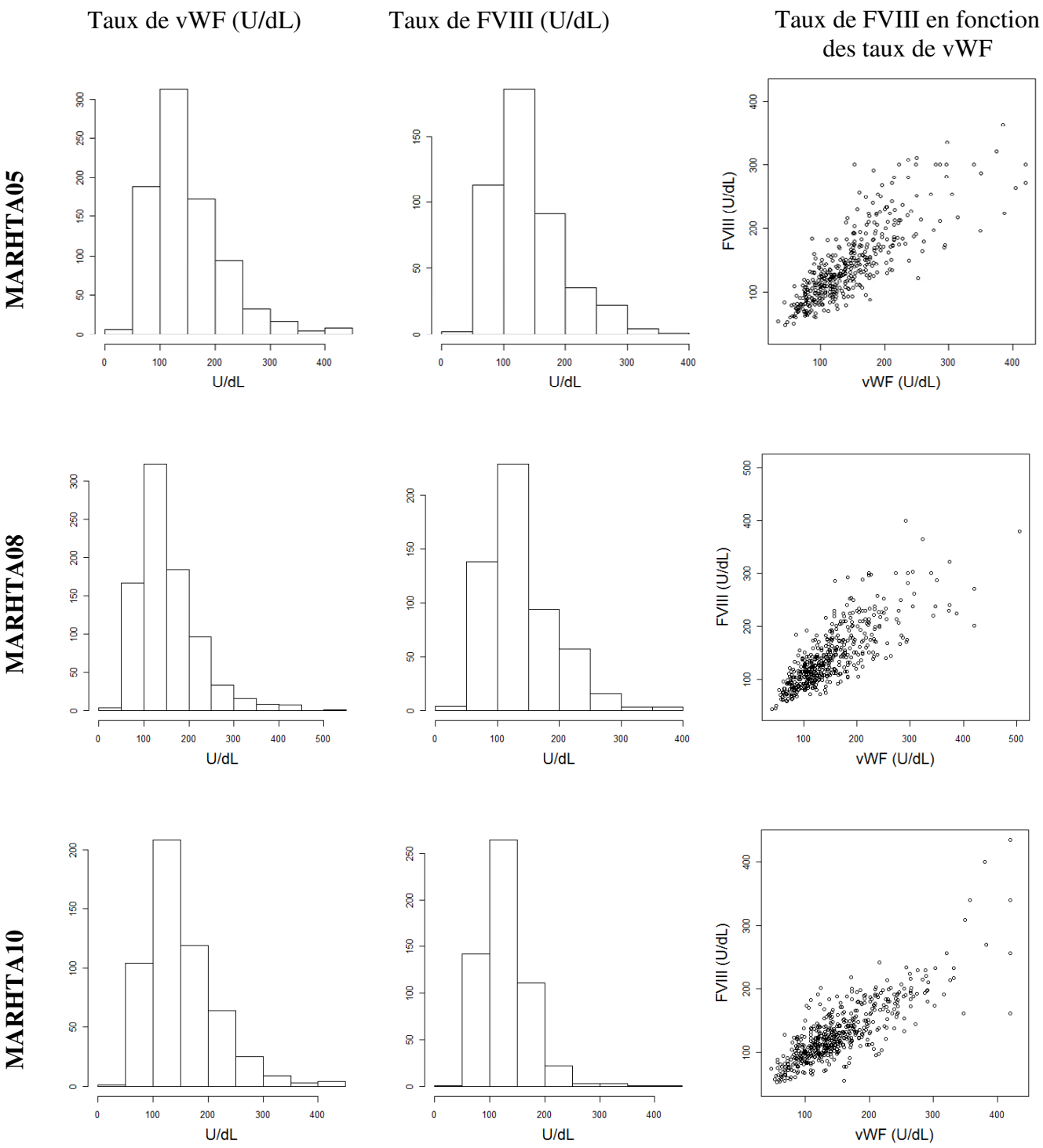


Figure 5: distribution et corrélation entre les taux plasmatiques de vWF et de FVIII dans les échantillons Familles-FVL, Stanislas et MARTHA





Les corrélations entre les taux de vWF et de FVIII sont de 0,78 dans l'échantillon familles-FVL, 0,72 chez les parents de l'échantillon Stanislas, 0,69 chez les enfants de l'échantillon Stanislas, 0,81 dans MARTHA05, 0,80 dans MARTHA10 et 0,82 dans MARTHA10.

Tableau 7. Caractéristiques des échantillons

	Familles-FVL	Stanislas	GWAS *in silico*	FARIVE	MARTHA05	MARTHA08	MRTHA10*
Type d'échantillon	5 grands pedigrees	108 familles nucléaires	Cas-témoins	Cas-témoins	Cas-témoins	Observatoire de cas*	Observatoire de cas*
Effectif	255 (dont 254 vWF et 253 FVIII)	451 (aucun phénotype manquant)	419 cas 1228 témoins	607 cas 607 témoins	1150 Cas (dont 1048 vWF et 548 FVIII) 801 témoins	972 (dont 834 vWF et 541 FVIII)	570 (dont 537 vWF et 548 FVIII)
Recrutement	Un probant présentant une MTEV et une mutation FVL	Familles en bonne santé, de la région Meuthe et Moselle	Cas : 4 CHU français (Grenoble, Marseille, Montpellier, Paris) Témoins : SUVIMAX	Cas et témoins : patients de centres hospitaliers parisiens	Centre de Thrombophilie de Marseille Témoins FITENAT: 40% mutations FVL ou FII	Patients du centre de Thrombophilie de Marseille	Patients du centre de Thrombophilie de Marseille
Caractéristiques de la MTEV	Absence de fdr acquis (1) ou constitutionnel (3)	-	Age<50ans Absence de fdr acquis (2) ou constitutionnel (4)	Absence de cancer	Absence de fdr constitutionnel (4)	Absence de fdr constitutionnel (4)	Absence de fdr constitutionnel (4)
Analyses réalisées	vWF et FVIII liaison et association pangénomique	vWF et FVIII association gènes candidats	MTEV association pangénomique	MTEV association gènes candidats	MTEV, vWF, FVIII association gènes candidats	vWF et FVIII association pangénomique	vWF et FVIII association pangénomique
Objectif	Identification de régions/gènes candidats	Etude du gène candidat *BAI3*	« fine mapping » des régions candidates	Etude de réplication de *BAI3*	Etude de réplication de *BAI3*	Identification de gènes candidats	Identification de gènes candidats
		Parents / enfants	**Témoins / Cas**	**Témoins / Cas (5)**	**Témoins / Cas (6)**		
Age : Moyenne [min-max]	40,4 [19-93]	45,8 [36-65] / 17,7 [8-30]	50,2 / 36,2	51,5 [18-91] / 32,9 [18-49]	47,7 [18-74] / 32,4 [18-49]	45,7 [18-91]	49,2 [18-88]
Sexe (% femmes)	50,6	50 / 48,1	69,9 / 44,8	57,3 / 82,0	52,2 / 69,9	70,8	58,2
Tabagisme (%)	24,4	24,1 / 15,4	Non Disponible	44,5 / 48,8	27,4 / 35,4	24,9	22,7
FVIII (UI/dL) Moyenne (Ecart-type)	118,6 (38,5)	99,0 (34,3) / 89,5 (29,0)	Non Disponible	Non Disponible	140,7 (56,5)	138,7 (55,3)	130,2 (46,4)
vWF (UI/dL) Moyenne (Ecart-type) DM	130,3 (53,2)	117,8 (43,2) / 112,0 (42,8)	Non Disponible	Non Disponible	148,9 (66,4)	152,3 (68,2)	152,9 (63,9)
FVL (%)	24,9	Non Disponible	Non Disponible	4.6 / 15,0	22,5 / 30.2	26.6	14,1
FII-G20210A (%)	0,4	Non Disponible	Non Disponible	3,2 / 11	18,2 / 13,7	15,9	10,6
ABO (%)O **A** **B** **AB**	40,6 57,8 1,6 0	42,1 / 37,1 47,2 / 48,9 8,4 / 7,6 2,3 / 6,3	Non Disponible	45,2 / 21,7 42,1 / 58,8 8,7 / 12,2 4,2 / 7,2	42,6 / 24,6 44,2 / 59,7 9,0 / 10,1 4.2 / 5.6	22,9 61,8 10,3 5	22,4 59,3 14,4 3,9

(1) fdr: facteur de risque; cancer, grossesse et post-partum, immobilisation, traumatisme, chirurgie, syndrome des anticorps antiphospholipides

(2) cancer, grossesse et post-partum, hospitalisation, chirurgie, contraception orale, maladie auto-immune

(3) déficit en AT, en PC ou en PS

(4) déficit en AT, en PC, en PS, homozygotie pour le FV Leiden, homozygotie pour la mutation FII 20210 A

(5) sélection des cas de moins de 50 ans (N=250)

(6) sélection des cas de moins de 50 ans (N=916)

VI. Méthodes d'analyse statistique

VI.1. Analyse de liaison dans l'échantillon Familles-FVL

VI.1.1 Présentation des méthodes classiques d'analyse de liaison-

Les analyses de liaison permettent d'estimer la distance génétique entre deux loci. Si deux variations génétiques (ou marqueurs) sont suffisamment "proches", elles seront transmises ensemble d'un parent à son enfant plus souvent que ne le voudrait le hasard. La distance génétique se mesure à partir de l'estimation du taux de recombinaison θ survenant entre ces deux loci au cours des méioses. Cette estimation repose sur le génotypage, dans une famille, des deux marqueurs dont on cherche à connaître la distance génétique, puis sur l'observation du nombre de gamètes recombinants et parentaux transmis des parents aux enfants. En absence de liaison, on observe autant de gamètes recombinants que parentaux. Le taux de recombinaison θ est alors égal à 0,5. Il est d'autant plus proche de 0 que les deux marqueurs sont liés. L'estimation de θ est obtenue par maximisation de la vraisemblance $L(\theta)$. Celle-ci est la probabilité de survenue des recombinaisons effectivement observées dans l'échantillon, et calculée pour différentes valeurs de θ variant entre 0 et 0,5. L'estimateur du maximum de vraisemblance, θ_{max}, permet de mesurer la force de la liaison par le calcul du Lod-Score Z [126]

$$Z_{\max} = \log_{10} \frac{L(\theta = \theta_{\max})}{L(\theta = 0.5)}$$

Cette méthode permet d'identifier le ou les loci influençant un phénotype binaire (présence de la maladie ou non) ou quantitatif, par l'estimation de la liaison génétique entre des marqueurs génétiques génotypés et le(s) polymorphisme(s) responsable(s) du phénotype étudié. Les génotypes du (des) polymorphisme(s) "fonctionnel(s)" étant bien évidemment inconnus, ils sont inférés à partir des observations phénotypiques et de la connaissance de plusieurs paramètres caractérisant le modèle génétique supposé: fréquence des allèles au locus-phénotype, moyenne et variance du phénotype conditionnellement au génotype (pour un phénotype quantitatif), pénétrance (pour un phénotype binaire). Le modèle génétique peut être issu de connaissances antérieures ou s'estimer par des analyses de ségrégation. Cette méthode est particulièrement puissante dans le cadre des maladies monogéniques. Elle est par contre moins adaptée aux maladies complexes ou aux traits quantitatifs dont la variabilité phénotypique est la

résultante de multiples facteurs génétiques et environnementaux qui empêchent une estimation correcte d'un modèle génétique calqué sur un modèle monogénique. La perte de puissance entraînée par cette mauvaise estimation est telle que des méthodes dites « modèle-indépendantes » se sont développées.

Il existe différentes méthodes « modèle-indépendantes » d'analyse de liaison d'un phénotype quantitatif. Ces méthodes ne nécessitent pas l'inférence du génotype au locus-phénotype. Elles étudient directement le lien entre le phénotype et le génotype du marqueur. Elles concluent à une liaison génétique entre les locus-phénotype et locus–marqueur lorsque des individus apparentés et de phénotype proche partagent, pour le marqueur, plus d'allèles hérités d'un même ancêtre (allèle identiques par descendance – IBD) que ne le voudrait le hasard des transmissions. Il existe classiquement deux grandes familles de méthodes d'analyses de liaison de traits quantitatifs. La première est une méthode régressive dont les bases ont été posées par Haseman et Elston [127]. La deuxième est une méthode de décomposition de la variance (VC pour Variance Component), initialement développée par Amos [128] et Goldar [129].

Dans sa formulation originale, la méthode régressive nécessite un échantillon constitué exclusivement de paires de germains. Par la méthode des moindres carrés, elle réalise la régression de la différence élevée au carré $Yi^2=(Z_{i1}-Z_{i2})^2$ sur π_i , où Z_{i1} et Z_{i2} sont les phénotypes de la paire de germains i, et π_i est la proportion d'allèles IBD partagés par cette paire. En cas d'absence de liaison, Y_i^2 est indépendant de π_i. La pente de la droite de régression est alors nulle. Elle est par contre négative en cas de liaison : plus le pourcentage d'allèles IBD partagés entre une paire de germains est grand, plus la différence phénotypique de cette paire est petite. Des développements de cette méthode permettent de l'appliquer à d'autres types de paires d'apparentés et à des familles étendues [130][131]. Toutefois, ces méthodes ne sont pas adaptées à l'étude de familles de structure très complexe, tel l'échantillon Famille-FVL. Elles nécessitent en effet de scinder chaque pedigree en plusieurs sous-familles, diminuant la puissance apportée par de tels échantillons [132].

Les bases de la méthode par décomposition de la variance sont développées au **§VI.2.1 p57**. Nous retiendrons pour l'instant qu'elle a bénéficié d'important développement, grâce aux travaux d'Almasy et Blangero [133], permettant de réaliser des analyses de liaison dans de très grandes familles de structure complexe. Elle repose sur un calcul de maximisation de la vraisemblance qui suppose que le phénotype suive

une loi multinormale. Elle s'avère être très puissante, à condition toutefois que l'hypothèse de normalité soit vérifiée. Dans le cas contraire, et notamment en présence d'un excès de valeurs extrêmes, cette méthodologie se traduit par une augmentation de l'erreur de type 1. De tels écarts à la normalité peuvent se rencontrer en cas de sélection de l'échantillon sur un critère phénotypique. Cet écueil pourrait concerner certains traits quantitatifs mesurés dans les Famille-FVL en raison du mode de recrutement *via* un cas de MTEV. En outre, le processus de maximisation de la vraisemblance devient démesurément long en présence de structures familiales complexes et d'un grand nombre de marqueurs.

Pour ces raisons, nous nous sommes tournés vers l'approche bayésienne fondée sur la méthode de Monte Carlo par Chaînes de Markov implémentée dans le logiciel "Loki" [134][135]. Cette approche envisage de nombreuses configurations génotypiques [nombre, position(s), effet(s) du ou des loci incriminés (ou QTL pour Quantitative Trait Loci)] potentiellement observables, puis étudie leur probabilité compte-tenu des données effectivement observées (structure familiale, marqueurs génétiques, phénotypes). Alors qu'un calcul de maximisation de la vraisemblance exigerait d'énumérer de façon exhaustive ces configurations, la MCMC n'en étudie qu'un échantillon, en les sélectionnant par une série de tirages aléatoires itératifs.

VI.1.2. Principe des statistiques bayésiennes par une méthode de Monte Carlo par Chaînes de Markov (MCMC)

Soit X, un vecteur d' observations suivant une loi de probabilité, P_θ , dépendante d'un paramètre d'intérêt θ. Contrairement aux méthodes fondées sur le maximum de vraisemblance, les approches bayésiennes ne visent pas à estimer le paramètre d'intérêt θ, mais à évaluer sa loi de probabilité, conditionnellement à l'observation des données, $\Pi(\theta|X)$. $\Pi(\theta|X)$ est appelé loi de probabilité *a posteriori* de θ. Soit $\Pi(\theta)$, une loi de probabilité de θ, définie *a priori.* On peut écrire :

$$\boxed{\Pi(\theta|X) = \frac{\Pi(\theta)P(X|\theta)}{P(X)}} \qquad (1)$$

Dans les cas simples, on peut calculer $P(X)$:

$$P(X) = \int \Pi(\theta)P(X|\theta)d\theta$$

Lorsqu'il n'est pas possible de calculer de cette manière $P(X)$, on s'aide alors d'une méthode de Monte Carlo par Chaînes de Markov (MCMC). Cette méthode s'affranchit de la nécessité de calculer $P(X)$. Elle réalise une succession de tirages

aléatoires de θ permettant d'obtenir $\theta_0, \theta_1, \ldots, \theta_n, \theta_{n+1}, \ldots$. Le premier tirage, θ_0, est réalisé à partir de la loi *a priori* $\Pi(\theta)$. Chaque nouveau tirage est dépendant du précédent. Le processus qui permet le passage de θ_n à θ_{n+1} se fait en plusieurs étapes. Il fait intervenir un algorithme décisionnel d'acceptation ou de rejet. Nous allons présenter ici l'algorithme de Metropolis-Hasting, qui est utilisé dans le logiciel Loki [136].

Cet algorithme propose à chaque étape $n+1$, une nouvelle valeur de θ, notée θ'. θ' est issu d'un tirage au sort à partir d'une loi de probabilité Q, qui dépend de la dernière valeur de θ, θ_n . Par exemple, Q peut être une loi normale de moyenne θ_n . Ensuite, l'algorithme décide si θ' est rejeté ou accepté comme nouvelle valeur de θ, θ_{n+1}. La probabilité d'acceptation est fonction du rapport des probabilités de θ' et de θ_n , conditionnellement aux observations (ou rapport des probabilités *a posteriori* de θ' et de θ_n):

$$R = \frac{\Pi(\theta'|X)}{\Pi(\theta_n|X)}$$

qui d'après (1) est égal à

$$R = \frac{\Pi(\theta')P(X|\theta')}{\Pi(\theta_n)P(X|\theta_n)} \qquad (2)$$

Ce rapport dépend donc du rapport des probabilités *a priori* de θ' et de θ, ainsi que de leur rapport de vraisemblance, $L_X(\theta')$ sur $L_X(\theta)$.

Une valeur α est tirée au sort d'une loi uniforme U(0,1). La valeur θ' est acceptée (c'est-à-dire $\theta_{n+1} = \theta'$) si α répond à la condition

$$\alpha < R\frac{Q(\theta_n|\theta')}{Q(\theta'|\theta_n)}$$

Dans le cas particulier d'une distribution Q symétrique, telle une loi normale de moyenne θ_n , avec laquelle la probabilité de tirer θ' en partant de θ_n est égale à la probabilité de tirer θ_n en partant de θ', θ' est accepté si $\alpha<R$. Il est ainsi toujours accepté si sa probabilité *a posteriori* est meilleure que celle de θ_n. Dans le cas contraire, θ' sera accepté avec une probabilité égale au rapport R des probabilités *a posteriori* de θ' et θ_n .

Cet algorithme est conçu de telle façon que, d'itération en itération, la distribution des valeurs θ_n suivent une loi qui se rapproche de la loi *a posteriori* $\Pi(\theta|X)$. Cette phase d'approche est appelée « période de burning ». A partir d'un certain nombre d'itérations, la période de burning est terminée, et tous les θ_n sont issus de la loi $\Pi(\theta|X)$.

On continue alors de la même façon un très grand nombre de tirages itératifs de θ_n . On peut alors déduire la loi de probabilité *a posteriori* de θ par la simple observation des valeurs de θ_n obtenues. Afin de s'assurer du bon déroulement de l'algorithme, deux caractéristiques doivent être vérifiées : le mixing et la convergence. Un bon mixing implique que l'algorithme balaie l'ensemble des possibilités offerte par la loi de probabilité. Il ne doit pas *a contrario* rester bloqué sur une valeur particulière du paramètre. Quant à la convergence, elle est obtenue quand l'algorithme propose de façon préférentielle un tirage particulier correspondant à un maximum local de probabilité.

VI.1.3. Application aux analyses conjointes de ségrégation et de liaison

Dans le cadre des analyses de liaison, on suppose un ou plusieurs polymorphismes bi-alléliques, dont les effets sont additifs, à la fois entre les deux allèles présents au même QTL, mais aussi entre les différents QTL. Les génotypes de ces polymorphismes n'étant pas observés, on les nomme des variables « latentes ». A chaque itération correspond un tirage (ou « échantillonnage ») au cours duquel le programme implémenté dans Loki impute des génotypes à chaque variable latente, *i.e* aux QTLs, d'abord aux fondateurs, puis à leurs descendants en respectant les lois de transmissions mendeliennes. Il découle de cet échantillionnage un certain nombre de paramètres θ d'intérêt (nombre de QTLs, localisation sur le génome, intensité de l'effet) que l'on peut estimer à l'aide d'un modèle linéaire explicité dans la section suivante. Comparativement au principe général des méthodes MCMC présentées précédemment, ce ne sont pas directement les valeurs θ qui sont tirées au sort ici, mais simplement les variables latentes. Pour chacun des modèles analysés, j'ai réalisé 500 000 itérations. J'ai ignoré les 5 000 premières, considérant qu'elles correspondaient à la période de burn-in.

Les modèles du phénotype et des marqueurs

A chaque nouvel échantillonnage, l'algorithme calcule le rapport R (2). Ce calcul nécessite l'estimation de la vraisemblance du modèle qui s'écrit:

P(Phénotypes|modèle phénotypique, modèle des marqueurs, observations des marqueurs, observations des covariables, observations des liens familiaux)

Le modèle phénotypique repose sur un certain nombre d'hypothèses que nous supposons vraies : les résidus sont distribués normalement ; les effets du modèle sont additifs ; les membres fondateurs sont issus d'une population panmictique où l'équilibre

d'Hardy-Weinberg est vérifié; il n'y a pas de déséquilibre de liaison entre les marqueurs.

Il s'agit d'un modèle linéaire du type :

$$y_j = \mu + \sum_{i=1}^{n} Q_{ji} + \sum_{i=1}^{s} \beta_i C_{ji} + \sum_{i=i}^{k} G_{ji} + e_j$$

y_j : le phénotype, *i.e* le trait quantitatif (TQ), observé pour l'individu j

μ : la moyenne du TQ

e_j : l'erreur résiduelle, que l'on suppose suivre une loi normale $N(0,V_e)$

k : le nombre de Loci (QTL) contribuant à la variabilité du TQ. Ce nombre varie d'une itération à l'autre, puisqu'il est estimé par une procédure bayésienne.

G_{ji} : effet du $i^{ème}$ QTL chez l'individu j. Chaque QTL est supposé biallélique. Il est caractérisé par la fréquence de son allèle mineur, ainsi que par les trois moyennes phénotypiques correspondant aux trois génotypes. L'effet génotypique moyen n'est autre que l'écart de ces moyennes à la moyenne globale.

Ce modèle inclut également deux types de covariables. Les covariables du premier type, dites « classiques », sont traitées comme elles le seraient dans un modèle linéaire d'épidémiologie générale. L'âge, le sexe, ou n'importe quelle variable qualitative ou quantitative, peuvent ainsi être inclus dans le modèle phénotypique.

s : le nombre de covariables classiques.

C_{ji} : valeur de la $i^{ème}$ covariable chez l'individu j

β_i : effet de la $i^{ème}$ covariable

Les covariables du second type sont des gènes majeurs. Il s'agit de gènes qui ont un effet connu sur le trait quantitatif étudié et dont le génotype est effectivement observé pour l'ensemble des sujets de l'étude, à l'exception des valeurs manquantes.

n : le nombre de gènes majeurs. Contrairement à k, ce nombre est fixé par l'utilisateur, qui choisit d'inclure ou non dans le modèle certains gènes majeurs.

Q_{ji} : effet du $i^{ème}$ gène majeur chez l'individu j, modélisé de la même manière que G_{ji}, à la différence que le gène majeur peut avoir plus de deux allèles.

NB : Un gène majeur peut être déclaré comme une covariable classique. S'il n'y avait aucune donnée manquante concernant ce gène, les résultats seraient identiques. En revanche, un sujet présentant une donnée manquante pour une covariable classique est ignoré dans les analyses. A contrario, s'il s'agit d'une donnée manquante pour le génotype d'un gène majeur, celui-ci peut être imputé à partir des autres données génotypiques (marqueurs de l'individu, marqueurs et gène majeur d'autres individus lui étant apparentés).

Au total, ce modèle phénotypique, appliqué à un échantillon familial dont on a recueilli un phénotype quantitatif, mais pas de marqueurs génétiques, permet de réaliser une analyse de ségrégation. Le nombre de QTL peut en effet être estimé, ainsi que leurs effets génétiques. Ces derniers correspondent au pourcentage de la variance du phénotype expliqué par chaque QTL. En présence de marqueurs génétiques, un paramètre supplémentaire λ_i est estimé, correspondant à la localisation sur le génome du $i^{ème}$ QTL ; ce qui permet de réaliser une analyse de liaison conjointement à l'analyse de ségrégation.

En plus du modèle phénotypique, on doit alors définir un second modèle, celui des marqueurs. Il nécessite la spécification des fréquences alléliques, de la position relatives des marqueurs les uns par rapport aux autres, et d'une fonction cartographique permettant de rétablir l'additivité des distances intermarqueurs. Pour cela, la carte Haldane est la plus usitée.

Les distributions *a priori*

Pour réaliser les analyses de liaison-ségrégation, on doit spécifier les distributions *a priori* de k , de G_{ji} , de λ_i et de p_{Ai} la fréquence de l'allèle A du $i^{ème}$ QTL. Les distributions *a priori* peuvent être assez éloignées des distributions réelles, à condition qu'elles soient assez larges (grande étendue pour une loi uniforme ou grand variance pour une loi Normale, par exemple). Une distribution très large entraîne une longue période de burning, avant de se stabiliser autour de la loi exacte $\Pi(\theta|X)$. En revanche, la loi exacte $\Pi(\theta|X)$ risque de n'être envisagée à aucun moment par l'algorithme, si la distribution de départ est à la fois éloignée de la distribution réelle, et très étroite.

J'ai utilisé :

-Pour λ_i, une loi uniforme sur l'ensemble du génome

-Pour k, une loi de poisson, de moyenne 1 ou 2, et tronquée à 17

-Pour p_{Ai}, une loi uniforme (0,1)

-Pour G_{ji} : une loi normale de moyenne 0 et de variance τ.

Les valeurs optimales de τ et de la moyenne de k sont déterminées par des analyses de ségrégations préalables, comme nous le verrons lors de la présentation des résultats des analyses de ségrégations.

VI.1.4. Exploitation des résultats obtenus à la suite d'une analyse conjointe de liaison et de ségrégation : graphiques, Facteur Bayésien, et valeur de p empirique

Après avoir réalisé une analyse conjointe de liaison et de ségrégation par MCMC, l'utilisateur obtient autant de lignes de résultats qu'il a demandé d'itérations. Chaque ligne indique la valeur estimée à partir du tirage *i* d'un certain nombre de paramètres, dont le nombre de QTL, leurs localisations, leurs effets, ainsi que ceux des covariables. C'est à l'utilisateur d'exploiter secondairement l'ensemble de ces lignes. Il s'agit là, d'une part de juger de la qualité du mixing et de la convergence, et d'autre part de connaître la loi de probabilité *a posteriori* des paramètres, et d'en déduire la valeur la plus probable de ces paramètres. L'exploitation graphique est souvent la plus informative.

Graphiques permettant de juger de la qualité de l'algorithme

Il est utile de représenter graphiquement à l'aide d'un simple nuage de points la valeur estimée du nombre de QTL en ordonnée et le numéro de l'itération en abscisse (exemple en **figure 7.A p73**). On peut de même examiner l'évolution de la variance résiduelle e d'une itération à l'autre (exemple en **figure 7.B p73**). Pour ces deux graphiques, on vérifie la qualité du mixing, en s'assurant visuellement que l'ensemble de l'espace possible est bien balayé, et qu'il n'y a pas de blocage sur une valeur particulière (absence de segment horizontal). Un troisième graphique permet de vérifier qu'il y a bien eu convergence, en regardant la variation de la position λ_i du QTL au cours des itérations (représentation de λ_i en fonction du numéro i de l'itération, exemple en **figure 9 p74**). L'ensemble de l'espace doit être criblé de points, avec

parfois, en cas de liaison, une préférence pour une valeur de λ_i tout au long des itérations, se traduisant par une ligne horizontale plus ou moins épaisse et nette.

Graphique permettant de juger de l'importance d'un QTL

Les graphiques les plus informatifs à réaliser sont probablement ceux qui, pour chaque chromosome, représentent la taille du QTL (*i.e* la racine carrée de la variance du phénotype attribuable au QTL) en fonction de sa position. En supposant qu'un QTL en moyenne soit identifié par chromosome et par itération, et que nous ayons réalisé une analyse de 100000 itérations, chaque graphique sera composé de 100000 points en moyenne. Un signal de liaison se caractérise par une concentration de points, étroite, distincte du bruit de fond. Ces images sont observées lorsqu'un QTL, de taille modérée ou importante, est identifié dans la même région, de manière récurrente au cours des itérations. Inversement, des points répartis de manière homogène sur la partie basse du graphique correspondent au bruit de fond attendu en cas d'absence de QTL.

Le Facteur Bayésien (BF)

Le résultat d'une analyse par MCMC, nous l'avons vu, est une distribution du (ou des) paramètre(s) d'intérêt, et non pas sa (ou leur) seule estimation. Pour une analyse de liaison, le paramètre de plus grand intérêt est la présence d'un signal de liaison à un locus particulier. Le pourcentage d'itérations au cours desquelles un QTL a été envisagé par tirage au sort dans un intervalle donné correspond à une estimation de la probabilité *a posteriori* (q_1) de liaison dans cet intervalle. Cette probabilité est difficilement interprétable, car elle est très faible si l'on s'intéresse à un petit intervalle, et cela même s'il existe un fort QTL dans la région. L'interprétation de l'importance d'un signal de liaison, à un locus donné, est facilitée par le calcul du rapport $\frac{q_1}{q_0}$ [137] avec q_0 la probabilité *a priori* de liaison. On peut également s'intéresser à la valeur du Bayes Factor (BF) [138], correspondant à l'odds ratio :

$$\boxed{BF = \frac{\frac{q_1}{1-q_1}}{\frac{q_0}{1-q_0}}}$$

Avec $q_1 = P(\ Liaison \mid Observations\ )$

Et $q_0 = P_{prior}\ (\ Liaison\ )$

Le pourcentage q_1 s'obtient tout simplement par le décompte des itérations dont un QTL a été placé dans l'intervalle en question, rapporté au nombre total d'itérations réalisées. Le pourcentage q_0 se déduit de la loi de probabilité définie *a priori*, soit une loi uniforme sur l'ensemble du génome. Soient L la longueur totale du génome, ν la longueur de l'intervalle étudié, et κ le nombre moyen de QTL envisagé par itération. On montre que [135]

$$\boxed{q_0 = 1 - \exp\left(\frac{-\kappa\nu}{L}\right)}$$

Explication de ce résultat :

Soit n le nombre de QTL envisagé au cours d'une itération donnée (noté k lors de la présentation du modèle phénotypique, renommé ici pour bien le différencier de κ, le nombre moyen de QTL envisagé au cours des itérations).

Si $n = 1$, la probabilité que le QTL soit dans l'intervalle est $\frac{\nu}{L}$, et la probabilité que le QTL soit hors de l'intervalle est $1 - \frac{\nu}{L}$.

Si $n > 1$, la probabilité que les QTL soient tous hors de l'intervalle est $\left(1 - \frac{\nu}{L}\right)^n$.

La probabilité qu'au moins un QTL soit dans l'intervalle est donc :

$$q_0(n) = P_{prior}\ (\ Liaison \mid n\ )$$

$$\boxed{q_0(n) = 1 - \left(1 - \frac{\nu}{L}\right)^n}$$

Pour connaître la probabilité *a priori* de liaison dans l'intervalle considéré, il faut maintenant réaliser la somme des $q_0(n)$ obtenus pour toutes les valeurs possibles de n (de 0 à l'infini), pondérée par la probabilité d'observer chacune de ces valeurs :

$$q_0 = P_{prior}\ (\ Liaison\ ) = \sum_{n=0}^{\infty} P_{prior}\ (\ Liaison \mid n\ ) * P\ (n\ )$$

$$q_0 = \sum_{n=0}^{\infty} q_0(n) \,.\, P(n)$$

Or n suit une loi de Poisson de moyenne κ, dont la densité de probabilité s'écrit :

$$\Pr(n) = \exp(-\kappa)\frac{\kappa^n}{n!}$$

Nous avons donc :

$$q_0 = \sum_{n=0}^{\infty}\left[1-\left(1-\frac{\nu}{L}\right)^n\right].\exp(-\kappa)\frac{\kappa^n}{n!}$$

$$q_0 = \exp(-\kappa)\sum_{n=0}^{\infty}\left[1-\left(1-\frac{\nu}{L}\right)^n\right].\frac{\kappa^n}{n!}$$

$$q_0 = \exp(-\kappa)\sum_{n=0}^{\infty}\left[\frac{\kappa^n}{n!}-\frac{\left(1-\frac{\nu}{L}\right)^n\kappa^n}{n!}\right]$$

On rappelle pour la suite des calculs que $\exp(x) = \sum_{n=0}^{\infty}\frac{x^n}{n!}$

$$q_0 = \exp(-\kappa)\left[\exp(\kappa)-\exp\left(\kappa\left(1-\frac{\nu}{L}\right)\right)\right]$$

$$\boxed{q_0 = 1-\exp\left(-\kappa\frac{\nu}{L}\right)}$$

Dans les analyses de liaison que j'ai menées, j'ai considéré des intervalles $\nu = 2$ cM, pour une longueur de génome L=3600 cM.

Le BF ainsi calculé ne peut être interprété de manière univoque, et on ne peut pas établir de corrélation avec un Lod-Score classique. En particulier, le Lod-Score augmente avec le nombre de transmissions observées, tandis que le BF est limité par $1/q_0$. Nous nous sommes cependant aidés pour nos interprétations de quelques repères, suggérés par [138][135].

BF>100 : forte évidence d'un signal de liaison.

BF>20 : évocation d'un signal de liaison.

Estimation d'une valeur de p empirique

Le principe général consiste à simuler un grand nombre de jeux de données, conditionnellement aux données originales (comme, par exemple, la structure familiale), sous l'hypothèse nulle H0 qu'il n'existe aucune liaison génétique. On réalise, pour chaque jeu de données, une analyse de liaison et ségrégation conjointes par MCMC du chromosome sur lequel un signal de liaison a été obtenu. On considère le BF maximum obtenu pour chacune de ces analyses. Ce BF_{max} ne pouvant être que le fruit du hasard, le pourcentage de simulations conduisant à un BF_{max} supérieur au BF observé à partir des données réelles correspond bien à une valeur de p empirique. La principale difficulté réside dans la manière de simuler sous H0.

Il existe schématiquement deux manières de simuler un échantillon familial, sous l'hypothèse nulle qu'il n'y a pas de liaison entre les marqueurs génotypiques et les phénotypes. Soit on conserve les marqueurs, et on simule de nouveaux traits quantitatifs par tirage au sort suivant une loi de distribution normale ou multimodale, soit on conserve les TQs, et on simule un nouveau jeu de marqueurs génétiques, sur la base de leurs fréquences alléliques, de la carte génétique, et bien sûr des lois de transmission mendéliennes. Les méthodes qui conservent les marqueurs, contrairement celles qui les simulent, permettent d'intégrer (ou de « capturer ») dans le calcul des BF_{max} certains biais éventuels. Ceux-ci peuvent être dus à des erreurs de génotypages, de carte génétique, ainsi qu'à la présence de déséquilibres de liaison entre les marqueurs, et à la présence de polymorphismes autres que les microsatellites (CNV et d'inversion)[139]. Ces erreurs ayant un impact similaire sur les BF observés et simulés, la valeur de p empirique tient compte de leur influence sur le calcul du BF. En revanche, et contrairement aux simulations qui conservent les TQs, ces méthodes s'exposent à des erreurs de type 1 dues à une distribution particulière du TQ. De plus, les TQs simulés ne reproduisent pas la non-indépendance des TQs réels. Plus précisément, ils ne reflètent ni la présence d'un environnement partagé, ni celle de facteurs génétiques.

La méthode que j'ai utilisée pour mes analyses de liaison est celle qui conserve les marqueurs, et simule des traits quantitatifs tout en conservant leur distribution réelle, y compris la non-indépendance des observations. Ainsi, l'hypothèse nulle testée est non pas l'absence de QTL, mais la présence d'un ou plusieurs QTL non lié(s) aux marqueurs génétiques étudiés. Pour cela, j'ai simulé les traits quantitatifs sous l'hypothèse nulle de telle sorte qu'ils suivent le modèle d'hérédité complexe (*i.e* incluant

une part environnementale et oligogénique) estimé par l'analyse de ségrégation des données réelles par la méthode MCMC[140]. Cette analyse de ségrégation nous donne les lois de probabilité *a posteriori* de différents paramètres d'intérêt utiles à la simulation des TQs : la variance attribuable aux facteurs génétiques, le nombre de QTL, l'effet de chacun, l'effet des covariables et la variance résiduelle du trait.

Pour constituer 3 000 échantillons, on tire au sort 3 000 itérations issues de l'analyse de ségrégations par MCMC, parmi celles survenant après la période de burn-in (donc issues de loi de probabilité *a posteriori*). Les valeurs des paramètres d'intérêt considérées à l'itération *i* sont transmises au programme GENEDROP du package MORGAN http://www.stat.washington.edu/thompson/Genepi/MORGAN/ morgan303-tut-html/. Ce programme constitue l'échantillon *i* en attribuant à chaque membre des cinq pedigrees une valeur de phénotype tel que leur distribution soit conforme aux paramètres de ségrégation de l'itération *i*. Nous obtenons ainsi 3 000 échantillons dont les phénotypes simulés sont bien issus de la loi de probabilité *a posteriori* du modèle d'hérédité du phénotypes réel, tout en étant indépendants des marqueurs. Nous sommes donc bien sous l'hypothèse nulle d'une absence de liaison aux marqueurs. Finalement, on réalise une analyse conjointe de liaison et de ségrégation avec les marqueurs situés sur les chromosomes ayant montré des signaux de liaison, et on retient le BF_{max} obtenu pour chacun des 3 000 échantillons.

VI.1.5. Conclusion sur les analyses conjointes de liaison et de ségrégation par MCMC

Cette méthode permet l'analyse d'un très grand nombre de marqueurs génotypés dans de larges pedigrees de structures complexes, ce qui constitue un plan d'étude particulièrement puissant dans le cadre d'analyse de liaison [132]. Encore en cours de développement, elle rencontre cependant peu de popularité. Cela s'explique en grande partie par le manque de convivialité du programme. En particulier, les résultats issus de Loki ne sont pas interprétables tels quels. Ils nécessitent une exploitation secondaire par l'utilisateur. De plus, seules les analyses de traits quantitatifs ont été bien développées. Des travaux sont en cours pour les étendre à d'autres types de phénotypes, notamment aux mesures censurées. Enfin, les analyses de liaison pangénomiques ont été quasiment abandonnées en faveur des analyses d'associations pangénomiques (GWAS) chez des

sujets non apparentés. Ce manque d'utilisation pourrait avoir comme conséquence que les effets de certaines violations des hypothèses sous-jacentes soient méconnus, bien qu'un certains nombre de travaux aient été publiés sur ce sujet (rev. dans [135]). En particulier, il est intéressant de noter que la violation de l'hypothèse de normalité des résidus n'aurait qu'un faible impact sur les résultats [135].

VI.2. Analyse d'association pangénomique (GWAS) en présence de données familiales par une méthode de décomposition de la variance

VI.2.1. Notions de décomposition de la variance phénotypique et d'héritabilité

La méthode de décomposition de la variance (« VC » en anglais) permet d'estimer la part imputable aux caractéristiques génétiques dans la variabilité d'un phénotypique quantitatif. L'idée fondamentale à la base de cette méthode est de décomposer la variance du phénotype en deux types de variances : la variance due à des caractéristiques génétiques, et la variance due à des expositions environnementales (au sens large du terme, soit tout ce qui n'est pas génétique). Comme le résument Laura Almasy et John Blangero : « Chercher à identifier les différentes sources de la variance phénotypique revient à se demander ce qui rend les individus différents les uns des autres »[141]. Ainsi, tout en restant critiques quant au caractère réducteur et à la dimension normalisatrice d'une telle question, nous pouvons la traduire en équation :

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2$$

avec σ_p^2 la variance phénotypique, σ_g^2 sa composante génétique, et σ_e^2 sa composante environnementale, et en supposant l'indépendance des effets dus aux caractères génétiques et des effets dus à l'exposition environnementale. L'héritabilité d'un phénotype est alors défini par :

$$h^2 = \frac{\sigma_g^2}{\sigma_p^2}$$

La variance génétique peut se décomposer en

(1) un terme de variance additive σ_a^2. Cette variance implique qu'un allèle à l'état homozygote a un effet double de ce même allèle à l'état hétérozygote. Elle implique également que l'effet conjoint de deux polymorphismes est égal à la somme des effets de chaque polymorphisme pris isolément.

(2) un terme de variance dominante qui permet de prendre en compte un écart à l'additivité de l'effet allélique.

(3) un terme de variance épistasique qui permet de prendre en compte une éventuelle interaction entre deux polymorphismes.

Pour des raisons de simplicité, bien souvent, seule est prise en compte la variance additive, et l'hérédité se calcule par :

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

La variance environnementale, quant à elle, peut se décomposer en deux types de variance : d'une part, la variance due à l'environnement partagé par l'ensemble de la famille, et d'autre part, celle qui est due à l'environnement non partagé. Cette dernière est une variance résiduelle. Elle inclut, comme son nom l'indique, toute variabilité due à une exposition propre à chaque individu de la famille. Elle agrège toute source de variabilité non modélisée, dont les erreurs de mesure. Lorsqu'on utilise un modèle qui ne distingue pas les variance dues à l'environnement partagé de celles dues à l'environnement non partagé, les variances environnementales, quelle que soit leur nature, les erreurs de mesure et toute variance résiduelle se retrouvent agrégées dans le même terme σ_e^2.

Quand on ne prend en compte que σ_a^2 pour caractériser la variance due aux caractères génétiques, les phénomènes de dominance et d'épistasie sont « relégués » dans les erreurs non modélisées, donc dans σ_e^2. L'héritabilité est ainsi sous-estimée en cas de fort phénomène de dominance ou d'épistasie. A l'inverse, si une variable environnementale mime une variable génétique en suivant grossièrement les lois de l'hérédité mendelienne, alors la variabilité qu'elle induit se trouve agrégée dans σ_a^2. L'héritabilité estimée est alors surestimée.

VI.2.2. Modélisation de la variable phénotypique

La variable phénotypique (notée p) peut être modélisée par une fonction linéaire s'écrivant :

$$p = \mu + g + e$$

μ étant la moyenne de la variable phénotypique, g l'effet des caractères génétiques supposés additifs (mesuré comme l'écart phénotypique par rapport à la moyenne induit par les caractères génétiques), et e un terme d'erreur dite « environnementale ». On suppose que g et e suivent une loi Normale de moyenne nulle et de variance σ^2_g et σ^2_e. On a donc bien $\sigma^2_p = \sigma^2_g + \sigma^2_e$. Ce modèle peut être affiné par l'intégration d'une covariable x, dont l'effet sur le phénotype est β lorsque x augmente d'une unité. On a alors :

$$p = \mu + \beta x + g + e \qquad (1)$$

On obtient toujours $\sigma^2_p = \sigma^2_g + \sigma^2_e$ pour une valeur donnée de x. Ce qui peut s'écrire $Var(p - \mu_x) = \sigma^2_g + \sigma^2_e$

avec μ_x, la valeur de p attendue sur la droite de régression $\mu_x = \mu + \beta x$.

Notons dès à présent que pour les analyses d'association génome-entier que j'ai réalisées dans les famille FVL avec ~500 000 SNPs, un modèle linéaire du type (1) a été appliqué pour chaque SNP codé par une variable x prenant les valeurs 0, 1 ou 2 selon le nombre d'allèles rares portés au niveau du SNP étudié. Il peut paraître paradoxal à première vue que l'effet d'un polymorphisme génétique soit pris en compte indépendamment des effets génétiques g. Il faut cependant comprendre que le terme g inclut globalement tous les effets génétiques résiduels une fois pris en compte l'effet du SNP étudié.

VI.2.3. Prise en compte des corrélations intra-familiales par une matrice de variance-covariance

On sait calculer la vraisemblance des paramètres du modèle (1), en supposant que le phénotype, conditionnellement aux variables explicatives, est distribué suivant une loi multinormale. Ce calcul fait intervenir une matrice de variance-covariance du phénotype, conditionnellement aux variables explicatives. Cette matrice dépend du degré de parenté entre les individus de l'échantillon, ainsi que des variances génétiques et environnementales.

Soit Ω la matrice de variance-covariance observée des phénotypes. Elle est de taille N*N, où N est le nombre de sujets de l'échantillon. Les covariances entre les individus peuvent à leur tour être décomposées en différentes sources, génétique et environnementale :

$$\Omega = 2\Theta\sigma_g^2 + I\sigma_e^2 \qquad (2)$$

Θ est la matrice constituée des coefficients de parentés entre les N sujets. Le coefficient de parenté vaut ½ entre un parent et son enfant ou entre deux frères/sœurs, ¼ entre un grand-parent et son petit-enfant, par exemple… Il se calcule aisément même pour des structures familiales complexes. Ce coefficient correspond au pourcentage de matériel génétique partagé en moyenne entre deux personnes apparentées. L'équation (2) traduit donc ce que l'on conçoit intuitivement, à savoir que les phénotypes d'apparentés au premier degré seront deux fois plus corrélés que ceux d'apparentés au deuxième degré, mais uniquement pour ce qui concerne les corrélations dues aux effets génétiques.

La matrice I est relative aux corrélations dues à l'environnement. Nous avons considéré qu'il s'agissait d'une matrice identité, qui vaut 1 sur sa diagonale et 0 ailleurs. Ceci implique que l'exposition environnementale est propre à chaque individu et non pas corrélée entre les membres d'une même famille. Il s'agit d'une hypothèse extrêmement forte, et qui reflète de manière très approximative la complexité de la réalité. Nous avons vu plus haut que de telles approximations ont pour effet de surestimer l'héritabilité, paramètre auquel nous ne nous intéresserons pas ici. On pourrait affiner la matrice structurelle relative à l'environnement en considérant, par exemple, les corrélations égales à 1 entre chaque personne vivant dans le même habitat.

VI.2.4. Estimation de l'effet d'un SNP sur le phénotype par maximisation de la vraisemblance

En supposant que les phénotypes, conditionnellement aux covariables, suivent une loi multinormale, un calcul du maximum de vraisemblance de β (et des autres paramètres du modèle) est possible. La vraisemblance $L_{phénotypes}(\mu, \sigma^2_g, \sigma^2_e, \beta)$, s'écrit en fonction de ces paramètres, de la matrice structurelle des coefficient de parenté Θ, et de la matrice structurelle des covariances dues au partage de l'environnement (le calcul de cette vraisemblance est explicité en **annexe pA13**).

Nous ne présenterons dans la partie consacrée aux résultats ni les estimations de σ_g^2, de σ_e^2, ni celles de l'héritabilité. Nous exposerons, en revanche, les résultats des tests d'association d'une analyse pangénomique de 490 083 SNPs, inclus individuellement en tant que covariable x dans autant de modèles. Ainsi, nous présenterons les paramètres β, et la valeur de p d'un test de Wald, testant l'absence

d'association entre chaque SNP et les taux de FVIII et de vWF. Chaque SNP ayant été codé 0, 1, ou 2, suivant le nombre d'allèles rares, β correspond à l'effet sur le phénotype de l'allèle rare en mode additif. L'effet d'un SNP sur le phénotype ne fait pas à proprement parler partie de la décomposition de la variance, mais il est plus exactement estimé conjointement à l'analyse par VC. Il s'agit d'un effet « fixe », qui ne serait pas différent de l'effet estimé par une régression classique. L'intérêt de l'estimer conjointement à une analyse de décomposition de la variance est de prendre en compte la non-indépendance des observations (capturée dans l'effet « aléatoire » du modèle par la composante génétique de la variance), et d'obtenir une variance de β, et de ce fait une valeur de p, non biaisées.

L'estimation de la vraisemblance repose sur l'hypothèse de normalité de la distribution des phénotypes. Le skewness et le kurtosis de la distribution décrivent deux manières différentes pour une distribution de s'écarter de la normalité. Le skewness est relatif à l'asymétrie de la distribution, c'est-à-dire à la situation où il y aurait plus de valeurs au dessus qu'en dessous de la moyenne, ou inversement. Le kurtosis est relatif à un excès ou un défaut de valeurs dans les deux queues de la distribution. Des études de simulations ont mis en évidence que l'écart à la normalité pose un problème en cas de « leftkurtosis », c'est-à-dire lorsqu'il y a un excès de valeurs dans les queues de la distribution. Cette distribution peut conduire à une surestimation des effets modélisés [142]. Afin de prévenir cet écueil, j'ai utilisé la z-transformation des deux phénotypes FVIII et vWF. Celle-ci attribue à chaque valeur du FVIII ou vWF le quantile de la loi normale correspondant à son rang.

Nous avons considéré comme significatives les associations dont la significativité p était inférieure au seuil de Bonferroni ($p<1.02\ 10^{-7}$) afin de conserver un risque α inférieur à 5%. Les associations évocatrices qui se détachaient du bruit de fond ($p<10^{-5}$) sont présentées succinctement en annexe. Le taux attendu de faux positif a été calculé au moyen du False Discovery Rate (voir annexe **pA7**).

J'ai utilisé pour ces analyses le logiciel SOLAR [143] dans lequel est implémentée la méthode de décomposition de la variance.

VI.3. Analyses d'association en présence de données familiales par la méthode des Equations d'Estimation (EE)

VI.3.1. Principe général des EE

Initialement développées pour l'analyse de données répétées longitudinales, les Equations d'Estimation constituent plus généralement une méthode pour l'analyse de données corrélées [144][145]. Ainsi, elles trouvent un intérêt considérable dans l'analyse de données familiales, notamment en raison de leur souplesse d'application et de leur robustesse par rapport aux écarts éventuels d'hypothèse [146]. Chaque famille constitue en effet un ensemble (« cluster ») indépendant des autres, au sein duquel les données sont corrélées, en raison du partage d'un même environnement et d'un même patrimoine génétique. La prise en compte de l'existence des corrélations intra-familiales est nécessaire pour estimer correctement les paramètres de régression.

J'ai utilisé les EE1 (« Estimation des moments d'ordre 1 ») pour étudier l'association entre des polymorphismes candidats et les taux plasmatiques de FVIII et vWF dans la cohorte STANISLAS. Les EE1 ont la propriété d'être robustes par rapport à une spécification incomplète ou imprécise des corrélations intra-familiales Les estimateurs obtenus des paramètres d'association sont asymptotiquement sans biais même si les corrélations intra-familiales sont mal spécifiées. Il n'est donc pas nécessaire d'utiliser la vraie matrice de corrélations, généralement inconnue, pour faire des inférences valides sur les paramètres de régression. Plusieurs types de matrices simplifiées ont été proposés. J'ai choisi une matrice diagonale (dite « d'indépendance »), dans laquelle seules les variances sont estimées, les covariances entre individus étant fixées à 0. Bien que les estimateurs de paramètres de régression obtenus par EE1 soient asymptotiquement équivalents à ceux obtenus par un modèle de régression linéaire généralisé (« Generalized Linear Model »), les deux méthodes ne sont pas équivalentes. En effet, les GLM considèrent chaque individu comme autant d'entités indépendantes. Au contraire, les EE1 agrègent les informations de tous les individus d'un même cluster, et considèrent que seuls les clusters sont indépendants. La prise en compte de la structure en cluster ne modifie que d'une façon modeste l'estimation β des coefficients de la régression. Elle conduit par contre à une estimation majorée de leur variance $var(\beta)$ qui minimise le risque d'erreur de type 1 [145] [147].

VI.3.2. Test de l'effet d'un SNP sur le phénotype

Le test d'hypothèse d'absence d'association entre le phénotype d'intérêt et une variable est obtenu par le test de Wald, qui compare à 0 le rapport $\beta^2/var(\beta)$. Lorsque l'on s'intéresse à l'effet conjoint de plusieurs variables, ou bien d'une seule variable en plusieurs classes (comme le test génotypique, dont la variable est en 3 classes : génotype A_1A_1, A_1A_2 ou A_2A_2), on fait appel au test de Wald généralisé. En effet, les EE ne font aucune hypothèse sur la distribution des observations. Elles ne permettent donc pas d'estimer la vraisemblance d'un modèle, ni par conséquent de réaliser les tests de modèles emboîtés, auxquels les utilisateurs de « GLM » sont habitués. Le test de Wald généralisé fait intervenir le vecteur des coefficients de la régression des variables explicatives et la matrice variance-covariance de ces coefficients [148].

Plusieurs modèles génétiques ont été appliqués pour tester les effets sur les taux de FVIII et de vWF des cinq SNPs qui ont été génotypés dans l'étude STANISLAS. Les modèles dominant (contraignant les génotypes présentant au moins un allèle rare à avoir le même effet), récessif (contraignant les génotypes présentant au moins un allèle fréquent à avoir le même effet) ou additif (contraignant le génotype présentant les deux allèles rares à avoir un effet double de celui n'en présentant qu'un) ont été testés grâce à un codage spécifique du génotype (**tableau 8**).

Tableau 8 : Codage du génotype d'un SNP en fonction du modèle de transmission

	Modèle				
	Additif	**Dominant**	**Récessif**	**Génotypique**	
Génotype	$\mathbf{X_{add}}$	$\mathbf{X_{dom}}$	$\mathbf{X_{rec}}$	$\mathbf{X_{gen1}}$	$\mathbf{X_{gen2}}$
$\mathbf{A_1A_1}$	0	0	0	0	0
$\mathbf{A_1A_2}$	1	1	0	1	0
$\mathbf{A_2A_2}$	2	1	1	0	1

$\mathbf{A_1}$**et** $\mathbf{A_2}$ sont les deux allèles du SNPs et **X** est la variable incluse dans le modèle.

Nous avons estimé l'importance de l'effet d'un SNP au moyen de la valeur R^2, qui est le pourcentage de la variance phénotypique expliqué par le SNP, dans un modèle génotypique. Soient Var_{e1} la variance résiduelle d'un modèle incluant des covariables et un SNP, et Var_{e0} la variance résiduelle d'un modèle n'incluant que les covariables :

$$R^2 = 1 - \frac{Var_{e1}}{Var_{e0}}$$

La méthode des EE1 est implémentée dans la fonction *geeglm()* du module « *GEEPACK* » disponible sur le site CRAN du logiciel « R » (http://www.R-project.org/). Toutes les analyses étaient systématiquement ajustées sur l'âge et le sexe.

VI.4. Méthodes d'analyse d'association utilisées dans les échantillons d'individus non apparentés

VI.4.1. Trait quantitatif

Une régression linéaire, systématiquement ajustée sur l'âge et le sexe, a permis de modéliser la relation entre les SNPs étudiés et les taux plasmatiques de vWF et de FVIII.

Deux SNPs ont été testés dans MARTHA05. Des analyses stratifiées sur le groupe ABO ont été réalisées. Des analyses d'interaction ont également été conduites pour vérifier que l'effet d'un SNP était homogène selon le niveau d'une autre variable d'intérêt. Pour cela, un terme d'interaction correspondant au produit des variables impliquées dans l'interaction a été introduit dans le modèle de régression. J'ai utilisé pour cela la fonction *glm()* du logiciel R.

Les analyses pangénomiques des échantillons MARTHA08 et MARTHA10 ont été réalisées à l'aide du logiciel PLINK [123] en considérant un modèle de transmission additif pour chaque SNP. Elles étaient ajustées sur les génotypes du groupe ABO (voir **tableau 6 p39**) Des analyses haplotypiques ont parfois été réalisées à partir du logiciel "THESIAS" [149]. Les analyses haplotypiques sont des analyses multivariées permettant de prendre en compte la phase des deux SNPs (**voir annexe pA4**)

VI.4.2. Trait qualitatif (GWAS *in silico*, MARTHA05 et FARIVE)

Les comparaisons des fréquences alléliques entre les cas et les témoins de la GWAS *in silico* ont été réalisées par le test de tendance de Cochran-Armittage. Soit un SNP présentant deux allèles A_1 et A_2 dont les génotypes se répartissent selon les notations suivantes :

	A_1A_1	A_1A_2	A_2A_2	Total
Témoin	N_{T0}	N_{T1}	N_{T2}	L_0
Cas	N_{C0}	N_{C1}	N_{C2}	L_1
Total	C_0	C_1	C_2	N

La statistique du test de tendance T et sa variance s'écrivent :

$$T = \sum_{i=0}^{2} w_i \left(N_{Ti} L_1 - N_{Ci} L_0 \right)$$

$$V(T) = \frac{L_0 L_1}{N} \left(\sum_{i=0}^{2} w_i^2 C_i \left(N - C_i \right) - 2 \sum_{i=0}^{1} \sum_{j=i+1}^{2} w_i w_j C_i C_j \right)$$

Les valeurs w_i sont des poids qui dans le cas du test d'une relation linéaire prennent les valeurs ($w_0=0$, $w_1=1$, $w_2=2$).

Sous l'hypothèse nulle d'absence d'association entre la maladie et le SNP, $T \Big/ \sqrt{\text{var}(T)}$ suit une loi normale centrée réduite.

Une modélisation plus fine a été réalisée pour les deux SNPs testés dans MARTHA05 et FARIVE au moyen d'une régression logistique avec ajustement sur l'âge et le sexe. Des analyses stratifiées sur la présence de mutation FVL ou du FII, ou selon le groupe ABO, ont été conduites. L'homogénéité des effets observés entre les études MARTHA05 et FARIVE, ainsi qu'entre les différentes strates, a été testée au moyen du test de Cochran-Mantel-Haenszel. Si k est le nombre de strates, β_i est le coefficient de la régression logistique de l'étude i, et ω_i l'inverse de la variance de β_i, la statistique Q du test d'homogénéité de Cochran-Mantel-Haenszel s'écrit :

$$Q = \sum \omega_i \beta_i^2 - \frac{\left(\sum \omega_i \beta_i \right)^2}{\sum \omega_i}$$

et suit une loi de Chi-2 à $k-1$ degrés de liberté.

Les analyses des échantillons cas-témoins ont été réalisées avec le logiciel R, notamment avec la fonction *glm(family=binomial)* pour les régressions logistiques.

VI.5. Méta-analyse des GWAS réalisées dans les Familles-FVL, MARTHA08 et MARTHA10

VI.5.1. Présentation générale

Le moyen d'augmenter la puissance d'une analyse qui vient en premier à l'esprit est d'augmenter la taille de l'échantillon. Il est également possible, grâce à une méta-analyse, de combiner les résultats obtenus dans plusieurs échantillons. Cette pratique est utile lorsque l'on n'a pas accès à la base de données de chaque étude, ou encore lorsque l'on ne peut pas les regrouper en une seule base de données. Afin d'augmenter la

puissance de découverte des GWAS des FVIII et vWF, j'ai réalisé une méta-analyse des GWAS réalisées dans les échantillons de Famille-FVL, MARTHA08 et MARTHA10. En effet, le recrutement très particulier des Familles-FVL impose une méthode d'analyse spécifique qui exclue le regroupement de ces trois échantillons en un seul.

Il y avait 442 728 SNPs communs aux trois échantillons après application de filtres de qualité (voir **tableau 5 p38**). Chacun de ces SNPs a fait l'objet, un par un, d'une méta-analyse utilisant les données des trois échantillons. J'ai utilisé une méthode de calcul fondée sur les estimations des paramètres $\hat{\beta}$ et de leurs variances [150]. Rappelons que, pour les trois échantillons, ces paramètres sont issus d'un modèle linéaire, dans lequel les génotypes du SNP sont codés selon le mode additif. Ils ont été estimés au moyen d'une régression linéaire dans MARTHA08 et MARTHA10, et dans le cadre d'une analyse par décomposition de la variance pour prendre en compte les corrélations familiales dans les Familles-FVL.

VI.5.2. Effet fixes, effets aléatoires, mesure Q de l'hétérogénéité

Soit un SNP dont nous connaissons, dans chaque étude i ($i = 1,...,k$), l'estimation $\hat{\beta}_i$ de son effet sur le phénotype d'intérêt, ainsi que la précision ω_i de cette estimation, ω_i étant l'inverse de la variance de $\hat{\beta}_i$. Nous pouvons obtenir une estimation $\hat{\mu}_F$ de l'effet du SNP en réalisant une moyenne des $\hat{\beta}_i$, pondérée par leur précision respective ω_i :

$$\hat{\mu}_F = \frac{\sum(\omega_i \hat{\beta}_i)}{\sum \omega_i}$$

La variance de $\hat{\mu}_F$, notée v_F, est alors égale à :

$$v_F = \frac{1}{\sum \omega_i}$$

Il est recommandé, avant d'estimer l'effet combiné $\hat{\mu}_F$, de tester l'homogénéité des effets $\hat{\beta}_i$ des k études au moyen de la statistique Q (dont le test est connu sous le nom du Chi-2 de Cochran [151]). Cette statistique dépend des écarts des $\hat{\beta}_i$ à $\hat{\mu}_F$:

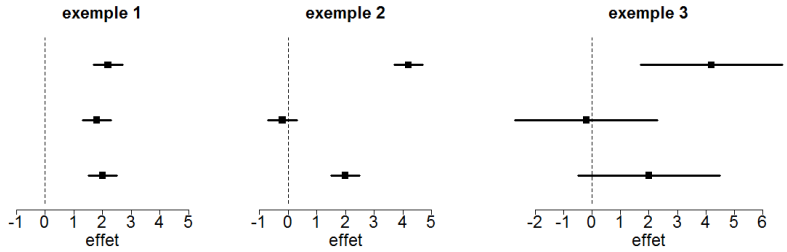
$$Q = \sum \omega_i (\hat{\beta}_i - \hat{\mu}_F)^2 \qquad (1)$$

Q suit une loi de χ^2 à $k-1$ degrés de liberté. Son espérance est égale à $k-1$ sous l'hypothèse nulle d'homogénéité entre les études. Ce test est connu pour être peu

puissant lorsque l'on analyse un petit nombre d'études, et à l'inverse, pour un grand nombre d'études, « trop » puissant (il est significatif en cas d'hétérogénéité, certes réelle, mais de peu d'intérêt clinique du fait de sa faible ampleur)[152].

$\hat{\mu}_F$ est appelé effet fixe car son calcul suppose que les k études donnent une estimation d'un même effet. La variabilité observée entre les études n'est, par conséquent, que le fruit de fluctuations d'échantillonnage. Par exemple, les calculs de $\hat{\mu}_F$ et de sa variance v_F donneraient exactement les mêmes résultats dans les deux premiers exemples de la **figure 6**. Le test d'égalité à 0 de l'effet combiné serait donc le même dans les exemples 1 et 2. Cependant, de manière intuitive, on aimerait « pénaliser » le test de l'effet combiné dans l'exemple 2, en augmentant la variance de cet effet. Il est donc intéressant de faire intervenir une estimation de la variabilité inter-étude, notée $\hat{\tau}^2$, dans le calcul de la précision de l'effet du SNP [150].

Figure 6 : Illustration de l'hétérogénéité de trois études fictives



Sur cette figure est représentée l'estimation d'un paramètre de régression (et son intervalle de confiance à 95%) dans trois situations différentes: exemple 1 : faibles variabilités intra et inter-études **; exemple 2 :** faible variabilité intra-étude et forte variabilité inter-études **; exemple 3 :** fortes variabilités intra et inter-études

Le calcul de $\hat{\tau}^2$ s'obtient en exprimant l'espérance de Q dans le cas général où il existe une variabilité inter-études :

$$E(Q) = \tau^2 \left(\sum \omega_i - \frac{\sum \omega_i^2}{\sum \omega_i} \right) + k - 1$$

puis en égalant l'espérance de Q à sa valeur observée, calculée d'après (1) :

$$\hat{\tau}^2 = \frac{Q - (k-1)}{\sum \omega_i - \frac{\sum \omega_i^2}{\sum \omega_i}} \text{ , si } Q \geq k-1 \qquad (2)$$

$$\text{et } \hat{\tau}^2 = 0, \text{ si } Q < k-1$$

On définit ainsi $\omega^*_i = (\omega_i^{-1} + \hat{\tau}^2)^{-1}$ qui nous permet d'estimer l'effet aléatoire (« Random ») $\hat{\mu}_R$:

$$\hat{\mu}_R = \frac{\sum \omega^*_i \hat{\beta}_i}{\sum \omega^*_i}$$

et sa variance v_R :

$$v_R = \frac{1}{\sum \omega^*_i}$$

VI.5.3. Autres mesures de l'hétérogénéité

Higgins et Thompson ont développé plusieurs alternatives à la statistique Q [153], permettant de s'affranchir de la dépendance au nombre k d'études. Nous les introduisons dans le cas simple où les variances des effets estimés $\hat{\beta}_i$ sont toutes égales à $1/\omega_i = \sigma^2$. Avec cette simplification, l'équation (2) devient

$$\hat{\tau}^2 = \sigma^2 \left(\frac{Q}{k-1} - 1 \right) \tag{3}$$

et l'on a : $\hat{\mu}_R = \hat{\mu}_F$.

$$v_F = \frac{\sigma^2}{k} \tag{4}$$

$$v_R = \frac{\sigma^2 + \tau^2}{k} \tag{5}$$

Les mesures d'hétérogénéité envisagées par Higgins et Thompson sont toutes, à l'instar de *Q*, des fonctions monotones croissantes de $\rho = \tau^2 / \sigma^2$. Contrairement à Q, elles ne dépendent pas du nombre d'études *k*. De plus, leur dépendance vis-à-vis de σ^2 ne doit pas être explicite, afin de pouvoir les calculer dans le cas plus général où les précisions $1/\omega_i$ ne seraient pas égales entre les études. La présence implicite de σ^2 est cependant désirée par les auteurs. Ainsi, ce n'est pas tant la variabilité inter-étude qui importe. A variabilité inter-étude égales, l'hétérogénité des études sera d'autant plus grande que la variabilité intra-étude sera faible (voir les exemples 2 et 3 de la **figure 6**).

Deux fonctions monotones croissantes de ρ sont envisagées : **(i)** $\rho + 1$, **(ii)** $\frac{\rho}{1+\rho}$. Voyons de quelles manières ces fonctions permettent de ne plus avoir à expliciter σ^2 :

(i) $$\rho + 1 = \frac{\tau^2}{\sigma^2} + 1$$

L'équation (3) nous donne

$$\frac{\hat{\tau}^2}{\sigma^2} = \frac{Q}{k-1} - 1$$

Ainsi, en subtilisant la valeur τ^2 par son estimation, nous avons :

$$\rho + 1 = \frac{Q}{k-1}$$

définissant la mesure d'hétérogénéité H^2 :

$$\boxed{H^2 = \frac{Q}{k-1}}$$

Une autre estimation de $\rho + 1$ est possible :

$$\rho + 1 = \frac{\tau^2 + \sigma^2}{\sigma^2}$$

L'équation (3) nous donne une estimation du dénominateur :

$$\sigma^2 = k v_F$$

tandis que l'équation (4) nous donne une estimation du numérateur :

$$\tau^2 + \sigma^2 = k v_R$$

Ainsi, nous avons :

$$\rho + 1 = \frac{v_R}{v_F}$$

définissant la mesure d'hétérogénéité R^2

$$\boxed{R^2 = \frac{v_R}{v_F}}$$

NB : H^2 et R^2 ne sont strictement identiques que dans le cas particulier où les précisions $1/\omega_i$ sont les mêmes dans toutes les études.

(ii) $$\frac{\rho}{1+\rho} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2}$$

définissant ainsi la mesure d'hétérogénéité

$$\boxed{I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2}}$$

A première vue, la dépendance de I vis-à-vis de σ^2 n'est pas implicite. Cependant, elle le devient après avoir exprimé I^2 en fonction de H^2. On montre en effet que

$$I^2 = \frac{H^2 - 1}{H^2}$$

I^2 est le pourcentage de variance de l'effet du SNP dû à l'hétérogénéité entre les études. Higgins et Thompson [153] , qui ont étudié différentes mesures de l'hétérogénéité, considèrent que des valeurs de I^2 comprises entre 31% et 56% sont le reflet d'une hétérogénéité modérée, tandis que des valeurs de I^2 supérieures à 56% signent une hétérogénéité importante, sans pour autant en faire une règle universelle.

En pratique, j'ai réalisé la méta-analyse des trois études (MARTHA08, MARTHA10, et Familles-FVL) à l'aide du logiciel GWAMA [154], qui m'a permis d'obtenir les effets fixes et aléatoires, la valeur de p du test Q, et les valeurs I^2 pour les 442 728 SNPs en commun dans les trois études. J'ai considéré comme significatives les associations dont le FDR était inférieur à 5% (voir encadré sur les tests multiples en annexes), et comme évocatrices celles dont la valeur de p était inférieure à $p<10^{-5}$. Les tests basés sur les effets aléatoires étant particulièrement conservateurs [155][143], j'ai mené les étapes exploratoires en me fondant sur les effets fixes. J'ai vérifié secondairement que la prise en compte, par les effets aléatoires, d'une éventuelle hétérogénéité entre études ne modifiait pas les conclusions.

RESULTATS ET DISCUSSIONS

VII- Identification des gènes *BAI3* et *STAB2* comme nouveaux déterminants génétiques des taux de FVIII et vWF à partir d'une analyse de liaison génétique

VII.1 Analyses de liaison pangénomiques des taux de vWF et FVIII dans l'échantillon Familles -FVL

Pour identifier de nouveaux QTLs pouvant contribuer à la variabilité des taux de FVIII et vWF, j'ai réalisé une série d'analyses de liaison pangénomiques conjointement à des analyses de ségrégation en appliquant aux données des Familles-FVL la méthode Bayésienne décrite au **§VI.1**. J'ai étudié les taux de vWF, de FVIII et enfin de FVIII ajustés sur les taux de vWF, après avoir vérifié que leurs distributions n'étaient pas trop éloignées d'une loi normale (**figure 5 p41**). Toutes ces analyses ont été ajustées sur l'âge et le sexe. Un ajustement supplémentaire sur le gène ABO a été réalisé ultérieurement grâce au codage explicité dans le **tableau 6 p39**.

VII.1.1. Analyse du taux de plasmatique de vWF

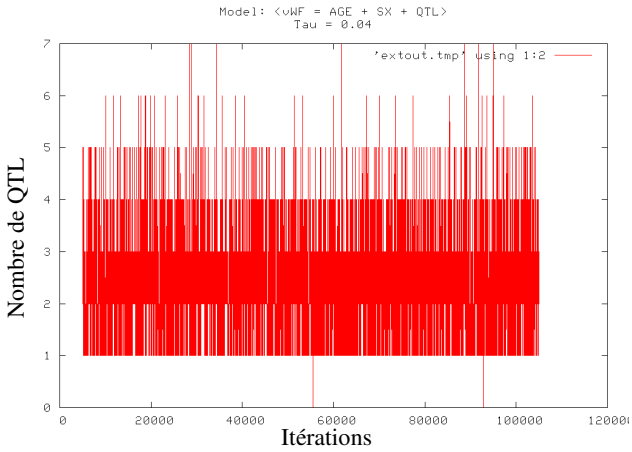
Afin de choisir les valeurs optimales des paramètres k (nombre de QTLs) et τ (variance de la distribution a priori des effets génétiques) qui seront utilisées pour définir les distributions *a priori* des analyses de liaison, j'ai d'abord réalisé plusieurs analyses de ségrégations, en faisant varier τ entre 0 et 0,25, une valeur proche de la variance de vWF observée dans l'échantillon FVL. Les autres distributions *a priori* ont été définies conformément à ce qui a été décrit au **§VI.1.3 p49**.

J'ai ainsi réalisé 14 analyses de ségrégation, chacune constituée de 110 000 itérations. J'ai ignoré les résultats des 10 000 premières itérations, ce qui autorise à l'algorithme de Metropolis-Hasting une période de burn-in confortable. Pour chacune des analyses, j'ai vérifié la qualité du mixing (exemple pour l'une d'entre elles en **figure 7, page suivante**). Après avoir calculé le nombre moyen de QTL (*k*) estimé sur l'ensemble des itérations d'une analyse de ségrégation, j'ai représenté *k* en fonction de τ pour les 14 analyses (**figure 8, page suivante**). La meilleure distribution *a priori* est celle qui maximise *k*. Ainsi, nous pouvons définir précisément les distributions *a priori* présentées au **§VI.1.3 p49** et que nous utiliserons pour les analyses de liaison suivantes. Les distributions *a priori* choisies pour les effets génétiques et pour *k* sont respectivement une loi normale de moyenne nulle et de variance $\tau = 0,04$ et une loi de Poisson de moyenne 2, tronquée à 17.

Une fois ces analyses de ségrégations préliminaires réalisées, j'ai pu procéder aux analyses de liaison. J'ai analysé chacun des 22 autosomes, au moyen de 500 000 itérations. J'ai vérifié pour ces 22 analyses la qualité du mixing au moyen de graphiques du type de ceux présentés en **figure 7.** J'ai également vérifié la convergence de l'algorithme à l'aide des graphiques de la **figure 9 page suivante**.

Figure 7 :Qualité du mixing de l'analyse de ségrégation du vWF ajusté sur l'âge et le sexe obtenue avec τ = 0.04

A. Nombre de QTL obtenu à chaque itération en fonction du numéro de l'itération



Le mixing semble très bon ici. Le nombre de QTL estimé à chaque itération varie de 0 à 7 en oscillant le plus souvent entre 2 et 3. Il balaye de manière continue et homogène l'ensemble de l'espace entre 0 et 7. Un trait horizontal tranchant avec l'aspect homogène observé s'observerait au contraire s'il restait bloqué sur une même valeur durant un grand nombre d'itérations successives. Cela serait le signe d'un mixing de mauvaise qualité

B. Variance résiduelle obtenue à chaque itération en fonction du numéro de l'itération

Là encore, on vérifie que les tirages au sort à partir des lois *a posteriori* explorent l'ensemble de celles-ci sans se focaliser sur un tirage en particulier. Si cela se produisait, la variance résiduelle resterait la même d'un tirage à l'autre, et l'on observerait un segment horizontal. On vérifie également que la variance résiduelle ne descend pas en-dessous d'une erreur de mesure raisonnable.

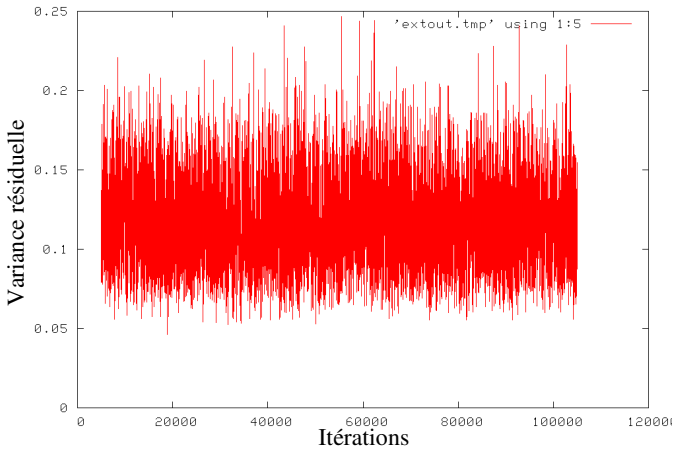
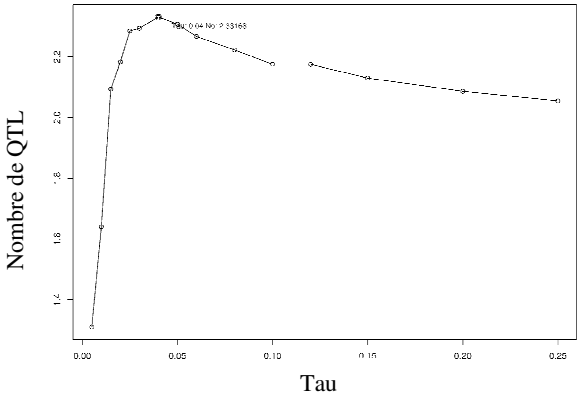
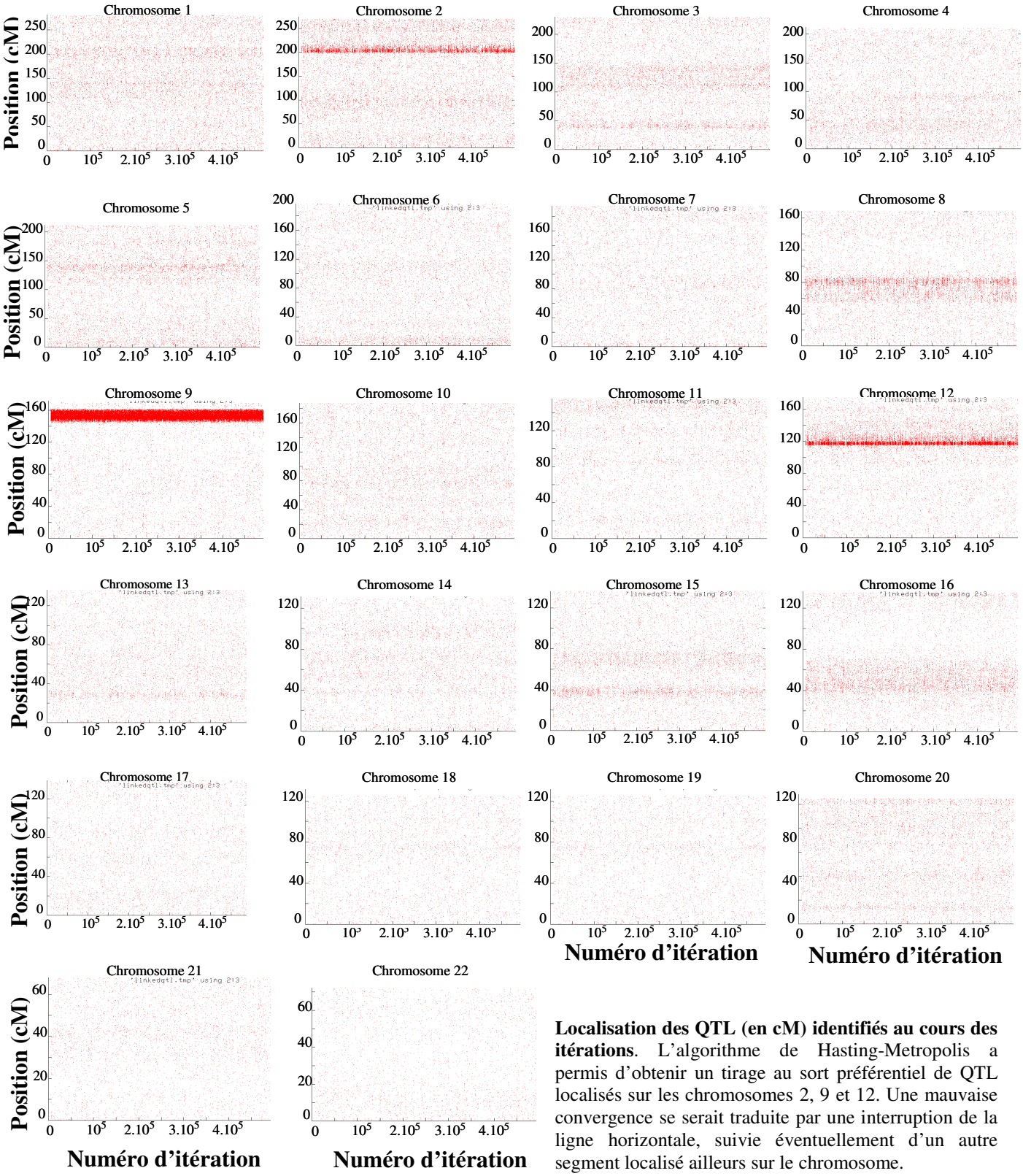


Figure 8 : Choix des paramètres optimaux *k* et τ des distributions *a priori* du nombre de QTL et de leurs effets sur les taux plasmatiques de vWF ajustés sur l'âge et le sexe



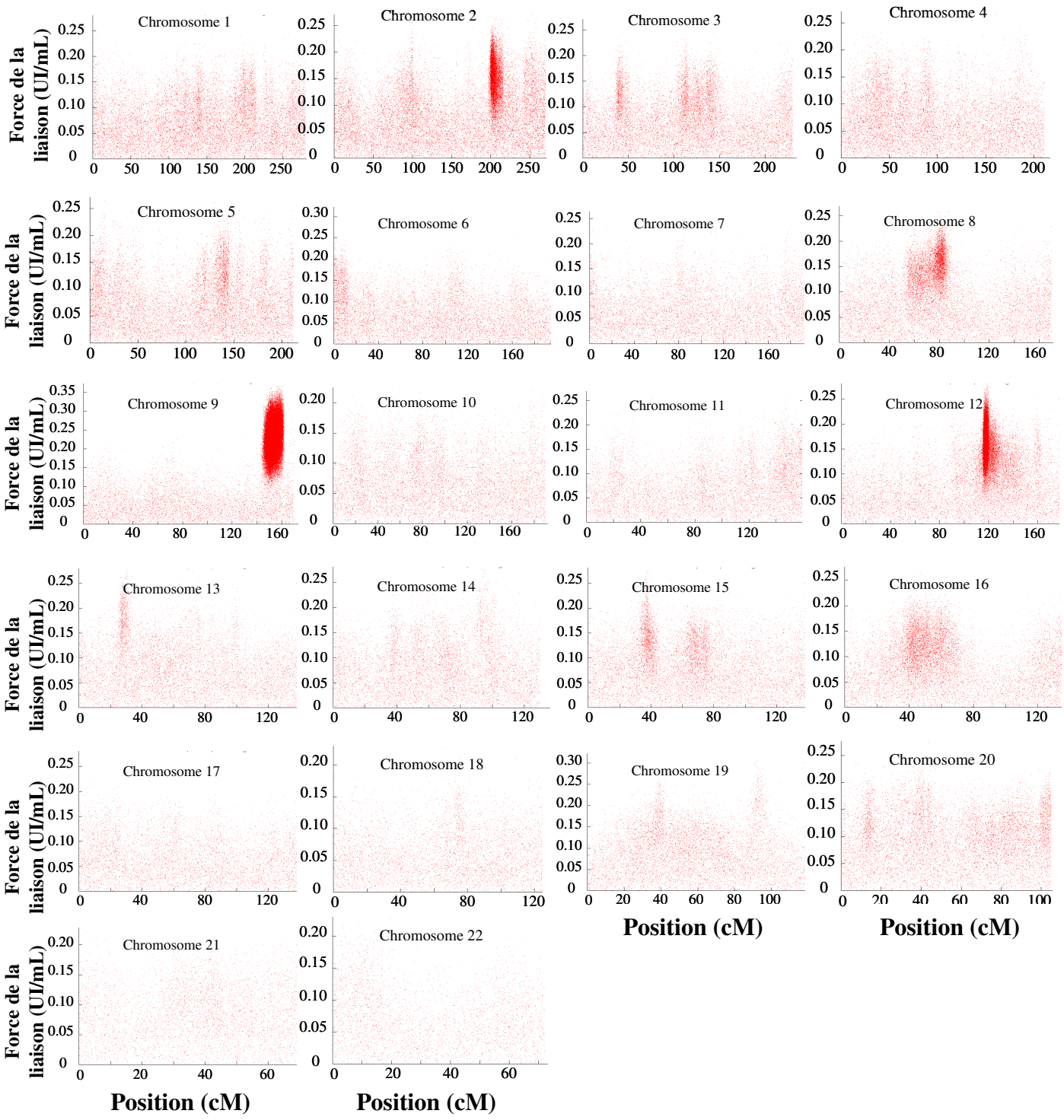
Nombre de QTL moyen obtenu pour chacune des 14 analyses de ségrégation en fonction de la variance τ de la distribution *a priori* des effets génétiques.
On retient pour les prochaines analyses de liaison k=2 et τ =0.04, correspondant respectivement au nombre entier le plus proche de la valeur maximale des moyennes de QTL et du paramètre τ -beta ayant conduit à cette estimation.

Figure 9 : Convergence de l'algorithme et localisation les signaux de liaison des taux plasmatiques de vWF ajusté sur l'âge et le sexe



Localisation des QTL (en cM) identifiés au cours des itérations. L'algorithme de Hasting-Metropolis a permis d'obtenir un tirage au sort préférentiel de QTL localisés sur les chromosomes 2, 9 et 12. Une mauvaise convergence se serait traduite par une interruption de la ligne horizontale, suivie éventuellement d'un autre segment localisé ailleurs sur le chromosome.

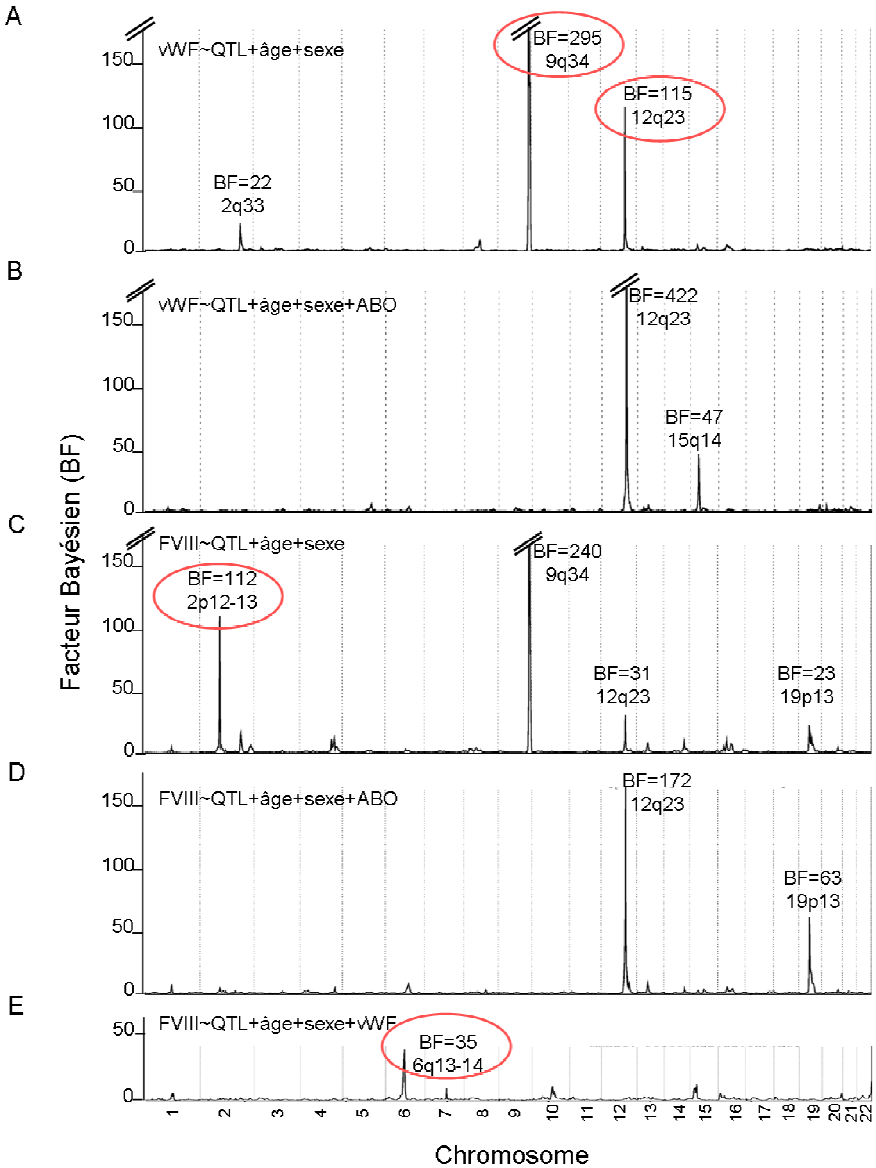
Figure 10: Signaux de liaison des taux plasmatiques de vWF ajustés sur l'âge et le sexe



Effet du QTL (mesuré par la racine carrée de la variance du taux de vWF expliquée) en fonction de sa localisation (en cM) . Chaque point correspond à une itération au cours de laquelle un QTL a été identifié par l'algorithme de Hasting-Metropolis. Un signal de liaison se caractérise par une forte concentration de points se détachant du bruit de fond. Trois signaux de liaison apparaissent sur les chromosomes 2, 9 et 12. Notons que l'échelle n'est pas la même sur toutes les figures. Ainsi, la force du signal de liaison du chromosome 9 est supérieure à celle des autres.

Trois signaux de liaison sont apparus (**figure 10 et 11.A**). L'un, situé sur le bras long chromosome 9, était beaucoup plus fort que les deux autres (BF=295), situés sur les bras longs des chromosomes 2 et 12 (BF=22 et BF=115, respectivement). Le gène du groupe ABO est situé au centre du signal de liaison du chromosome 9. J'ai réanalysé ces données avec un ajustement supplémentaire sur le groupe ABO. Cette analyse n'a pu être menée que deux années plus tard, lorsque le génotypage de plus de 500 000 SNPs m'a permis de déduire les allèles A1, A2, B et O à partir de rs8176704, rs8176746, et rs505922 (voir **tableau 6 p39**). Tandis que les signaux des chromosomes 2 et 9 disparaissaient totalement (BF<2), le signal du chromosome 12 augmentait fortement (BF=422). Un plus faible signal est apparu sur le bras chromosome 15 (BF=47) (**figure 11 B**).

Figure 11 : Résultat des analyses de liaison pangénomique dans les familles-FVL



Facteur Bayésien en fonction de la localisation chromosomique obtenu pour cinq analyses ajustées sur l'âge et le sexe : **A.** vWF **B.** vWF ajusté sur ABO **C.** FVIII **D.** FVIII ajusté sur ABO **E.** FVIII ajusté sur vWF. Les cercles rouges correspondent aux signaux ayant fait l'objet d'une recherche de gènes candidats.

VII.1.2. Analyse des taux de FVIII non ajustés, puis ajustés, sur les taux de vWF.

La séquence d'analyses suivie a été exactement la même que celle de l'analyse du taux de vWF : plusieurs analyses de ségrégation préliminaires, permettant de choisir les paramètres des distributions *a priori*, suivies des analyses de liaison des 22 autosomes, avec vérification de la qualité du mixing et de la convergence de l'algorithme pour chacune des analyses de ségrégation et de liaison. Seuls les résultats des analyses de liaison sont présentés ici. Les paramètres des distributions *a priori* était $k = 2$ *et* $\tau = 0.015$.

Analyse des taux de FVIII non ajustés sur les taux de vWF

Les analyses ont fait apparaître deux signaux de liaison très nets sur le bras court du chromosome 2 (BF=112) et le bras long du chromosome 9 (BF=240). On pouvait déceler deux signaux nettement plus faibles sur les bras longs des chromosomes 12 (BF=31), et 19 (BF=23). Le signal du chromosome 12 était situé exactement au même locus que celui observé avec le taux de vWF (**figure 11 C**). L'ajustement sur le groupe ABO réalisé plus tardivement au cours de ma thèse absorbait intégralement les signaux observés sur les chromosomes 2 (BF<5) et 9 (BF<2), tandis qu'il renforçait les signaux des chromosomes 12 (BF=172) et 19 (BF=63) de façon très sensible (**figure 11 D**).

Analyse des taux de FVIII ajustés sur les taux de vWF

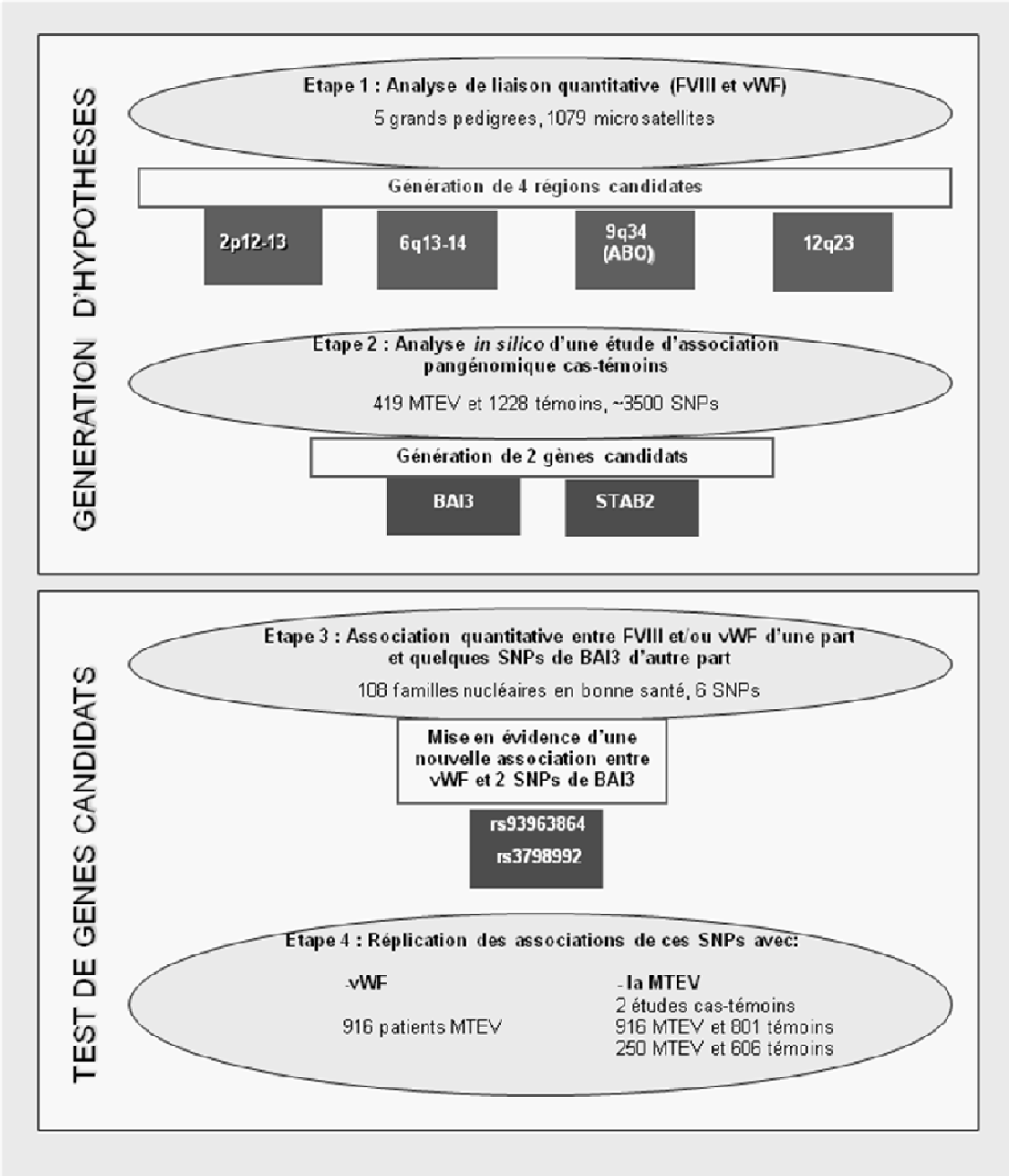
L'ajustement sur le vWF, quant à lui, faisait émerger un nouveau signal de liaison sur le bras long du chromosome 6 (BF=35), tandis que disparaissaient les autres signaux (**figure 11 E**), notamment celui au niveau du gène ABO. Cette observation nous permet d'émettre l'hypothèse qu'ABO n'aurait pas d'influence directe sur FVIII, mais seulement par l'intermédiaire de vWF. Les analyses de FVIII ajustés sur vWF et ABO menaient à des résultats sensiblement identiques en termes de BF.

VII.1.3. Stratégie pour l'étude plus fine des signaux de liaison observés

Au vu des premières analyses (**figure 11 A, C, et E**), j'ai porté mon attention sur quatre loci. J'ai calculé la valeur p empirique à partir de 3 000 simulations. Outre le locus 9q34 ($p<3.10^{-4}$) qui incluait le gène *ABO*, le locus 2p12-13 (p=0.002) contenait 84 gènes répartis sur 6 Mb ; le locus 6q12-13 (p=0.009) en contenait 47 répartis sur 18 Mb ; enfin, le locus 12q23 (p=0.005) en contenait 33 répartis sur 4 Mb. L'étape suivante aurait logiquement été d'explorer ces régions au moyen d'études d'association de SNPs avec les taux de vWF et FVIII. Malheureusement, au moment où je réalisais ce travail,

aucun génotypage de SNPs n'avait été réalisé dans l'étude FVL et il n'était pas envisageable de réaliser un criblage en SNPs au sein de chacune de ces régions de liaison. Il était donc nécessaire d'adopter une autre stratégie, schématisée **en figure 12,** pour identifier en leur sein un nombre limité de gènes candidats.

Figure 12: Stratégie adoptée pour explorer plus finement les signaux de liaison détectés



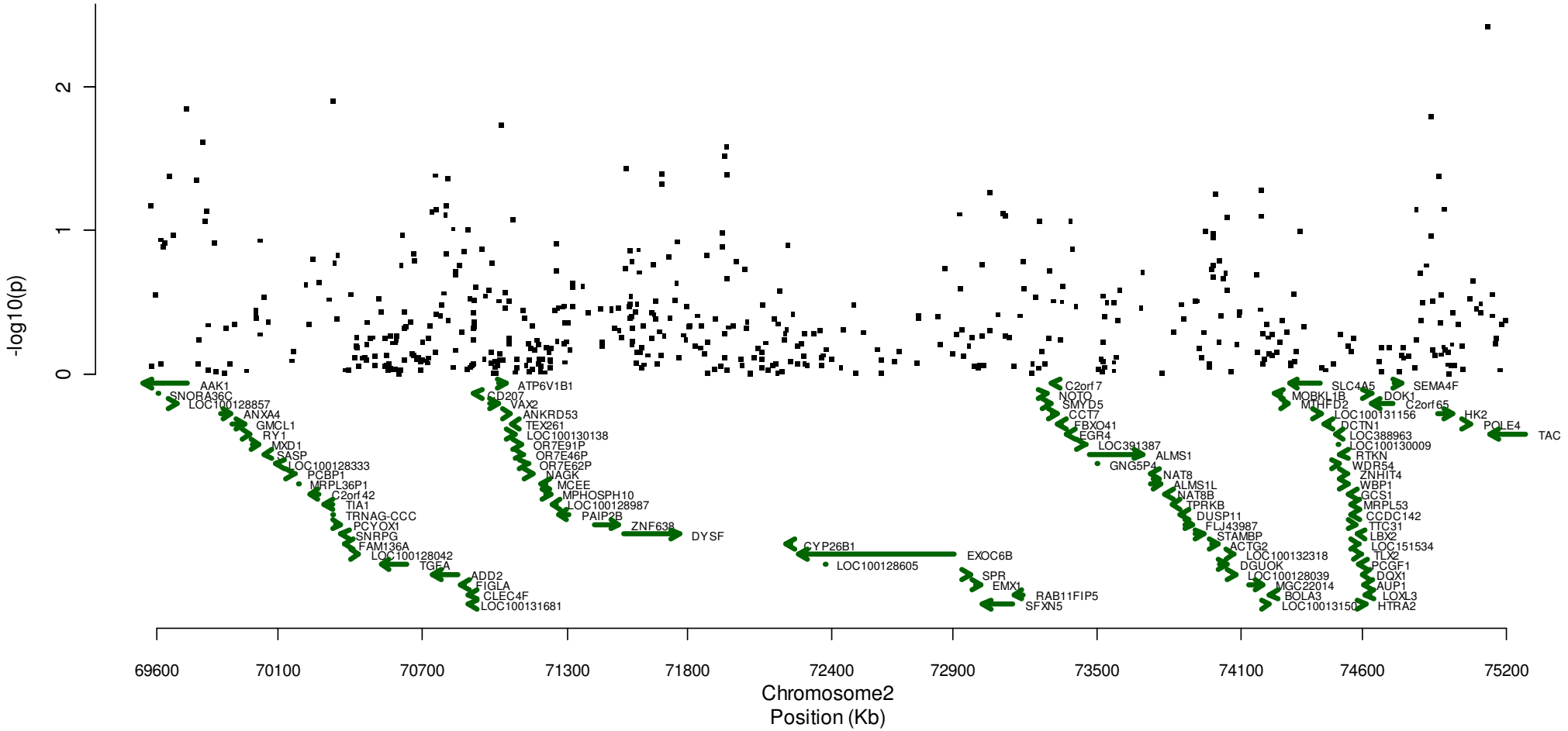
La recherche de SNPs situés dans les signaux de liaison de FVIII et/ou vWF, et associés au risque de MTEV a identifié, outre le gène *ABO*, deux gènes *BAI3* en 6q13 et *STAB2* en 12q23. Le gène *STAB2* venait juste d'être découvert par une grande étude GWAS (résultat présenté alors en congrès et publié ultérieurement [101]**). La suite de mon travail a consisté à étudier l'association entre des SNPs de BAI3 et les taux de vWF et de FVIII d'une part, et le risque de MTEV d'autre part.**

Chacune des régions des signaux de liaison a fait l'objet d'une recherche *in silico* d'associations entre des SNPs situés en leur sein et le risque de MTEV. Nous disposions en effet des résultats d'une GWAS portant sur 317 000 SNPs, génotypés chez 419 cas de MTEV survenue avant l'âge de 50 ans et 1228 témoins [71]. Nous avons ré-examiné ces résultats en nous focalisant sur les régions nouvellement mises en lumière grâce aux analyses de liaison. Nous nous reposions sur l'hypothèse qu'un gène modulant les taux de FVIII et vWF au point de générer un signal de liaison pourrait, quoique de manière modeste, influencer le risque de survenue de MTEV. Les SNPs de ce gène, ne pouvant passer le seuil de significativité, auraient donc été négligés lors de l'analyse pangénomique. Ainsi, sans nous fixer un seuil de significativité particulier, nous avons recherché des associations entre la MTEV et des SNPs situés dans les régions liées aux FVIII et/ou vWF, se détachant du bruit de fond de l'analyse pangénomique. Les résultats obtenus pour les régions correspondant aux signaux de liaison situés en 2p12-13, 6q13-14, 9q34 et 12q23 sont présentés en **figures 13 A, B, C et D**. Les résultats correspondants à des signaux de liaison plus discrets (2q33) ou révélés plus tardivement dans le courant de ma thèse (15q14, 19p13) sont présentés en annexe **pA10-12** à titre informatif.

C'est ainsi que nous avons identifié deux gènes : *BAI3* dans la région 6q13 et *STAB2* dans la région 12q23. Tous les SNPS dont l'association avec la MTEV avait une significativité inférieure à 0.001 sont décrits dans le **tableau 9 p84**. Ces SNPs étaient de bons candidats pour la recherche de polymorphismes associés aux FVIII et vWF. Cependant, le consortium CHARGE présentait les résultats d'une étude d'association pangénomique (« *GWAS* ») des taux de FVIII et vWF [101](voir §**IV.3.2.2**). Parmi les six gènes nouvellement découverts se trouvait *STAB2*. Il a donc été décidé que j'étudie en priorité les associations entre des polymorphismes de *BAI3* et les taux de FVIII et vWF. Une analyse haplotypique (voir **annexe pA4**) à partir des SNPs dont les associations avec la MTEV présentaient une significativité p<0,001 (voir **figure 13B**) a été réalisée à l'aide d'un programme développé par David Trégouët [156]. Il apparut que cinq SNPs permettaient de capturer l'ensemble des associations observées. Ils ont été génotypés dans l'échantillon Stanislas afin de tester leur association avec les taux de vWF et FVIII. Les échantillons MARTHA05 et FARIVE étaient disponibles pour des études de réplication.

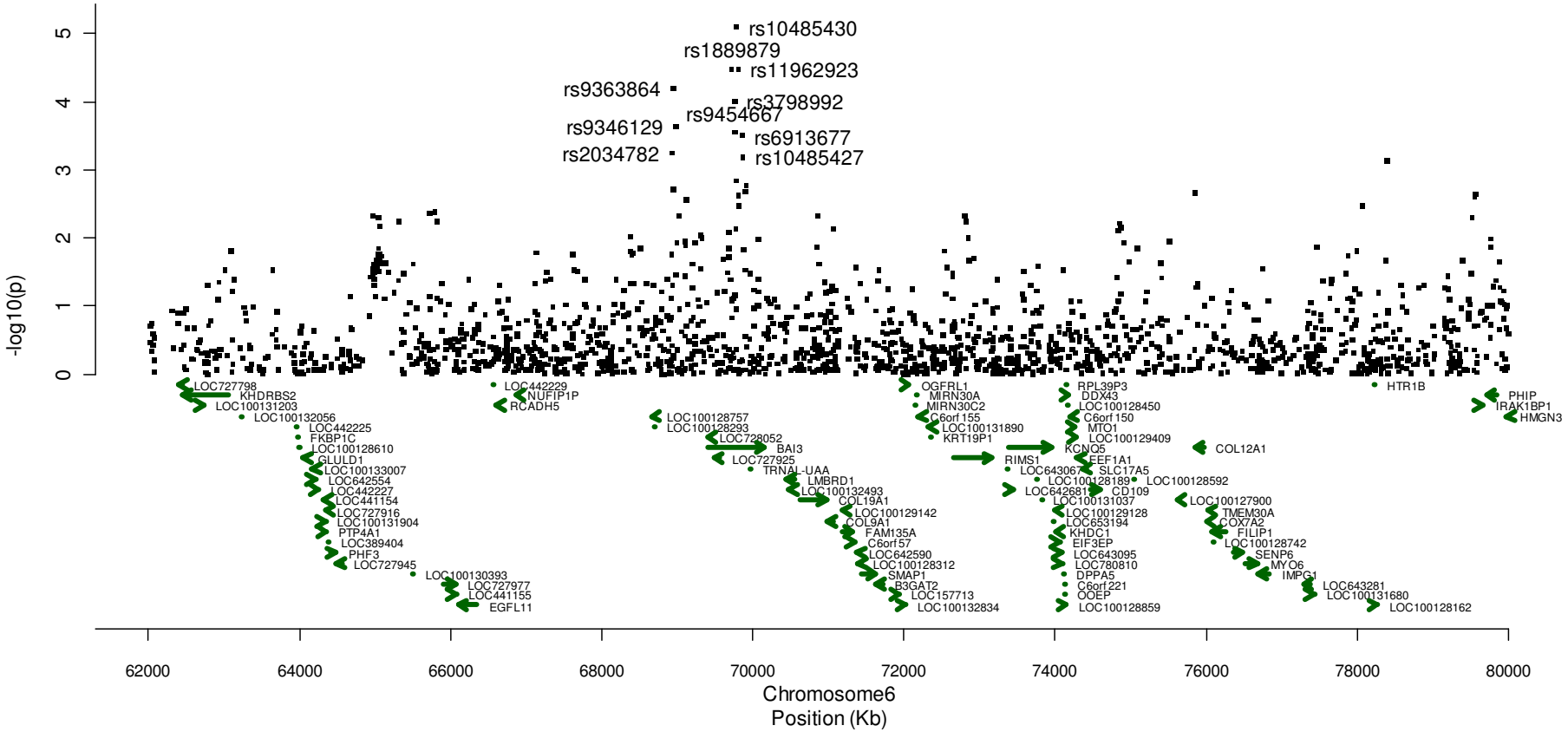
Figure 13 : Associations avec la MTEV observées *in silico* dans une étude d'association pangénomique cas-témoins (453 cas de MTEV de moins de 50 ans et 1327 témoins)

13.A- GWAS *in silico* : résultats obtenus dans la région 2p12-13



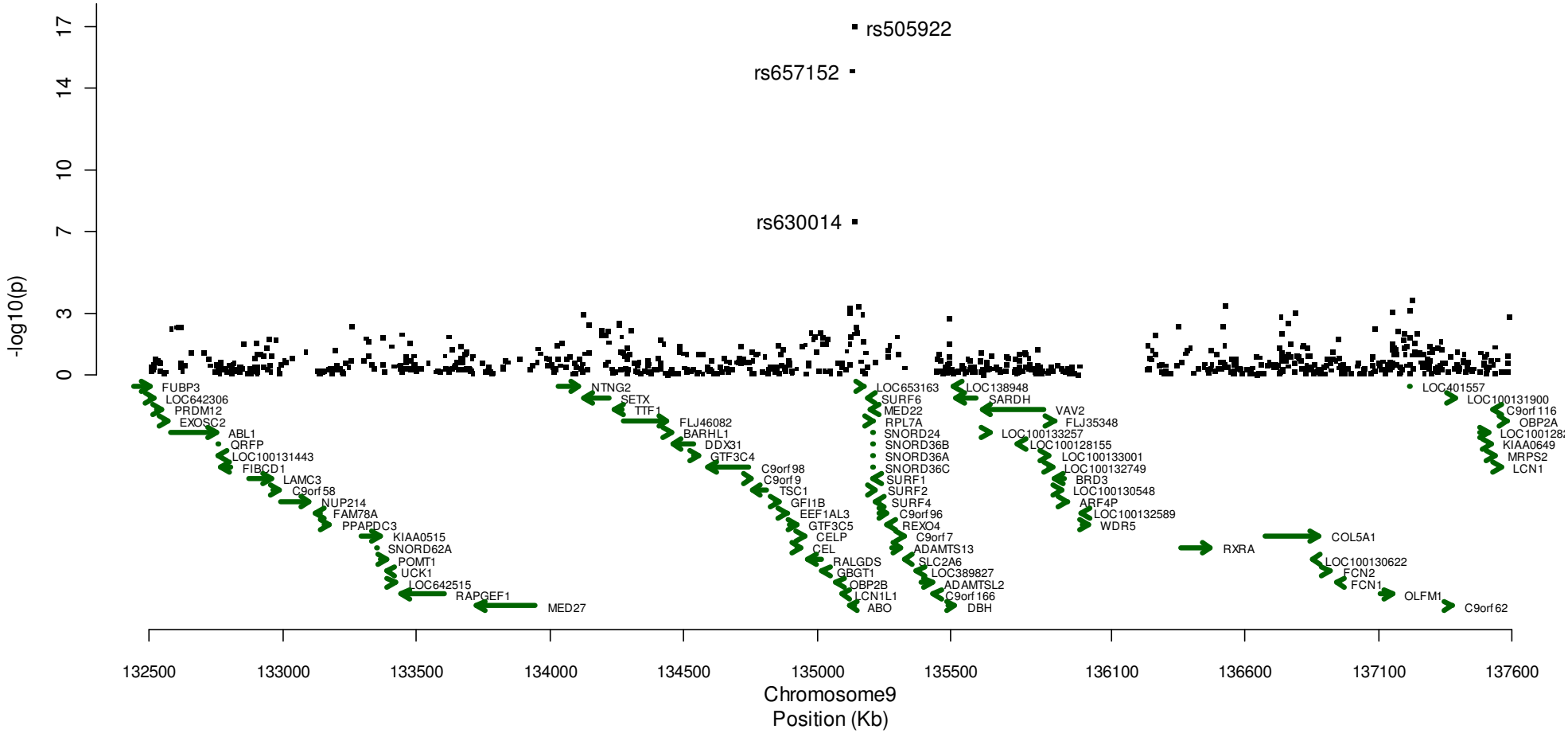
Aucune association avec la MTEV ne se détachait du bruit de fond de la GWAS *in silico* dans la région 2p12-13

13.B. GWAS *in silico* : résultats obtenus dans la région 6q13-14



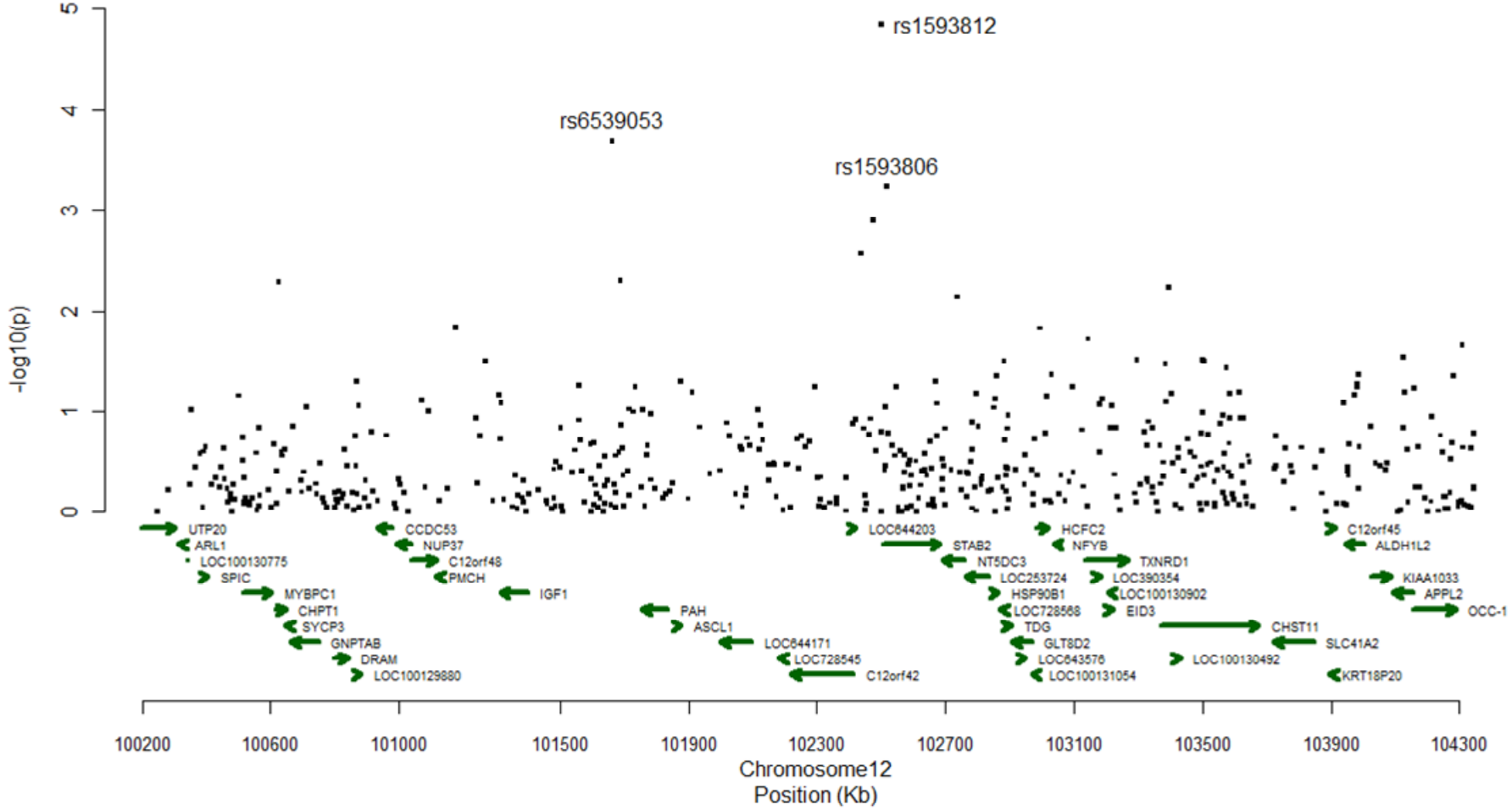
Un ensemble de SNPs situés dans le promoteur ou au sein du gène *BAI3,* lui-même localisé en 6q13, présentaient des associations avec la MTEV qui se détachaient du bruit de fond dans la GWAS *in silico*, avec une significativité p proche de 10^{-5}. La suite de mon travail s'est concentrée sur l'étude de polymorphismes du gène *BAI3*

13.C. GWAS *in silico* : résultats obtenus dans la région 9q34



Trois SNPs situés dans le gène *ABO* (9q34) présentaient des associations avec la MTEV très significatives. Ce gène était déjà connu pour influencer le risque de survenue de MTEV par l'intermédiaire de son rôle dans la clairance du vWF.

13.D. GWAS *in silico* : résultats obtenus dans la région 12q23



Un SNP situé dans le gène *STAB2,* lui-même localisé en 12q23, présentait une association avec la MTEV qui se détachait du bruit de fond dans la GWAS *in silico*, avec une significativité proche de 10^{-5}. Ce gène a également été découvert par le consortium CHARGE au moyen d'une étude GWAS portant sur les taux plasmatiques de FVIII et vWF . Je ne l'ai pas étudié plus en détail.

Tableau 9. Associations observées *in silico* dans une étude d'association pangénomique cas-témoins (419 MTEV de moins de 50 ans et 1228 témoins) : SNPs situés dans les signaux de liaison précédemment identifiés et associés à la MTEV avec une p-value<0.001 par un test de Cochran-Armitage

	Position	Localisation intra-génique	Allèles[1]	p HW[2]	MAF[3] témoins	MAF[3] cas	p[4]
Région 6q12							
BAI3							
rs2034782	68932330	Promoteur	C>**T**	0,550	0,051	0,084	$5{,}62\ 10^{-4}$
rs9363864*	68949277	Promoteur	**G**>A	0,501	0,483	0,403	$6{,}23\ 10^{-5}$
rs9346129	68977898	Promoteur	C>**T**	0,832	0,396	0,470	$2{,}33\ 10^{-4}$
rs1889879*	69720601	Intron	**A**>C	0,776	0,406	0,325	$3{,}26\ 10^{-5}$
rs3798992*	69755259	Intron	**T**>G	0,933	0,492	0,414	$9{,}64\ 10^{-5}$
rs9454667*	69759137	Intron	G>**A**	0,182	0,279	0,344	$2{,}81\ 10^{-4}$
rs10485430*	69785173	Intron	G>**A**	0,851	0,139	0,205	$8{,}01\ 10^{-6}$
rs11962923	69805321	Intron	T>**G**	0,891	0,139	0,199	$3{,}28\ 10^{-5}$
rs6913677	69860183	Intron	G>**A**	0,007	0,317	0,383	$3{,}18\ 10^{-4}$
rs10485427	69873172	Intron	T>**C**	0,525	0,242	0,302	$6{,}50\ 10^{-4}$
Région 12q13							
IGF1-PAH							
rs6539053	101663294	Intergénique	**G**>T	0,617	0,059	0,026	$2{,}05\ 10^{-4}$
STAB2							
rs1593812	102499779	Promoteur	A>**G**	0,132	0,123	0,183	$1{,}15\ 10^{-5}$
rs1593806	102517696	Intron	A>**G**	0,794	0,220	0,279	$5{,}80\ 10^{-4}$

[1] Les allèles fréquents et rares sont notés respectivement à gauche et à droite du signe >. L'allèle sur-représenté parmi les cas est en gras.
[2] Significativité du test d'Hardy-Weinberg
[3] Fréquence de l'allèle rare
[4] Significativité du test de Cochran-Armitage
* D'après une analyse haplotypique, ces SNPs capturaient l'ensemble des associations de ce locus.

VII.2. Etude d'association de polymorphismes du gène *BAI3* dans l'échantillon familial de la cohorte STANISLAS.

Cinq SNPs ont été génotypés dans l'échantillon Stanislas : rs9363864, rs1889879, rs3798992 rs9454667, rs10485430, afin d'étudier leurs associations avec les taux de FVIII et de vWF. Dans un premier temps, l'effet génotypique de chaque SNP sur les taux de FVIII, de vWF et de FVIII ajustés sur vWF a été étudié dans une analyse univariée ajustée sur l'âge et le sexe (**tableau 10 p85**). Les SNPs rs9363864 et rs3798992 étaient associés aux taux plasmatiques de vWF, expliquant respectivement R^2=2.30% (p=0.003) et R^2=2.13 (p=0.028) de la variance totale de vWF. Les allèles A de rs9363864 et G de rs3798992, qui étaient tous deux sous-représentés parmi les cas de la GWAS *in silico* (**tableau 9 p84**), sont ici, de manière cohérente, associés à des taux plus faibles de vWF. On observait un effet récessif de rs9363864-A et un effet dominant de rs3798992-G sur les taux de vWF. Ces SNPs ne semblent pas avoir d'influence sur les taux de FVIII, indépendamment de l'effet du vWF, puisque les associations observées avec le FVIII disparaissent après ajustement sur vWF.

Tableau 10: Association entre les SNPs du gène *BAI3* et les taux plasmatiques de vWF et FVIII (échantillon STANISLAS)

			FVIII			vWF			FVIII ajusté sur vWF		
	p_{HW}[1]	MAF[2]	Effet	R^2	p	Effet	R^2	p	Effet	R^2	p
rs9363864	0,89	0,51		1,97%	0,033		**2,3%**	**0,003**		0,87%	0,204
GG (107)			réf			**réf**			réf		
GA (225)			-3,3 (4,1)			**2,4 (5,8)**			-4,6 (2,7)		
AA (109)			-11,9 (5)			**-13,1 (6,3)**			-5 (3,2)		
rs1889879	0,17	0,37		0,57%	0,354		0,28%	0,530		0,3%	0,573
AA (163)			réf			réf			réf		
AC (221)			-4 (3,7)			-4,2 (4,5)			-1,8 (2,5)		
CC (24)			2,2 (6)			1 (7,8)			1,7 (4,1)		
rs3798992	0,73	0,45		1,82%	0,091		**2,13%**	**0,028**		0,37%	0,565
TT (129)			réf			**réf**			réf		
TG (221)			-7,6 (4)			**-13,2 (5,1)**			-0,7 (2,7)		
GG (95)			-11,3 (5,6)			**-14,6 (7,1)**			-3,6 (3,7)		
rs9454667	0,96	0,25		0,01%	0,977		0%	0,993		0,02%	0,968
GG (257)			réf			réf			réf		
GA (150)			0,8 (3,8)			0,4 (4,2)			0,6 (2,4)		
AA (26)			0,3 (7,7)			-0,3 (6,7)			0,5 (5,4)		
rs10485430	0,98	0,15		0,59%	0,257		0,43%	0,194		0,19%	0,665
GG (326)			réf			réf			réf		
GA (111)			4,8 (3,9)			5,3 (4,6)			2 (2,6)		
AA (10)			-7,1 (8,5)			-9,5 (9,8)			-2,1 (5,3)		

[1] : p value du test de Hardy-Weinberg réalisé chez les parents
[2] : fréquence de l'allèle rare chez les parents

Ces premiers résultats nécessitent une étude plus approfondie afin d'établir l'indépendance des effets de rs9363864 et rs3798992, mais aussi de s'assurer que la prise en compte de ces deux effets ne révèle aucune association avec les autres SNPs de *BAI3*. Les analyses multivariées permettent de répondre à ces questions. Une étude du déséquilibre de liaison entre ces SNPs est utile avant d'appréhender ces dernières (**voir annexe pA5 « déséquilibre de liaison »**). Elle a été réalisée uniquement chez les parents afin de s'assurer de l'indépendance des observations. Le déséquilibre de liaison entre les SNPs est faible. Notamment, la valeur r^2 est nulle entre rs3798992 et rs9363864 (**tableau 11**).

Tableau 11. Mesure du déséquilibre de liaison entre les cinq SNPs de *BAI3* (échantillon STANISLAS)

	rs9363864	**rs1889879**	**rs3798992**	**rs9454667**	**rs10485430**
rs9363864	-	0,02	-0,06	-0,02	-0,43
rs1889879	0	-	0,68	-0,95	-1
rs3798992	0	0,32	-	-1	-1
rs9454667	0	0,17	0,28	-	0,79
rs10485430	0,03	0,10	0,14	0,31	-

Les valeurs D' et r^2 , estimées chez les parents de l'échantillon STANISLAS, sont indiquées respectivement au-dessus et en-dessous de la diagonale.

L'indépendance des associations avec les taux de vWF observées pour rs3798992 et rs9363864 a été confirmée par des analyses multivariées, réalisées selon un mode récessif pour rs9363864-A, dominant pour rs3798992-G, et additif pour les trois autres SNPs qui n'étaient pas associés significativement aux phénotypes. Ces modèles permettent d'augmenter la puissance des analyses en faisant l'économie d'un degré de liberté, mais aussi de simplifier les interprétations des interactions testées. Les premiers modèles testés incluaient rs9363864 et rs3798992, avec, puis sans leur interaction. Ils ont permis d'établir l'absence d'interaction entre ces SNPs (p=0.24), de même que l'indépendance de leurs effets sur vWF. En effet, les allèles rares rs9363864-A rs3798992-G diminuent en moyenne respectivement de 14 (p=0.002), et de 13 (p=0.011) UI/dL les taux de vWF.

J'ai recherché ensuite si la prise en compte de ces deux effets était à même de révéler un nouvel effet de l'un des trois autres SNPs disponibles. Ces analyses incluaient également l'estimation de l'effet du gène *ABO*, dont les allèles A1 et B (par rapport aux allèles A2 et O) sont le déterminant génétique majeur d'un taux élevé de vWF. Le but recherché de l'ajustement sur *ABO* n'était pas tant de prendre en compte un éventuel facteur de confusion que d'augmenter la puissance de l'analyse en diminuant la variance résiduelle du modèle. Cependant, de façon surprenante et inexpliquée, les génotypes de *BAI3* se répartissaient différemment selon la présence d'allèles A1 et/ou B, ou bien uniquement d'allèles O et A2 de manière significative pour plusieurs SNP, dont rs3798992 (**tableau 12**).

Tableau 12. Répartition des génotypes des SNPs de *BAI3* en fonction du groupe sanguin

	Groupe sanguin non à risque*	**Groupe sanguin à risque****	**p**
rs9363864			0,076
GG	41 (18,47%)	66 (30,14%)	
GA	117 (52,7%)	108 (49,32%)	
AA	64 (28,83%)	45 (20,55%)	
rs1889879			0,076
AA	67 (30,73%)	96 (43,64%)	
AC	120 (55,05%)	101 (45,91%)	
CC	31 (14,22%)	23 (10,45%)	
rs3798992			0,006
TT	49 (21,88%)	80 (36,2%)	
TG	112 (50%)	109 (49,32%)	
GG	63 (28,12%)	32 (14,48%)	
rs9454667			0,032
GG	146 (66,97%)	111 (51,63%)	
GA	64 (29,36%)	86 (40%)	
AA	8 (3,67%)	18 (8,37%)	
rs10485430			0,014
GG	179 (79,56%)	147 (66,22%)	
GA	40 (17,78%)	71 (31,98%)	
AA	6 (2,67%)	4 (1,8%)	

*** Les groupes sanguins non à risque ne sont constitués que d'allèles O et/ou A2**
**** les groupes sanguins à risque sont les groupes présentant au moins un allèle A1 ou B.**

Ce résultat obtenu par une EE1 avec l'ensemble de l'échantillon a été confirmé par une régression linéaire classique en restreignant l'analyse aux parents de l'échantillon. Il m'a alors fait considérer le gène *ABO* comme un éventuel facteur de confusion pour l'étude du gène *BAI3*. De fait, l'ajustement sur ABO affaiblissait l'effet de rs3798992. Sans ajustement sur ABO, le taux de vWF, parmi les porteurs de G par rapport au homozygote TT, était inférieur de 13 UI/dL (p=0.011) (cf ci-dessus). Ce chiffre était de 7 UI/dL (p=0.112) après ajustement sur ABO. Cependant, la prise en compte des deux SNPs rs9363864 et rs3798992, ainsi que d'ABO, dévoile une nouvelle association significative entre rs9454667 et les taux de vWF. Le modèle finalement retenu **(tableau 13**), qui satisfait au mieux les critères d'Akaike, inclut outre l'âge, le sexe, et le gène *ABO*, trois SNPs de *BAI3* aux effets significatifs et indépendants : rs9363864 (p=0.018), rs3798992 (p=0.023) et rs9454667 (p=0.009).

Tableau 13. Analyse multivariée des effets des SNPs de *BAI3* sur les taux de vWF (échantillon STANISLAS)

	Effectif	Effet (sd)	p
Age (ans)		0,27 (0,12)	**0,029**
Sexe			
Pères et Fils	230		
Mères et Filles	221	-0,43 (3,78)	0,911
Nombre d'allèles A1 et/ou B			
0	226		
1 ou 2	225	34,3 (4,1)	**$<10^{-16}$**
rs9363864 (récessif)			
GG+GA	332		
AA	109	-9,0 (3,8)	**0,018**
rs3798992 (dominant)			
TT	129		
TG+GG	316	-11,2 (5,0)	**0,025**
rs9454667 (additif)			
GG	257		
GA	150	-7,9 (3,0)	**0,009**
AA	26		

Finalement, une étude stratifiée sur le lien de parenté (parent ou enfant) ainsi que sur le gène *ABO*, suggère que dans cet échantillon, les effets de BAI3 pourraient être plus marqués chez les enfants **(tableau 14)** et chez les porteurs de A1 ou B **(tableau 15)**. Cependant aucune des interactions testées entre les trois SNPs et ABO d'une part, les trois SNPs et le lien de parenté d'autre part, n'était significative.

Tableau 14. Analyse multivariée des effets des SNPs de *BAI3* sur les taux de vWF, stratifiée sur le lien de parenté (échantillon STANISLAS)

	Parents			Enfants		
	Effectif	**Effet (sd)**	**p**	**Effectif**	**Effet (sd)**	**p**
Age (ans)		1,5 (0,5)	0,004		-0,9 (0,6)	0,127
Sexe						
Pères ou Fils	107			123		
Mères ou Filles	107	4,5 (5,7)	0,426	114	-1,6 (4,8)	0,748
Nombre d'allèles A1 et/ou B						
0	112			114		
1 ou 2	102	32,1 (5,4)	**$3\ 10^{-9}$**	123	37,5 (5,2)	**$9\ 10^{-13}$**
rs9363864 (récessif)						
GG+GA	155			177		
AA	55	-7,3 (5,9)	0,219	54	-11,7 (5,0)	**0,020**
rs3798992 (dominant)						
TT	63			66		
TG+GG	150	-5,5 (6,2)	0,37	166	-14,7 (7,7)	**0,058**
rs9454667 (additif)						
GG	116			141		
GA	77	-6,8 (4,5)	0,135	73	-7,6 (4,3)	**0,079**
AA	13			13		

Tableau 15. Analyse multivariée des effets des SNPs de BAI3 sur les taux de vWF, stratifiée sur la présence d'au moins un allèle A1 ou B (échantillon STANISLAS).

	Aucun allèle A1 ou B			Au moins un allèle A1 ou B		
	Effectif	**Effet (sd)**	**p**	**Effectif**	**Effet (sd)**	**p**
Age (ans)		0,39 (0,16)	0,015		0,13 (0,18)	0,470
Sexe						
Pères et Fils	118			112		
Mères et Filles	108	2,3 (4,0)	0,564	113	-3,2 (6,4)	0,619
rs9363864 (récessif)						
GG+GA	158			174		
AA	64	-5,2 (5,0)	0,298	45	-12,2 (6,1)	**0,045**
rs3798992 (dominant)						
TT	49			80		
TG+GG	175	-4,4 (6,9)	0,519	141	-14,5 (6,3)	**0,023**
rs9454667 (additif)						
GG	146			111		
GA	64	-1,4 (4,7)	0,759	86	-10,8 (4,6)	
AA	8			18		**0,020**

Au total, parmi les cinq SNPs-candidats à une association aux taux de vWF et/ou FVIII, nous retiendrons rs9363864 et rs3798992. Leurs associations avec les taux de vWF étaient significatives et indédendantes. Elles étaient atténuées avec les taux de FVIII et disparaissaient après ajustement sur vWF. Un troisième SNP, rs9454667, était associé aux taux de vWF, mais le sens de l'association n'était pas homogène celui de la GWAS-*in silico*. En effet, l'allèle A, sur-représenté parmi les cas de cette dernière, était associé à une diminution des taux de vWF. La raison de sa présence dans le modèle finalement retenu était de permettre un meilleur ajustement des autres effets de rs9363864 et rs3798992. Dans les sections suivantes, nous concentrerons notre recherche sur la réplication de ces derniers.

VII.3. Réplication de l'association entre les SNPs rs9363864 et rs3798992 du gène *BAI3* et les taux de vWF (cas de l'échantillon MARTHA05)

Parmi les 1148 patients de l'étude MARTHA05 pour lesquels je disposais des taux de vWF, j'ai sélectionné les 916 patients dont la thrombose veineuse était survenue avant l'âge de 50 ans. J'avais pour but d'homogénéiser cet échantillon à celui de la GWAS *in silico* à l'origine de l'hypothèse d'une influence de *BAI3* sur les taux de vWF et/ou FVIII. De plus, les effets des SNPs de *BAI3* semblaient plus importants parmi les sujets les plus jeunes de l'échantillon STANISLAS. Ainsi, la perte de puissance induite par la diminution de la taille de l'échantillon MARTHA05 pouvait être compensée par une amélioration de la spécificité de l'effet recherché. Les deux polymorphismes rs9363864 et rs3798992 du gène *BAI3* ont été génotypés dans ce sous-ensemble de 916 sujets. L'influence des SNPs sur les taux de vWF a été modélisée au moyen d'une régression linéaire ajustée sur l'âge et le sexe. Les modes de dominance et récessivité précédemment étudiés ont été conservés. Les allèles rs9363864-A et rs3798992-G, étudiés séparemment dans deux modèles, était associés respectivement à une baisse de 1,6 (p=0,78) et de 13,8 (p=0,007).

Le gène *ABO* a été intégré aux analyses, mais je ne disposais pas de la même précision de génotypage que pour l'étude Stanislas. En effet, les allèles A1 et A2 ne pouvaient être distingués. J'ai constitué deux classes : d'une part le groupe O présentant en moyenne des taux faibles de vWF, d'autres parts les groupes A, B et AB présentant en moyenne des taux élevés. Le pourcentage de personnes A2 dans ce groupe est probablement faible. Contrairement à ce qui était observé dans Stanislas, mais conformément à ce qui était attendu, la distribution des génotypes de rs9363864 et rs3798992 ne semblait pas dépendre significativement du groupe ABO (**tableau 16**). Cependant, une interaction entre rs9363864 et ABO s'étant avérée significative, j'ai considéré un modèle incluant, outre l'âge et le sexe, rs3798992, rs9363864, le groupe ABO, et l'interaction rs9363864*ABO (**tableau 17**).

Tableau 16. Répartition des génotypes des SNPs de *BAI3* en fonction du groupe sanguin (échantillon MARTHA05).

	Groupe O	Groupes A, B ou AB	P
rs9363864			0,19
GG	44 (24,2%)	147 (26,0%)	
GA	90 (49,5%)	305 (54,0%)	
AA	48 (26,4%)	113 (20,0%)	
rs3798992			0,52
TT	45 (24,5%)	163 (28,7%)	
TG	101 (54,9%)	290 (51,1%)	
GG	38 (20,7%)	114 (20,1%)	

Tableau 17. Analyse multivariée des effets des SNPs de *BAI3* sur les taux de vWF (échantillon MARTHA05)

	Effectif	Effet (écart-type)	p
Age (ans)		0,79 (0,29)	**0,007**
Sexe			
Hommes	231	réf,	
Femmes	685	-15,5 (6,1)	**0,012**
Groupe ABO[1]			
O	188	Réf.	
A, B ou AB	577	28,5 (6,6)	**1,89 10^{-5}**
rs9363864-récessif			
GG+GA	688	réf.	
AA	199	-27,2 (11,1)	**0,015**
rs3798992-dominant			
TT	253	réf.	
TG+GG	640	-15,8 (5,5)	**0,004**
Interaction ABO*rs9363864[1]			
O et/ou rs9363864-GG+GA	634	réf.	
A+B+AB et rs9363864-AA	113	34,5 (13,2)	**0,009**

[1] Il y avait 151 patients dont le groupe sanguin était inconnu. Si les analyses ajustées sur le groupe ABO font gagner de la puissance grâce à une meilleure modélisation, elles en font perdre par diminution de la taille de l'échantillon analysé.

Ainsi, nous avons reproduit, au sein d'un échantillon de patients atteints de MTEV, les associations observées parmi les familles en bonne santé de Stanislas entre le taux de vWF et deux SNPs de BAI3, rs9363864 et rs3798992. Cependant, les analyses suggèrent des phénomènes d'interaction entre le gène ABO et le gène BAI3, soit artefactuels, soit particulièrement complexes. En effet, si la modulation des taux de vWF par rs9363864 semble plus importante parmi les porteurs d'allèles A1 et B d'un échantillon de familles en bonne santé, elle paraît en revanche l'apanage des patients atteints de MTEV de groupe sanguin O.

VII.4. Association entre les polymorphismes rs9363864 et rs3798992 du gène *BAI3* et le risque de MTEV (échantillons MARTHA05 et FARIVE)

L'influence de rs9363864 et rs3798992 sur le risque de MTEV a été testée dans deux études cas-témoins, MARTHA05 et FARIVE. Pour les raisons précédemment avancées d'homogénéisation des échantillons et de puissance, j'ai restreint mes analyses aux MTEV survenues avant l'âge de 50 ans. En revanche, aucune sélection de témoins n'a été effectuée. Les analyses ont été conduites sur 916 cas et 801 témoins de MARTHA05, et 250 cas et 606 témoins de FARIVE. De manière analogue à ce qui avait été réalisé pour la GWAS *in silico*, les fréquences alléliques de rs9363864 et rs3798992 des cas et des témoins ont d'abord été comparées par un test de tendance de Cochran-Armitage. Les résultats obtenus dans chaque échantillon sont résumés dans le **tableau 18**. La seule association proche de la significativité a été observée dans FARIVE pour rs9363864 (p=0.052). L'allèle rare A est sous-représenté

parmi les cas par rapport aux témoins (42,8% *versus* 48,2%). Le sens de l'association est là encore concordant avec celui des associations observées dans la GWAS *in silico*, et dans les deux études quantitatives réalisées avec les familles de STANISLAS et les cas de MARTHA05.

Tableau 18. Association entre les polymorphismes de BAI3 et le risque de MTEV (échantillons MARTHA05 et FARIVE)

	pHW	Génotype			MAF	p*
MARTHA05						
rs9363864		**GG**	**GA**	**AA**		
Controls	0,64	231 (29,1%)	389 (48,9%)	175 (22,0%)	46,5%	0,386
Cases	-	235 (26,5%)	453 (51,1%)	199 (22,4%)	48,0%	
rs3798992		**TT**	**TG**	**GG**		
Controls	0,43	244 (30,6%)	384 (48,2%)	169 (21,2%)	45,3%	0,671
Cases	-	253 (28,3%)	458 (51,3%)	182 (20,4%)	46,0%	
FARIVE						
rs9363864		**GG**	**GA**	**AA**		
Controls	0,09	168 (28,5%)	274 (46,5%)	147 (25,0%)	48,2%	**0,052**
Cases	-	82 (33,6%)	115 (47,1%)	47 (19,3%)	42,8%	
rs3798992		**TT**	**TG**	**GG**		
Controls	0,04	148 (25,4%)	315 (54,1%)	119 (20,4%)	47,5%	0,122
Cases		77 (31,3%)	124 (50,4%)	45 (18,3%)	43,4%	

***test de tendance de Cochran-Armitage**

Cependant, les cas et les témoins de MARTHA05 étaient, par construction de l'échantillon, composés respectivement de 44% et de 41% porteurs de mutation du FII ou de la mutation FVL. Ils n'étaient que de 24% et de 8% dans FARIVE. Des analyses stratifiées sur la présence ou non de ces mutations ont été réalisées dans le but de déterminer si la discordance des résultats de ces deux études pouvait être incriminée à leur mode de recrutement différent. J'ai utilisé un modèle logistique ajusté sur l'âge et le sexe, prenant en compte les modes de dominance et récessivité précédemment étudiés. De façon remarquable, on observe des résultats très homogènes entre les études MARTHA05 et FARIVE pour le polymorphisme rs9363864, à la fois dans la strate des non-porteurs de mutations thrombogènes, et dans la strate des porteurs de ces mutations **(tableau 19)**. En effet, aucune association significative n'a pu être observée parmi les porteurs de mutation, tandis que l'allèle A est significativement moins fréquent parmi les cas de MARTHA05 (OR=0,62 ; p=0,02) et de FARIVE (OR=0,56 ; p=0,02). Le test d'interaction de Cochran-Mantel-Haenszel a confirmé l'absence d'hétérogénéité des résultats de ces deux études chez les porteurs de mutations (p=0,38) et chez les non-porteurs (p=0,74).

Tableau 19. Association entre les polymorphismes de *BAI3* et le risque de MTEV stratifiée sur la présence d'une mutation thrombogène (échantillons MARTHA05 et FARIVE)

NON PORTEURS DE MUTATIONS DU FII OU DE FVL						
	MARTHA			FARIVE		
rs9363864	**GG/GA**	**AA**	**OR**	**GG/GA**	**AA**	**OR**
Témoins	361 (76,3%)	112 (23,7%)	0,62 [0,43 – 0,90]	401 (74,3%)	139 (25,7%)	0,56 [0,35-0,90]
Cas	394 (78,5%)	108 (21,5%)	**p = 0,012**	151 (82,5%)	32 (17,5%)	**p=0,016**
rs3798992	**TT**	**TG/GG**		**TT**	**TG/GG**	
Témoins	146 (30,9%)	327 (69,1%)	1,06 [0,75-1,5]	127 (23,8%)	406 (76,2%)	0,63 [0,42-0,97]
Cas	131 (26,1%)	370 (73,9%)	p=0,751	60 (32,4%)	125 (67,6%)	**p=0,034**
PORTEURS DE MUTATIONS DU FII OU DE FVL						
	MARTHA			FARIVE		
rs9363864	**GG/GA**	**AA**	**OR**	**GG/GA**	**AA**	**OR**
Témoins	259 (80,4%)	63 (19,6%)	1,19 [0,77-1,82]	38 (84,4%)	7 (15,6%)	2,30 [0,57-9,30]
Cas	294 (76,4%)	91 (23,6%)	p=0,433	44 (75,9%)	14 (24,1%)	p=0,244
rs3798992	**TT**	**TG/GG**		**TT**	**TG/GG**	
Témoins	98 (30,2%)	226 (69,8%)	0,99 [0,68-1,45]	20 (44,4%)	25 (55,6%)	1,73 [0,58-5,14]
Cas	122 (31,1%)	270 (68,9%)	p=0,968	16 (27,6%)	42 (72,4%)	p=0,323

Analyse multivariée incluant rs9363864, rs3798992, l'âge et le sexe

En raison de l'homogénéité des effets de rs9363864 dans les deux études, j'ai pu combiner tous les sujets porteurs de mutations thrombogènes d'un côté, et tous les sujets non porteurs de mutations thrombogènes de l'autre. L'OR mesurant la force de l'association du génotype rs9363864-AA avec la MTEV était alors de 0,59 [0,44-0,80] ($p = 5,04\ 10^{-4}$) parmi les 1 698 non-porteurs de mutations, tandis qu'aucune association significative n'était observée parmi les 810 porteurs de mutations (OR=1,26 [0,83-1,89], p = 0,28). Les ORs estimés dans les deux strates étaient significativement différents (p= 0,004, testé par un test d'interaction de Cochran-Mantel-Haenszel). Quant à rs3798992, dont l'allèle G tendait à être moins fréquent parmi les cas de FARIVE par rapport aux témoins, son association avec la MTEV apparaît, là encore, uniquement significative parmi les non porteurs de mutations thrombogènes (OR=0,63 [0,42-0,97] p=0,034). Néanmoins, une telle association n'a pas pu être observée dans l'échantillon MARTHA05.

Compte-tenu des interactions entre rs9363864 et le gène *ABO* intervenant dans la modulation des taux de vWF observées **tableau 15 p88 et tableau 17 p90**, il était intéressant d'étudier si ABO interagissait avec rs9386864 dans le risque de survenue de MTEV. Ces analyses ont été conduites dans les deux échantillons combinant les sujets de MARTHA05 et de FARIVE: celui des non-porteurs et celui des porteurs de mutations thrombogènes. Parmi les non-porteurs de mutations, encore une fois, l'interaction entre ABO et rs9363864 s'est avérée significative (p=0.008). En effet, chez les personnes présentant l'allèle A et/ou B, le génotype rs9386864-AA est associé à la MTEV avec un OR de 0,46 [0,33-0,66] ($p = 2,52\ 10^{-5}$) **(tableau 20).** Chez les personnes du groupe O, rs9363864 ne semble pas modifier le risque

de survenue d'une MTEV. Parmi les porteurs de mutations thrombogènes, rs9363864 ne semble toujours pas intervenir dans le risque de survenue de MTEV, quel que soit le groupe sanguin.

Tableau 20. Association entre le polymorphisme *BAI3*-rs9363864 et le risque de MTEV stratifiée sur la présence d'une mutation thrombogène et le groupe sanguin (échantillons MARTHA05 et FARIVE combinés)

NON PORTEURS DE MUTATIONS DU FII OU DE FVL						
	Groupe O			Groupe A, B et AB		
rs9363864	**GG/GA**	**AA**	**OR**	**GG/GA**	**AA**	**OR**
Témoins	316 (76.5%)	97 (23.5%)	1.04 [0.62-1.74]	376 (73.2%)	138 (26.8%)	0.46 [0.33-0.66]
Cas	115 (73.7%)	41 (26.3%)	p=0.879	404 (81.5%)	92 (18.5%)	**$p=2.52\ 10^{-5}$**
PORTEURS DE MUTATIONS DU FII OU DE FVL						
	Groupe O			Groupe A, B et AB		
rs9363864	**GG/GA**	**AA**	**OR**	**GG/GA**	**AA**	**OR**
Témoins	117 (80.7%)	28 (19.3%)	1.31 [0.55-3.12]	157 (80.5%)	38 (19.5%)	1.13 [0.66-1.95]
Cas	59 (80.8)	14 (19.2%)	p=0.540	178 (76.1%)	56 (23.9%)	p=0.661

Analyse ajustée sur rs3798992, l'âge et le sexe

VII.5. Discussion

Synthèse des principaux résultats

Ce travail consistait à réaliser des analyses de liaison pangénomique des taux vWF et FVIII, suivies d'explorations des signaux de liaison dans le but de rechercher des polymorphismes associés à ces phénotypes. Les analyses ont été conduites à partir de plusieurs échantillons de conception (*design*) très variée, afin d'éprouver la robustesse des résultats. Les analyses menées dans un échantillon de grandes familles enrichies en mutation FVL ont mis en lumière, outre un signal de liaison dans la région du gène *ABO,* trois autres en 2p12-13, 6q13-14 et 12q23. Parmi les quelques 200 gènes situés dans les signaux de liaison, *BAI3* et *STAB2* étaient particulièrement intéressants. Ils étaient en effet associés au risque de MTEV dans une étude pangénomique, quoique de façon non significative compte-tenu du grand nombre de tests réalisés par de telles études.

BAI3. Dans un échantillon de familles nucléaires en bonne santé (cohorte Stanislas), deux SNPs du gène *BAI3* en faible déséquilibre de liaison l'un avec l'autre (rs9386864 situé dans la région promotrice et rs3798992 situé dans un intron) étaient associés de manière indépendante aux taux de vWF. Les associations étaient plus fortes chez les personnes présentant au gène *ABO* les allèles A1 et B que chez celles ne présentant que les allèles A2 ou O. Ces deux associations ont été reproduites dans un second échantillon incluant des patients ayant un antécédent de MTEV (les cas de l'étude MARTHA05). Cette fois-ci, cependant,

l'association entre rs9386864 et les taux de vWF s'observait uniquement chez les personnes de groupe O. Enfin, l'allèle de rs9386864 associé aux taux élevés de vWF était également associé à un excès de risque de MTEV dans deux échantillons cas-témoins (les études MARTHA05 et FARIVE) lorsque seuls étaient considérés les non porteurs des mutations FVL et FII. De façon cohérente avec les analyses de l'échantillon Stanislas, rs9386864 était spécifiquement associé au risque de survenue d'une MTEV lorsque les patients étaient du groupe sanguin A, B ou AB. L'association entre rs3798992 et la MTEV, quant à elle, n'a été reproduite que dans l'un de ces deux échantillons (FARIVE). La méconnaissance du mode d'action de *BAI3* sur les taux de vWF ne nous permet pas d'expliquer la manière dont il pourrait interagir avec *ABO*. Cette interaction est particulièrement complexe puisqu'elle est en sens inverse chez les sujets sains de Stanislas, et les sujets atteints de MTEV de MARTHA05.

STAB2. Le deuxième gène de grand intérêt était *STAB2,* situé dans la région 12q13. Il présentait non seulement un SNP (rs1593812) associé à la MTEV dans la GWAS *in silico*, avec une significativité proche de 10^{-5}, mais aussi deux autres SNPs (rs8981022 et rs12229292) associés aux taux de FVIII et vWF d'après une étude pangénomique publiée récemment par le Consortium CHARGE [101]. Les groupes de David Trégouët et de Pierre Morange ont étudié plus en détail l'influence sur le risque de MTEV des polymorphismes identifiés par CHARGE. Les SNPs rs8981022 et rs12229292 n'étaient pas associés au risque de MTEV ni dans la GWAS *in silico* ni dans MARTHA05 [157]. Cependant, dans une nouvelle GWAS de la MTEV (constituée des échantillons MARTHA08 et MARTHA10 pour les cas, et de la cohorte des Trois Cités pour les témoins) disponible à la fin de ma thèse, rs12229292 était associé au risque de MTEV. De plus, une analyse haplotypique a permis de mettre en évidence un effet synergique entre rs1593812 et rs12229292 à la fois dans la GWAS *in silico* et dans la nouvelle GWAS[158]. Ainsi, l'haplotype GT constitué par ces deux SNPs augmentait significativement le risque de MTEV par rapport l'haplotype AC qui était le plus fréquent L'OR était de 2,18 [1,66-2,87] ($p = 2{,}2.10^{-8}$) dans les deux GWAS combinées.

Limites des études statistiques quant à la découverte d'un polymorphisme fonctionnel

Plusieurs SNPs de *BAI3* étaient associés au risque de survenue de MTEV dans l'étude pangénomique, tandis que seuls deux d'entre eux étaient associés aux taux de vWF dans STANISLAS. Ces deux SNPs étaient associés au risque de survenue de MTEV dans l'étude cas-témoin FARIVE, mais seulement l'un d'entre eux l'était dans l'étude MARTHA05. De

façon analogue, le gène *STAB2* est mis en lumière par plusieurs études aux approches méthodoliques variées, mais les polymorphismes associés à la MTEV ou aux traits quantitatifs ne sont pas toujours les mêmes. Cela souligne le caractère non fonctionnel des SNPs initialement découverts. Les SNPs non fonctionnels sont des SNPs qui n'ont pas d'action directe sur le caractère étudié. En revanche, ils peuvent éventuellement se trouver en déséquilibre de liaison avec le SNP biologiquement responsable du contrôle de ce caractère. Ainsi, il est certainement prématuré de proposer l'étude fonctionnelle de *BAI3*-rs9386864 ou *STAB2*-rs1593812 à des équipes de biologistes, bien que ces associations aient pu être reproduites dans plusieurs échantillons indépendants. Il conviendrait auparavant d'identifier l'ensemble des SNPs en déséquilibre de liaison avec ces SNPs au moyen des données de Hapmap obtenues pour les individus d'origine européenne. Un moyen de « resserrer l'étau » autour du SNP fonctionnel serait alors de tester l'association de chacun de ces SNPs avec les taux de vWF et/ou la MTEV dans plusieurs échantillons issus de populations dont la structure de déséquilibre de liaison diffère. Cela permettrait ainsi d'éliminer un certain nombre de SNPs, qui, n'étant plus en déséquilibre de liaison avec le SNP fonctionnel, ne présenteraient alors plus d'association avec vWF.

Il est bien établi qu'on ne peut pas préjuger de la fonctionnalité d'un polymorphisme à partir de sa force d'association avec le caractère étudié. Un SNP non fonctionnel peut en effet montrer une association plus forte et plus significative qu'un SNP fonctionnel du même gène. Il suffit pour cela qu'il soit en déséquilibre de liaison avec un autre SNP fonctionnel, absent du panel de SNP étudié et présentant une association plus importante que le premier. Un autre cas de figure est celui d'un SNP non fonctionnel, en déséquilibre de liaison avec deux polymorphismes fonctionnels du panel. En capturant une part de l'information apportée par chacun, il présente une association plus forte que l'un ou l'autre des SNPs fonctionnels. Blangero *et al* (2005) proposent une méthode de statistique Bayesienne qui, à partir des associations observées entre un panel de SNP et un caractère donné, permet d'établir la probabilité que l'association observée pour un SNP donné soit due à un déséquilibre de liaison avec un autre SNP [159]. Cette démarche est intéressante dans la mesure où elle permet d'établir un ordre de priorité entre les différents SNPs montrant des associations significatives, dans le but de réaliser secondairement des études fonctionnelles. Elle est pertinente à condition que les polymorphismes fonctionnels soient tous présents au sein du panel testé. Ceci implique d'avoir réalisé un séquençage du gène à l'étude, et ne pas se contenter des SNPs répertoriés dans les bases publiques, notamment pour ne pas évincer un

polymorphisme rare. La taille importante de *BAI3* (725 Kb, et plus de 1000 SNPs dont les fréquences alléliques sont connus) n'encourage pas à considérer comme prioritaire cette voie de recherche. Toutefois, cette méthodologie sera peut-être plus attractive dans les prochaines années grâce au développement de techniques de séquençage à haut débit.

Implication potentielle de *BAI3* et de *STAB2* dans le risque de survenue de MTEV *via* une action sur les taux de vWF

Il a été observé dans la littérature un rapport de risque (Hasard Ratio) de MTEV proche de 4 entre le dernier et le premier quartile de vWF [86]. Or, d'après ce que j'ai estimé à partir des données de la cohorte STANISLAS, rs9386864 et rs3798992, bien que significativement associés à vWF, n'expliquent qu'une part infime (1 ou 2%) de la variabilité des taux de vWF. Il en est de même pour la variabilité des taux de vWF expliquée par les polymorphismes de *STAB2*. Il est difficile d'imaginer comment ces mêmes SNPs pourraient entraîner un OR de MTEV aussi important que 2 par leur seule action sur les taux de vWF. Ce phénomène pourrait être la conséquence d'un mécanisme pleiotropique des gènes découverts, qui pourraient moduler, de manière très modeste, les taux plasmatiques de plusieurs protéines intervenant dans la cascade hémostatique, aboutissant finalement à un important excès de risque de survenue de MTEV. Sans la remettre en cause, ces résultats n'illustrent pas l'hypothèse de travail de cette thèse qui était que l'étude d'un phénotype intermédiaire de la MTEV permettrait de découvrir de faibles facteurs de risque de MTEV. En effet, en présence d'un effet pléïotropique, un polymorphisme peut avoir un faible effet sur l'un des phénotypes intermédiaires tout en ayant un effet important sur le risque de MTEV. La découverte d'un tel mécanisme a pu être favorisée par la stratégie de sélection des gènes candidats *via* l'étude *in silico* d'une GWAS de la MTEV.

Une autre explication pourrait être que l'effort de modélisation a été moins important lors des étapes de découvertes des effets des SNPs sur les taux de vWF qu'il ne l'a été lors des étapes de validations, au cours desquelles l'effet sur le risque de MTEV est apparu après diverses stratifications (sur les mutations thrombogènes, sur *ABO*) pour l'étude de *BAI3*, ou après des analyses haplotypiques pour l'étude de *STAB2*. Afin de savoir si l'association d'un SNP avec le risque de MTEV est due uniquement à son action sur les taux de vWF, il serait intéressant de tester s'il reste une association résiduelle après ajustement sur les taux de vWF, mais je ne disposais pas d'une étude cas-témoins dont les traits quantitatifs étaient connus pour les cas et les témoins.

L'hypothèse d'un rôle de *BAI3* dans la modulation des taux de vWF ou FVIII est peu étayée par des connaissances biologiques. *BAI3* (Brain Angiogenesis Inhibitor 3) appartient à une famille de récepteurs transmembranaires, et présente de fortes similitudes avec *BAI1* et *BAI2* [160][161]. Le rôle de cette protéine dans la physiologie humaine est très peu décrit. Exprimée dans le cerveau durant l'embryogenèse, on suspecte qu'elle intervient dans des phénomènes d'ischémie induits par l'angiogenèse [161]. D'après une étude récente, elle pourrait être associée à la schizophrénie [162]. On peut peut-être établir un lien avec *ADAMTS13*, une protéase qui clive vWF, et qui est associée au risque d'accident vasculaire cérébral ischémique [163][164][165]. Ces diverses informations ne permettent pas d'émettre d'hypothèse quant aux mécanismes physiopathologiques par lesquels certains polymorphismes de *BAI3* pourraient favoriser la survenue d'une MTEV.

Quant à *STAB2*, il code pour un récepteur transmembranaire qui pourrait jouer un rôle dans l'angiogénèse [166]. Il est exprimé par les cellules endothéliales du foie, lieu du catabolisme du FVIII et du vWF. De façon très intéressante, il code pour un récepteur *scavenger* (ou « éboueur ») qui reconnaît les LDL, à l'instar de *LDLR* et *LRP1*. Ces derniers sont les seuls gènes (en dehors de *ABO* et des gènes structuraux de FVIII et vWF) qui jusqu'à présent étaient connus pour influencer l'un ou l'autre des phénotypes à l'étude dans ce travail. De plus, le récepteur codé par *STAB2* semble capable de reconnaître les protéines présentant des résidus sucrés complexes [167], argument supplémentaire en faveur d'un rôle fonctionnel de *STAB2*.

Les associations découvertes sont-elles responsables des signaux de liaison obtenus initialement dans les Familles-FVL ?

Si l'objectif des analyses de liaison est simplement de localiser une région chromosomique d'intérêt, il est classique de s'interroger sur l'origine du signal de liaison. Quels polymorphismes, en raison de leurs associations au phénotype, ont généré le signal de liaison ? Peut-on considérer que les signaux de liaison en 6q13-14 et 12q23 sont entièrement expliqués par les gènes nouvellement découverts (respectivement *BAI3* et *STAB2*) ? Faut-il continuer à explorer ces régions ?

Il existe plusieurs arguments invalidant l'hypothèse que des polymorphismes de *BAI3* puissent être à l'origine du signal de liaison dans les familles-FVL. Tout d'abord, rappelons que les polymorphismes de *BAI3* semblent agir sur les taux de FVIII seulement de manière indirecte. En effet, dans l'échantillon STANILAS , ils s'observent nettement sur les taux de

vWF, de manière atténuée sur les taux de FVIII, et disparaissent après ajustement sur les taux de vWF. Or, le signal de liaison ne s'observe que pour l'étude des taux de FVIII ajustés sur vWF. De plus, il apparaît, dans les échantillons MARTHA05 et FARIVE, que *BAI3* influence le risque de MTEV des non-porteurs de mutation FVL ou FII exclusivement. Or, justement, l'échantillon Famille-FVL est enrichi en mutation FVL. Ces diverses incohérences n'excluent pas de manière formelle le rôle de *BAI3* dans le signal de liaison en 6q13-14. Elles imposent néanmoins, si on admet la réalité d'une implication de *BAI3*, d'envisager des mécanismes physiopathologiques particulièrement complexes.

Plus tardivement dans ma thèse, j'ai tenté d'établir d'une manière assez objective la responsabilité de *BAI3* dans la genèse du signal de liaison. J'avais en effet la possibilité, grâce à des données issues d'une biopuce, de réanalyser la liaison avec les marqueurs du chromosome 6 des taux de FVIII ajustés sur les taux de vWF et sur chacun des 200 polymorphismes de *BAI3* de la biopuce. Mon hypothèse était que si j'ajustais sur un polymorphisme à l'origine du signal de liaison alors ce dernier disparaîtrait entièrement. Les BF maximaux obtenus dans la région oscillaient tous autour du BF maximal obtenu sans ajustement sur les SNPs (entre 25 et 45). Ainsi, j'échouai dans ma tentative de mettre en évidence un polymorphisme de *BAI3* responsable du signal de liaison. Cette démarche était cependant assez grossière. En effet, elle ne permettait pas de modéliser l'accumulation de faibles effets apportés par plusieurs polymorphismes, ni encore un éventuel effet haplotypique. De plus, il est possible que les signaux de liaison aient été générés par un polymorphisme très rare absent de la puce. En effet, le mode de recrutement des familles-FVL *via* un cas de MTEV avec présence d'une mutation FVL avait pour but de limiter l'hétérogénéité génétique au sein de l'échantillon. En contrepartie, la constitution d'un échantillon composé d'un petit nombre de grandes familles très sélectionnées peut favoriser l'enrichissement dans l'échantillon de polymorphismes rares, et rendre malaisée la réplication des résultats en population générale.

Je n'ai pas eu le temps de réaliser d'analyses poussées avec les polymorphismes de *STAB2*, notamment en vue d'établir leur responsabilité dans la genèse du signal de liaison en 12q23. A la lumière de mes travaux, de ceux de l'ensemble de l'équipe de David Trégouët et Pierre Morange, et de ceux du consortium CHARGE, les associations obtenues avec les polymorphismes de *STAB2* semblent robustes. Il est donc fort probable qu'elles soient à l'origine du signal de liaison à l'origine de ces travaux. Comme discuté précédemment, nous ne sommes vraisemblablement pas en présence des SNPs responsables biologiquement des

modulations de vWF. L'observation d'un effet porté par un haplotype particulier [Germain M] plaide en faveur de la présence d'un allèle fonctionnel rare. Ceci est cohérent avec le mode de recrutement de l'échantillon-FVL comme cela a déjà été signalé au paragraphe précédent. L'ensemble de ces résultats est suffisamment convaincant pour justifier un travail de recherche précis sur les différents polymorphismes de *STAB2*, impliquant en particulier le séquençage de ce gène et une étude du déséquilibre de liaison dans plusieurs échantillons.

Conclusion

Finalement, cette étude a permis de mettre en lumière *STAB2* et *BAI3*, qui pourraient être des gènes de susceptibilité à la MTEV, dont le mécanisme d'action passerait au moins en partie par la modulation des taux de vWF. Nous avons été aiguillés vers ces gènes par la confrontation d'une analyse de liaison sur les taux de vWF et FVIII, et d'une étude pangénomique cas-témoin sur la MTEV. On aurait pu imaginer une stratégie plus directe. En effet, le génotypage de manière intensive de SNPs situés dans l'un des signaux de liaison aurait permis d'étudier directement leur association avec les taux de vWF et de FVIII. La priorité aurait été donnée au criblage du signal de liaison en 12q13 qui présentait le BF le plus élevé, et qui expliquait la plus grande part de variabilité des TQs, tout en étant le plus étroit. Afin de couvrir efficacement les signaux de liaison, le génotypage de plusieurs centaines de SNPs, préalablement sélectionnés de manière à intégrer le maximum de SNPs situés dans les exons, non synonymes, et en faible déséquilibre de liaison avec les SNPs déjà sélectionnés, aurait certainement été nécessaire. Un tel génotypage est particulièrement coûteux, et s'est vu détrôné par l'avènement du génotypage à haut débit, réalisable grâce à la technologie des biopuces. En effet, celui-ci permet, pour un prix comparable, le génotypage d'environ 500 000 SNPs couvrant l'intégralité du génome. Ainsi, non seulement il permet d'explorer les signaux de liaison, mais aussi, il apporte des informations sur le reste du génome.

VIII. Identification de nouveaux déterminants génétiques des taux de vWF et FVIII par des analyses d'association pangénomique

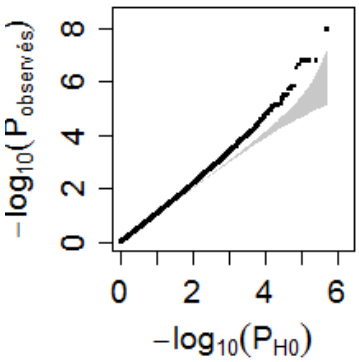
Au moment où j'achevais l'exploitation des résultats des analyses de liaison menées sur les taux de vWF et FVIII, France Gagnon recevait le génotypage de l'échantillon Familles-FVL avec la puce Illumina 660-Quad. Ce chapitre retrace le travail que j'ai mené à partir de ces nouvelles données dans le but d'identifier de nouveaux déterminants génétiques des taux de vWF et FVIII. J'ai analysé les données de l'échantillon Familles-FVL au moyen d'une méthode par décomposition de la variance avec prise en compte des corrélations intra-familiales (voir **§VI.2 p56**). Parallèlement à ces analyses, Tiphaine Oudot-Mellakh, chercheuse postdoctorale à l'UMRS-937, menait des analyses d'association pangénomique sur les données des échantillons MARTHA08 et MARTHA10. Pour rendre comparables les résultats obtenus sur ces trois échantillons, nous avons appliqué aux taux de vWF et FVIII une transformation par quantile de loi normale. Les analyses étaient systématiquement ajustées sur le sexe et l'âge. Deux approches stratégiques ont ensuite été adoptées et sont présentées ici. Dans un premier temps, j'ai recherché si les associations les plus significatives obtenues dans les Familles-FVL s'observaient aussi dans MARTHA08 et MARTHA10. Dans un second temps, j'ai réalisé une méta-analyse des résultats obtenus à partir des trois échantillons.

VIII.1. Analyse d'association pangénomique de l'échantillon Familles-FVL

VIII.1.1. Analyse des taux plasmatiques de vWF

Les associations avec le taux de vWF de 490 083 SNPs répondant à nos critères de qualité (voir **tableau 5 p38**) ont été testées dans l'échantillon Familles-FVL. Les résultats des tests sont représentés sous forme d'un Quantile-Quantile plot (QQ plot) (**figure 14**) des valeurs p. Les p observées s'écartaient des p attendues sous l'hypothèse nulle d'absence d'association. Le facteur d'inflation λ_{50} calculé au 50ème percentile était de 1,09.

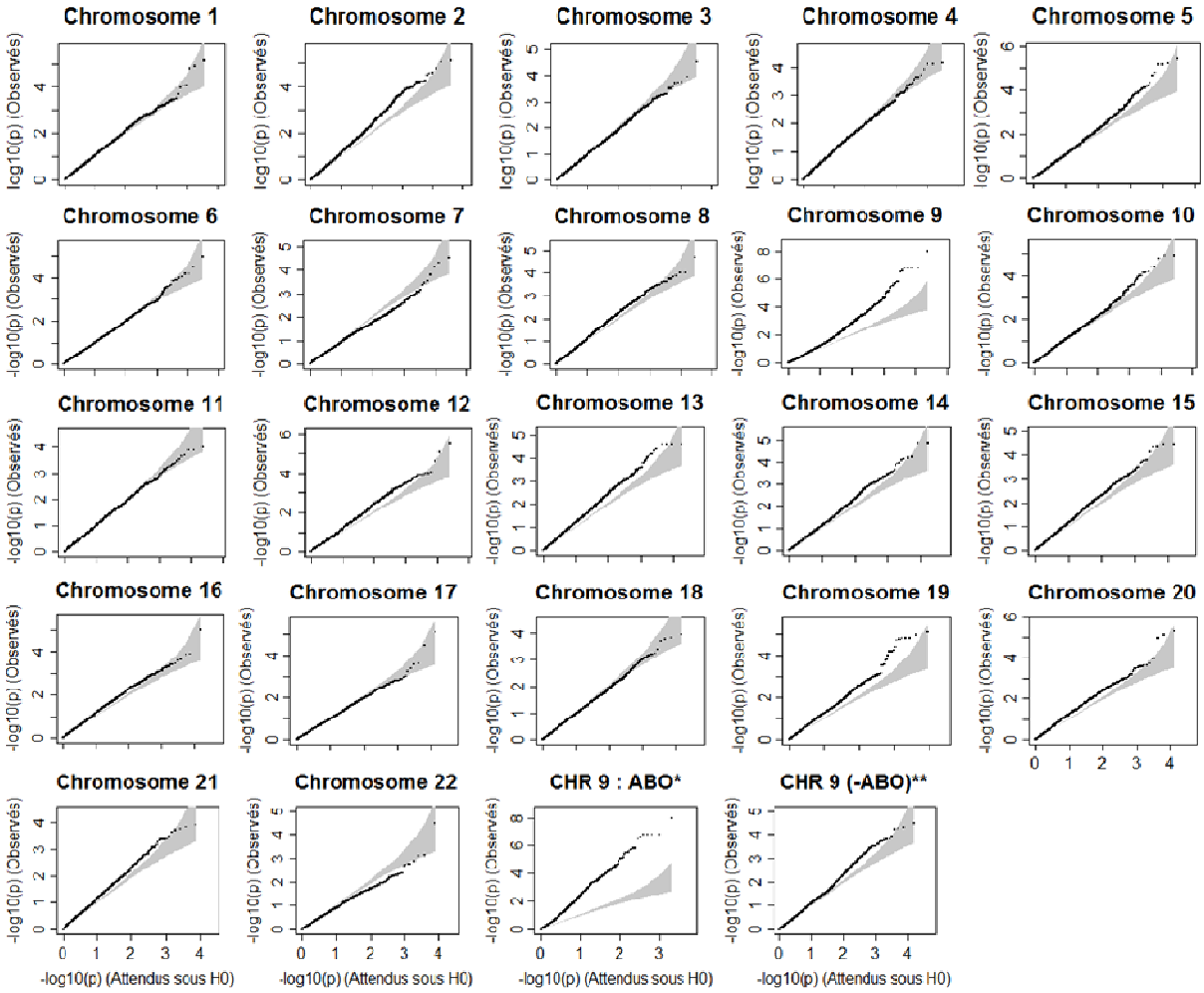
Figure 14. QQ-Plot de l'analyse d'association pangénomique des taux de vWF (Familles-FVL)



Quantiles des valeurs de p observées en fonction des quantiles de valeurs de p suivant une loi uniforme U(0,1), correspondant à l'hypothèse nulle d'absence d'association. L'intervalle de confiance des quantiles attendus lorsque tous les tests sont réalisés sous l'hypothèse nulle est grisé. Sachant que l'immense majorité des tests est effectuée sous H0, on s'attend à ce que le QQ plot suive la bissectrice, et ne s'en écarte éventuellement que pour les p-values proches de 0 (donc les grandes valeurs de -log(p)). Le facteur d'inflation calculé au 50ème percentile est de 1,09.

Une inflation systématique de la statistique de test utilisée était donc à craindre. Cependant, l'écart entre les valeurs de p observées et attendues n'était sensible qu'à partir de $-log_{10}(p)=2$. Il ne concernait donc pas les 99% plus grandes valeurs de p. Par ailleurs, la réalisation de QQ plots chromosome par chromosome a mis en évidence que seuls étaient concernés par ce phénomène certains chromosomes, particulièrement le chromosome 9. De manière plus précise, la réalisation d'un QQ plot à partir des valeurs de p observées dans une fenêtre de 10 Mb autour du gène *ABO* (1996 SNPs) montrait une amplification du phénomène : l'écart aux valeurs de p attendues apparaissait dès les plus grandes valeurs de p (proches de 1). A l'inverse, l'ensemble du chromosome 9 à l'exclusion de cette fenêtre ne montrait pas d'inflation pour les 99% plus grandes valeurs de p (**figure 15**).

Figure 15. QQ-plots des analyses des taux plasmatiques de vWF réalisées chromosome par chromosome

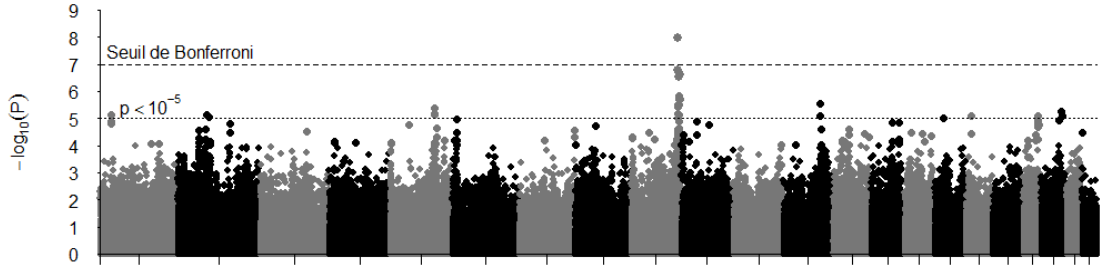


* QQ-plot réalisé avec 1996 SNPs situés dans une fenêtre de 10 Mb autour du gène *ABO*.
** QQ-plot réalisé avec l'intégralité du chromosome 9, à l'exclusion des 1996 SNPs proches du gène *ABO*.

L'interprétation à donner à ces observations n'est donc probablement pas une inflation systématique du test utilisé. Une hypothèse est que la constitution d'un échantillon de grandes familles induit de forts déséquilibres de liaison à travers le génome. Toute association se traduit donc par un certain nombre de tests significatifs, non indépendants. Ainsi, l'ensemble de la région peut montrer des valeurs de p s'écartant de leur distribution attendue sous H0. Si cette interprétation est bonne, il n'est alors pas nécessaire de corriger cette inflation. Elle est la conséquence d'un ensemble d'associations bien réelles, qui sont le reflet de l'importance du déséquilibre de liaison de notre échantillon.

L'analyse pangénomique des taux de vWF a révélé un unique SNP, rs17553234, dont l'allèle G était associé à une élévation des taux de vWF avec une significativité ($p=1,11.10^{-8}$) passant le seuil de significativité de Bonferroni ($p<1,02.10^{-7}$) (**figure 16**). Il était toujours significatif après correction de la statistique par $\lambda_{50}=1.09$ ($p=4,38.10^{-8}$).

Figure 16. Analyse pangénomique des taux de vWF (Familles-FVL)

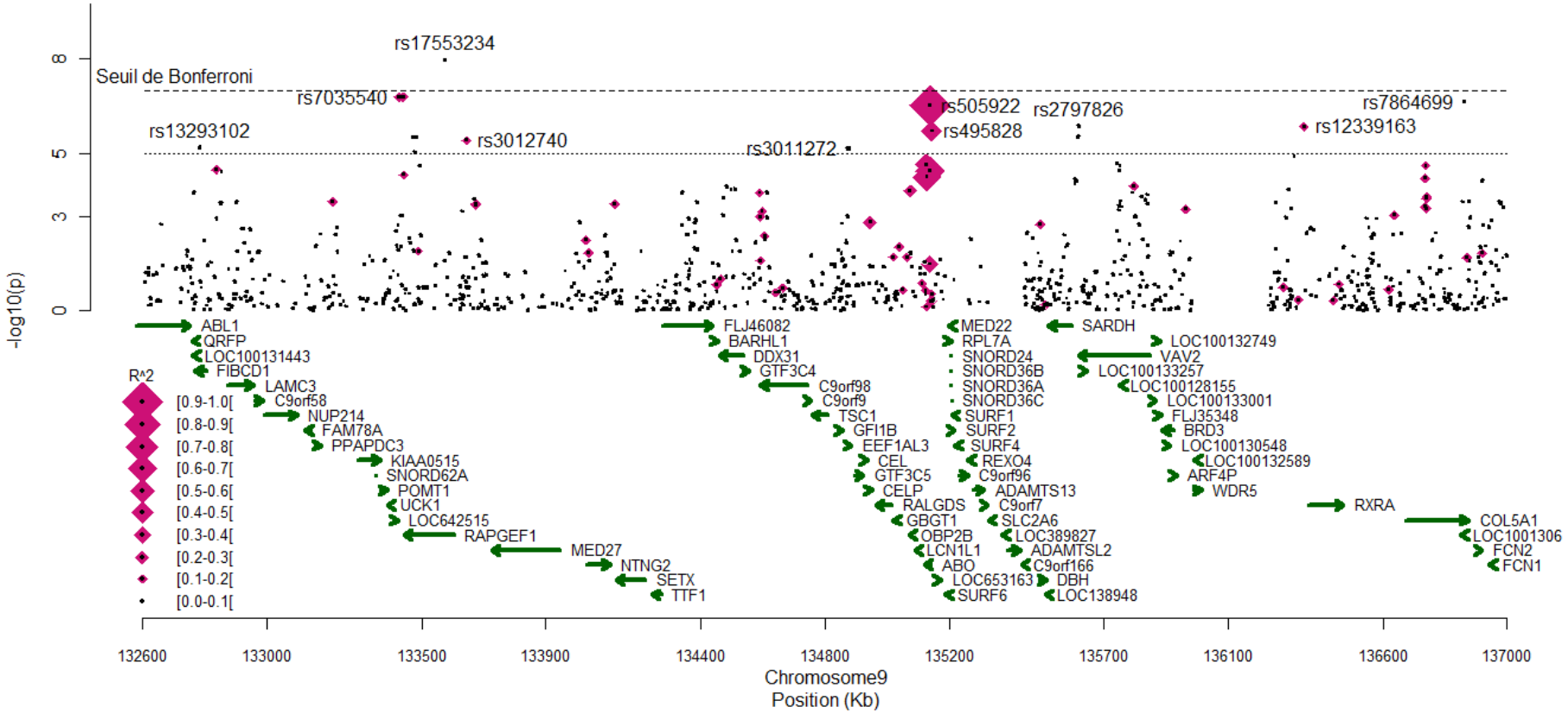


Les significativités p des tests d'association de chaque SNP sont représentées en fonction de la position de ces derniers. Le seuil de Bonferroni est égal à $1,02.10^{-7}$

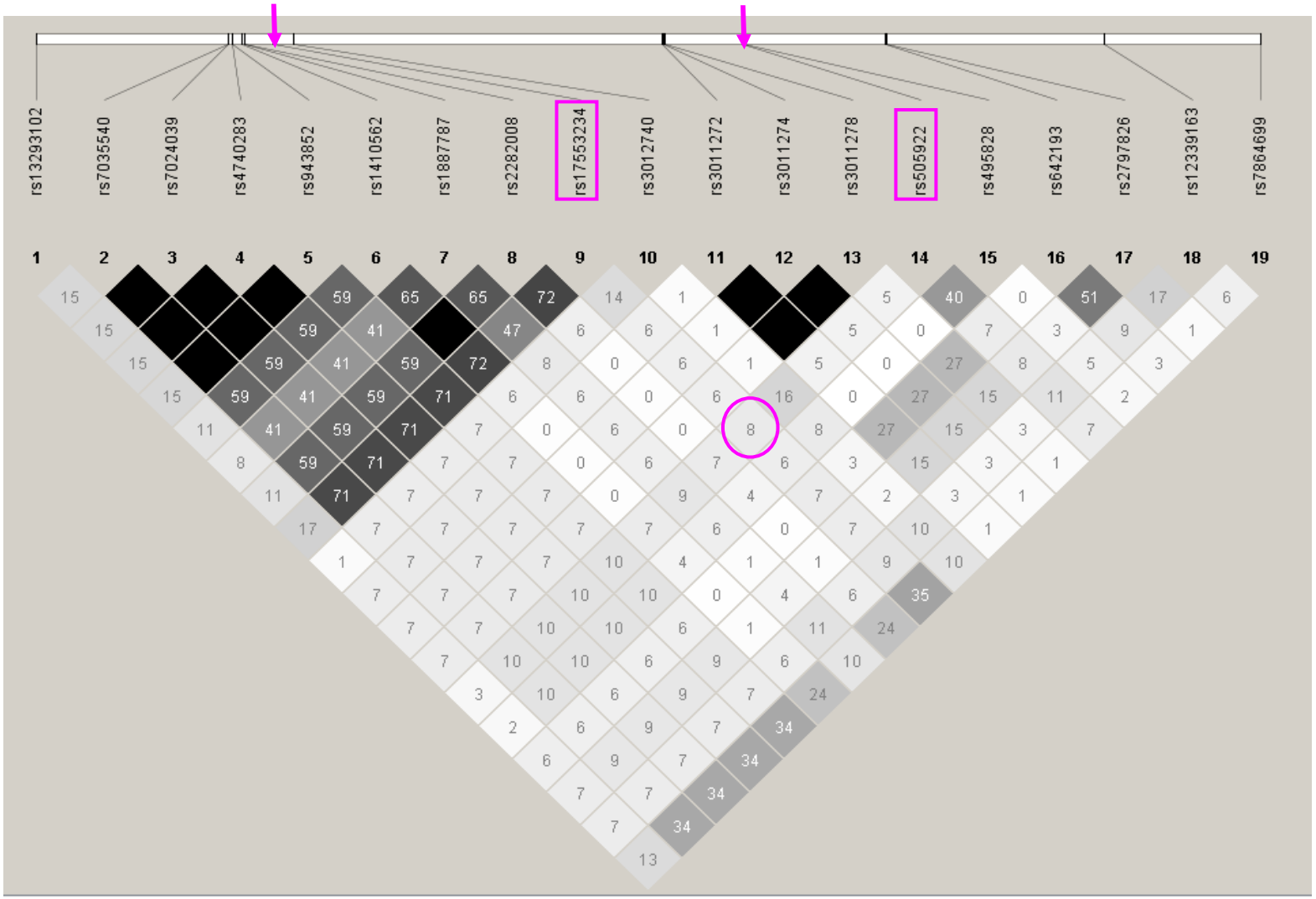
Ce SNP était situé en 9q34, dans un intron du gène *RAPGEF1*. Un ensemble de 20 SNPs de cette même région était associé aux taux de vWF à $p<10^{-5}$. Cet ensemble s'étendait sur 4,4 MB et une cinquantaine de gènes. Il était centré autour du gène *ABO* (**figure 17.A**). L'étendue de ce signal d'association ne semblait pas pouvoir s'expliquer directement par un fort déséquilibre de liaison ($r^2>0.80$) sur de longues distances, notamment entre rs17553234 et un polymorphisme du gène *ABO*. En particulier, la valeur du r^2 entre rs17553234 et le SNP le plus significatif du gène *ABO* (rs505922, $p=3,04.10^{-7}$) n'était que de 0,08 (**figure 17.B**). Ailleurs dans le génome, aucun SNP n'était associé aux taux de vWF avec une significativité en-deçà du seuil de Bonferroni.

Figure 17. Association entre les polymorphismes du locus ABO et les taux plasmatiques de vWF (échantillon Familles-FVL)

16.A : les p-values obtenues autour du gène *ABO* sont représentées en fonction de la position du SNP. La taille des losanges rouges correspond à la valeur r^2 du déséquilibre de liaison entre rs505922 et les autres SNPs de la région. Rs505922 est le SNP du gène ABO présentant l'association la plus significative. L'allèle A du rs505922 caractèrise le groupe sanguin O. Le seuil de Bonferroni est égal à $1{,}02.10^{-7}$



16.B



Matrice (« demi-matrice ») du déséquilibre de liaison, mesuré par la valeur r^2, entre tous les SNPs de la région d'*ABO* présentant des associations dont la significativité p était inférieure à 10^{-5}. La position des SNPs est à la même échelle que sur la figure A. Les deux SNPs encadrés sont *RAPGEF1*-rs17553234, le seul SNP dont la significativité était inférieure au seuil de Bonferroni, et *ABO*-rs505922, le SNP d'*ABO* dont l'association avec les taux de vWF était la plus significative.

J'ai étudié plus finement le déséquilibre de liaison entre rs17553234 (le SNP le plus significatif de notre analyse pangénomique) et le gène *ABO*, afin de savoir si l'élévation des taux de vWF observée avec rs17553234-G pouvait s'expliquer par un déséquilibre de liaison avec le gène *ABO*, dont les allèles A1 et B sont connus pour être associés à des taux élevés de vWF. Le faible r^2 entre rs17553234 et rs505922 était un premier élément (négatif) de réponse, mais insuffisant. En effet, la région dans son ensemble présentait un déséquilibre de liaison important (voir **annexe pA17 et A20**). La présence de valeurs de D'=1 entre SNPs très éloignés incitait à regarder de près les structures haplotypiques. J'ai estimé les fréquences des haplotypes constitués par rs17553234 et les trois SNPs « taguant » les allèles A1, A2, B, O (**tableau 21**). Supposons que rs17553234-G soit systématiquement associé à A1 et B et qur rs17553234-A soit systématiquement associé à A2 et O. Alors l'association observée avec rs17553234 serait alors plus significative qu'avec n'importe lequel des allèles taguant *ABO*, tout en n'étant que le fruit du déséquilibre de liaison. Cette situation ne semble pas être le cas ici même si une tendance en ce sens est observée : les allèles A2 et O étaient associés à l'allèle rs17553234-A, de manière systématique pour A2 et quasi-systématique pour O. L'allèle A1 était associé dans 2/3 des cas à rs17553234-A et dans 1/3 des cas à rs17553234-G. Les probabilités de l'allèle rs17553234-G conditionnellement aux allèles d'*ABO,* P(G| A1), P(G| A2) et P(G| O), étaient respectivement 33%, 0% et 6%.

Tableau 21: Haplotypes constitués des SNPs taguant *ABO* et de rs17553234

Allèle ABO	Tags d'*ABO*			*RAPGEF1*	Fréquence
	rs8176704	rs8176746	rs505922	rs17553234	haplotypique
A1	G	C	G	A	20%
	G	C	G	G	10%
A2	**A**	C	G	A	7%
B	G	**A**	G	A	<1%
O	G	C	**A**	A	58%
	G	C	**A**	G	4%

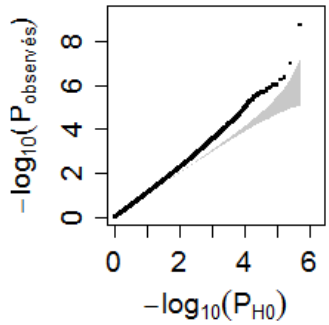
Un moyen de savoir si l'association observée avec rs17553234 était, au moins en partie, le reflet de l'effet du gène *ABO*, ou si *a contrario* elle existait indépendamment de ce dernier, est de réaliser une analyse haplotypique. Cependant, la struture particulièrement complexe de l'échantillon Familles-FVL était une limite à cette approche. Une alternative était de réaliser une analyse ajustée sur les SNPs taguant *ABO*. J'ai donc analysé l'intégralité de la puce avec un ajustement sur rs8176704, rs8176746, et rs505922. L'objectif de cette analyse était double. Il s'agissait, d'une part, de s'affranchir d'un éventuel facteur de confusion du gène ABO, et d'autre part, d'augmenter la puissance des tests des associations en diminuant la variabilité résiduelle de vWF.

VIII.1.2 Analyse des taux de vWF ajusté ssur le gène *ABO*

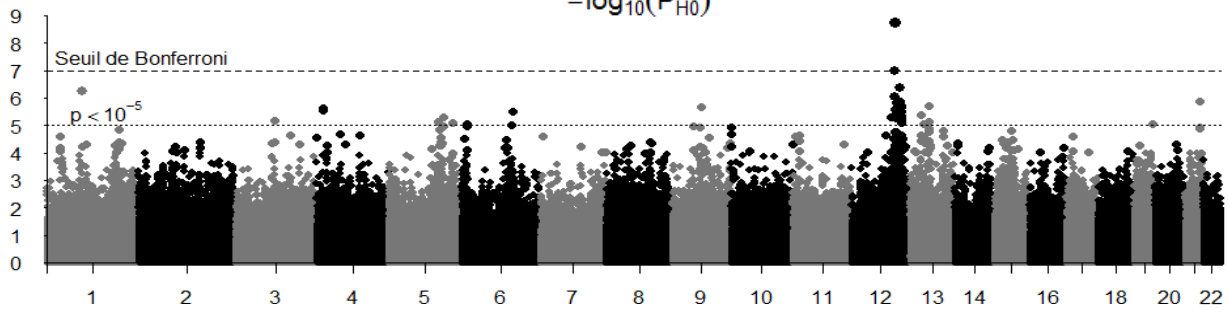
L'inflation du QQ plot à partir des valeurs de p inférieures à 0,01 était similaire à celle du QQ plot précédent. Le facteur d'inflation λ_{50} calculé au 50ème percentile était de 1,14. Cependant, la plupart des chromosomes était, là encore, entièrement exempte de toute inflation. Par conséquent, je n'ai pas appliqué de correction des p-values, pour les raisons évoquées précédemment. L'ensemble des résultats est présenté en **figure 18**.

Figure 18. Analyse pangénomique des taux de vWF ajustés sur le gène *ABO* (Familles-FVL)

A.



B

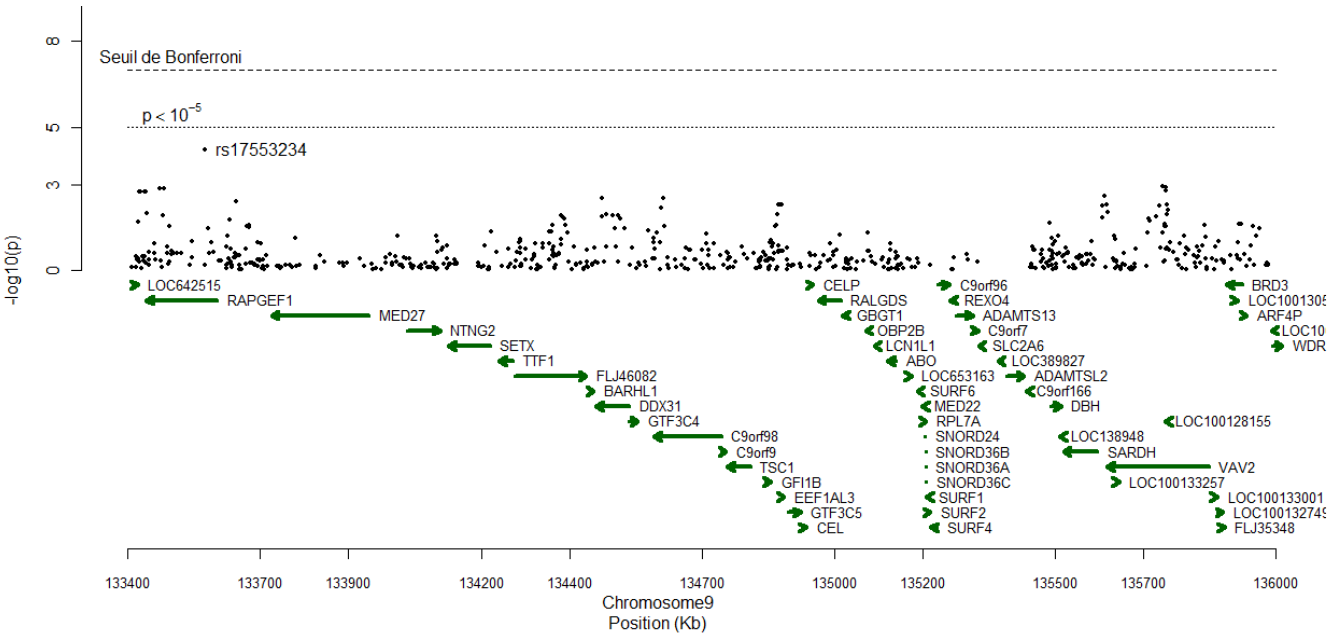


A. QQ-plot des valeurs p montrant un facteur d'inflation calculé au 50ème percentile égale à de 1,14

B. Les significativités p des tests d'association de chaque SNP sont représentée en fonction de la position de ces derniers. Le seuil de Bonferroni est égal à 1,02 10^{-7}

Après prise en compte du groupe *ABO*, les associations avec l'ensemble des SNPs du gène *ABO* ne se distinguaient plus du bruit de fond. L'association observée avec rs17553234 du gène *RAPGEF1* était quant à elle amoindrie, mais toujours assez significative (p=6,3.10^{-5}) (**figure 19**). Parmi l'ensemble des résultats de cette analyse pangénomique, 164 associations présentaient des valeurs p<6,3.10^{-5}. Elles se répartissaient de manière homogène à travers le génome. Le FDR correspondant à cette valeur de p était ~20%. La découverte de rs17553234 est probablement due à la conjonction de deux phénomènes : d'une part, une faible association avec vWF, compatible avec une fluctuation d'échantillonnage, et d'autre part, un déséquilibre de liaison, certes modéré, mais néanmoins suffisamment influent, avec le gène *ABO*.

Figure 19. Association entre les polymorphismes du locus *ABO* et les taux plasmatiques de vWF ajustés sur le groupe sanguin ABO (échantillon Familles-FVL)

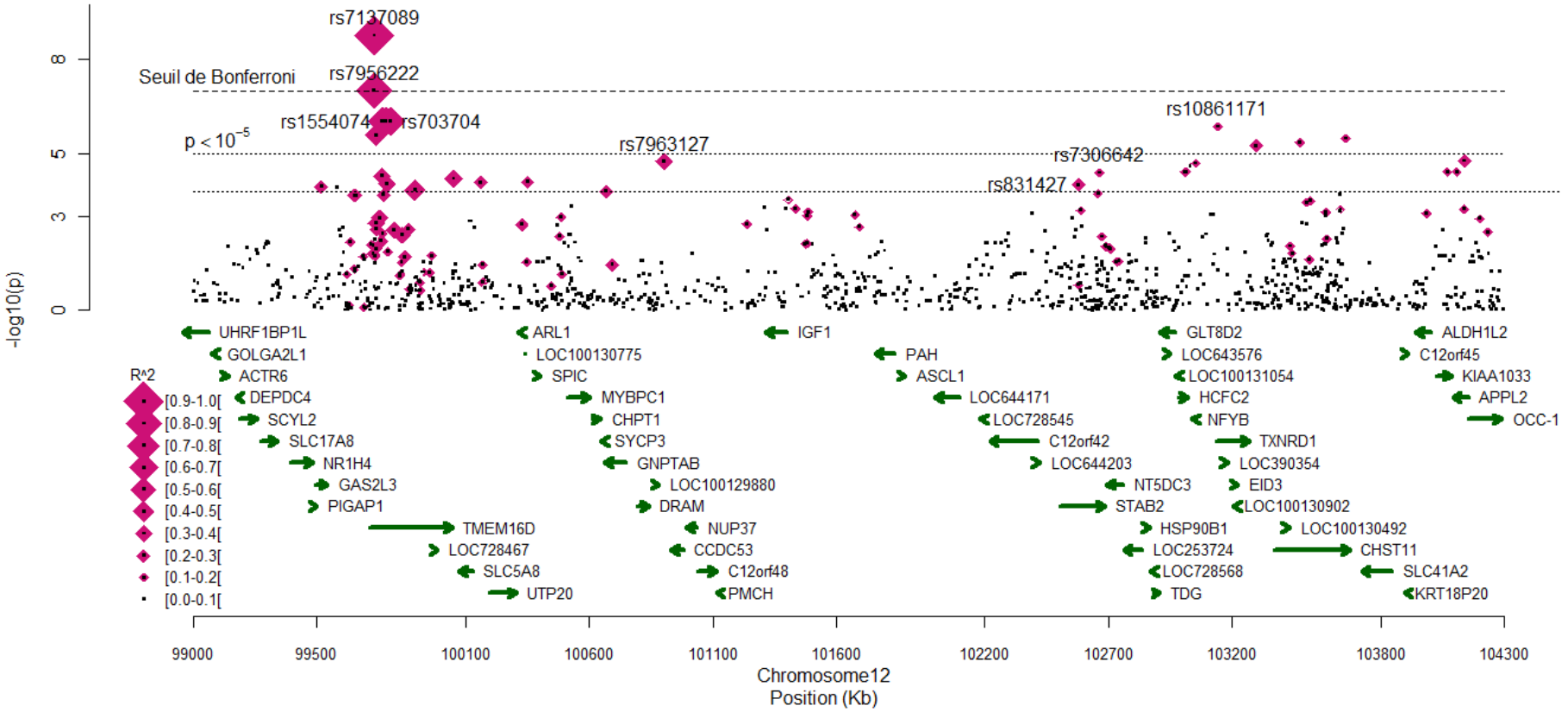


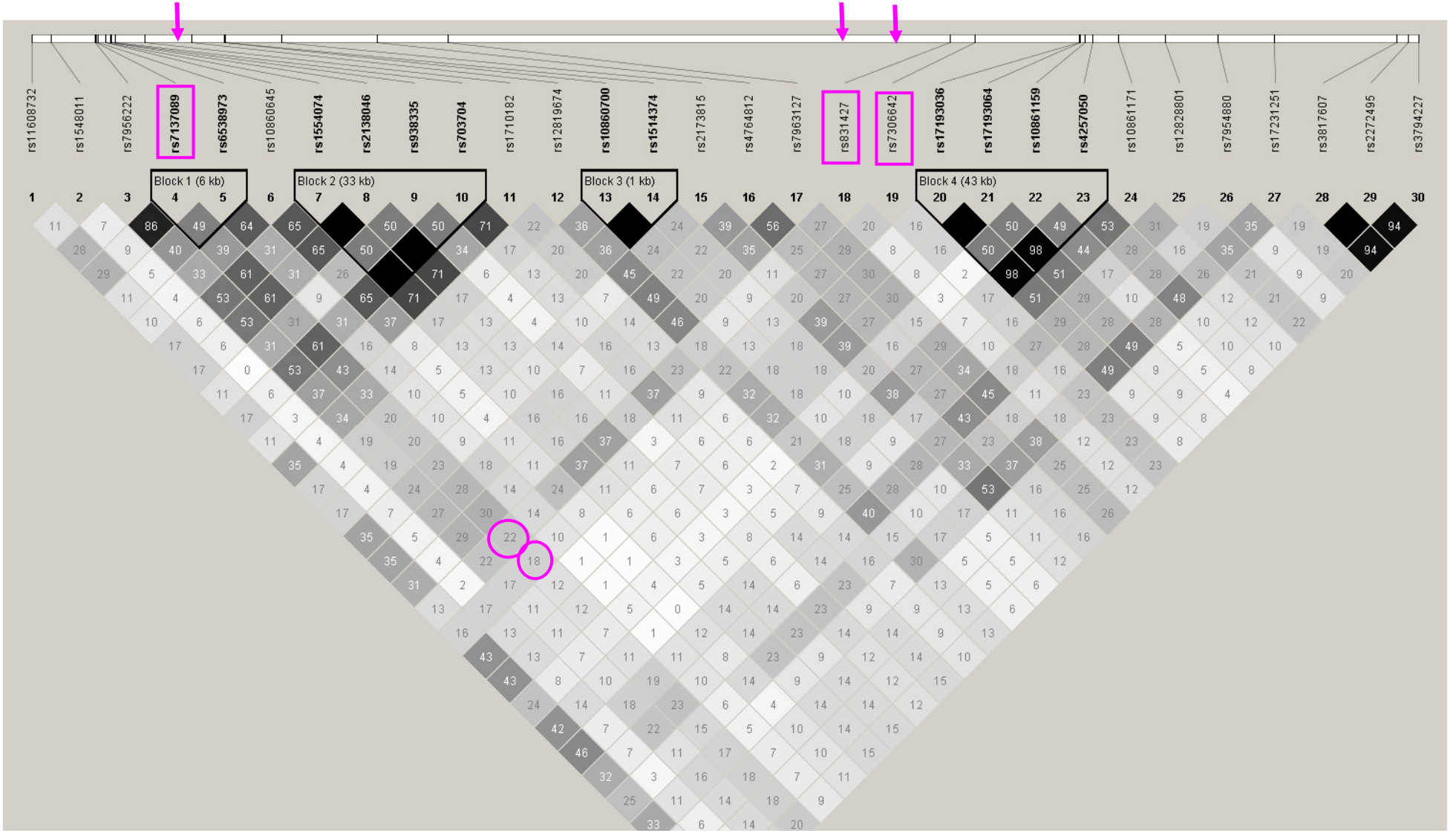
Les significativités p des tests d'association de chaque SNP sont représentées en fonction de la position de ces derniers. Après ajustement sur les trois SNP taguant le groupe ABO, on n'observe aucun signal d'association au niveau du gène ABO, tandis que l'association observée avec RAPGEF1-rs17553234 n'était plus significative.

L'ajustement sur les trois SNPs taguant le groupe ABO a permis, par ailleurs, de faire apparaître une association dépassant le seuil de Bonferroni. En effet, l'allèle rare rs7137089-T, situé dans un intron du gène *TMEM16D* en 12q23, était associé à une diminution des taux de vWF ($p = 1{,}8.10^{-9}$). Ce gène est situé à un peu plus d'une mégabase du gène *STAB2*. Ce dernier se situait dans le signal de liaison des taux de vWF ajusté sur ABO (voir **figure 13.D p83**) et présentait des polymorphismes associés au risque de MTEV (voir **tableau 9 p84** et **§VII.5** p94). Au sein de *STAB2*, les polymorphismes les plus fortement associés aux taux de vWF étaient rs831427 ($p = 1{,}0.10^{-4}$) et rs7306642 ($p = 4{,}1.10^{-5}$) (**figure 20A**). Ces deux polymorphismes étaient en faible déséquilibre de liaison avec le polymorphisme rs7137089 de *TMEM16D* : $r^2 = 0{,}22$ et $r^2 = 0{,}18$, respectivement pour rs831427 et rs7306642 (**figure 20B**)

Figure 20. Association entre les polymorphismes du locus 12q23 et les taux plasmatiques de vWF ajustés sur le gène *ABO* (échantillon Familles-FVL)

A : les valeurs de p obtenues en 12q13 sont représentées en fonction de la position du SNP. La taille des losanges rouges correspond à la valeur r^2 du déséquilibre de liaison entre rs7137089 et les autres SNPs de la région. Rs7137089 est le SNP de la région présentant l'association la plus significative. Il est situé au sein du gène *TMEM16D*. Les deux lignes horizontales supérieures correspondent au seuil de Bonferroni (p= $1{,}02.10^{-7}$) et à p=10^{-5}. La ligne horizontale inférieure délimite les SNPs présentant les 30 plus petites valeurs de p de la région (p <$1{,}64\ 10^{-4}$)





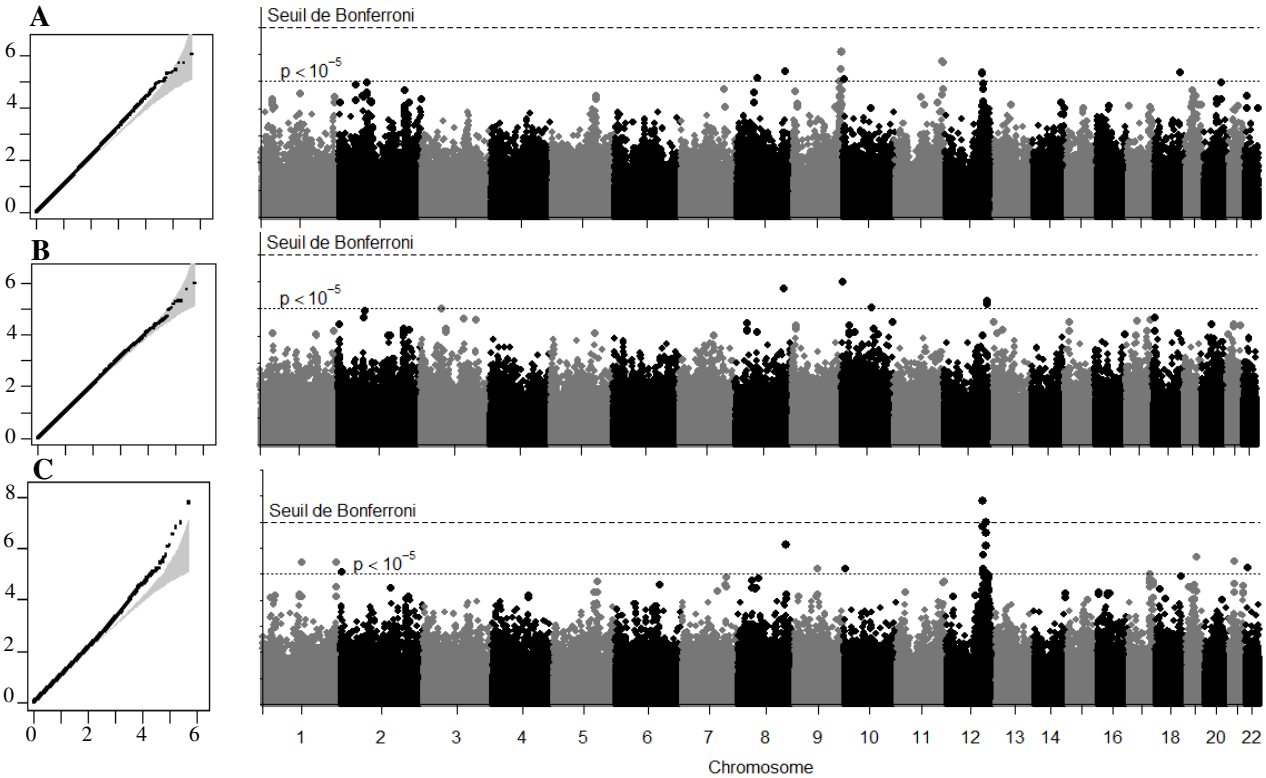
B : Matrice (« demi-matrice ») du déséquilibre de liaison, mesuré par la valeur r^2, entre les SNPs de la région présentant les 30 plus petites valeurs de p de la région. Ce choix, arbitraire, est fondé sur un compromis entre l'étendue de l'information et la lisibilité de la matrice. Pour des raisons de lisibilité, la position des SNPs des figures A et B ne sont pas exactement à la même échelle. Les trois SNPs encadrés en rose sont *TMEM16D*-rs7137089, *STAB2*-rs831427 et *STAB2*-rs7306642

VIII.1.3. Analyse des taux plasmatiques de FVIII

Les trois analyses d'association génome-entier que j'ai réalisées sur les taux de FVIII, puis sur les taux de FVIII ajustés sur les taux de vWF, et enfin sur les taux de FVIII ajustés sur le groupe sanguin ABO n'ont pas apporté d'éléments nouveaux à ma recherche.

Les deux premières analyses n'ont révélé aucune association significative au seuil de Bonferroni. L'analyse des taux de FVIII non ajustés sur les taux de vWF révèle tout au plus quelques valeurs de p se détachant légèrement du bruit de fond. L'association la plus significative était observée avec rs505922 qui tague le groupe O ($p=3,4.10^{-6}$) (**figure 21.A**). Il n'y avait aucune association notable après ajustement de FVIII sur vWF, notamment dans le gène *ABO* (**figure 21.B**). De façon analogue à ce qui était observé pour l'analyse de vWF, l'ajustement sur le groupe sanguin ABO a révélé un signal d'association au niveau du locus 12q23 (**figure 21.C**). Deux SNPs présentaient des associations significatives. Le premier était là encore rs713089 dans le gène *TMEM16D* ($p=1,63.10^{-8}$). Le deuxième, à envrion 10 Mb du premier, était rs11615047 dans le gène *PTPN11* ($p=9,90.10^{-8}$) (**figure 22**). Les valeurs de r^2 et D' entre ces deux SNPs étaient modérées (respectivement 0.19 et 0.45), mais cependant étonnamment importantes compte-tenu de leur distance physique.

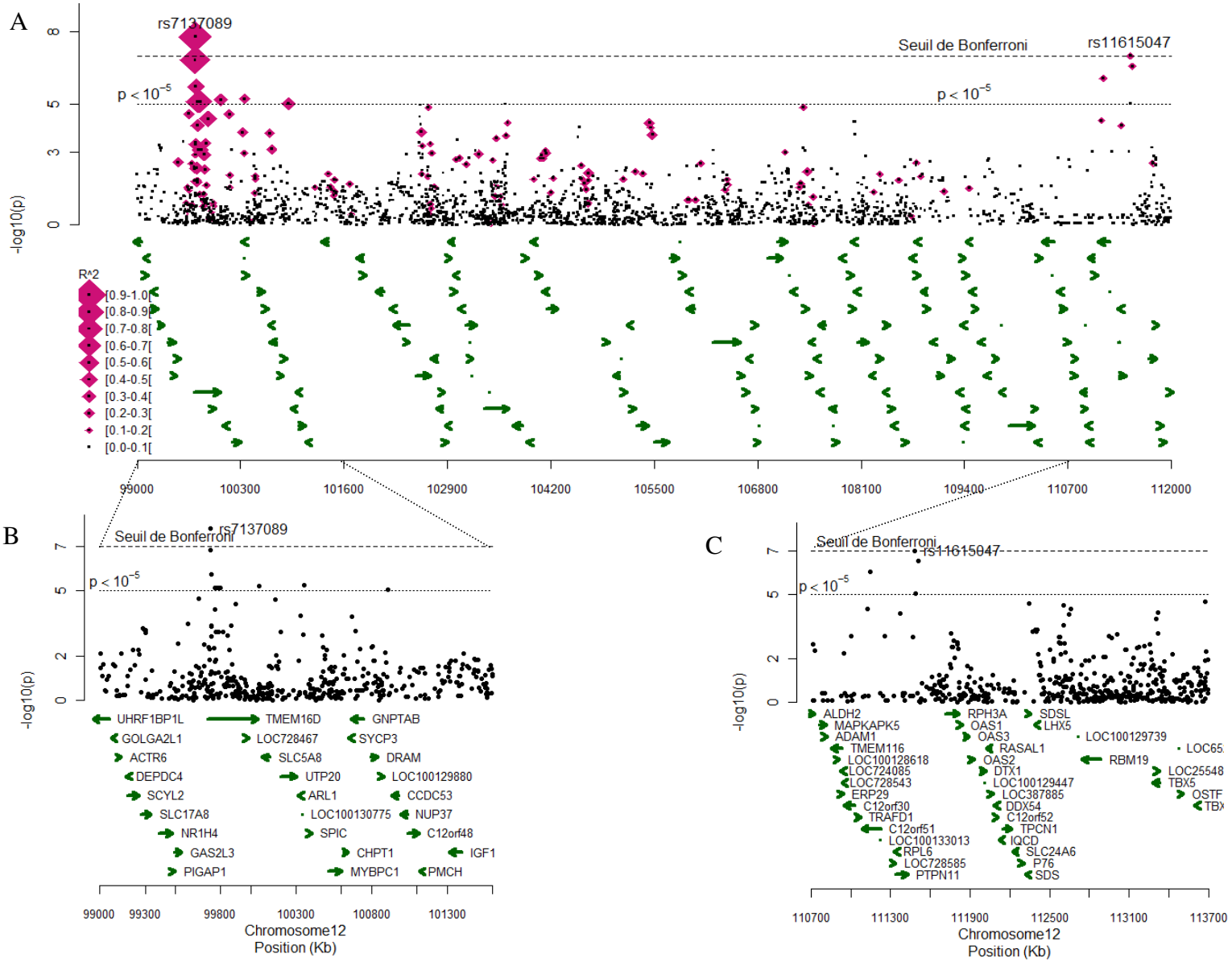
Figure 21. Analyses pangénomiques des taux de FVIII (Familles-FVL)



A. FVIII. Le QQ-plot montre un facteur d'inflation λ_{50} à 1,07. Les associations les plus significatives sont en 9q34**.**
B. FVIII ajusté sur vWF (λ_{50}=1,06). La région 9q34 ne présente plus d'associations notables.
C. FVIII ajusté sur ABO (λ_{50}=1,08). De nouvelles assocaitions sont révélées en 12q23.

Le seuil de Bonferroni est égal à $1,02.10^{-7}$

Figure 22. Association entre les polymorphismes du locus 12q23 et les taux plasmatiques de FVIII ajustés sur le gène *ABO* (échantillon Familles-FVL)



A. Deux SNPs, à environ 10 Mb l'un de l'autre, sont significativement associés aux taux de FVIII ajustés sur *ABO*.
B. Le premier est situé dans le gène *TMEM16D*. **C.** Le deuxième est situé dans le gène *PTPN11*.

Résultats présentés en annexes

1) Les associations se détachant du bruit de fond général ($p< 10^{-5}$) sans nécessairement passer le seuil de Bonferroni sont présentées **en annexe pA28-A31** pour chacun des phénotypes étudiés. Le taux de faux positifs attendus (FDR) est relativement élevé. Il est de 13% parmi les SNPs associés au taux de WF (mais s'élève à 28% en excluant les SNPs de la région d'ABO), et de 9% parmi les SNPs retenus après ajustement sur le groupe ABO. Parmi les SNPs associés aux taux de FVIII, le FDR est de 27% (mais s'élève à 40% en excluant les SNPs de la région d'ABO), de 19% pour les analyses ajustées sur le groupe ABO, et de 56% pour les analyses ajustées sur le taux de vWF.

2) Les deux signaux d'association significatifs, le premier au niveau des gènes *RAPGEF1* et *ABO* en 9q34, et le deuxième au niveau des gènes *TMEM16, STAB2* et *PTPN11* en 12q23*,* sont superposables aux signaux de liaison obtenus par l'analyse des microsatellites. J'ai étudié de plus près les résultats des tests d'association au niveau des autres signaux en 2p12-13, 2q33, 6q13-14, 15q14 et 19p13 (voir **figure 10 p75**). Aucune association n'était significative au seuil de Bonferroni calculé pour le nombre de SNPs situés dans chacune des régions d'études (**figures en annexe pA32-A36**)

VIII.1.4. Réplication dans les études MARTHA08 et MARTHA10

J'ai porté mon attention sur les deux régions du génome qui avaient montré des associations significatives dans l'échantillon Familles-FVL (9q34 et 12q23). L'étude du déséquilibre de liaison au sein de ces régions a confirmé que celui-ci était bien moindre dans les échantillons MARTHA que dans les Familles-FVL (**annexe p15-A27**). En conséquence, le signal d'association en 9q34, plutôt que de s'étaler sur plusieurs Mb autour du gène *ABO* à l'instar de celui obtenu avec les Familles-FVL, était au contraire bien circonscrit au niveau d'*ABO* (**figure 23 et 24**). En particulier, il n'y avait aucune association significative au niveau du gène *RAPGEF1* (voir *éventuellement* tableau en **annexe pA28**). Ceci nous conforte donc dans l'hypothèse que le « hit-SNP » des analyses réalisées avec les Familles-FVL (rs17553234 du gène *RAPGEF1*) n'était peut-être qu'un reflet du déséquilibre de liaison avec le gène *ABO*, ou, plus précisément, avec une structure haplotypique particulière du gène *ABO*.

.

Figure 23. Analyse des associations entre les polymorphismes du locus *ABO* et le taux plasmatique de vWF dans l'échantillon MARTHA08.

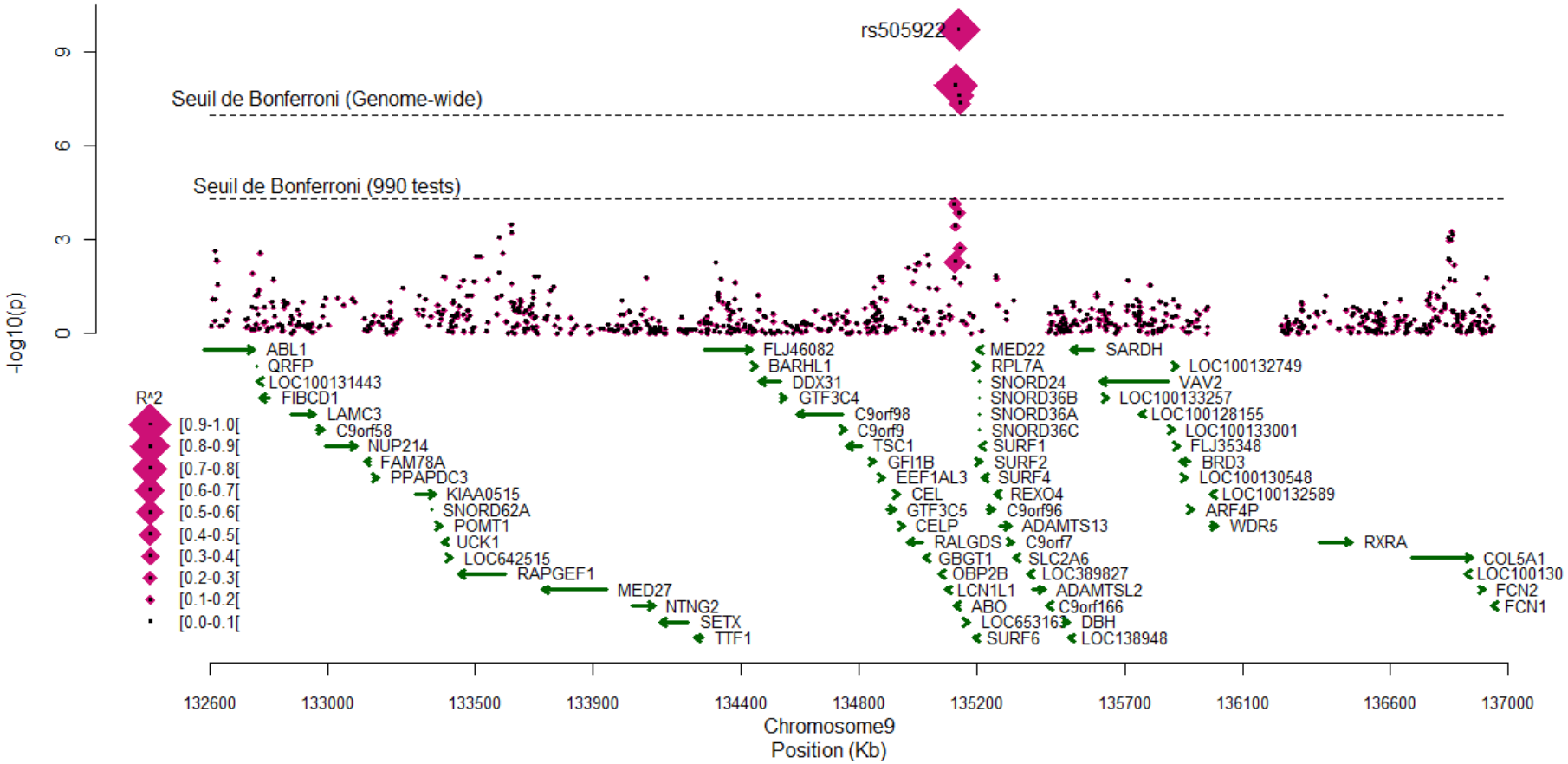
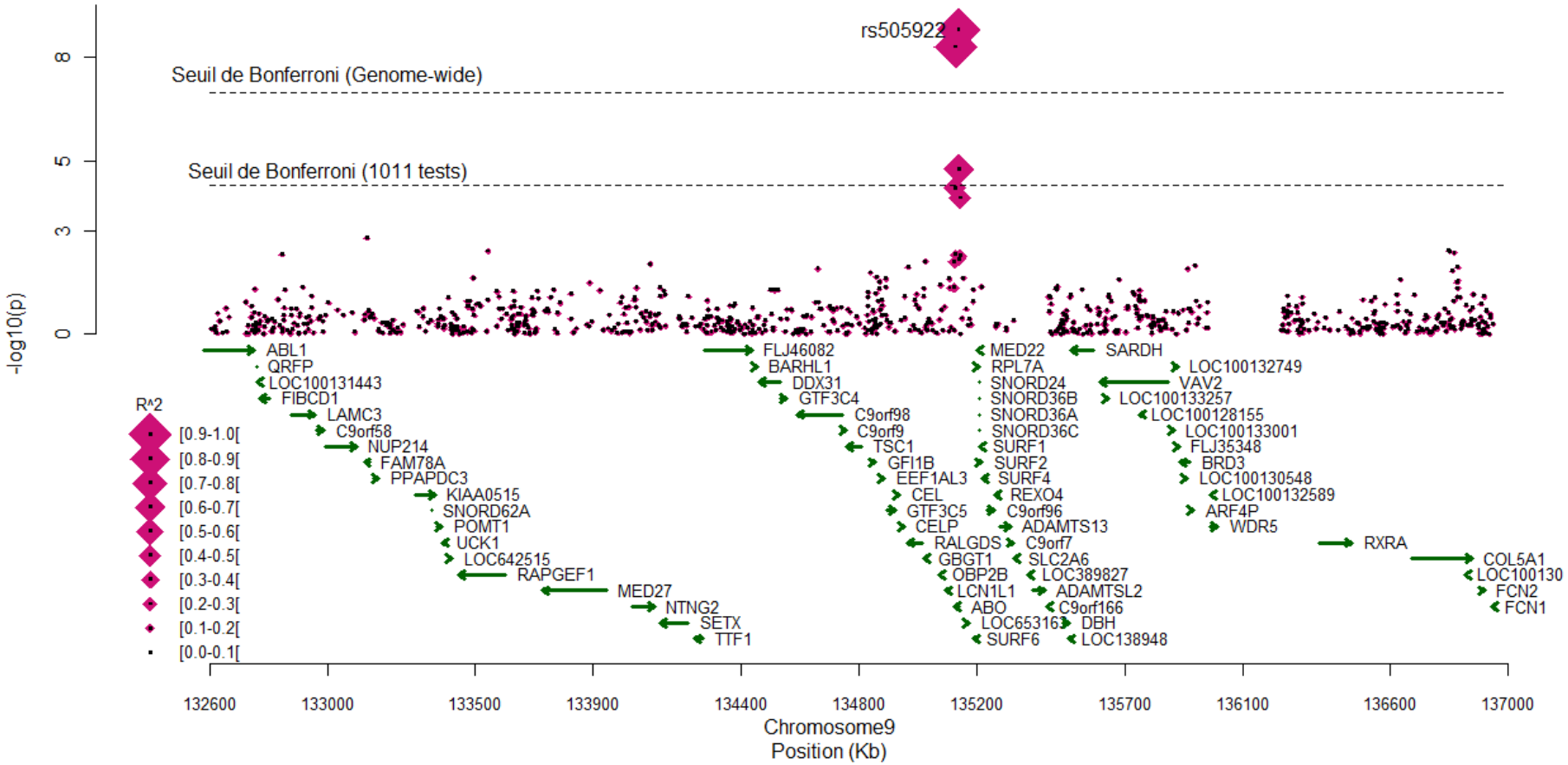
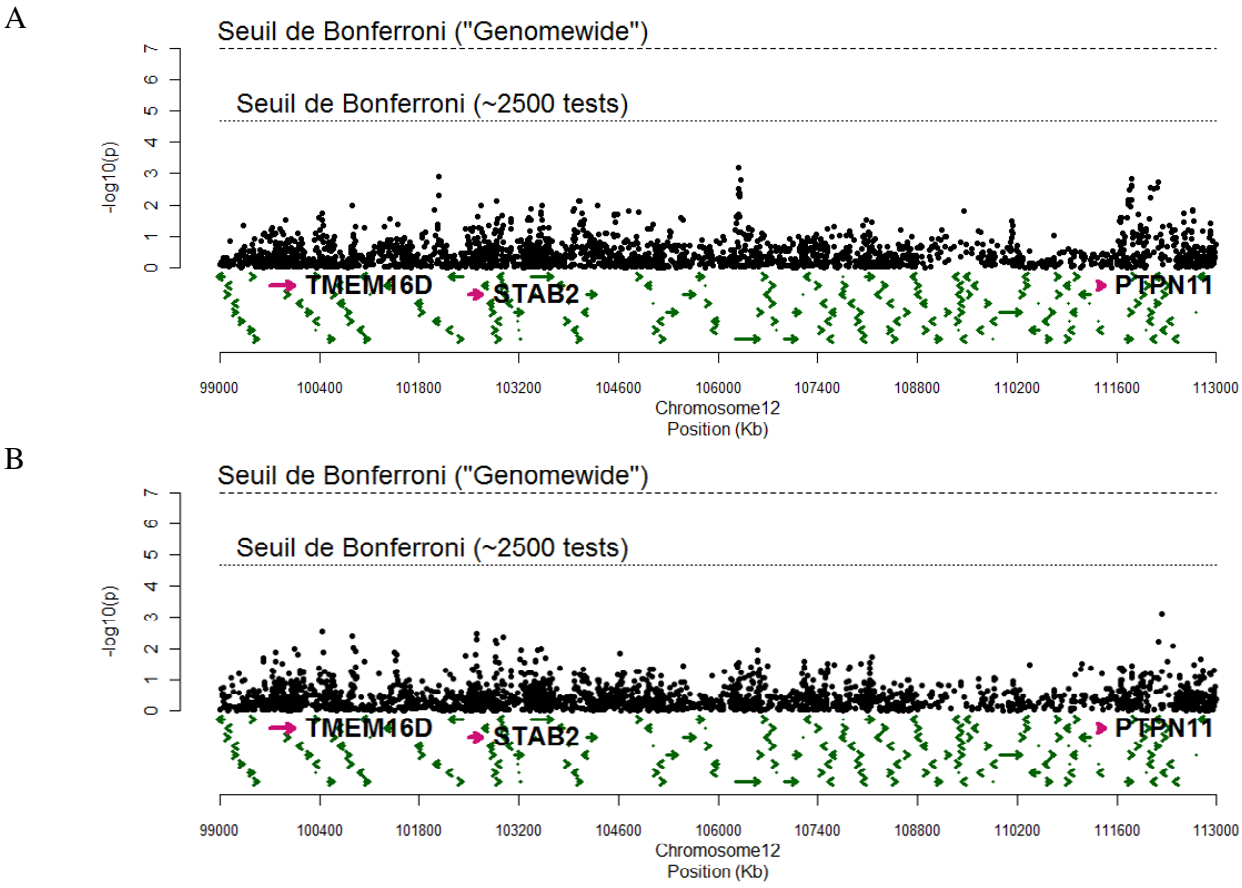


Figure 24. Analyse des associations entre les polymorphismes du locus *ABO* et les taux plasmatiques de vWF dans l'échantillon MARTHA10



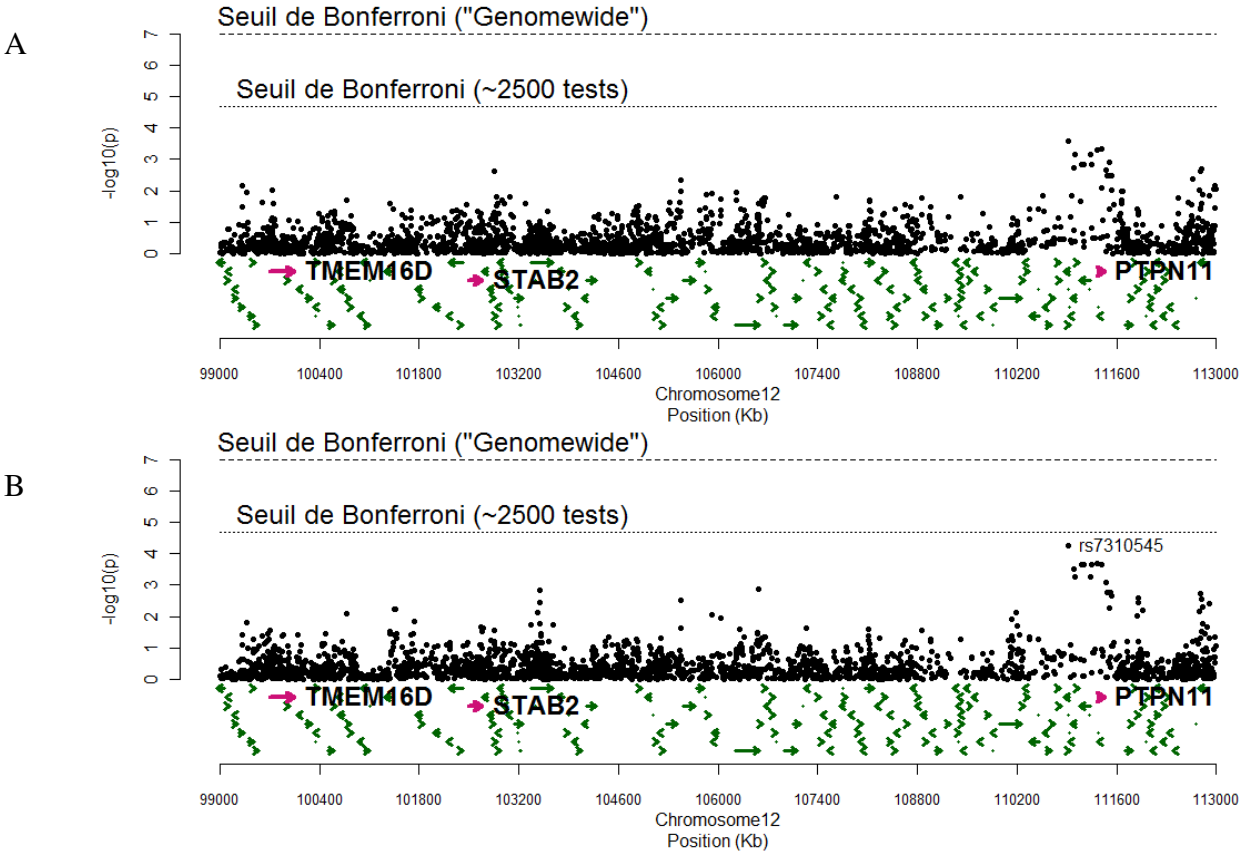
La recherche d'une réplication des associations observées en 12q23 s'est faite sur l'ensemble de la région, et non pas seulement sur les deux SNPs présentant les associations les plus significatives (rs713089 dans *TMEM16D* et rs11615047 dans *PTPN11*). Ce choix a eu pour conséquence une perte de puissance due à la prise en compte de la multiplicité des tests, correspondant aux quelques 2 500 SNPs de la région étudiée. Il était dicté par la supposée imprécision de la localisation des associations, fruit du déséquilibre de liaison, dans les Familles-FVL. Aucun des deux modèles étudiés (vWF et FVIII, tous deux ajustés sur les trois SNPs taguant le groupe *ABO*) n'a révélé d'association significative au seuil de Bonferroni $p=2.10^{-5}$(**figure 25 et 26 page suivante**).

Figure 25. Association des SNPs du locus 12q23 avec les taux plasmatiques de vWF et de FVIII dans MARTHA08



A. taux de vWF ajustés sur le gène *ABO*. :.
B. taux de FVIII ajustés sur le gène *ABO*.
Trois gènes sont signalés en rouge : *STAB2* ainsi que *TMEM16D* et *PTPN11*, les deux gènes associés aux taux de vWF et/ou FVIII dans les analyses menées dans l'échantillon Familles-FVL

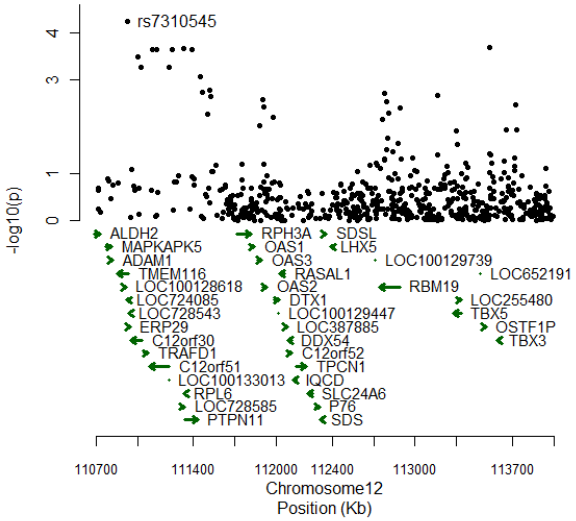
Figure 26. Association des SNPs du locus 12q23 avec les taux plasmatiques de vWF et de FVIII dans MARTHA10



A. taux de vWF ajustés sur le gène *ABO*. :.
B. taux de FVIII ajustés sur le gène *ABO*.
Trois gènes sont signalés en rouge : *STAB2* ainsi que *TMEM16D* et *PTPN11*, les deux gènes associés aux taux de vWF et/ou FVIII dans les analyses menées dans l'échantillon Familles-FVL

On observe toutefois (**figure 26.B**) que l'analyse du FVIII dans MARTHA10 a décelé un petit groupe de SNPs qui s'approchaient de la significativité. Certains d'entre eux étaient localisés au sein de *PTPN11* (**figure 27**). L'association la plus significative ($p=2,7.10^{-4}$) était observée pour rs7310545 dans le gène *TMEM116*. Cependant, aucune association similaire n'était observée pour ce SNP dans MARTHA08.

Figure 27. Association des SNPs du locus de PTPN11 avec les taux plasmatiques de FVIII dans MARTHA10



On observe dans MARTHA10 quelques associations s'approchant de la significativité avec des polymorphismes de *TMEM116* et *PTPN11*

Résultats présentés en annexes

Chaque association dont la significativité était inférieure à 10^{-5} dans l'échantillon Familles-FVL a été testée dans MARTHA08 et MARTHA10 (résultats en **annexe pA28-A31**). Cent un SNPs étaient concernés par ces analyses, abaissant ainsi le seuil de significativité à $5\ 10^{-4}$. Aucun des tests réalisés ne passait ce seuil. Les associations les plus significatives (p=0,03 ou 0,04) n'étaient observées que dans l'un des deux échantillons MARTHA.

VIII.1.5. Discussion

Le génotypage de plus de 500 000 SNPs de l'échantillon Famille-FVL m'a permis de rechercher des associations avec les taux plasmatiques de vWF et FVIII, à travers tout le génome, sans hypothèse *a priori*.

Synthèse et interprétation des principaux résultats

L'analyse des taux de vWF a révélé un signal d'association dans la région du gène *ABO*. Ce signal, dont le pic se trouvait pour un polymorphisme du gène *RAPGEF1,* concernait plusieurs gènes. Il était nettement plus étalé que ceux observés habituellement à partir d'individus non apparentés. En particulier, dans MARTHA08 et MARTHA10, seul le gène *ABO* présentait des polymorphismes associés taux de vWF. Ce phénomène s'explique par l'importance du déséquilibre de liaison dans un échantillon constitué d'un petit nombre de grandes familles. Cependant, l'association observée au niveau *RAPGEF1* ne disparaissait pas entièrement après ajustement sur les trois SNPs à même de distinguer les allèles A1, A2, B et O. Deux hypothèses peuvent expliquer cette observation. Il se peut que des polymorphismes de *RAPGEF1* soient effectivement associés aux taux de vWF indépendamment du groupe ABO (association réelle ou fortuite en raison d'une fluctuation d'échantillonnage). Il se peut aussi que les polymorphismes caractérisant le groupe ABO ne soient pas les seuls du gène *ABO* à avoir une influence sur les taux de vWF. Il pourrait alors rester une association résiduelle avec un polymorphisme (éventuellement haplotypique) d'*ABO*, absent de nos données, et en déséquilibre de liaison avec *RAPGEF1*. De fait, Heit et al [64] viennent de mettre en évidence un SNP d'*ABO* (rs2519093), associé au risque de MTEV et indépendant des SNPs qui taguent A1, A2, B et O.

La région 12q23, après ajustement sur le gène *ABO*, présentait un ensemble de SNPs associés aux taux de FVIII et de vWF. Le plus significatif était situé dans le gène *TMEM16D.* Celui-ci code pour une protéine transmembranaire dont la fonction est inconnue, mais qui présente 38.6% de similitude avec *TMEM16A*[168]. Ce dernier code pour un canal chlorure calcium dépendant, présent dans les cellules musculaires lisses des artères cérébrales[169]. Malgré ces éléments biologiques intéressants dans le contexte de la MTEV, des réserves peuvent être émises quant à son implication dans la modulation des taux de vWF et FVIII. En effet, de façon remarquablement similaire à la région 9q34, le signal d'association s'étendait sur plusieurs gènes. De plus, les associations observées n'ont pas été reproduites dans MARTHA08 et MARTHA10. L'échec de la réplication pourrait s'expliquer par un recrutement très spécifique de l'échantillon Familles-FVL. Plus vraisemblablement, l'importance du déséquilibre de liaison généré par un échantillon constitué d'un petit nombre de grandes familles induit une localisation imprécise du gène causal. La région 12q23 contient en particulier le gène *STAB2* dont certains polymorphismes sont asociés aux taux de vWF et de FVIII [101] et à la MTEV [158]. Afin de clarifier le signal d'association de cette région, il serait nécessaire, de manière similaire au travail que j'ai réalisé pour la région 9q34, d'étudier, dans les Familles-FVL, les structures haplotypiques entre les polymorphismes de *TMEM16D* et *STAB2*. Cette étude pourrait être complétée par une analyse des associations entre les SNPs de la région et les taux de vWF ajustés sur *ABO* et *STAB2*.

Complémentarité des analyses d'association et de liaison

Les analyses de liaison présentées au **§VII.1.** se sont avérées plus puissantes que les analyses d'associations réalisées à partir du même échantillon de grandes familles. En effet, il y avait plusieurs signaux de liaison très significatifs, alors que seuls deux gènes (*RAPGEF1* et *TMEM16A*) présentaient des polymorphismes dont les associations étaient significatives après avoir contrôlé le risque d'erreur de type à 5% par une correction de Bonferroni. Il était à ce propos assez décevant de ne pas pouvoir observer une significativité p inférieure au seuil de Bonferroni au sein du gène *ABO*. L'échantillon Familles-FVL est parfaitement adapté aux analyses de liaison, méthodologie pour laquelle il a été conçu initialement. Par contre, il n'est pas optimal lorsqu'il s'agit d'analyses d'association. En effet, les corrélations intra-familiales diminuent le nombre d'observations indépendantes, réduisant ainsi l'effectif efficace, et par là, la puissance de l'analyse. De plus, la normalisation des traits quantitatifs requise par la méthode de décomposition de la variance utilisée a pu contribuer à la perte de puissance. Par ailleurs, l'ampleur du déséquilibre de liaison amoindrit le bénéfice attendu des analyses

d'association par rapport aux analyses de liaison en terme de précision de la localisation du QTL. Les analyses d'association ont toutefois permis de pointer quelques gènes dont les polymorphismes étaient plus significativement associés aux phénotypes que les autres. Ainsi, cette stratégie est à même de générer quelques gènes candidats parmi les quelques dizaines de gènes inclus dans les signaux de liaison.

Afin d'optimiser cette stratégie, il est souhaitable de chercher à augmenter la puissance des analyses d'association pangénomiques. Dans cette optique, la méthode de contrôle de l'erreur de type 1 au moyen d'un FDR pondéré est intéressante [170]. Il s'agit d'une méthode qui ne corrige pas de manière homogène toutes les valeurs de p. En effet, elle fait intervenir un poids, spécifique à chaque test, issu de connaissances antérieures. Dans le cadre des analyses d'associations pangénomiques réalisées dans cette section, un choix judicieux de poids aurait pu être l'inverse de la valeur du BF obtenue en chaque point par les analyses de liaison. Ainsi seraient favorisées les découvertes de nouvelles associations situées au sein des signaux de liaison, tout en contrôlant globalement la proportion de faux positifs. L'utilisation d'un FDR stratifié est également possible [171]. Dans ce cas, les tests réalisés sont regroupés en différentes strates selon que les polymorphismes correspondants sont situés ou non dans un signal de liaison. Un calcul de FDR est ensuite réalisé dans chaque strate séparément. Le seuil sera donc beaucoup moins conservateur dans la strate correspondant aux signaux de liaison puisqu'elle contient moins de SNP.

Je n'ai pas exploré les possibilités offertes par ces méthodes. Il est cependant extrêmement vraisemblable que les associations observées avec les polymorphismes du gène *ABO* affleurant le seuil de Bonferroni et situées dans un signal de liaison très fort auraient franchi le seuil de significativité. Par contre, ces méthodes n'auraient probablement pas été à même de révéler des associations significatives au sein des signaux de liaison en 2p12-13, 2q33, 6q13-14, 15q14, 19p13. En effet, comme on l'observe en annexe pA32-A36, aucune association n'était significative au seuil de Bonferroni calculé séparément pour chacune des régions étudiées. Il n'y a donc aucune association significative pour un FDR à 5%. Or, cette façon de procéder revient à réaliser un FDR stratifié favorisant au maximum les découvertes dans les régions de liaison, puisqu'elle attribue un poids égal à 1 à la région considérée et un poids nul au reste du génome. C'est pourquoi, bien que ces méthodes soient intéressantes, je n'ai pas cherché à les utiliser. De plus, elles sont discutables lorsque l'échantillon qui a permis d'établir les poids *a priori* est le même que celui qui est l'objet de l'étude. En effet, s'il existe une simple fluctuation d'échantillonnage à l'origine d'un faux signal d'association,

elle pourrait également être à l'origine d'un faux signal de liaison. L'application de cette méthode favoriserait alors cette fausse découverte.

Conséquence sur la suite du travail

La stratégie d'analyse présentée dans cette section n'a pas permis de conclure formellement quant à l'implication d'autres gènes que le gène *ABO* dans le contrôle des taux de FVIII et vWF. Les résultats convergent toutefois vers la région 12q23 sans pour autant permettre d'identifier précisément un polymorphisme identique dans les trois études. Dans le but d'augmenter la puissance de découverte de polymorphismes modulant les taux de FVIII et/ou de vWF, et particulièrement de ceux dont les effets sont faibles, j'ai combiné les résultats des trois échantillons au moyen d'une méta-analyse. Toujours dans une optique d'augmentation de la puissance, j'ai considéré les modèles avec ajustement sur le groupe ABO.

VIII.2. Méta-analyse de trois études d'association pangénomique sur les taux plasmatiques de vWF et FVIII

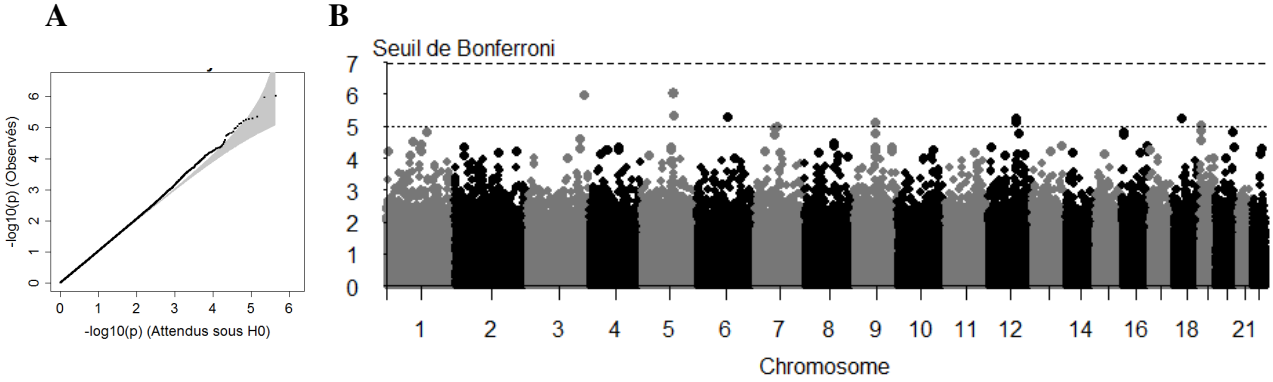
VIII.2.1. Présentation globale des résultats

Parmi l'ensemble des polymorphismes qui ont passé les contrôles de qualité (**tableau 5 p38**) propres à chaque échantillon (*i.e* Familles-FVL, MARTHA08 et MARTHA10), 442 728 étaient communs aux trois études. J'ai ainsi pu mener une méta-analyse des résultats obtenus avec les trois échantillons pour ces 442 728 polymorphismes. Les résultats que j'ai utilisés pour cette méta-analyse sont ceux issus des études des transformations par quantile de loi normale des taux plasmatiques de vWF et de FVIII, ajustées sur l'âge, le sexe et le groupe sanguin ABO défini par les trois SNPs caractérisant les allèles A1, A2 B et O (**tableau 6** p39).

Comme on peut l'observer sur la **figure 24.A**, le QQ-plot des valeurs de p obtenues pour l'analyse des taux de vWF présente une légère "boursoufflure" à partir des 1‰ plus petites valeurs de p. Elle est probablement due à la prise en compte dans la méta-analyse des résultats de l'échantillon Famille-FVL, et au phénomène déja mentionné pour ce dernier (voir p100). Le QQ-plot des valeurs de p obtenues pour l'analyse des taux de FVIII ne montre pas d'écart important entre les valeurs observées et les valeurs attendues sous l'hypothèse nulle d'absence d'association (**figure 25.A**). Aucune association n'était significative au seuil de

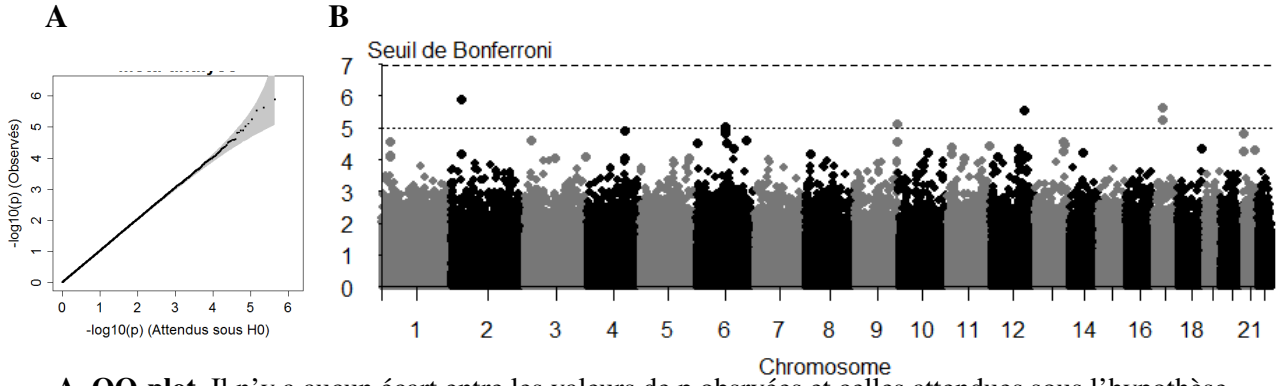
Bonferroni ($p<1,12.10^{-7}$) ni avec l'analyse des taux de vWF ni avec celle de FVIII (**figures 28.B et 29.B**).

Figure 28. Méta-analyse de trois études d'association pangénomique des taux de vWF



A. QQ-plot. La légère boursoufflure est probablement due à la présence de déséquilibre de liaison sur de grandes distances dans l'échantillon Famille-FVL.
B. Manhattant plot. Aucune association n'est significative au seuil de Bonferroni.

Figure 29. Méta-analyse de trois études d'association pangénomique des taux de FVIII



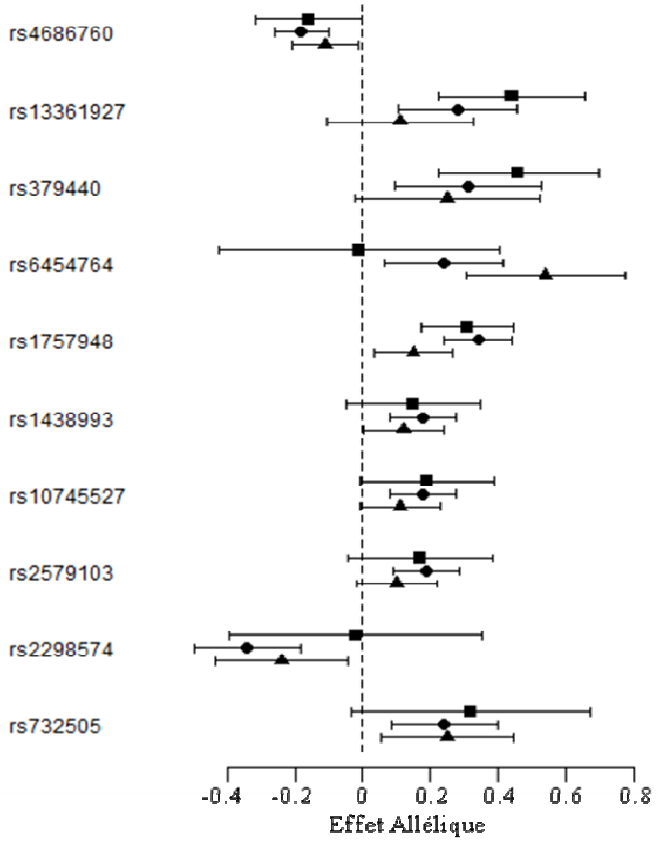
A. QQ-plot. Il n'y a aucun écart entre les valeurs de p obsrvées et celles attendues sous l'hypothèse nulle d'absence d'association.
B. Manhattant plot. Aucune association n'est significative au seuil de Bonferroni.

Constatant l'absence de signaux statistiques dépassant le seuil de Bonferroni, je me suis intéressée aux polymorphismes dont l'association avec les taux de vWF ou de FVIII présentaient une significativité inférieure à 10^{-5}. En effet, si la correction de Bonferroni permet de limiter le nombre de faux positifs, elle a l'inconvénient d'induire une perte de puissance pour détecter d'éventuels effets modestes qui pourraient se cacher parmi les signaux statistiques légèrement moins forts.

VIII.2.2. Examen des signaux d'association ($p<10^{-5}$) détectés pour les taux de vWF

Dix polymorphismes, répartis dans sept loci, étaient associés à vWF au seuil $p <10^{-5}$. Le test de Cochran-Mantel-Haenszel d'hétérogénéité Q n'était significatif que pour l'un d'entre eux, *ANKDR*-rs645764 (p=0,036), dont l'effet n'était pas observé dans l'échantillon Familles-FVL. Aucun test d'hétérogénéité ne serait significatif si on appliquait une correction de Bonferroni pour ces 10 tests (p<0,005). Cependant, ce test est peu puissant lorsqu'il est réalisé à partir d'un faible nombre d'études. L'examen de la statistique I^2 pointe deux autres polymorphismes, *EPB41L4A*-rs13361927 et *KRT18P24*-rs1757948, dont les effets étaient hétérogènes entre les études (I^2 vaut respectivement 53% et 62%). A l'exception de ces trois polymorphismes, les significativités des effets fixes et aléatoires étaient similaires (**figure 30** et **tableau 22**).

Figure 30: Méta-analyse des 10 polymorphismes les plus fortement associés ($p<10^{-5}$) aux taux de vWF



L'effet de chaque polymorphisme est représenté par un carré dans les Familles-FVL, un rond dans MARTHA08, et un triangle dans MARTHA10

Tableau 22 – Polymorphismes associés aux taux de vWF dans une méta-analyse de trois études pangénomiques, avec une significativité $p<10^{-5}$

Gene	SNP	Alleles*		MAF+	β (SE)	p	I^2	p_{het}	Effet aléatoire$ β (Ecart-Type)	Effet aléatoire$ p	Effet fixe$ β (Ecart-Type)	Effet fixe$ p
VPS8	**rs4686760**	A/G	FVL	0,47	-0,16 (0,08)	0,044	0	0,549	-0,15 (0,03)	$1,10\ 10^{-6}$	-0,15 (0,03)	$1,08\ 10^{-6}$
			MARTHA08	0,46	-0,18 (0,04)	$4,11\ 10^{-5}$						
			MARTHA10	0,45	-0,11 (0,05)	0,047						
EPB41L4A	rs13361927	*G/A*	FVL	0,15	0,44 (0,11)	$3,08\ 10^{-4}$	0,53	0,119	0,28 (0,09)	0,002	0,28 (0,06)	$4,51\ 10^{-6}$
			MARTHA08	0,06	0,28 (0,09)	0,003						
			MARTHA10	0,05	0,11 (0,11)	0,316						
	rs379440	A/G	FVL	0,12	0,46 (0,12)	$8,35\ 10^{-4}$	0	0,502	0,34 (0,07)	$9,99\ 10^{-7}$	0,34 (0,07)	$9,82\ 10^{-7}$
			MARTHA08	0,04	0,31 (0,11)	0,004						
			MARTHA10	0,03	0,25 (0,14)	0,071						
ANKRD6	rs6454764	*C/T*	FVL	0,04	-0,01 (0,21)	0,977	0,70	**0,036**	0,29 (0,14)	0,035	0,31 (0,07)	$5,12\ 10^{-6}$
			MARTHA08	0,06	0,24 (0,09)	0,007						
			MARTHA10	0,05	0,54 (0,12)	$8,97\ 10^{-6}$						
KRT18P24	**rs1757948**	*T/G*	FVL	0,27	0,34 (0,09)	$2,82\ 10^{-4}$	0,62	0,071	0,18 (0,06)	0,003	0,15 (0,03)	$7,37\ 10^{-6}$
			MARTHA08	0,27	0,10 (0,05)	0,030						
			MARTHA10	0,30	0,15 (0,06)	0,009						
désert	rs1438993	G/A	FVL	0,19	0,15 (0,10)	0,127	0	0,666	0,16 (0,03)	$6,34\ 10^{-6}$	0,16 (0,03)	$6,25\ 10^{-6}$
			MARTHA08	0,28	0,18 (0,05)	$1,11\ 10^{-4}$						
			MARTHA10	0,27	0,12 (0,06)	0,052						
	rs10745527	T/G	FVL	0,20	0,19 (0,10)	0,062	0	0,663	0,16 (0,03)	$5,51\ 10^{-6}$	0,16 (0,03)	$5,43\ 10^{-6}$
			MARTHA08	0,28	0,18 (0,05)	$1,63\ 10^{-4}$						
			MARTHA10	0,27	0,11 (0,06)	0,056						
	rs2579103	T/G	FVL	0,18	0,17 (0,11)	0,098	0	0,533	0,16 (0,04)	$7,72\ 10^{-6}$	0,16 (0,04)	$7,61\ 10^{-6}$
			MARTHA08	0,26	0,19 (0,05)	$8,24\ 10^{-5}$						
			MARTHA10	0,25	0,10 (0,06)	0,090						
CDH2	rs2298574	A/G	FVL	0,04	-0,02 (0,19)	0,905	0,19	0,290	-0,26 (0,07)	$1,81\ 10^{-4}$	-0,27 (0,06)	$5,67\ 10^{-6}$
			MARTHA08	0,08	-0,34 (0,08)	$2,77\ 10^{-5}$						
			MARTHA10	0,07	-0,24 (0,10)	0,022						
SAFB2	**rs732505**	G/A	FVL	0,05	0,32 (0,18)	0,080	0	0,929	0,25 (0,06)	$9,50\ 10^{-6}$	0,25 (0,06)	$9,38\ 10^{-6}$
			MARTHA08	0,09	0,24 (0,08)	0,001						
			MARTHA10	0,08	0,25 (0,10)	0,013						

*Allèle commun/rare

+ Fréquence de l'allèle rare

$ Effet de l'allèle rare

L'association la plus forte était observée avec *EPB41L4A*-rs379440 ($p = 9{,}82\ 10^{-7}$). Un second polymorphisme de ce même gène, *EPB41L4A*-rs13361927, présentait également un effet fixe assez fort ($p = 4{,}51\ 10^{-6}$). Cette association était probablement la conséquence du déséquilibre de liaison de rs13361927 et rs379440. En effet, les r^2 entre ces deux SNPs valaient respectivement 0,78, 0,69 et 0,62 dans les échantillons familles-FVL, MARTHA08 et MARTHA10. La **figure 30 p122** illustre par ailleurs une atténuation de l'effet de rs13361927, par rapport à rs379440, dans chacune des trois études. Cette atténuation est nettement plus importante dans MARTHA10, où le déséquilibre de liaison était le moins fort. En conséquence, alors que l'effet de rs379440 était parfaitement homogène entre les études ($I^2 = 0$), l'effet de rs13361929 ne l'était pas ($I^2 = 53\%$), conduisant à un effet aléatoire nettement moins significatif ($p = 0{,}002$).

Outre *ANKRD6*-rs6454764 et *KRT18P24*-rs1757948 dont nous avons déjà mentionné l'importante hétérogénéité des effets entre études, et dont les effets aléatoires étaient plus faiblement significatifs (respectivement $p = 0{,}035$, et $p = 0{,}003$), les autres SNPs associés à vWF étaient *CDH2*- rs2298574 ($p = 5{,}67\ 10^{-6}$), *SAFB2*-rs732505 ($p = 9{,}38\ 10^{-6}$), *VPS8*-rs4686760 ($p = 1{,}08\ 10^{-6}$), ainsi que trois SNPs situés au locus 12q21.33, dans une région ne comportant aucun gène. Ces trois polymorphismes étaient par ailleurs en association quasi complète dans les trois études (r^2 proche de 1), signifiant qu'il ne s'agissait pas de trois associations indépendantes, mais le reflet d'un phénomène unique. Bien que le test d'hétérogénéité des effets de *CDH2* soit non significatif, et la valeur I^2 faible, notons que l'effet global de *CDH2* est dû uniquement aux échantillons MARTHA.

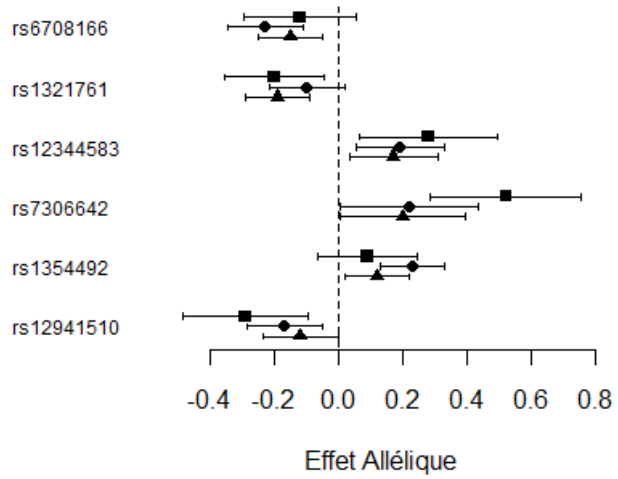
Le pourcentage de variance R^2 de vWF expliqué par ces associations a été calculé dans les échantillons MARTHA. Seules ont été gardées les associations indépendantes. Ainsi rs13361927, rs1438993 et rs2579103 n'ont pas été intégrés au modèle, de même que rs6454764 et rs2298574 dont les effets étaient quasiment nuls dans l'échantillon Familles-FVL. Les polymorphismes rs4686760, rs379440, rs1757948, rs10745527, rs732505 expliquaient 5,7% ($p = 1{,}30\ 10^{-10}$) et 3,8% ($p = 8{,}39\ 10^{-5}$) de la variabilité plasmatique de vWF dans respectivement MARTHA08 et MARTHA10. En combinant ces deux échantillons ce pourcentage était de 5,3% ($p = 5{,}55\ 10^{-16}$).

VIII.2.3. Examen des signaux d'association ($p < 10^{-5}$) détectés pour les taux de FVIII

Aucun des dix polymorphismes précédemment identifiés n'était associé au seuil $p < 10^{-5}$ aux taux de FVIII. En revanche, six nouvelles association ont pu être décelées, au seuil de $p<10^{-5}$ **(figure 31 et tableau 23**). Aucun des tests d'hétérogénéité fondés sur la statistique Q n'était significatif au seuil p = 0.05. Cependant, deux polymorphismes présentaient une hétérogénéité notable par la statistique I^2 : hétérogénéité modérée pour *ACCN1* rs1354492 (I^2 = 39%), hétérogénéité importante pour *STAB2* rs7306642 (I^2 = 59%). Les effets aléatoires étaient peu éloignés des effets fixes en ce qui concernait *ACCN1*-rs1354492, sensiblement plus faibles pour *STAB2*-rs7306642, et similaires pour les autres SNPs.

L'association la plus forte était observée pour *STAB2*-rs7306642. Il est situé dans un exon et entraîne une modification d'acide aminé Pro2039Thr dans la protéine. Cette association est particulièrement marquée dans les familles-FVL, tandis qu'elle est deux fois moindre dans les échantillons MARTHA. La significativité de l'effet aléatoire est par conséquent relativement faible (p=0.002). Cette association reste toutefois très intéressante, d'autant plus qu'une récente étude d'association génome-entier, menée par le consortium CHARGE [101], avait mis en évidence deux SNPs, rs4981022 et rs12229292, situés dans *STAB2* et associés aux taux de vWF et FVIII. L'étude du déséquilibre de liaison entre rs7306642 et les deux polymorphismes identifiés par CHARGE, réalisée séparément dans nos trois échantillons, nous apprend que rs7306642 est nettement indépendant de rs4981022 (r^2<0.01 dans les trois études), et de rs4981021 (r^2<0.07 dans les trois études).

Figure 31: Méta-analyse des 10 polymorphismes les plus fortement associés ($p<10^{-5}$) aux taux de FVIII



L'effet de chaque polymorphisme est représenté par un carré dans les Familles-FVL, un rond dans MARTHA08, et un triangle dans MARTHA10

Tableau 23. Polymorphismes associés aux taux plasmatiques de FVIII dans une méta-analyse de trois études pangénomiques, avec u significativité $p<10^{-5}$

Gene	SNP	Alleles*		MAF+	β$ (SE)	p	I^2	p_{het}	Effet aléatoire		Effet fixe	
									β$ (Ecart-Type)	p	β$ (Ecart-Type)	p
LBH	rs6708166	G/A	FVL	0,41	-0,12 (0,09)	0,156	0	0,478	-0,17 (0,04)	$1,32\ 10^{-6}$	-0,17 (0,04)	$1,30\ 10^{-6}$
			MARTHA08	0,40	-0,23 (0,06)	8,98 10-^{5}						
			MARTHA10	0,42	-0,15 (0,05)	0,007						
FAM46A	rs1321761	T/C	FVL	0,42	-0,20 (0,08)	0,014	0	0,451	-0,15 (0,04)	$9,67\ 10^{-6}$	-0,15 (0,04)	$9,54\ 10^{-6}$
			MARTHA08	0,45	-0,10 (0,06)	0,074						
			MARTHA10	0,47	-0,19 (0,05)	$5,93\ 10^{-4}$						
VAV2	rs12344583	A/G	FVL	0,17	0,28 (0,11)	0,012	0	0,716	0,20 (0,04)	$8,03\ 10^{-6}$	0,20 (0,04)	$7,92\ 10^{-6}$
			MARTHA08	0,20	0,19 (0,07)	0,006						
			MARTHA10	0,18	0,17 (0,07)	0,012						
STAB2	rs7306642	C/A	FVL	0,16	0,52 (0,12)	$1,36\ 10^{-5}$	0,59	0,086	0,31 (0,10)	0,002	0,30 (0,06)	$2,95\ 10^{-6}$
			MARTHA08	0,07	0,22 (0,11)	0,057						
			MARTHA10	0,07	0,20 (0,10)	0,052						
ACCN1	rs1354492	G/A	FVL	0,53	0,09 (0,08)	0,293	0,39	0,192	0,16 (0,04)	$5,47\ 10^{-6}$	0,16 (0,03)	$2,41\ 10^{-6}$
			MARTHA08	0,49	0,23 (0,05)	$1,20\ 10^{-5}$						
			MARTHA10	0,47	0,12 (0,05)	0,027						
	rs12941510	G/A	FVL	0,22	-0,29 (0,10)	0,004	0,12	0,321	-0,17 (0,04)	$2,18\ 10^{-5}$	-0,17 (0,04)	$5,67\ 10^{-6}$
			MARTHA08	0,31	-0,17 (0,06)	0,002						
			MARTHA10	0,33	-0,12 (0,06)	0,029						

*Allèle commun/rare, + Fréquence de l'allèle rare $ Effet de l'allèle rare

Les autres polymorphismes associés aux taux de FVIII incluaient rs6708166 ($p = 1{,}30\ 10^{-6}$) près de *LBH*, rs1321761 près de *FAM46A* ($p = 9{,}54\ 10^{-6}$), *VAV2*-rs12344583 ($p = 7{,}92\ 10^{-6}$), ainsi que deux polymorphismes, rs1354492 et rs12941510, au sein de *ACCN1*. L'allèle rare A de *ACCN1*-rs1354492 était associé à une élévation des taux de FVIII ($\beta = +0.16$, $p = 2{,}42\ 10^{-6}$), à l'inverse de l'allèle rare A de *ACCN1*-rs12941510 ($\beta = -0.17$, $p = 5{,}67\ 10^{-6}$). Ces deux polymorphismes présentaient un déséquilibre de liaison négatif complet, constituant trois haplotypes. Des analyses haplotypiques conduites dans MARTHA08 et MARTHA10 ont mené à la conclusion que seul rs1354492 (sur fond génétique rs12941510-G) était associé aux taux plasmatiques de FVIII tandis que ces derniers n'étaient pas significativement modifiés par rs12941510 (sur fond génétique rs1354492-G) (**tableau 24**: comparaison des haplotypes 1 et 3 d'une part, 1 et 2 d'autre part).

Tableau 24. Association haplotypique des polymorphismes rs1354492 et rs12941510 avec les taux plasmatiques de FVIII dans les échantillons MARTHA08 et MARTHA10

			MARTHA08		MARTHA10	
Haplotype	**rs1354492**	**rs12941510**	**Fréquence Haplotypique**	**effet[(1)]**	**Fréquence Haplotypique**	**effet[(1)]**
1	**G**	**G**	0,20	-0,22 [-0,36 - -0,08]	0,20	-0,07 [-0,21 - 0,07]
2	**G**	**A**	0,31	-0,24 [-0,34 - -0,12]	0,33	-0,14 [-0,26 - -0,02]
3	**A**	**G**	0,49	référence	0,47	référence

(1) effet haplotypique [95%IC] sur les taux de FVIII associé à une une copie de chaque haplotype sous l'hypothèse d'additivité des effets haplotypiques. Les analyses sont ajustées sur l'âge, le sexe, et le groupe sanguin ABO.

Finalement, les pourcentages de la variance des taux plasmatiques de FVIII expliquée par les cinq polymorphismes dont les effets étaient indépendants (rs6708166, rs1321761, rs12344583, rs7306642, rs1354492) valaient 8,2% ($p = 4{,}27\ 10^{-10}$) dans MARTHA08, 4,6% ($p = 8{,}36\ 10^{-6}$) dans MARTHA10 et 6,3% ($p = 2{,}66\ 10^{-15}$) dans les deux échantillons combinés.

VIII.2.4. Validation de travaux antérieurs

J'ai ensuite regardé comment se comportaient dans nos trois échantillons les polymorphismes qui, dans d'autres études, étaient associés aux taux de vWF et/ou FVIII. Lorsque ces polymorphismes n'étaient pas présents dans notre étude, j'ai recherché leurs proxy (*i.e.* un polymorphisme en déséquilibre de liaison avec un $r^2 \geq 0{,}8$) à l'aide du programme SNAP [172]. J'ai ainsi étudié les deux polymorphismes de *BAI3* mis en évidence

dans l'échantillon STANISLAS (voir **§VII.2**.), deux polymorphismes du gène *LDLR* et deux polymorphismes du gène *VWF* associés à l'un ou l'autre phénotype dans des études par approche « gène candidat » [110][100]. De plus, j'ai étudié les associations nouvellement mises en évidence par le consortium CHARGE [101] (voir **§IV.3.2.2**) : deux polymorphismes de *STAB*2 et un polymorphisme de chacun des gènes *VWF*, *STX*, *STXBP5*, *SCARA5*, *TC2N* et *CLEC4M*. Le seuil de Bonferroni de $1.12\ 10^{-7}$ n'avait plus de raison d'être ici, puisque je testais la réplication de gènes candidats. Le seuil $p<0.05$ retenu pour ces analyses se discute cependant, puisque 13 SNPs ont été testés. Par ailleurs, des tests unilatéraux auraient été parfaitement justifiés. En effet, l'hypothèse nulle d'absence de réplication n'a pas été rejetée lorsqu'un allèle était associé à une élévation des taux alors qu'il était associé dans la littérature à une diminution de ces derniers. Cependant, les tableaux présentent les valeurs de p de tests bilatéraux, plus conventionnels, particulièrement dans le domaine de la recherche clinique. Les résultats sont présentés dans le **tableau 25.A** pour vWF et **tableau 25.B** pour FVIII.

Les associations entre les taux de vWF et les polymorphismes des gènes *STXBP5* (rs9390459, p = 0,005), *VWF (*rs1063856*,* p = $1,30\ 10^{-4}$), *STX2* (rs4334059, p = 0,003), *TC2N* (rs2402074, p = 0,033), et *CLEC4M* (rs868875, p = 0,026). étaient reproduites au seuil p<5%. Elles permettaient d'expliquer 1,4% et 3,2% de la variance de vWF dans respectivement MARTHA08 et MARTHA10, après prise en compte des associations mises en évidence par l'approche pangénomique. Notons que l'effet fixe sur le taux de vWF du polymorphisme rs10866867 de *SCARA5* était également significatif (p = 0,015), mais qu'une hétérogénéité importante entre les études (I^2 = 71%) rendait l'effet aléatoire non significatif. Les associations avec les taux de FVIII ont été reproduites pour les polymorphismes des gènes *VWF* (rs1063856, p = 0,02) et *SCARA5* (rs9644133, p = 0,009). Elles expliquaient encore 0,7% et 0,2% de la variance des taux de FVIII dans MARTHA08 et MARTHA10. Finalement, si l'on tient compte du nombre de test réalisés (N = 17) et en considérant que nous faisons des tests unilatéraux, seules les associations avec *VWF (*rs1063856*), STXBP5 (*rs9390459*)*, et *STX2 (*rs4334059*)* étaient significatives (p<0.005).

Tableau 25 A– Polymorphismes associés aux taux de vWF dans des études antérieures : résultat d'une méta-analyse de trois études pangénomiques

Gène	SNP	Allèles*		MAF+	β (Ecart-Type)	p	I^2	p_{het}	Effet aléatoire β (Ecart-Type)	Effet aléatoire p	Effet fixe β (Ecart-Type)	Effet fixe p
			FVL	0,42	0,04 (0,08)	0,618						
BAI3	rs9363864	A/G	MARTHA08	0,52	0,03 (0,04)	0,421	0	0,838	0,02 (0,03)	0,461	0,02 (0,03)	0,461
			MARTHA10	0,49	-0,002 (0,05)	0,973						
			FVL	0,42	0,02 (0,08)	0,764						
	rs3798992	T/G	MARTHA08	0,44	-0,01 (0,04)	0,809	0	0,759	-0,02 (0,03)	0,595	-0,02 (0,03)	0,595
			MARTHA10	0,46	-0,05 (0,05)	0,401						
			FVL	0,43	-0,08 (0,08)	0,366						
STXBP5[101]	rs9390459	**G**/A	MARTHA08	0,42	-0,06 (0,04)	0,197	0	0,545	-0,09 (0,03)	0,005	-0,09 (0,03)	**0,005**
			MARTHA10	0,43	-0,13 (0,05)	0,011						
			FVL	0,20	-0,08 (0,10)	0,446						
SCARA5[101]	rs10866867(1)	G/**T**	MARTHA08	0,25	0,17 (0,05)	$4,88\ 10^{-4}$	0,71	0,03	0,05 (0,07)	0,466	0,09 (0,04)	**0,015**
			MARTHA10	0,25	0,01 (0,06)	0,830						
			FVL	0,06	-0,28 (0,19)	0,141						
[100]	rs216335(2)	**G**/A	MARTHA08	0,08	-0,23 (0,08)	0,003	0	0,945	-0,23 (0,06)	$1,31\ 10^{-4}$	-0,23 (0,06)	**$1,30\ 10^{-4}$**
			MARTHA10	0,06	-0,21 (0,11)	0,059						
			FVL	0,45	0,07 (0,08)	0,371						
VWF[101]	**rs1063856**(3)	A/**G**	MARTHA08	0,37	0,08 (0,05)	0,094	0	0,889	0,09 (0,03)	0,006	0,09 (0,03)	**0,006**
			MARTHA10	0,38	0,11 (0,05)	0,041						
			FVL	0,48	-0,04 (0,08)	0,612						
[100]	rs7306706	**A**/G	MARTHA08	0,45	0,02 (0,04)	0,634	0	0,754	0,01 (0,03)	0,664	0,01 (0,03)	0,664
			MARTHA10	0,46	0,03 (0,05)	0,604						
			FVL	0,30	-0,05 (0,09)	0,601						
STAB2[101]	rs4981022	**T**/C	MARTHA08	0,30	0,02 (0,05)	0,652	0	0,541	-0,01 (0,03)	0,664	-0,01 (0,03)	0,664
			MARTHA10	0,28	-0,06 (0,06)	0,333						
			FVL	0,33	0,01 (0,09)	0,863						
STX2[101]	rs4334059(4)	C/**T**	MARTHA08	0,37	0,08 (0,04)	0,067	0,01	0,363	0,1 (0,03)	0,004	0,1 (0,03)	**0,003**
			MARTHA10	0,36	0,15 (0,06)	0,008						
			FVL	0,52	0,05 (0,08)	0,548						
TC2N[101]	rs2402074(5)	G/**A**	MARTHA08	0,48	0,04 (0,04)	0,382	0	0,509	0,07 (0,03)	0,033	0,07 (0,03)	**0,033**
			MARTHA10	0,47	0,12 (0,05)	0,030						
			FVL	0,22	-0,07 (0,10)	0,515						
CLEC4M[101]	rs868875	**A**/G	MARTHA08	0,32	-0,10 (0,05)	0,036	0	0,762	-0,08 (0,03)	0,026	-0,08 (0,03)	**0,026**
			MARTHA10	0,35	-0,05 (0,06)	0,424						

25.B – Polymorphismes associés aux taux de FVIII dans des études antérieures : résultat d'une méta-analyse de trois études pangénomiques

Gene	SNP	Alleles*		MAF+	β$ (Ecart-Type)	p	I^2	p_{het}	Effet Aléatoire β (Ecart-Type)	p	Effet Fixe β (Ecart-Type)	p
STXBP5[101]	rs9390459	**G**/A	FVL	0,43	0,15 (0,08)	0,083	0,65	0,059	-0,02 (0,06)	0,795	-0,04 (0,03)	0,310
			MARTHA08	0,42	-0,08 (0,06)	0,158						
			MARTHA10	0,43	-0,07 (0,05)	0,199						
SCARA5[101]	rs9644133	**C**/T	FVL	0,24	-0,08 (0,10)	0,433	0	0,753	-0,12 (0,05)	0,009	-0,12 (0,05)	**0,009**
			MARTHA08	0,17	-0,16 (0,07)	0,029						
			MARTHA10	0,18	-0,10 (0,07)	0,152						
VWF[101]	rs1063856	A/**G**	FVL	0,45	0,11 (0,08)	0,170	0	0,843	0,08 (0,03)	0,020	0,08 (0,03)	**0,020**
			MARTHA08	0,37	0,09 (0,06)	0,114						
			MARTHA10	0,38	0,06 (0,05)	0,249						
STAB2[101]	rs4981021(6)	G/**A**	FVL	0,27	-0,13 (0,09)	0,146	0	0,389	-0,02 (0,04)	0,521	-0,02 (0,04)	0,521
			MARTHA08	0,32	-0,02 (0,06)	0,737						
			MARTHA10	0,29	0,02 (0,06)	0,782						
LDLR [110]	rs2228671	C/**T**	FVL	0,14	-0,03 (0,11)	0,816	0,46	0,157	-0,01 (0,07)	0,890	-0,01 (0,05)	0,894
			MARTHA08	0,11	0,11 (0,09)	0,193						
			MARTHA10	0,10	-0,13 (0,09)	0,161						
[110]	rs688	C/**T**	FVL	0,38	-0,25 (0,09)	0,005	0,79	0,010	-0,05 (0,08)	0,531	-0,02 (0,03)	0,652
			MARTHA08	0,45	0,06 (0,05)	0,235						
			MARTHA10	0,45	-0,007 (0,05)	0,901						

*Allèle commun/rare ; l'allèle associé une élévation des taux de FVIII dans une étude antérieure est en gras

+ Fréquence de l'allèle rare

$ Effet de l'allèle rare

(1) proxy pour rs2726953 (r^2=0.92), (2) proxy pour rs216318 (r^2=1), (3) proxy pour rs1063857 (r^2=1), (4) proxy pour rs7978987 (r^2=1), (5) proxy pour rs10133762 (r^2=0.96), (6) proxy pour rs12229292 (r^2=0.88), Pas de proxy pour *VWF*-rs4764478

VIII.2.5. Influence sur le risque de MTEV des polymorphismes mis en évidence par cette méta-analyse

J'ai étudié l'influence sur le risque de MTEV des polymorphismes mis en évidence par mes travaux de méta-analyse en utilisant les données *in silico* de l'étude *GWAS* (voir **§III.3.2** p18-19). Lorsque ces polymorphismes n'étaient pas présents dans cette étude, j'ai recherché le meilleur proxy à l'aide du programme SNAP [172]. Je n'ai malheureusement pas pu identifier de proxy (avec $r^2 > 0.8$) pour les polymorphismes rs6708166 (*LBH*), rs1321761 (*FAM46A*) et rs7306642 (*STAB2*). Les résultats de l'association des neuf autres polymorphismes avec le risque de MTEV sont indiqués dans le **tableau 26**. Les valeurs p correspondent à des tests bilatéraux bien que, là encore, des tests unilatéraux seraient parfaitement légitimes. J'ai en effet considéré que les associations entre un allèle donné et une augmentation du risque de MTEV étaient cohérentes avec nos hypothèses seulement si cet allèle était associé à une augmentation des taux de vWF et/ou FVIII. Deux polymorphismes, *VPS8*-rs4686760 et *ACCN1*-rs12941510, tendaient à être associés au risque de MTEV (p = 0,10 et p = 0,05, respectivement), dans un sens cohérent avec leurs associations aux taux de FVIII et vWF.

Tableau 26 - Influence sur le risque de MTEV des polymorphismes mis en évidence par une méta-analyse des taux de vWF et FVIII.

		Allèles*	**MAF**		**p**
			Cas	Témoins	**(Cochran Armitage)**
Polymorphismes associés aux taux de vWF					
VPS8	rs4686760	**A**/G	0,441	0,475	0,101
EPB41L4A	rs13361927	G/**A**	0,065	0,062	0,797
KRT18P24	rs1634352†	A/**G**	0,284	0,318	0,055
12q2.33	rs1438933	G/**A**	0,256	0,294	0,051
CDH2	rs2298574	**A**/G	0,084	0,093	0,444
SAFB2	rs732505	G/**A**	0,061	0,064	0,713
Polymorphismes associés aux taux de FVIII					
VAV2	rs12344583	A/**G**	0,217	0,193	0,133
ACCN1	rs1354492	G/**A**	0,476	0,469	0,740
ACCN1	rs12941510	**G**/A	0,310	0,350	0,046

*Allèle commun/rare ; l'allèle associé à une élévation des taux de vWF ou FVIII dans la méta-analyse est en gras
† proxy pour rs1757948 ($r^2 = 1$).

VIII.2.6. Discussion

Ma recherche à travers l'ensemble du génome de nouveaux polymorphismes associés aux taux de FVIII et de vWF, au moyen d'une méta-analyse rassemblant un total de 1624 personnes, n'a révélé aucune association significative au seuil $p < 1{,}12\ 10^{-7}$, correspondant à la correction de Bonferroni. Une telle taille d'échantillon de sujets indépendants a une puissance de 95% de détecter , au seuil de $p<1{,}12.10^{-7}$, l'effet d'un SNP expliquant au moins 3% de la variabilité d'un trait quantitatif [173]. Cette puissance serait encore de 86% et 66% pour un SNP expliquant respectivement 2.5% et 2% de la variabilité, mais tomberait à 10% pour un SNP expliquant 1% de la variabilité. Ainsi, s'il existe des SNPs, tagués par l'un des SNPs de la puce Illumina et influençant les taux de FVIII et vWF, alors ils n'expliqueraient qu'une part infime de la variabilité de FVIII et vWF.

Parmi les onze gènes qui présentaient des polymorphismes dont la significativité de l'association avec l'un ou l'autre phénotype était inférieure à 10^{-5}, il était particulièrement intéressant de noter la présence de *STAB2***.** En premier lieu, la révélation de ce gène par la méta-analyse a permis de préciser l'origine probable du signal d'association observé en 12q23 avec l'échantillon des Familles-FVL (**VIII.1.5. p108**). Mais surtout, elle ajoute une nouvelle observation indépendante d'une association entre l'un des phénotypes étudiés et un polymorphisme de *STAB2*. Ce nouveau polymorphisme était en faible déséquilibre de liaison avec les autres polymorphismes de *STAB2* décrits au cours de cette thèse (deux SNPs associés aux taux de vWF et FVIII dans l'étude du consortium CHARGE [101], et un autre, associé au risque de MTEV [158]. Cette convergence de résultats encourage à réaliser une étude approfondie de l'ensemble des polymorphismes de *STAB2*, grâce au séquençage du gène, suivie d'une étude du déséquilibre de liaison et des effets haplotypiques. Ce travail aurait pour objectif d'identifier de manière exhaustive les polymorphismes potentiellement responsables de l'association, avant d'envisager des études de fonctionnalité biologique.

Les associations avec les taux de vWF et de FVIII dont la significativité était inférieure à 10^{-5} correspondent à un FDR respectivement de 29% et 62%. En dehors de *STAB2*, il s'agissait d'associations nouvellement décrites. Compte-tenu du faible niveau de preuve statistique et de l'absence d'échantillon de réplication, seuls des arguments en faveur d'une plausibilité biologique pourraient conforter l'un ou l'autre de ces résultats. En cela, le gène *VPS8* (Vacuolar Protein Sorting 8 homolog gene) semble peut-être mériter une attention particulière. Il intervient dans le transport intracellulaire de protéines [174]. Il pourrait participer à la régulation de l'urokinase, un activateur du plasminogène [175], élément majeur

de la physiologie de l'hémostase (**voir figure 3** et **tableau 1 p7**). Parmi les autres gènes décrits, *LBH* et *VAV2* pourraient tous deux jouer un rôle dans l'angiogénèse, comme le gène *BAI3*. *LBH* (Limb-Bud-and-Heart) est un cofacteur de transcription dont la surexpression a pour conséquence une diminution importante de l'expression de *VEGF* [176] (Vascular Endothelial Growth Factor), lui-même étant un facteur de croissance de l'endothélium vasculaire. Quant à *VAV2*, il s'agit d'un échangeur de nucléotide, dont l'expression supprime celle de *VEPTP* (Vascular Endothelial-Protein Tyrosine Phosphatise) [177], modifiant ainsi la régulation de l'angiogénèse. Le lien avec les taux de FVIII et de VWF est certes éloigné. Cependant, la coïncidence entre les fonctions connues de ces deux gènes et de *BAI3* est notable.

Parmi les polymorphismes décrits dans cette section, neuf ont pu faire l'objet d'une étude de leur éventuelle répercussion sur le risque de MTEV. Seuls deux (*VPS8*-rs4686760 et *ACCN1*-rs12941510) tendaient à être associés au risque de MTEV. Ceci n'est pas incompatible avec l'hypothèse de travail de ma thèse. En effet, la recherche de facteurs génétiques influençant des phénotypes intermédiaires de la MTEV avait pour but de pallier le manque de puissance d'une étude étudiant directement la MTEV. Ces deux associations méritent d'être validées par l'étude d'un échantillon indépendant, d'autant plus que ni l'une ni l'autre n'était significative, si l'on tenait compte de la correction de Bonferroni pour les tests multiples.

Enfin, cette méta-analyse a pu confirmer le rôle de plusieurs polymorphismes décrits dans la littérature car associés soit aux taux de vWF soit aux taux de FVIII. Parmi eux, trois étaient toujours significatifs dans cette méta-analyse après application de la correction de Bonferroni pour dix-sept tests. Ces polymorphismes étaient situés dans les gènes *VWF* (rs216335), *STXBP5* (rs9390459) et *STX2* (rs4334059). Il était très intéressant de reproduire le polymorphisme de *VWF*, bien que plusieurs études aient déjà montré de façon robuste l'influence de plusieurs polymorphismes de *VWF* (voir **tableau 3 p30**). En effet, rs216318, un proxy de rs216335, avait été découvert par une approche systématique explorant l'intégralité des polymorphismes de *VWF*. Il atteignait juste la significativité de 5%, de sorte qu'il n'était plus significatif après correction pour tests multiples [100]. A ma connaissance, il n'avait pas encore fait l'objet d'une étude de réplication. Quant aux récentes découvertes du consortium CHARGE, elles étaient presque toutes robustes puisqu'elles avaient pu être confirmées dans l'échantillon de réplication prévu lors de la conception du projet [101]. Seule l'association

observée avec le polymorphisme de *STX2* n'avait pas pu être reproduite. Ainsi, le travail présenté dans cette section a permis d'accréditer cette découverte.

En conclusion, la méta-analyse réalisée à partir des Familles-FVL et des échantillons MARTHA08 et MARTHA10 n'avait pas la puissance nécessaire pour mettre en évidence de manière significative de nouvelles associations avec les taux de vWF et FVIII. Cependant, elle a permis de confirmer l'association avec les taux de vWF d'un polymorphisme de *VWF* qui n'avait encore jamais fait l'objet d'étude de réplication. De même, elle corrobore la participation des gènes de *STAB2, STXBP5* et *STX2* à la variabilité des taux de vWF et FVIII.

DISCUSSION GENERALE :

BILAN, PERSPECTIVES, CONCLUSION

L'objectif de ma thèse était de contribuer à la recherche de nouveaux facteurs de risque génétique de la Maladie ThromboEmbolique Veineuse (MTEV) en étudiant deux de ses phénotypes intermédiaires : le FVIII, et sa protéine de transport, le vWF. La stratégie adoptée était une approche pangénomique sans *a priori* biologique. Deux voies méthodologiques ont été suivies. La première reposait sur des *analyses de liaison pangénomiques* réalisées dans un échantillon composé de cinq grandes familles franco-canadiennes. Elles ont été suivies par l'exploration d'un des signaux de liaison au moyen d'études d'association menées dans un échantillon de familles nucléaires et d'études cas-témoins. La deuxième reposait sur des *analyses d'association pangénomiques (GWAS)* réalisées dans trois échantillons : l'échantillon des grandes familles franco-canadiennes, et deux échantillons de personnes non apparentées ayant un antécédent de MTEV. La multiplicité des échantillons et la diversité de leur mode de recrutement, tout en étant une source de difficultés pour les étapes indispensables de réplication des résultats, étaient un gage de robustesse en cas de succès de ces dernières.

Bilan et perspectives en vue d'étudier plus finement les deux gènes découverts au cours de ce travail, *BAI3* et *STAB2*

Ce travail a permis la découverte de deux gènes, *STAB2* et *BAI3*, qui répondaient parfaitement à l'objectif de ce travail. En effet, tous deux présentaient, d'une part, des polymorphismes associés aux taux de vWF et/ou de FVIII, et d'autre part, des polymorphismes associés à la MTEV. Chacune de ces associations a pu être constatée dans au moins deux échantillons indépendants. Les observations concernant *STAB2* sont d'autant plus convaincantes qu'elles ont été faites par des équipes indépendantes et sont désormais décrites dans plusieurs publications [101][178][158]. Les associations de certains polymorphismes de *BAI3* avec les taux de vWF et le risque de MTEV ont été observées dans plusieurs échantillons indépendants. Néanmoins, les modèles nécessaires pour révéler ces associations nécessitaient la prise en compte d'une interaction avec le groupe ABO qui n'était pas toujours cohérente entre les études. Ainsi, l'acceptation de l'hypothèse que *BAI3* participe au risque de MTEV par l'intermédiaire d'une modulation des taux de vWF impose d'envisager des mécanismes physiopathologiques particulièrement complexes. Il se pourrait en particulier que *BAI3* joue un rôle de gène modulateur de la MTEV en favorisant le développement d'une sous-entité clinique particulière. Dans cette perspective, une étude visant à rechercher des facteurs favorisant le développement d'une embolie pulmonaire chez des patients atteints de MTEV est menée actuellement par Marine Germain sous la direction David-Alexandre

Trégouët. Elle inclut six échantillons GWAS, rassemblant 3225 patients ayant présenté une MTEV dont 1103 avec embolie pulmonaire et dont la méta-analyse identifie, de façon intéressante, plusieurs associations au sein du gène *BAI3* avec des valeurs de $p < 6\ 10^{-5}$.

La possibilité d'un mécanisme pléïotropique des polymorphismes de *BAI3* et de *STAB2* a été discutée au cours de cette thèse. La pléïotropie pourrait en effet expliquer l'importance relative de l'influence sur le risque de MTEV de ces polymorphismes comparativement à la faible part de la variance des taux de vWF expliquée par ceux-ci. Il se pourrait ainsi que *BAI3* et *STAB2* soient impliqués dans la modulation de multiples phénotypes intermédiaires de la MTEV. Une explication supplémentaire, et non incompatible avec la précédente, serait que les polymorphismes de *BAI3* et *STAB2* déterminent les valeurs particulièrement extrêmes des phénotypes intermédiaires, qui sont fortement à risque de MTEV. Afin d'avancer dans cette hypothèse, il conviendrait d'étudier les associations entre ces polymorphismes et les taux plasmatiques des nombreux facteurs impliqués dans l'hémostase et mesurés dans les échantillons STANISLAS et MARTHA. Un écart à la linéarité pourra être recherché en considérant différentes classes ordonnées constituées à partir des taux plasmatiques.

Toujours en vue d'accréditer l'hypothèse de pléïotropie, il serait intéressant d'évaluer l'effet résiduel de *BAI3* et *STAB2* sur le risque de MTEV, en ajustant sur les taux de FVIII et/ou vWF. Je n'ai pas pu réaliser de telles analyses durant ma thèse. En effet, aucun des échantillons cas-témoins que j'ai analysés ne disposaient des mesures des taux de FVIII et vWF à la fois chez les cas et les témoins. Un nouvel échantillon composé de MARTHA08 et MARTHA10, pour les cas, et de sujets issus de la cohorte des Trois Cités, pour les témoins, pourrait permettre d'appréhender cette question, puisque la cohorte des Trois Cités dispose d'une banque de plasma à laquelle il est possible d'avoir accès. On pourrait alors envisager l'étude de modèles structuraux, dans lesquels chaque variable peut être considérée à la fois comme une variable « à expliquer » et « explicative ». Ces modèles pourraient inclure les polymorphismes à l'étude, le groupe ABO, les mutations FVL et FII, les taux de FVIII, les taux de vWF, éventuellement d'autres phénotypes intermédiaires de la MTEV, et enfin le statut cas-témoin. On pourrait peut-être tenter d'élaborer ainsi un modèle biologique compatible avec les observations statistiques.

Une question reste en suspens : les signaux de liaison en 6q13-14 et en 12q23 qui nous ont mis originellement sur la voie de *BAI3* et *STAB2* ont-ils été générés, partiellement ou entièrement, par la présence d'associations d'un ou plusieurs polymorphismes de ceux-ci avec

les phénotypes étudiés ? Une réponse négative ne remettrait aucunement en cause les conclusions concernant ces gènes, mais inciterait cependant à poursuivre les recherches afin de découvrir un autre gène dont les polymorphismes participeraient eux aussi à la modulation des taux de vWF et FVIII. Une réponse par l'affirmative apporterait une satisfaction d'esprit certainement appréciable et réconfortante. J'ai tenté de répondre à cette question en ré-estimant la force du signal de liaison en introduisant dans le modèle, un à un, tous les génotypes de *BAI3* génotypés par la puce à ADN. Je n'ai observé aucune modification notable du Facteur Bayésien qui mesure la force du signal de liaison. Cela peut laisser présager que l'origine du signal de liaison doit être recherchée ailleurs que dans le gène *BAI3*. Un travail analogue pourrait être entrepris avec la région 12q23 en étudiant des modèles ajustés sur les polymorphismes de *STAB2* associés aux taux de vWF. Cependant, ce travail pourrait s'avérer ardu. En effet, comme il a été discuté au cours de cette thèse, il est extrêmement vraisemblable que le variant fonctionnel de *STAB2* soit une configuration haplotypique particulière ou un variant rare. Pour cette raison, il pourrait être pertinent de réaliser un séquençage du gène *STAB2*, à partir d'individus présentant des valeurs extrêmes de taux plasmatiques de FVIII et/ou vWF.

Ce dernier point est bien sûr également valable pour *BAI3*, mais ni le signal de liaison, ni les associations ne m'apparaissent aussi robustes que celles de *STAB2*. Il pourrait donc être prématuré d'entreprendre le séquençage de *BAI3*. La relative faiblesse du signal de liaison en 6q13-14 pourrait être expliquée par une hétérogénéité génétique. Cette dernière, éventuellement favorisée par la rareté du polymorphisme causal, aurait comme conséquence qu'une seule des familles contribuerait à la liaison en 6q13-14. Afin d'explorer cette hypothèse, il faudrait estimer la force de la liaison apportée par chaque famille individuellement.

Rappelons enfin que le signal de liaison en 6q13-14 avait été obtenu par l'étude des taux de FVIII ajustés sur les taux de vWF. Ce modèle était justifié par des connaissances biologiques : la protéolyse du FVIII est diminuée lorsque celui-ci est porté par le vWF. On observait d'ailleurs dans nos données de très fortes corrélations entre les taux de vWF et de FVIII. Il était donc utile d'étudier ce modèle afin de rechercher les facteurs génétiques influençant les taux de FVIII indépendamment des taux de vWF. Dans cette même optique, il serait intéressant d'étudier si le QTL en 6q13-14 influence également le phénotype constitué du rapport *taux de FVIII / taux de vWF*. Ce phénotype pourrait en effet être plus proche d'une certaine réalité biologique [179][180]. L'étude de ce rapport pourrait permettre de déterminer

les facteurs expliquant pourquoi les taux de FVIII peuvent être très élevés chez certaines personnes alors qu'elles présentent de faibles taux de vWF (et inversement). Il semblerait que des taux élevés de FVIII associés à un rapport *FVIII / taux de vWF* proche de 1 soit le reflet d'une diminution de la clairance du complexe FVIII-vWF, tandis que des taux élevés de FVIII associés à un rapport *taux de FVIII / taux de vWF* élevé soit le reflet d'une synthèse accrue de FVIII [181].

Bilan et perspectives des analyses de liaison et d'association pangénomiques des taux plasmatiques de vWF et de FVIII

J'ai étudié de nombreux modèles incluant les taux de FVIII et vWF (taux de vWF, taux de FVIII, taux de FVIII ajustés sur les taux de vWF, chacun de ces modèles ayant été ajusté ou non sur les polymorphismes du gène ABO). On pourrait entreprendre de nouvelles analyses pangénomiques de liaison et d'association à partir du rapport *taux de FVIII / taux de vWF*, comme il a été discuté précédemment. Il pourrait également être intéressant de réaliser des analyses bivariées. Ces dernières recherchent des QTLs liés ou associés à la fois aux taux de FVIII et de vWF considérés de manière conjointe par le modèle. Cependant, la faisabilité de telles analyses sera certainement un facteur limitant. En effet, ce type de modélisation n'est implémenté dans aucun des logiciels que j'ai utilisés pour ce travail. Elle pourrait éventuellement s'adapter à la méthode de décomposition de la variance telle qu'elle est implémentée dans le logiciel SOLAR.

Un autre ajustement pourrait être pris en considération. En effet, je n'ai pas tiré partie au cours de ma thèse du recrutement *via* une mutation FVL des Familles-FVL. Il serait très intéressant d'étudier la présence d'une interaction entre les QTLs et la mutation FVL. Des analyses d'association pangénomiques avec le logiciel SOLAR (pour l'étude des Familles-FVL) ou PLINK (pour l'étude des échantillons MARTHA) pourraient être réalisées aisément en ajoutant une covariable égale au produit du code de la mutation FVL et du code (0, 1 ou 2 pour un modèle additif) de chacun des 500 000 SNPs. Concernant les analyses de liaison pangénomiques, il n'est pas possible à l'heure actuelle de modéliser une interaction entre un QTL et une covariable. Il serait possible cependant d'appréhender cette question de manière indirecte. En effet, on pourrait regarder de quelle manière est modifié un signal de liaison après prise en compte d'une interaction entre un polymorphisme situé dans le signal de liaison et la mutation FVL. Cela sous-entend bien sûr d'avoir une hypothèse sur l'origine du signal

de liaison. Par exemple, on pourrait étudier le signal en 6q13-14 après ajustement sur les génotypes de *BAI3,* la mutation FVL, et le produit des deux.

Enfin, je n'ai pas étudié les chromosomes X et Y. En particulier, le chromosome X contient le gène structural du FVIII. Les régions fonctionnelles de ce gène avaient fait l'objet d'une recherche exhaustive de polymorphismes parmi lesquels un était associé aux taux de FVIII [102]. Là encore, les analyses d'association pourraient être réalisées assez facilement. Par contre, un développement méthodologique serait nécessaire avant d'envisager de telles analyses de liaison dans l'échantillon de Familles-FVL.

Bilan et perspectives méthodologiques en vue d'optimiser les prochaines analyses de liaison et d'association de l'échantillon de grandes familles (Familles-FVL)

Ce travail m'a permis de comparer l'intérêt respectif des analyses de liaison et d'association. Depuis l'avènement du génotypage à haut débit à l'origine des puces à ADN, les *GWAS* tendent à supplanter les approches fondées sur des analyses de liaison. Cependant, la complémentarité des deux méthodes apparaît clairement dans ce travail. En particulier, le caractère très discriminant des analyses de liaison est remarquable. En effet, les analyses de liaison ont permis de mettre en lumière quelques signaux très significatifs qui tranchaient nettement avec le bruit de fond observé ailleurs dans le génome. Au contraire, les signaux d'association obtenus dans la GWAS, bien que localisés de manière plus précise sur le génome, avaient peine à s'extraire du bruit de fond et n'étaient pas significatifs après prise en compte d'une correction pour tests multiples. Le signal de liaison obtenu en 9q34, centré autour du gène *ABO*, et hautement significatif, était d'autant plus intéressant qu'il est parfois reproché aux analyses de liaison ne pas être adaptées à la découverte de polymorphismes fréquents. Il était par contre assez décevant de ne pas pouvoir observer, dans les Familles-FVL d'association significative avec l'un des polymorphismes du gène *ABO*. Une raison à cela est bien sûr que l'échantillon familial, pour des raisons sur lesquelles je reviendrai, n'est pas optimal pour de telles analyses. Cependant, ces considérations sont également valables pour les échantillons MARTHA qui ne présentaient aucune association significative après ajustement sur ABO contrairement à l'échantillon des familles-FVL qui a permis de mettre en évidence un signal de liaison particulièrement net et significatif en 12q23.

Analyses de liaison

La méthode bayésienne d'analyse de liaison par MCMC implémentée dans Loki utilisée pour ce travail présente l'avantage de modéliser au plus près la composante oligogénique d'un phénotype quantitatif. De plus, elle requiert un temps de calcul raisonnable. Elle reste cependant une méthode rarement choisie par les équipes travaillant sur des données familiales. Le manque d'expérience partagée par les quelques utilisateurs de cette méthode a pour conséquence regrettable que l'importance des signaux de liaison générés est difficilement appréciable. Les premiers temps, la seule échelle de mesure dont je disposais était le degré d'enthousiasme de France Gagnon à la vue des diverses figures que je lui présentais. Le calcul du Facteur Bayésien (BF) permet toutefois de comparer entre eux les signaux de liaison. Mais, contrairement au Lod-Score calculé par les méthodes fondées sur le Maximum de Vraisemblance, le BF ne permet pas de déterminer un intervalle de confiance de la localisation du signal de liaison. J'ai donc délimité les signaux de liaison de manière particulièrement large afin d'être sûre de ne pas négliger le locus causal. Cela aurait pu induire une importante perte de puissance si la stratégie d'exploration des signaux de liaison choisie avait été un criblage complet des QTL.

Des analyses supplémentaires seraient nécessaires pour progresser sur cette question. En effet, il est désormais possible de mener des analyses de liaison à partir de SNPs indépendants [182][183] génotypés au moyen de biopuces. Il serait particulièrement intéressant de confronter les résultats des analyses de liaison que j'ai obtenus à partir de l'information apportée par un millier de microsatellites à ceux qui seraient produits grâce à une sélection des SNPs génotypés dans l'échantillon FVL. Peut-être cette analyse permettrait-elle d'affiner les signaux de liaison que j'ai observés? De plus, les méthodes fondées sur le Maximum de Vraisemblance avaient initialement été écartées en raison de leur exigence en termes de temps de calcul, dans le cadre de grandes familles aussi complexes que celles des Familles-FVL. L'ajout de plusieurs dizaines de CPU au cluster d'ordinateurs du laboratoire de France Gagnon permet maintenant d'utiliser la méthode de décomposition de la variance telle qu'elle est implémentée dans le logiciel SOLAR. Il serait très intéressant de confronter les figures des signaux de liaison obtenus par la méthode MCMC avec l'intervalle de confiance de la localisation des signaux de liaison déterminé par le calcul de Lod-Scores.

Analyses d'association

L'échantillon de Familles-FVL avait été conçu pour la réalisation d'analyse de liaison. Il a toutefois permis de mener des analyses d'association. La relativement faible puissance des analyses, due notamment à l'absence d'indépendance des observations et peut-être à la nécessaire transformation par quantile de loi normale, a déjà été discutée dans ce mémoire, de même que l'imprécision de localisation des signaux d'association en raison du déséquilibre de liaison sur de grandes distances. Ce dernier, dont la réalité a bien été constatée, a été incriminé sans preuve réelle d'avoir provoqué la déviation de la statistique de test par rapport à sa distribution attendue sous l'hypothèse nulle d'absence d'association. Des simulations sous diverses hypothèses nulles et alternatives seraient donc nécessaires afin d'étudier le comportement de la statistique utilisée, notamment en présence de liaison et en absence d'association. En cas d'inflation, il serait intéressant d'évaluer le nombre de fondateurs et/ou le nombre de générations à même de faire disparaître ce phénomène. La présence d'une famille bien plus grande que les autres est peut-être responsable d'une accentuation de l'inflation. Là-encore des simulations permettraient de tester cette hypothèse. Si une correction de l'inflation de la statistique de tests s'avérait nécessaire, alors le calcul d'une valeur de p empirique, obtenu par permutations selon la méthode déjà suivie pour les analyses de liaison, serait probablement plus pertinent qu'une correction par le facteur d'inflation λ_{50} calculé au $50^{ème}$ percentile.

La raison habituellement invoquée en cas d'inflation de la statistique de test dans le cadre des GWAS est la présence d'une stratification de population. Celle-ci nécessite une correction par le facteur d'inflation λ_{50}. Il est peu vraisemblable que les familles franco-canadiennes proviennent de deux sous-populations. Il faut garder toutefois à l'esprit que si des tests d'associations en familles ont été développés afin de contrôler le biais induit par une stratification de population, la méthode par décomposition de la variance que j'ai utilisée n'en fait pas partie. Le test du QTDT proposé par Abecasis et al pour l'analyse de familles nucléaires [184] et développé secondairement pour des familles de structures quelconques [185] constitue une alternative intéressante. Non seulement le calcul d'une association intrafamiliale permet de contrôler un éventuel biais de stratification, mais de plus il est possible d'obtenir une p-value empirique par permutations. Cependant, il faudrait que le programme implémenté dans le logiciel QTDT soit retravaillé de façon importante afin d'être à même d'intégrer une famille dont la structure est aussi complexe que la famille 5.

Enfin, une approche intégrant à la fois les informations apportées par la liaison et l'association a déjà été évoquée. Il s'agit des méthodes de FDR pondéré (WFDR) ou stratifié (SFDR). Ces méthodes contrôlent efficacement le risque global de première espèce à condition toutefois que l'échantillon ayant généré les hypothèses *a priori* à l'origine de la constitution des poids ou des strates soit indépendant de l'échantillon sur lesquels ces derniers seront appliqués. En effet, dans le cas particulier des analyses de liaison et d'association, il ne faudrait pas qu'un même phénomène artéfactuel puisse créer à la fois un signal de liaison et d'association. La puissance de cette stratégie est déterminée par la pertinence du choix des strates ou des poids. Le gain de puissance est en effet d'autant plus important que les associations réelles ont généré un fort signal de liaison. Toutefois, dans le cas inverse, la puissance ne sera pas diminuée par l'application d'un SFDR comparativement à un FDR classique.

Conclusion

Les analyses présentées dans ce mémoire étaient les premières réalisées à partir des données de l'échantillon des Familles-FVL. Grâce aux concours des échantillons de familles nucléaires et cas-témoins, elles ont abouti à la découverte de deux gènes susceptibles d'augmenter le risque de MTEV par l'intermédiaire, au moins en partie, d'une modulation des taux plasmatiques de vWF et de FVIII. De nombreux autres phénotypes intermédiaires de la MTEV ont été mesurés dans ces échantillons. L'expérience acquise au cours de ce travail pourra éventuellement permettre d'optimiser les prochaines études. Pour les analyses de liaison, les utilisations des SNPs et d'une méthode par décomposition de la variance pourraient peut-être permettre de mieux définir les limites des signaux de liaison. Pour les analyses d'association, le contrôle de biais éventuels pourrait être obtenu éventuellement (après modification du programme « QTDT ») par la méthode du QTDT. Alternativement, il pourrait être intéressant de réaliser un calcul de valeurs de p empiriques par permutations. Enfin, l'augmentation de la puissance de détection de nouvelles associations pourrait être obtenue au moyen de l'utilisation d'un FDR pondéré ou stratifié intégrant aux analyses d'association l'information apportée par les analyses de liaison.

Une très faible part de l'héritabilité des taux de FVIII et de vWF a pu être expliquée par les nouveaux polymorphismes décrits au cours de mon travail (*STAB2*-rs7306643, *STXBP5*-rs9390459, *STX2*-4334059, et *VWF*-216335, pour les associations les plus robustes). La part de la variabilité des taux de FVIII et vWF attribuable à des facteurs génétiques encore

inconnus est cependant estimée aux alentours de 20%. Une part de cette héritabilité pourrait être expliquée par des facteurs génétiques non modélisés par les méthodes que j'ai employées, telles des interactions gène-gène (notamment avec la mutation FVL) ou gène-environnement. La présence de mécanismes épigénétiques, comme les variations dans le taux de méthylation de l'ADN, pourrait également contribuer à expliquer une partie de l'héritabilité non expliquée. Il se peut aussi celle-ci soit due à un très grand nombre de polymorphismes aux effets tellement faibles que nos échantillons ne seraient pas à même de les détecter. Dans quelle mesure faut-il augmenter les tailles des échantillons en vue de découvrir de tels polymorphismes ? Les applications cliniques seront en effet nulles en routine, sans parler des applications « sauvages » et lucratives de ce genre d'information, qui proposent un pronostic personnalisé du risque de développer de graves maladies. En particulier, d'après une étude réalisée par un chercheur post doctoral de l'UMR_S 937, la valeur prédictive de MTEV apportée par l'ensemble des polymorphismes connus pour influencer les phénotypes intermédiaires de la MTEV est très faible. En revanche, ces découvertes participent à une meilleure connaissance de la physiopathologie de la MTEV, élément indispensable au développement de nouvelles thérapeutiques, qui seraient idéalement plus efficaces tout en réduisant les risques iatrogènes hémorragiques.

BIBLIOGRAPHIE

[1] P. O. Hansson, L. Welin, G. Tibblin, et H. Eriksson, « Deep vein thrombosis and pulmonary embolism in the general population. "The Study of Men Born in 1913" », *Archives of Internal Medicine*, vol. 157, n°. 15, p. 1665-1670, août. 1997.
[2] F. A. Anderson Jr et al., « A population-based perspective of the hospital incidence and case-fatality rates of deep vein thrombosis and pulmonary embolism. The Worcester DVT Study », *Archives of Internal Medicine*, vol. 151, n°. 5, p. 933-938, mai. 1991.
[3] M. D. Silverstein, J. A. Heit, D. N. Mohr, T. M. Petterson, W. M. O'Fallon, et L. J. Melton 3rd, « Trends in the incidence of deep vein thrombosis and pulmonary embolism: a 25-year population-based study », *Archives of Internal Medicine*, vol. 158, n°. 6, p. 585-593, mars. 1998.
[4] M. Cushman et al., « Deep vein thrombosis and pulmonary embolism in two cohorts: the longitudinal investigation of thromboembolism etiology », *The American Journal of Medicine*, vol. 117, n°. 1, p. 19-25, juill. 2004.
[5] R. H. White, « The Epidemiology of Venous Thromboembolism », *Circulation*, vol. 107, n°. 23 suppl 1, p. I-4 -I-8, juin. 2003.
[6] S. Murin, P. S. Romano, et R. H. White, « Comparison of outcomes after hospitalization for deep venous thrombosis or pulmonary embolism », *Thrombosis and Haemostasis*, vol. 88, n°. 3, p. 407-414, sept. 2002.
[7] P. Prandoni et al., « The long-term clinical course of acute deep venous thrombosis », *Annals of Internal Medicine*, vol. 125, n°. 1, p. 1-7, juill. 1996.
[8] J. A. Heit, D. N. Mohr, M. D. Silverstein, T. M. Petterson, W. M. O'Fallon, et L. J. Melton 3rd, « Predictors of recurrence after deep vein thrombosis and pulmonary embolism: a population-based cohort study », *Archives of Internal Medicine*, vol. 160, n°. 6, p. 761-768, mars. 2000.
[9] P. O. Hansson, J. Sörbo, et H. Eriksson, « Recurrent venous thromboembolism after deep vein thrombosis: incidence and risk factors », *Archives of Internal Medicine*, vol. 160, n°. 6, p. 769-774, mars. 2000.
[10] S. R. Kahn, « The post-thrombotic syndrome: the forgotten morbidity of deep venous thrombosis », *Journal of Thrombosis and Thrombolysis*, vol. 21, n°. 1, p. 41-48, févr. 2006.
[11] M. Carrier, G. Le Gal, P. S. Wells, et M. A. Rodger, « Systematic review: case-fatality rates of recurrent venous thromboembolism and major bleeding events among patients treated for venous thromboembolism », *Annals of Internal Medicine*, vol. 152, n°. 9, p. 578-589, mai. 2010.
[12] R. M. Bertina, « The role of procoagulants and anticoagulants in the development of venous thromboembolism », *Thrombosis Research*, vol. 123 Suppl 4, p. S41-45, 2009.
[13] B. Furie et B. C. Furie, « Mechanisms of thrombus formation », *The New England Journal of Medicine*, vol. 359, n°. 9, p. 938-949, août. 2008.
[14] T. B. Larsen, H. T. Sørensen, A. Skytthe, S. P. Johnsen, J. W. Vaupel, et K. Christensen, « Major genetic susceptibility for venous thromboembolism in men: a study of Danish twins », *Epidemiology (Cambridge, Mass.)*, vol. 14, n°. 3, p. 328-332, mai. 2003.
[15] J. C. Souto et al., « Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia », *American Journal of Human Genetics*, vol. 67, n°. 6, p. 1452-1459, déc. 2000.
[16] J. A. Heit, M. A. Phelps, S. A. Ward, J. P. Slusser, T. M. Petterson, et M. De Andrade, « Familial segregation of venous thromboembolism », *Journal of Thrombosis and Haemostasis: JTH*, vol. 2, n°. 5, p. 731-736, mai. 2004.

[17] E. Oger, « Incidence of venous thromboembolism: a community-based study in Western France. EPI-GETBP Study Group. Groupe d'Etude de la Thrombose de Bretagne Occidentale », *Thrombosis and Haemostasis*, vol. 83, n°. 5, p. 657-660, mai. 2000.
[18] R. H. White, H. Zhou, et P. S. Romano, « Incidence of idiopathic deep venous thrombosis and secondary thromboembolism among ethnic groups in California », *Annals of Internal Medicine*, vol. 128, n°. 9, p. 737-740, mai. 1998.
[19] P. M. Ridker, J. P. Miletich, C. H. Hennekens, et J. E. Buring, « Ethnic distribution of factor V Leiden in 4047 men and women. Implications for venous thromboembolism screening », *JAMA: The Journal of the American Medical Association*, vol. 277, n°. 16, p. 1305-1307, avr. 1997.
[20] J. P. Gregg, A. J. Yamane, et W. W. Grody, « Prevalence of the factor V-Leiden mutation in four distinct American ethnic populations », *American Journal of Medical Genetics*, vol. 73, n°. 3, p. 334-336, déc. 1997.
[21] F. H. Herrmann et al., « Prevalence of factor V Leiden mutation in various populations », *Genetic Epidemiology*, vol. 14, n°. 4, p. 403-411, 1997.
[22] S. H. Cohen, G. E. Ehrlich, M. S. Kauffman, et C. Cope, « Thrombophlebitis following knee surgery », *The Journal of Bone and Joint Surgery. American Volume*, vol. 55, n°. 1, p. 106-112, janv. 1973.
[23] R. D. Hull et G. E. Raskob, « Prophylaxis of venous thromboembolic disease following hip and knee surgery », *The Journal of Bone and Joint Surgery. American Volume*, vol. 68, n°. 1, p. 146-150, janv. 1986.
[24] W. H. Geerts, K. I. Code, R. M. Jay, E. Chen, et J. P. Szalai, « A prospective study of venous thromboembolism after major trauma », *The New England Journal of Medicine*, vol. 331, n°. 24, p. 1601-1606, déc. 1994.
[25] A. N. Nicolaides, E. S. Field, V. V. Kakkar, A. J. Yates-Bell, S. Taylor, et M. B. Clarke, « Prostatectomy and deep-vein thrombosis », *The British Journal of Surgery*, vol. 59, n°. 6, p. 487-488, juin. 1972.
[26] M. E. Mayo, T. Halil, et N. L. Browse, « The incidence of deep vein thrombosis after prostatectomy », *British Journal of Urology*, vol. 43, n°. 6, p. 738-742, déc. 1971.
[27] J. J. Walsh, J. Bonnar, et F. W. Wright, « A study of pulmonary embolism and deep leg vein thrombosis after major gynaecological surgery using labelled fibrinogen-phlebography and lung scanning », *The Journal of Obstetrics and Gynaecology of the British Commonwealth*, vol. 81, n°. 4, p. 311-316, avr. 1974.
[28] F. R. Rosendaal, « Venous thrombosis: the role of genes, environment, and behavior », *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, p. 1-12, 2005.
[29] R. Beasley, N. Raymond, S. Hill, M. Nowitz, et R. Hughes, « eThrombosis: the 21st century variant of venous thromboembolism associated with immobility », *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology*, vol. 21, n°. 2, p. 374-376, févr. 2003.
[30] J. M. Cruickshank, R. Gorlin, et B. Jennett, « Air travel and thrombotic episodes: the economy class syndrome », *Lancet*, vol. 2, n°. 8609, p. 497-498, août. 1988.
[31] F. Lapostolle et al., « Severe pulmonary embolism associated with air travel », *The New England Journal of Medicine*, vol. 345, n°. 11, p. 779-783, sept. 2001.
[32] A. J. M. Schreijer et al., « Explanations for coagulation activation after air travel », *Journal of Thrombosis and Haemostasis: JTH*, vol. 8, n°. 5, p. 971-978, mai. 2010.
[33] J. W. Blom, C. J. M. Doggen, S. Osanto, et F. R. Rosendaal, « Malignancies, prothrombotic mutations, and the risk of venous thrombosis », *JAMA: The Journal of the American Medical Association*, vol. 293, n°. 6, p. 715-722, févr. 2005.

[34] J. W. Blom, S. Osanto, et F. R. Rosendaal, « The risk of a venous thrombotic event in lung cancer patients: higher risk for adenocarcinoma than squamous cell carcinoma », *Journal of Thrombosis and Haemostasis: JTH*, vol. 2, n°. 10, p. 1760-1765, oct. 2004.
[35] K. H. Zurborn, H. Duscha, J. Gram, et H. D. Bruhn, « Investigations of coagulation system and fibrinolysis in patients with disseminated adenocarcinomas and non-Hodgkin's lymphomas », *Oncology*, vol. 47, n°. 5, p. 376-380, 1990.
[36] C. J. van Rooden et al., « Central venous catheter related thrombosis in haematology patients and prediction of risk by screening with Doppler-ultrasound », *British Journal of Haematology*, vol. 123, n°. 3, p. 507-512, nov. 2003.
[37] N. I. Weijl et al., « Thromboembolic events during chemotherapy for germ cell cancer: a cohort study and review of the literature », *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 18, n°. 10, p. 2169-2178, mai. 2000.
[38] A. G. Holst, G. Jensen, et E. Prescott, « Risk factors for venous thromboembolism: results from the Copenhagen City Heart Study », *Circulation*, vol. 121, n°. 17, p. 1896-1903, mai. 2010.
[39] E. R. Pomp, F. R. Rosendaal, et C. J. M. Doggen, « Smoking increases the risk of venous thrombosis and acts synergistically with oral contraceptive use », *American Journal of Hematology*, vol. 83, n°. 2, p. 97-102, févr. 2008.
[40] A. W. Tsai, M. Cushman, W. D. Rosamond, S. R. Heckbert, J. F. Polak, et A. R. Folsom, « Cardiovascular risk factors and venous thromboembolism incidence: the longitudinal investigation of thromboembolism etiology », *Archives of Internal Medicine*, vol. 162, n°. 10, p. 1182-1189, mai. 2002.
[41] M. Abdollahi, M. Cushman, et F. R. Rosendaal, « Obesity: risk of venous thrombosis and the interaction with coagulation factor levels and oral contraceptive use », *Thrombosis and Haemostasis*, vol. 89, n°. 3, p. 493-498, mars. 2003.
[42] M.-C. Alessi et I. Juhan-Vague, « Metabolic syndrome, haemostasis and thrombosis », *Thrombosis and Haemostasis*, vol. 99, n°. 6, p. 995-1000, juin. 2008.
[43] L. M. Steffen et al., « Metabolic syndrome and risk of venous thromboembolism: Longitudinal Investigation of Thromboembolism Etiology », *Journal of Thrombosis and Haemostasis: JTH*, vol. 7, n°. 5, p. 746-751, mai. 2009.
[44] P. G. de Groot, B. Lutters, R. H. W. M. Derksen, T. Lisman, J. C. M. Meijers, et F. R. Rosendaal, « Lupus anticoagulants and the risk of a first episode of deep venous thrombosis », *Journal of Thrombosis and Haemostasis: JTH*, vol. 3, n°. 9, p. 1993-1997, sept. 2005.
[45] O. EGEBERG, « INHERITED ANTITHROMBIN DEFICIENCY CAUSING THROMBOPHILIA », *Thrombosis Et Diathesis Haemorrhagica*, vol. 13, p. 516-530, juin. 1965.
[46] J. H. Griffin, B. Evatt, T. S. Zimmerman, A. J. Kleiss, et C. Wideman, « Deficiency of protein C in congenital thrombotic disease », *The Journal of Clinical Investigation*, vol. 68, n°. 5, p. 1370-1373, nov. 1981.
[47] H. P. Schwarz, M. Fischer, P. Hopmeier, M. A. Batard, et J. H. Griffin, « Plasma protein S deficiency in familial thrombotic disease », *Blood*, vol. 64, n°. 6, p. 1297-1300, déc. 1984.
[48] R. C. Tait et al., « Prevalence of antithrombin deficiency in the healthy population », *British Journal of Haematology*, vol. 87, n°. 1, p. 106-112, mai. 1994.
[49] R. C. Tait et al., « Prevalence of protein C deficiency in the healthy population », *Thrombosis and Haemostasis*, vol. 73, n°. 1, p. 87-93, janv. 1995.
[50] F. R. Rosendaal, « Risk factors for venous thrombotic disease », *Thrombosis and Haemostasis*, vol. 82, n°. 2, p. 610-619, août. 1999.

[51] T. Koster et al., « Protein C deficiency in a controlled series of unselected outpatients: an infrequent but clear risk factor for venous thrombosis (Leiden Thrombophilia Study) », *Blood*, vol. 85, n°. 10, p. 2756-2761, mai. 1995.
[52] C. Demers, J. S. Ginsberg, J. Hirsh, P. Henderson, et M. A. Blajchman, « Thrombosis in antithrombin-III-deficient persons. Report of a large kindred and literature review », *Annals of Internal Medicine*, vol. 116, n°. 9, p. 754-761, mai. 1992.
[53] H. H. van Boven, J. P. Vandenbroucke, E. Briët, et F. R. Rosendaal, « Gene-gene and gene-environment interactions determine risk of thrombosis in families with inherited antithrombin deficiency », *Blood*, vol. 94, n°. 8, p. 2590-2594, oct. 1999.
[54] H. Heijboer, D. P. Brandjes, H. R. Büller, A. Sturk, et J. W. ten Cate, « Deficiencies of coagulation-inhibiting and fibrinolytic proteins in outpatients with deep-vein thrombosis », *The New England Journal of Medicine*, vol. 323, n°. 22, p. 1512-1516, nov. 1990.
[55] R. M. Bertina et al., « Mutation in blood coagulation factor V associated with resistance to activated protein C », *Nature*, vol. 369, n°. 6475, p. 64-67, mai. 1994.
[56] T. Koster, F. R. Rosendaal, H. de Ronde, E. Briët, J. P. Vandenbroucke, et R. M. Bertina, « Venous thrombosis due to poor anticoagulant response to activated protein C: Leiden Thrombophilia Study », *Lancet*, vol. 342, n°. 8886-8887, p. 1503-1506, déc. 1993.
[57] F. R. Rosendaal, T. Koster, J. P. Vandenbroucke, et P. H. Reitsma, « High risk of thrombosis in patients homozygous for factor V Leiden (activated protein C resistance) », *Blood*, vol. 85, n°. 6, p. 1504-1508, mars. 1995.
[58] S. Middeldorp et al., « The incidence of venous thromboembolism in family members of patients with factor V Leiden mutation and venous thrombosis », *Annals of Internal Medicine*, vol. 128, n°. 1, p. 15-20, janv. 1998.
[59] S. R. Poort, F. R. Rosendaal, P. H. Reitsma, et R. M. Bertina, « A common genetic variation in the 3'-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis », *Blood*, vol. 88, n°. 10, p. 3698-3703, nov. 1996.
[60] H. Jick et al., « Venous thromboembolic disease and ABO blood type. A cooperative study », *Lancet*, vol. 1, n°. 7594, p. 539-542, mars. 1969.
[61] O. Wu, N. Bayoumi, M. A. Vickers, et P. Clark, « ABO(H) blood groups and vascular disease: a systematic review and meta-analysis », *Journal of Thrombosis and Haemostasis: JTH*, vol. 6, n°. 1, p. 62-69, janv. 2008.
[62] P. V. Jenkins et J. S. O'Donnell, « ABO blood group determines plasma von Willebrand factor levels: a biologic function after all? », *Transfusion*, vol. 46, n°. 10, p. 1836-1844, oct. 2006.
[63] F. R. Rosendaal, « Venous thrombosis: a multicausal disease », *Lancet*, vol. 353, n°. 9159, p. 1167-1173, avr. 1999.
[64] J. A. Heit, J. M. Cunningham, T. M. Petterson, S. M. Armasu, D. N. Rider, et M. DE Andrade, « Genetic variation within the anticoagulant, procoagulant, fibrinolytic and innate immunity pathways as risk factors for venous thromboembolism », *Journal of Thrombosis and Haemostasis: JTH*, vol. 9, n°. 6, p. 1133-1142, juin. 2011.
[65] S. Uitte de Willige, M. C. H. de Visser, J. J. Houwing-Duistermaat, F. R. Rosendaal, H. L. Vos, et R. M. Bertina, « Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen γ' levels », *Blood*, vol. 106, n°. 13, p. 4176 -4183, déc. 2005.
[66] N. L. Smith et al., « Association of genetic variations with nonfatal venous thrombosis in postmenopausal women », *JAMA: The Journal of the American Medical Association*, vol. 297, n°. 5, p. 489-498, févr. 2007.

[67] A. P. Reiner, L. A. Lange, N. L. Smith, N. A. Zakai, M. Cushman, et A. R. Folsom, « Common hemostasis and inflammation gene variants and venous thrombosis in older adults from the Cardiovascular Health Study », *Journal of Thrombosis and Haemostasis: JTH*, vol. 7, n°. 9, p. 1499-1505, sept. 2009.
[68] S. J. Hasstedt et al., « Genome scan of venous thrombosis in a pedigree with protein C deficiency », *Journal of Thrombosis and Haemostasis: JTH*, vol. 2, n°. 6, p. 868-873, juin. 2004.
[69] S. J. Hasstedt et al., « Cell adhesion molecule 1: a novel risk factor for venous thrombosis », *Blood*, vol. 114, n°. 14, p. 3084-3091, oct. 2009.
[70] I. D. Bezemer et al., « Gene variants associated with deep vein thrombosis », *JAMA: The Journal of the American Medical Association*, vol. 299, n°. 11, p. 1306-1314, mars. 2008.
[71] D.-A. Trégouët et al., « Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach », *Blood*, vol. 113, n°. 21, p. 5298-5303, mai. 2009.
[72] H. Schunkert et al., « Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease », *Nature Genetics*, vol. 43, n°. 4, p. 333-338, avr. 2011.
[73] T. Koster, A. D. Blann, E. Briët, J. P. Vandenbroucke, et F. R. Rosendaal, « Role of clotting factor VIII in effect of von Willebrand factor on occurrence of deep-vein thrombosis », *Lancet*, vol. 345, n°. 8943, p. 152-155, janv. 1995.
[74] A. van Hylckama Vlieg, I. K. van der Linden, R. M. Bertina, et F. R. Rosendaal, « High levels of factor IX increase the risk of venous thrombosis », *Blood*, vol. 95, n°. 12, p. 3678-3682, juin. 2000.
[75] J. C. Meijers, W. L. Tekelenburg, B. N. Bouma, R. M. Bertina, et F. R. Rosendaal, « High levels of coagulation factor XI as a risk factor for venous thrombosis », *The New England Journal of Medicine*, vol. 342, n°. 10, p. 696-701, mars. 2000.
[76] A. Y. Nossent, J. C. J. Eikenboom, et R. M. Bertina, « Plasma coagulation factor levels in venous thrombosis », *Seminars in Hematology*, vol. 44, n°. 2, p. 77-84, avr. 2007.
[77] A. Dahm, A. Van Hylckama Vlieg, B. Bendz, F. Rosendaal, R. M. Bertina, et P. M. Sandset, « Low levels of tissue factor pathway inhibitor (TFPI) increase the risk of venous thrombosis », *Blood*, vol. 101, n°. 11, p. 4387-4392, juin. 2003.
[78] N. H. van Tilburg, F. R. Rosendaal, et R. M. Bertina, « Thrombin activatable fibrinolysis inhibitor and the risk for deep vein thrombosis », *Blood*, vol. 95, n°. 9, p. 2855-2859, mai. 2000.
[79] S. S. Kang, J. Zhou, P. W. Wong, J. Kowalisyn, et G. Strokosch, « Intermediate homocysteinemia: a thermolabile variant of methylenetetrahydrofolate reductase », *American Journal of Human Genetics*, vol. 43, n°. 4, p. 414-421, oct. 1988.
[80] M. Den Heijer, S. Lewington, et R. Clarke, « Homocysteine, MTHFR and risk of venous thrombosis: a meta-analysis of published epidemiological studies », *Journal of Thrombosis and Haemostasis*, vol. 3, n°. 2, p. 292-299, 2005.
[81] I. D. Bezemer, C. J. M. Doggen, H. L. Vos, et F. R. Rosendaal, « No association between the common MTHFR 677C->T polymorphism and venous thrombosis: results from the MEGA study », *Archives of Internal Medicine*, vol. 167, n°. 5, p. 497-501, mars. 2007.
[82] J. C. Souto et al., « Genetic determinants of hemostasis phenotypes in Spanish families », *Circulation*, vol. 101, n°. 13, p. 1546-1551, avr. 2000.
[83] M. de Lange, H. Snieder, R. A. Ariëns, T. D. Spector, et P. J. Grant, « The genetics of haemostasis: a twin study », *Lancet*, vol. 357, n°. 9250, p. 101-105, janv. 2001.

[84] P. E. Morange et al., « Biological and genetic factors influencing plasma factor VIII levels in a healthy family population: results from the Stanislas cohort », *British Journal of Haematology*, vol. 128, n°. 1, p. 91-99, janv. 2005.
[85] K. M. Brinkhous et al., « Purified human factor VIII procoagulant protein: comparative hemostatic response after infusions into hemophilic and von Willebrand disease dogs », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, n°. 24, p. 8752-8756, déc. 1985.
[86] A. W. Tsai et al., « Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE) », *The American Journal of Medicine*, vol. 113, n°. 8, p. 636-642, déc. 2002.
[87] K. H. Orstavik, P. Magnus, H. Reisner, K. Berg, J. B. Graham, et W. Nance, « Factor VIII and factor IX in a twin population. Evidence for a major effect of ABO locus on factor VIII level », *American Journal of Human Genetics*, vol. 37, n°. 1, p. 89-101, janv. 1985.
[88] M. G. Conlan et al., « Associations of factor VIII and von Willebrand factor with age, race, sex, and risk factors for atherosclerosis. The Atherosclerosis Risk in Communities (ARIC) Study », *Thrombosis and Haemostasis*, vol. 70, n°. 3, p. 380-385, sept. 1993.
[89] R. P. Tracy et al., « The distribution of coagulation factors VII and VIII and fibrinogen in adults over 65 years. Results from the Cardiovascular Health Study », *Annals of Epidemiology*, vol. 2, n°. 4, p. 509-519, juill. 1992.
[90] D. F. Geffken, M. Cushman, G. L. Burke, J. F. Polak, P. A. Sakkinen, et R. P. Tracy, « Association between physical activity and markers of inflammation in a healthy elderly population », *American Journal of Epidemiology*, vol. 153, n°. 3, p. 242-250, févr. 2001.
[91] G. D. Lowe et al., « Epidemiology of coagulation factors, inhibitors and activation markers: the Third Glasgow MONICA Survey. I. Illustrative reference ranges by age, sex and hormone use », *British Journal of Haematology*, vol. 97, n°. 4, p. 775-784, juin. 1997.
[92] J. O'Donnell et M. A. Laffan, « The relationship between ABO histo-blood group, factor VIII and von Willebrand factor », *Transfusion Medicine (Oxford, England)*, vol. 11, n°. 4, p. 343-351, août. 2001.
[93] A. M. Keightley, Y. M. Lam, J. N. Brady, C. L. Cameron, et D. Lillicrap, « Variation at the von Willebrand factor (vWF) gene locus is associated with plasma vWF:Ag levels: identification of three novel single nucleotide polymorphisms in the vWF gene promoter », *Blood*, vol. 93, n°. 12, p. 4277-4283, juin. 1999.
[94] P. J. Harvey, A. M. Keightley, Y. M. Lam, C. Cameron, et D. Lillicrap, « A single nucleotide polymorphism at nucleotide -1793 in the von Willebrand factor (VWF) regulatory region is associated with plasma VWF:Ag levels », *British Journal of Haematology*, vol. 109, n°. 2, p. 349-353, mai. 2000.
[95] C. Hough et al., « Influence of a GT repeat element on shear stress responsiveness of the VWF gene promoter », *Journal of Thrombosis and Haemostasis: JTH*, vol. 6, n°. 7, p. 1183-1190, juill. 2008.
[96] V. Daidone et al., « Microsatellite (GT)(n) repeats and SNPs in the von Willebrand factor gene promoter do not influence circulating von Willebrand factor levels under normal conditions », *Thrombosis and Haemostasis*, vol. 101, n°. 2, p. 298-304, févr. 2009.
[97] L. A. O'Brien et al., « Founder von Willebrand factor haplotype associated with type 1 von Willebrand disease », *Blood*, vol. 102, n°. 2, p. 549-557, juill. 2003.

[98] J. A. Davies, P. W. Collins, L. S. Hathaway, et D. J. Bowen, « Effect of von Willebrand factor Y/C1584 on in vivo protein level and function and interaction with ABO blood group », *Blood*, vol. 109, n°. 7, p. 2840-2846, avr. 2007.
[99] J. A. Davies, P. W. Collins, L. S. Hathaway, et D. J. Bowen, « von Willebrand factor: evidence for variable clearance in vivo according to Y/C1584 phenotype and ABO blood group », *Journal of Thrombosis and Haemostasis: JTH*, vol. 6, n°. 1, p. 97-103, janv. 2008.
[100] M. C. van Schie et al., « Variation in the von Willebrand factor gene is associated with von Willebrand factor levels and with the risk for cardiovascular disease », *Blood*, vol. 117, n°. 4, p. 1393-1399, janv. 2011.
[101] N. L. Smith et al., « Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium », *Circulation*, vol. 121, n°. 12, p. 1382-1392, mars. 2010.
[102] K. R. Viel et al., « A sequence variation scan of the coagulation factor VIII (FVIII) structural gene and associations with plasma FVIII activity levels », *Blood*, vol. 109, n°. 9, p. 3713-3724, mai. 2007.
[103] D. Scanavini, C. Legnani, B. Lunghi, F. Mingozzi, G. Palareti, et F. Bernardi, « The factor VIII D1241E polymorphism is associated with decreased factor VIII activity and not with activated protein C resistance levels », *Thrombosis and Haemostasis*, vol. 93, n°. 3, p. 453-456, mars. 2005.
[104] A. Y. Nossent et al., « Haplotypes encoding the factor VIII 1241 Glu variation, factor VIII levels and the risk of venous thrombosis », *Thrombosis and Haemostasis*, vol. 95, n°. 6, p. 942-948, juin. 2006.
[105] P. J. Lenting et al., « The light chain of factor VIII comprises a binding site for low density lipoprotein receptor-related protein », *The Journal of Biological Chemistry*, vol. 274, n°. 34, p. 23734-23739, août. 1999.
[106] E. L. Saenko, A. V. Yakhyaev, I. Mikhailenko, D. K. Strickland, et A. G. Sarafanov, « Role of the low density lipoprotein-related protein receptor in mediation of factor VIII catabolism », *The Journal of Biological Chemistry*, vol. 274, n°. 53, p. 37685-37692, déc. 1999.
[107] N. Bovenschen et al., « Elevated plasma factor VIII in a mouse model of low-density lipoprotein receptor-related protein deficiency », *Blood*, vol. 101, n°. 10, p. 3933-3939, mai. 2003.
[108] G. Marchetti et al., « Contribution of low density lipoprotein receptor-related protein genotypes to coagulation factor VIII levels in thrombotic women », *Haematologica*, vol. 91, n°. 9, p. 1261-1263, sept. 2006.
[109] R. Vormittag et al., « Low-density lipoprotein receptor-related protein 1 polymorphism 663 C > T affects clotting factor VIII activity and increases the risk of venous thromboembolism », *Journal of Thrombosis and Haemostasis: JTH*, vol. 5, n°. 3, p. 497-502, mars. 2007.
[110] N. Martinelli et al., « Polymorphisms at LDLR locus may be associated with coronary artery disease through modulation of coagulation factor VIII activity and independently from lipid profile », *Blood*, vol. 116, n°. 25, p. 5688-5697, déc. 2010.
[111] J. M. Soria et al., « A quantitative-trait locus in the human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease », *American Journal of Human Genetics*, vol. 70, n°. 3, p. 567-574, mars. 2002.
[112] I. Tirado et al., « Association after linkage analysis indicates that homozygosity for the 46C-->T polymorphism in the F12 gene is a genetic risk factor for venous thrombosis », *Thrombosis and Haemostasis*, vol. 91, n°. 5, p. 899-904, mai. 2004.

[113] C. Y. Johnson, A. Tuite, P. E. Morange, D. A. Tregouet, et F. Gagnon, « The factor XII -4C>T variant and risk of common thrombotic disorders: A HuGE review and meta-analysis of evidence from observational studies », *American Journal of Epidemiology*, vol. 173, n°. 2, p. 136-144, janv. 2011.
[114] J. M. Soria et al., « A new locus on chromosome 18 that influences normal variation in activated protein C resistance phenotype and factor VIII activity and its relation to thrombosis susceptibility », *Blood*, vol. 101, n°. 1, p. 163-167, janv. 2003.
[115] J. C. Souto et al., « Genome-wide linkage analysis of von Willebrand factor plasma levels: results from the GAIT project », *Thrombosis and Haemostasis*, vol. 89, n°. 3, p. 468-474, mars. 2003.
[116] M. Berger et al., « High factor VIII levels in venous thromboembolism show linkage to imprinted loci on chromosomes 5 and 11 », *Blood*, vol. 105, n°. 2, p. 638-644, janv. 2005.
[117] M. Berger et al., « Association of ADAMDEC1 haplotype with high factor VIII levels in venous thromboembolism », *Thrombosis and Haemostasis*, vol. 99, n°. 5, p. 905-908, mai. 2008.
[118] Q. Yang, S. Kathiresan, J.-P. Lin, G. H. Tofler, et C. J. O'Donnell, « Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study », *BMC Medical Genetics*, vol. 8 Suppl 1, p. S12, 2007.
[119] E. Mazoyer et al., « Prevalence of factor V Leiden and prothrombin G20210A mutation in a large French population selected for nonthrombotic history: geographical and age distribution », *Blood Coagulation & Fibrinolysis: An International Journal in Haemostasis and Thrombosis*, vol. 20, n°. 7, p. 503-510, oct. 2009.
[120] S. Hercberg et al., « The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals », *Archives of Internal Medicine*, vol. 164, n°. 21, p. 2335-2342, nov. 2004.
[121] M. Boehnke et N. J. Cox, « Accurate inference of relationships in sib-pair linkage studies. », *American Journal of Human Genetics*, vol. 61, n°. 2, p. 423-429, août. 1997.
[122] M. S. McPeek et L. Sun, « Statistical tests for detection of misspecified relationships by use of genome-screen data », *American Journal of Human Genetics*, vol. 66, n°. 3, p. 1076-1094, mars. 2000.
[123] S. Purcell et al., « PLINK: a tool set for whole-genome association and population-based linkage analyses », *American Journal of Human Genetics*, vol. 81, n°. 3, p. 559-575, sept. 2007.
[124] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, et D. Reich, « Principal components analysis corrects for stratification in genome-wide association studies », *Nature Genetics*, vol. 38, n°. 8, p. 904-909, août. 2006.
[125] G. Paré et al., « Novel association of HK1 with glycated hemoglobin in a non-diabetic population: a genome-wide evaluation of 14,618 participants in the Women's Genome Health Study », *PLoS Genetics*, vol. 4, n°. 12, p. e1000312, déc. 2008.
[126] N. E. Morton, « Sequential tests for the detection of linkage », *American Journal of Human Genetics*, vol. 7, n°. 3, p. 277-318, sept. 1955.
[127] J. K. Haseman et R. C. Elston, « The investigation of linkage between a quantitative trait and a marker locus », *Behavior Genetics*, vol. 2, n°. 1, p. 3-19, mars. 1972.
[128] C. I. Amos, « Robust variance-components approach for assessing genetic linkage in pedigrees », *American Journal of Human Genetics*, vol. 54, n°. 3, p. 535-543, mars. 1994.
[129] D. E. Goldgar, « Multipoint analysis of human quantitative genetic variation », *American Journal of Human Genetics*, vol. 47, n°. 6, p. 957-967, déc. 1990.

[130] C. I. Amos et R. C. Elston, « Robust methods for the detection of genetic linkage for quantitative data from pedigrees », *Genetic Epidemiology*, vol. 6, n°. 2, p. 349-360, 1989.
[131] P. C. Sham, S. Purcell, S. S. Cherny, et G. R. Abecasis, « Powerful regression-based quantitative-trait linkage analysis of general pedigrees », *American Journal of Human Genetics*, vol. 71, n°. 2, p. 238-253, août. 2002.
[132] E. M. Wijsman et C. I. Amos, « Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions », *Genetic Epidemiology*, vol. 14, n°. 6, p. 719-735, 1997.
[133] L. Almasy et J. Blangero, « Multipoint quantitative-trait linkage analysis in general pedigrees », *American Journal of Human Genetics*, vol. 62, n°. 5, p. 1198-1211, mai. 1998.
[134] S. C. Heath, « Markov chain Monte Carlo segregation and linkage analysis for oligogenic models », *American Journal of Human Genetics*, vol. 61, n°. 3, p. 748-760, sept. 1997.
[135] E. M. Wijsman et D. Yu, « Joint oligogenic segregation and linkage analysis using bayesian Markov chain Monte Carlo methods », *Molecular Biotechnology*, vol. 28, n°. 3, p. 205-226, nov. 2004.
[136] W. K. HASTINGS, « Monte Carlo sampling methods using Markov chains and their applications », *Biometrika*, vol. 57, n°. 1, p. 97 -109, avr. 1970.
[137] E. W. Daw, E. M. Wijsman, et E. A. Thompson, « A score for Bayesian genome screening », *Genetic Epidemiology*, vol. 24, n°. 3, p. 181-190, avr. 2003.
[138] R. E. Kass et A. E. Raftery, « Bayes Factors », *Journal of the American Statistical Association*, vol. 90, n°. 430, p. 773-795, juin. 1995.
[139] E. W. Daw, E. A. Thompson, et E. M. Wijsman, « Bias in multipoint linkage analysis arising from map misspecification », *Genetic Epidemiology*, vol. 19, n°. 4, p. 366-380, déc. 2000.
[140] R. P. Igo et E. M. Wijsman, « Empirical significance values for linkage analysis: trait simulation using posterior model distributions from MCMC oligogenic segregation analysis », *Genetic Epidemiology*, vol. 32, n°. 2, p. 119-131, févr. 2008.
[141] L. Almasy et J. Blangero, « Variance component methods for analysis of complex phenotypes », *Cold Spring Harbor Protocols*, vol. 2010, n°. 5, p. pdb.top77, mai. 2010.
[142] J. Blangero, J. T. Williams, et L. Almasy, « Robust LOD scores for variance component-based linkage analysis », *Genetic Epidemiology*, vol. 19 Suppl 1, p. S8-14, 2000.
[143] L. Almasy et J. Blangero, « Multipoint quantitative-trait linkage analysis in general pedigrees », *American Journal of Human Genetics*, vol. 62, n°. 5, p. 1198-1211, mai. 1998.
[144] S. L. Zeger et K. Y. Liang, « Longitudinal data analysis for discrete and continuous outcomes », *Biometrics*, vol. 42, n°. 1, p. 121-130, mars. 1986.
[145] K. Y. Liang et S. L. Zeger, « Regression analysis for correlated data », *Annual Review of Public Health*, vol. 14, p. 43-68, 1993.
[146] D. A. Tregouet et L. Tiret, « Applications of the estimating equations theory to genetic epidemiology: a review », *Annals of Human Genetics*, vol. 64, n°. 1, p. 1-14, janv. 2000.
[147] K. Y. Liang et A. E. Pulver, « Analysis of case-control/family sampling design », *Genetic Epidemiology*, vol. 13, n°. 3, p. 253-270, 1996.
[148] D. A. Trégouët, B. Herbeth, I. Juhan-Vague, G. Siest, P. Ducimetière, et L. Tiret, « Bivariate familial correlation analysis of quantitative traits by use of estimating

equations: application to a familial analysis of the insulin resistance syndrome », *Genetic Epidemiology*, vol. 16, n°. 1, p. 69-83, 1999.
[149] D. A. Tregouet et V. Garelle, « A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies », *Bioinformatics (Oxford, England)*, vol. 23, n°. 8, p. 1038-1039, avr. 2007.
[150] R. DerSimonian et N. Laird, « Meta-analysis in clinical trials », *Controlled Clinical Trials*, vol. 7, n°. 3, p. 177-188, sept. 1986.
[151] W. G. Cochran, « The Combination of Estimates from Different Experiments », *Biometrics*, vol. 10, n°. 1, p. 101-129, mars. 1954.
[152] R. J. Hardy et S. G. Thompson, « Detecting and describing heterogeneity in meta-analysis », *Statistics in Medicine*, vol. 17, n°. 8, p. 841-856, avr. 1998.
[153] J. P. T. Higgins et S. G. Thompson, « Quantifying heterogeneity in a meta-analysis », *Statistics in Medicine*, vol. 21, n°. 11, p. 1539-1558, juin. 2002.
[154] R. Magi et A. Morris, « GWAMA: software for genome-wide association meta-analysis », *BMC Bioinformatics*, vol. 11, n°. 1, p. 288, 2010.
[155] D. B. Pettiti, *Meta-analysis, decision analysis and cost-effectiveness analysis*, Oxford University Press. New York, NY: , 1994.
[156] D.-A. Tregouet et al., « In-depth haplotype analysis of ABCA1 gene polymorphisms in relation to plasma ApoA1 levels and myocardial infarction », *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 24, n°. 4, p. 775-781, avr. 2004.
[157] P.-E. Morange, N. Saut, G. Antoni, J. Emmerich, et D.-A. Trégouët, « Impact on venous thrombosis risk of newly discovered gene variants associated with FVIII and VWF plasma levels », *Journal of Thrombosis and Haemostasis: JTH*, vol. 9, n°. 1, p. 229-231, janv. 2011.
[158] M. Germain et al., « Genetics of venous thrombosis: insights from a new genome wide association study », *PloS One*, vol. 6, n°. 9, p. e25581, 2011.
[159] J. Blangero et al., « Quantitative trait nucleotide analysis using Bayesian model selection », *Human Biology*, vol. 77, n°. 5, p. 541-559, oct. 2005.
[160] J. Ito et al., « Anatomical and histological profiling of orphan G-protein-coupled receptor expression in gastrointestinal tract of C57BL/6J mice », *Cell and Tissue Research*, vol. 338, n°. 2, p. 257-269, nov. 2009.
[161] H. J. Kee et al., « Expression of brain-specific angiogenesis inhibitor 3 (BAI3) in normal brain and implications for BAI3 in ischemia-induced brain angiogenesis and malignant glioma », *FEBS Letters*, vol. 569, n°. 1-3, p. 307-316, juill. 2004.
[162] P. DeRosse, T. Lencz, K. E. Burdick, S. G. Siris, J. M. Kane, et A. K. Malhotra, « The genetics of symptom-based phenotypes: toward a molecular classification of schizophrenia », *Schizophrenia Bulletin*, vol. 34, n°. 6, p. 1047-1053, nov. 2008.
[163] T. N. Bongers et al., « Lower levels of ADAMTS13 are associated with cardiovascular disease in young patients », *Atherosclerosis*, vol. 207, n°. 1, p. 250-254, nov. 2009.
[164] M. Fujioka et al., « ADAMTS13 gene deletion aggravates ischemic brain damage: a possible neuroprotective role of ADAMTS13 by ameliorating postischemic hypoperfusion », *Blood*, vol. 115, n°. 8, p. 1650-1653, févr. 2010.
[165] B.-Q. Zhao et al., « von Willebrand factor-cleaving protease ADAMTS13 reduces ischemic brain injury in experimental stroke », *Blood*, vol. 114, n°. 15, p. 3329-3334, oct. 2009.
[166] H. Adachi et M. Tsujimoto, « FEEL-1, a novel scavenger receptor with in vitro bacteria-binding and angiogenesis-modulating activities », *The Journal of Biological Chemistry*, vol. 277, n°. 37, p. 34264-34270, sept. 2002.

[167] Y. Tamura et al., « FEEL-1 and FEEL-2 are endocytic receptors for advanced glycation end products », *The Journal of Biological Chemistry*, vol. 278, n°. 15, p. 12613-12617, avr. 2003.
[168] M. Katoh et M. Katoh, « FLJ10261 gene, located within the CCND1-EMS1 locus on human chromosome 11q13, encodes the eight-transmembrane protein homologous to C12orf3, C11orf25 and FLJ34272 gene products », *International Journal of Oncology*, vol. 22, n°. 6, p. 1375-1381, juin. 2003.
[169] C. Thomas-Gatewood et al., « TMEM16A channels generate Ca2+-activated Cl- currents in cerebral artery smooth muscle cells », *American Journal of Physiology. Heart and Circulatory Physiology*, août. 2011.
[170] K. Roeder, S.-A. Bacanu, L. Wasserman, et B. Devlin, « Using linkage genome scans to improve power of association in genome scans », *American Journal of Human Genetics*, vol. 78, n°. 2, p. 243-252, févr. 2006.
[171] L. Sun, R. V. Craiu, A. D. Paterson, et S. B. Bull, « Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies », *Genetic Epidemiology*, vol. 30, n°. 6, p. 519-530, sept. 2006.
[172] A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell, et P. I. W. de Bakker, « SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap », *Bioinformatics*, vol. 24, n°. 24, p. 2938 -2939, déc. 2008.
[173] Gauderman WJ et Morrison JM, « QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe », 2006.
[174] Y. J. Chen et T. H. Stevens, « The VPS8 gene is required for localization and trafficking of the CPY sorting receptor in Saccharomyces cerevisiae », *European Journal of Cell Biology*, vol. 70, n°. 4, p. 289-297, août. 1996.
[175] M. Agaphonov et al., « Defect of vacuolar protein sorting stimulates proteolytic processing of human urokinase-type plasminogen activator in the yeast Hansenula polymorpha », *FEMS Yeast Research*, vol. 5, n°. 11, p. 1029-1035, nov. 2005.
[176] K. L. Conen, S. Nishimori, S. Provot, et H. M. Kronenberg, « The transcriptional cofactor Lbh regulates angiogenesis and endochondral bone formation during fetal bone development », *Developmental Biology*, vol. 333, n°. 2, p. 348-358, sept. 2009.
[177] M. Mori et al., « Promotion of cell spreading and migration by vascular endothelial-protein tyrosine phosphatase (VE-PTP) in cooperation with integrins », *Journal of Cellular Physiology*, vol. 224, n°. 1, p. 195-204, juill. 2010.
[178] G. Antoni et al., « A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism », *Journal of Thrombosis and Haemostasis: JTH*, vol. 8, n°. 12, p. 2671-2679, déc. 2010.
[179] A. J. Vlot, S. J. Koppelman, M. H. van den Berg, B. N. Bouma, et J. J. Sixma, « The affinity and stoichiometry of binding of human factor VIII to von Willebrand factor », *Blood*, vol. 85, n°. 11, p. 3150-3157, juin. 1995.
[180] P. Lollar, D. C. Hill-Eubanks, et C. G. Parker, « Association of the factor VIII light chain with von Willebrand factor », *The Journal of Biological Chemistry*, vol. 263, n°. 21, p. 10451-10455, juill. 1988.
[181] J. C. J. Eikenboom, G. Castaman, P. W. Kamphuisen, F. R. Rosendaal, et R. M. Bertina, « The factor VIII/von Willebrand factor ratio discriminates between reduced synthesis and increased clearance of von Willebrand factor », *Thrombosis and Haemostasis*, vol. 87, n°. 2, p. 252-257, févr. 2002.
[182] E. L. Webb, G. S. Sellick, et R. S. Houlston, « SNPLINK: multipoint linkage analysis of densely distributed SNP data incorporating automated linkage disequilibrium removal », *Bioinformatics (Oxford, England)*, vol. 21, n°. 13, p. 3060-3061, juill. 2005.

[183] G. Lin, Z. Wang, L. Wang, Y.-L. Lau, et W. Yang, « Identification of linked regions using high-density SNP genotype data in linkage analysis », *Bioinformatics (Oxford, England)*, vol. 24, n°. 1, p. 86-93, janv. 2008.
[184] G. R. Abecasis, L. R. Cardon, et W. O. Cookson, « A general test of association for quantitative traits in nuclear families », *American Journal of Human Genetics*, vol. 66, n°. 1, p. 279-292, janv. 2000.
[185] G. R. Abecasis, W. O. Cookson, et L. R. Cardon, « Pedigree tests of transmission disequilibrium », *European Journal of Human Genetics: EJHG*, vol. 8, n°. 7, p. 545-551, juill. 2000.

ANNEXES

Annexe 1. Notion d'Epidémiologie génétique

Les polymorphismes utilisés en épidémiologie génétique

Le génome humain est constitué de 3,3 milliards de bases nucléotidiques. Ces bases sont de quatre types : Adénine (A), Thymine (T) Guanine (G) et Cytosine (C). L'immense majorité du génome est identique d'un individu à l'autre (99,5%). Les polymorphismes sont des séquences variables d'un individu à l'autre, localisées à des endroits précis du génome. Il en existe différents types :

***Single Nucleotide Polymorphisms (SNP*)**

Ils résultent de la substitution d'une base par une autre, survenue généralement lors de la réplication de l'ADN. Ils sont le plus souvent bi-alléliques : deux allèles sont observés dans la population (par exemple A et C). Plus de 15 millions de SNPs sont répertoriés par les projets Hapmap et 1000 génome, au niveau de séquences non-codantes (intergéniques ou introniques) ou codantes de l'ADN (exons). Dans ce dernier cas, ils peuvent entraîner une modification de l'acide aminé traduit depuis la séquence polymorphique. On parle alors de SNP non-synonyme. Les SNPs sont les polymorphismes de choix des analyses d'association.

Les microsatellites et minisatellites (*Variable Number Tandem Repeat-VNTR*)

Les microsatellites et les minisatellites consistent en une petite séquence de quelques bases nucléotidiques (moins de dix pour les micro-, plus pour les minisatellites) répétée en tandem. Le nombre de répétitions est hautement polymorphique (par exemple $(CATG)_n$ où n varie d'un individu à l'autre en moyenne entre 10 et 100). On en connaît plus de 10 000 VNTR. Ils sont répartis régulièrement dans le génome, majoritairement situés dans des régions non-codantes. Les microsatellites étaient jusqu'à tout récemment les polymorphismes de choix des analyses de liaison. En effet, en raison de leur grande variabilité, ils permettent de déterminer dans la plupart des cas l'origine parentale des allèles. Depuis l'avènement du génotypage à haute densité, les analyses de liaison sont également réalisées avec les SNPs dont le nombre très important compense la faible hétérogénéité

Les variants structuraux (Copy Number Variant - CNV)

Il s'agit d'une grande séquence d'ADN de plus de 1000 bases qui, à la suite de délétion et/ou d'insertion, est en nombre variable d'un individu à l'autre (0, 1, 2, 3, …). Découverts récemment, ils sont impliqués dans la susceptibilité à diverses maladies

Les différents types d'études en épidémiologie génétique

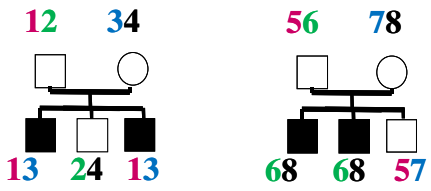
Etudes ne nécessitant pas le génotypage de polymorphismes (ou marqueurs génétiques).

Les études d'**agrégation familiale** permettent d'estimer le risque de récurrence intra-familiale d'une maladie. Elles estiment le risque d'être malade lorsqu'on a un apparenté malade rapporté au risque de la population générale. Une forte concentration familiale n'est pas nécessairement la conséquence d'une composante génétique mais aussi du partage d'un même environnement. Les **études d'héritabilité** permettent d'estimer la part respective imputable aux caractéristiques génétiques et environnementales dans la variabilité phénotypique. Les **études de ségrégation** cherchent à préciser la nature des facteurs génétiques, sans les localiser. Elles déterminent en particulier si les observations familiales sont compatibles (ou non) avec la présence d'un gène majeur. Le cas échéant, elles estiment les fréquences alléliques et le risque de développer la maladie sachant chaque génotype au gène majeur.

Etudes utilisant des marqueurs génétiques

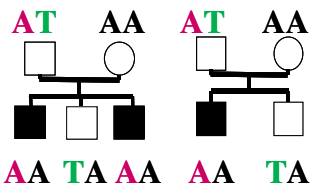
Les **études de liaison** cherchent à localiser des régions chromosomiques qui ségrègent de façon non aléatoire avec la maladie *au sein d'une même famille*. Elles identifient des marqueurs situés suffisemment proches du locus causal pour qu'il n'y ait eu que très peu de recombinaisons entre le locus causal et les marqueurs. Les **études d'association** visent quant à elle à évaluer le rôle potentiel de certains polymorphismes en reposant sur l'observation que le même allèle est partagé entre sujets atteints, au sein *d'une même famille et d'une famille à l'autre*. Ce marqueur peut soit être le locus causal lui-même ou bien simplement en déséquilibre de liaison avec le locus causal (voir **pA5**). La méthode classique consiste à comparer la distribution d'un polymorphisme entre des malades et des non malades non apparentés. En présence de données familiales, il est possible de prendre en compte les corrélations entre individus au moyen d'une statistique idoine. Il existe aussi une méthodologie spécifique (Test de Déséquilibre de Transmission) qui teste si les enfants atteints ont reçu de leurs parents hétérozygotes un allèle donné avec une fréquence significativement différente de ce que voudrait le hasard des transmissions.

LIAISON:
GENERATION D'UNE REGION
CANDIDATE



Coségrégation (co-transmission)
des locus marqueur et maladie

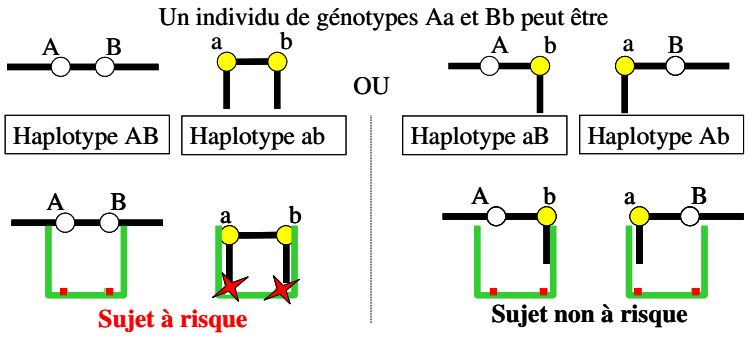
ASSOCIATION:
GENERATION D'ALLELES
POTENTIELLEMENT FONCTIONNELS



Association entre un
allèle particulier et la maladie

Analyses haplotypiques

Un haplotype est défini par la combinaison d'allèles présents à différentes positions d'une même séquence chromosomique. Considérons deux polymorphismes di-alléliques (A>a et B>b). Leur combinaison est potentiellement à l'origine de quatre haplotypes : AB, Ab, aB et ab. Les analyses haplotypiques étudient l'association de chaque haplotype avec un phénotype donné. Leur intérêt est multiple. En premier lieu, les deux polymorphismes peuvent être en déséquilibre de liaison (*i.e* la probabilité d'observer B plutôt que b dépend de la présence ou non de A, voir **pA5**). Dans ce cas, les analyses univariées ne peuvent pas estimer l'effet propre à chacun. De plus, les polymorphismes peuvent interagir uniquement lorsque leurs allèles délétères sont présents *sur le même chromosome (i.e* en phase*)*, et non pas lorsqu'ils sont présents sur deux chromosomes différents (*i.e* en opposition de phase) d'un même individu (voir un exemple sur la figure ci-dessous) :



Dans cet exemple, une analyse multivariée classique (incluant les effets des deux polymorphismes et de leur interaction) ne distinguerait pas les individus AB/ab des individus Ab/aB (qui sont tous double hétérozygotes A/a et B/b). L'interaction des deux polymorphismes, qui s'exprime spécifiquement lorsque a et b sont sur le même chromosome, serait « noyée » par l'absence d'interaction lorsqu'ils sont sur deux chromosomes différents. Un autre intérêt des analyses haplotypiques est que l'information apportée par un polymorphisme rare et non génotypé peut être « capturée » par un haplotype constitué de plusieurs polymorphismes communs. Notons enfin que la modélisation des effets haplotypiques est nettement plus économe en nombre de paramètres estimés qu'un modèle multivarié classique, ce qui est une source de satisfaction pour les statisticiens .[1]

La difficulté réside dans la détermination des haplotypes des individus double hétérozygotes. S'il existe des méthodes de biologie moléculaire permettant d'observer directement les haplotypes, elles sont trop longues et coûteuses pour être utilisées dans le cadre d'études d'épidémiologie génétiques. Il est donc nécessaire d'avoir recours à des méthodes d'inférence, utilisant un algorithme d'Expectation Maximisation, fondées sur des estimateurs de Maximum de Vraisemblance. Elles estiment la probabilité de chaque haplotype sachant les génotypes, tandis que l'effet de chaque haplotype est estimé par une régression des phénotypes observés sur les haplotypes inférés, pondérés par leur probabilité respective sachant les génotypes [2].

[1] Daniel J Schaid, « Evaluating associations of haplotypes with traits », Genetic Epidemiology 27, no. 4 (décembre 2004): 348-364.
[2] D A Tregouet et al., « A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm », Annals of Human Genetics 68, no. 2 (mars 2004): 165-177.

Déséquilibre gamétique et déséquilibre de liaison (DL)

Deux polymorphismes di-alléliques (A>a et B>b) sont en déséquilibre gamétique lorsque leurs allèles respectifs ne s'associent pas aléatoirement dans chaque gamète des individus d'une population. A l'inverse, en cas d'équilibre gamétique, la présence concomitante des allèles A et B a une fréquence f_{AB} égale au produit $f_A.f_B$. Le déséquilibre gamétique traduit souvent la présence de deux sous-populations dont les fréquences alléliques diffèrent. Lorsqu'un déséquilibre gamétique survient dans une unique population panmictique, entre des loci en liaison génétique, on parle de déséquilibre de liaison.

Le déséquilibre de liaison (DL) est alors l'association préférentielle d'allèles *sur un même brin d'ADN*. Théoriquement, quatre haplotypes peuvent être observés : AB, Ab, aB et ab. Le DL est complet si, par exemple, l'allèle B est exclusivement associé à l'allèle A. On n'observe alors que trois haplotypes (AB, Ab et ab), voire que deux (AB et ab) dans le cas où particulier où $f_A = f_B$. Dans ce dernier cas, la redondance entre les SNPs A>a et B>b est totale. On parle alors d'association complète ou de DL parfait.

Il existe différentes mesures du DL communément utilisées :

(1) $$D = f_{AB} - f_A.f_B = f_{AB}.f_{ab} - f_{Ab}.f_{aB}$$

D mesure l'écart entre les fréquences haplotypiques observées et attendues à l'équilibre. Sa valeur maximale dépend des fréquences alléliques. On ne peut donc pas comparer entre elles deux valeurs D.

(2) $D' = D/D_{\max}$ ($D_{\max}$ étant la valeur que prendrait si D le DL était complet)

$$D_{\max} = \min(f_A.f_b, f_a.f_B) \qquad \text{si } D' > 0$$

$$D_{\max} = \min(f_A.f_B, f_a.f_b) \qquad \text{si } D' < 0$$

D' vaut 0 à l'équilibre de liaison, 1 en cas de DL complet et une association préférentielle de A avec B, -1 en cas de DL complet et une association préférentielle de A avec b.

(3) $$r^2 = \frac{D^2}{f_A.f_a.f_B.f_b}$$

r^2 est une mesure de corrélation entre A>a et B>b. Il vaut 0 à l'équilibre de liaison ; il prend des valeurs comprise entre 0 et 1 d'autant plus grande que le DL est important ; et contrairement à D', n'est égal à 1 qu'en cas d'association complète (DL parfait). La valeur r^2 reflète mieux la redondance entre deux polymorphismes que D'.

Modèle d'Hardy-Weinberg (HW)

On appelle population panmictique toute population dans laquelle les individus (et leurs gamètes) s'unissent aléatoirement. On peut alors se représenter la constitution d'une nouvelle génération sous la forme d'un tirage au sort, réalisé à partir d'une immense urne contenant les gamètes de tous les individus. De cette façon, si l'on considère un polymorphisme bi-allélique, dont les allèles M et m ont pour fréquence p et q=1-p, le modèle d'Hardy-Weinberg (HW) permet de déterminer les fréquences génotypiques p_{MM}, p_{Mm} et p_{mm} en fonction de p et q :

$$P_{MM}=p^2$$

$$P_{Mm}=2pq$$

$$P_{mm}=q^2$$

La différence entre les fréquences génotypiques observées et celles attendues par le modèle d'HW se teste par un Chi-2 à un degré de liberté. Si l'hypothèse de panmixie est valide, une différence significative est souvent le signe d'erreurs de génotypage.

Remarques

(1) L'hypothèse de panmixie est très forte, particulièrement chez l'espèce humaine. Néanmoins, elle est raisonnable pour tout marqueur qui n'est pas en déséquilibre de liaison avec un polymorphisme entraînant une variabilité sensible d'un phénotype extérieur.

(2) Une population composée de deux sous-populations panmictiques n'est pas panmictique. L'équilibre d'HW ne sera donc pas obtenu en cas de stratification.

(3) Si un marqueur est associé à la maladie étudiée, les fréquences génotypes observées dans la sous-population des malades s'écarteront du modèle d'HW, alors même que la panmixie sera réalisée dans la population générale. Pour cette raison, dans le cas d'un échantillon cas-témoin, il est recommandé de réaliser le test du modèle d'HW chez les témoins.

(4) On parle d'équilibre (et non plus seulement de « modèle ») d'HW lorsque les fréquences alléliques et génotypiques sont stables au cours des générations. Cet équilibre nécessite, outre la panmixie, l'absence de migration, de sélection, de mutation, et de croisement entre générations.

Contrôle de l'erreur de type 1 en cas de tests multiples

Considérons un test unique de l'hypothèse nulle, H_0, d'absence d'association entre un génotype et un phénotype. Le risque d'erreur de type 1, α, est la probabilité de rejeter à tort H_0. Il s'agit du risque de faux positif. Habituellement, on cherche à contrôler α à un niveau fixé, traditionnellement 5%. Ce test nécessite le calcul d'une statistique T dont la distribution est connue sous H_0. Dans le cas où T suit un loi de Chi-2, la *p-value*, p, est la probabilité d'observer sous H_0 une valeur T au moins aussi grande que celle obtenue. On contrôle alors α en rejetant H_0 si $p < \alpha$.

Si on répète l'expérience sous H_0, en conservant un seuil α fixé, la probabilité d'obtenir au moins un faux positif augmente considérablement. De même, si on teste n hypothèses dont certaines sont nulles en proportion π_0, la proportion de faux positifs parmi tous les tests positifs augmente avec n pour tendre vers π_0. Il convient donc de définir un nouveau seuil α', tel que $p < \alpha'$ définisse une règle de rejet de H_0 qui contrôle à un niveau raisonnable le taux de faux positifs. Les méthodes pour définir α' sont fondées soit sur le calcul du taux d'erreur global (Family Wise Error Rate $(FWER)$), soit sur celui du taux de faux positifs (False discovery Rate (FDR)).

- Le $FWER$ est la probabilité d'obtenir au moins un faux positif si toutes les hypothèses sont nulles. En supposant tous les tests indépendants, $FWER(\alpha') = 1 - (1-\alpha')^n$. On peut alors calculer α', tel que $FWER(\alpha') = \alpha$. On obtient la correction de Sidak :

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

Bonferroni a montré dans le cas général, sans supposer les tests indépendants, que $FWER(\alpha') \leq n\alpha'$. Afin de contrôler $FWER(\alpha') = \alpha$, il suffit de prendre

$$\alpha'_B = \alpha / n .$$

Les méthodes fondées sur le calcul d'un $FWER$ sont extrêmement conservatrices et entraînent une forte perte de puissance.

- Le FDR est la probabilité d'obtenir un faux positif parmi tous les tests déclarés positifs. C'est donc la probabilité qu'une hypothèse soit nulle sachant qu'elle a été déclarée positive ($p < \alpha'$). En appliquant le théorème de Bayes des probabilités conditionnelles, on obtient :

$$FDR(\alpha') = P(H_0 / p \leq \alpha') = \frac{P(H_0)P(p \leq \alpha' / H_0)}{P(p \leq \alpha')} = \frac{\pi_0 \alpha'}{P(p \leq \alpha')}$$

En pratique, on ordonne les p-values par ordre croissant. On calcule pour chaque test, de rang $k(1,2,...,n)$, le FDR correspondant, ainsi qu'une q-value, q :

$$FDR(p_{(k)}) = \frac{\pi_0 p_{(k)}}{k/n}$$

$$q_{(k)} = \min(FDR(q_{(k)}), FDR(q_{(k+1)}))$$

En rejetant les hypothèses pour lesquelles $q_{(k)} < \alpha$, on contrôle le FDR au niveau α.

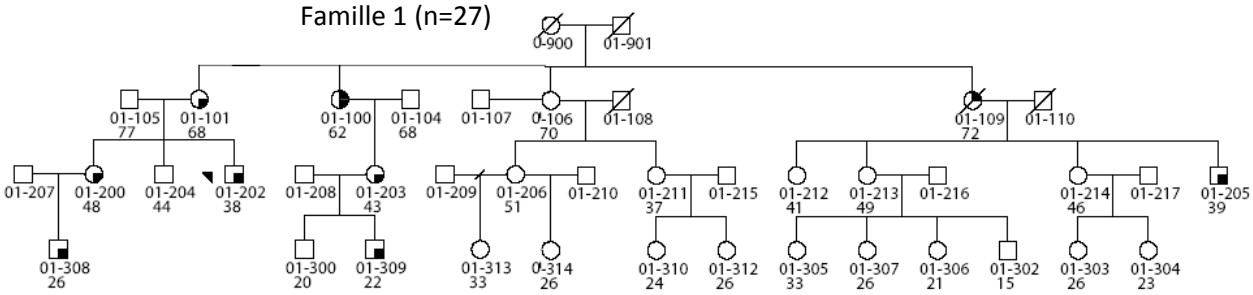
Annexe 2 : Arbres généalogiques de l'échantillon Familles-FVL

Légende
Cas Index
Décédé

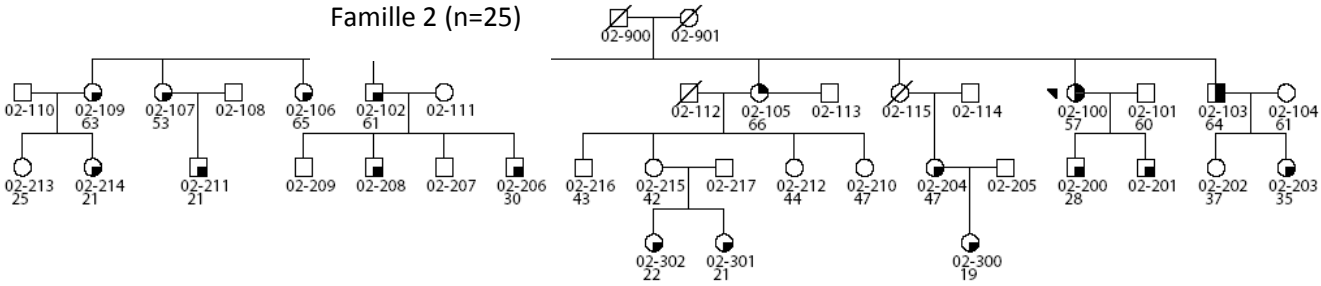
FVL
MTEV

La taille de la famille, n, inclut uniquement les individus dont on connaît les génotypes et les phénotypes

Famille 1 (n=27)

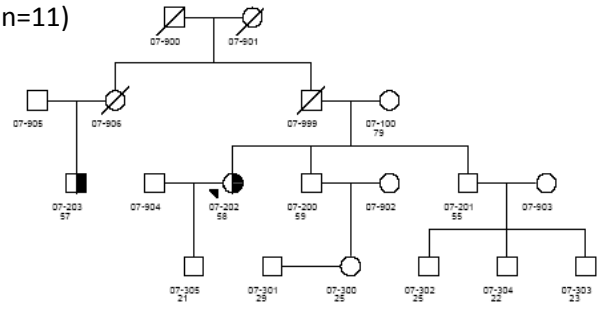


Famille 2 (n=25)

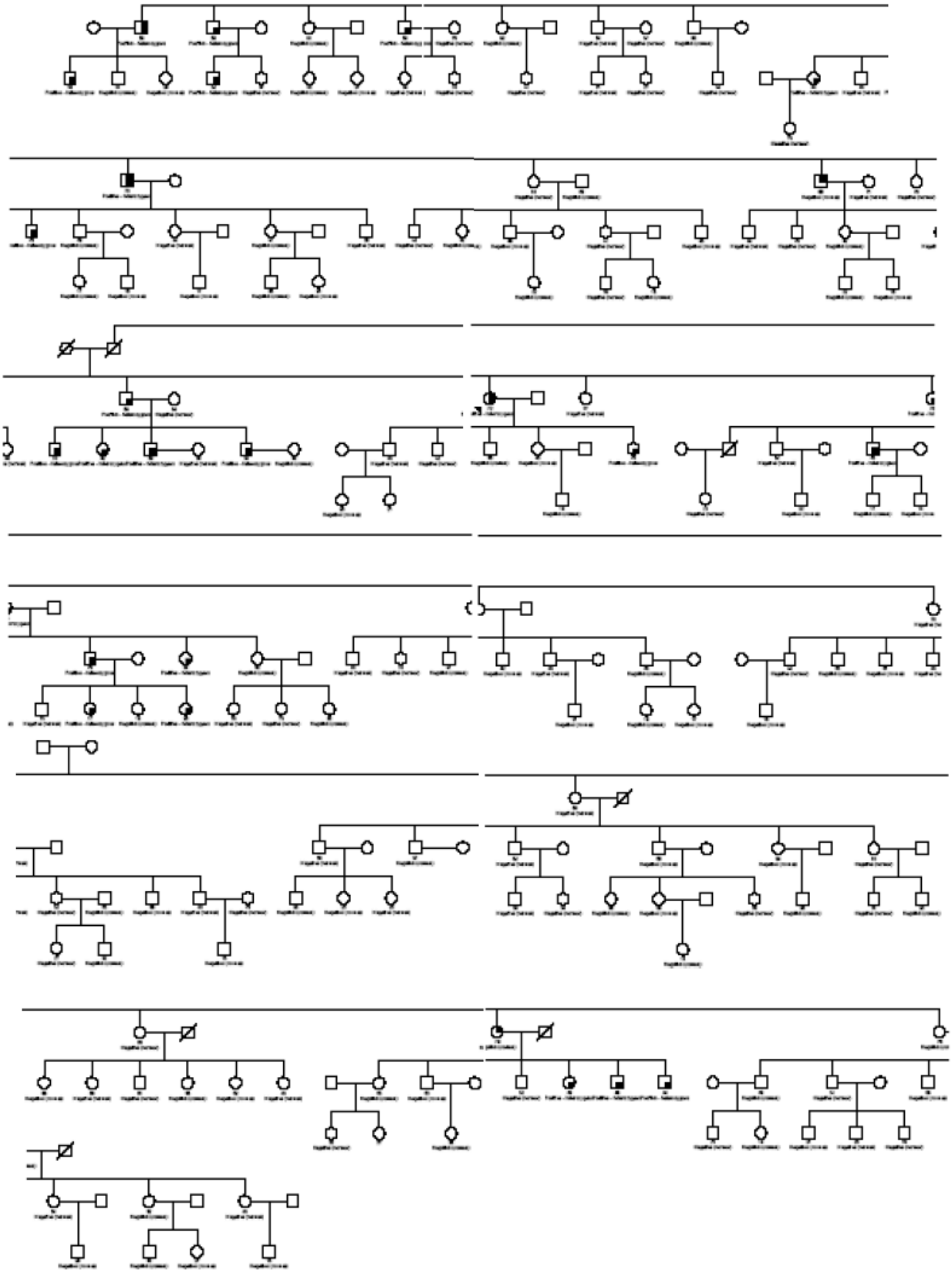


Famille 3 (n=41)

Famille 4 (n=11)

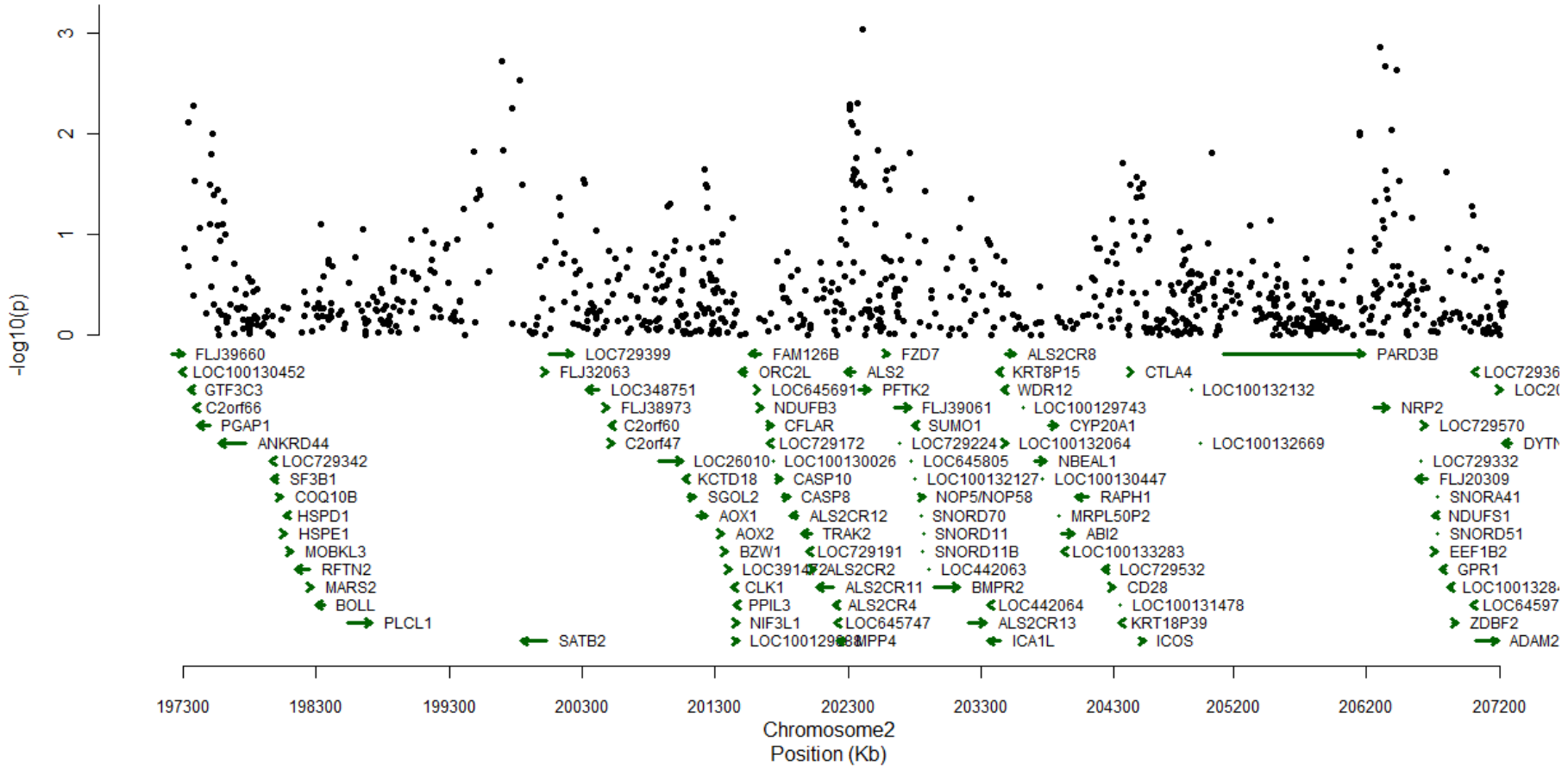


Famille 5 (n=165)

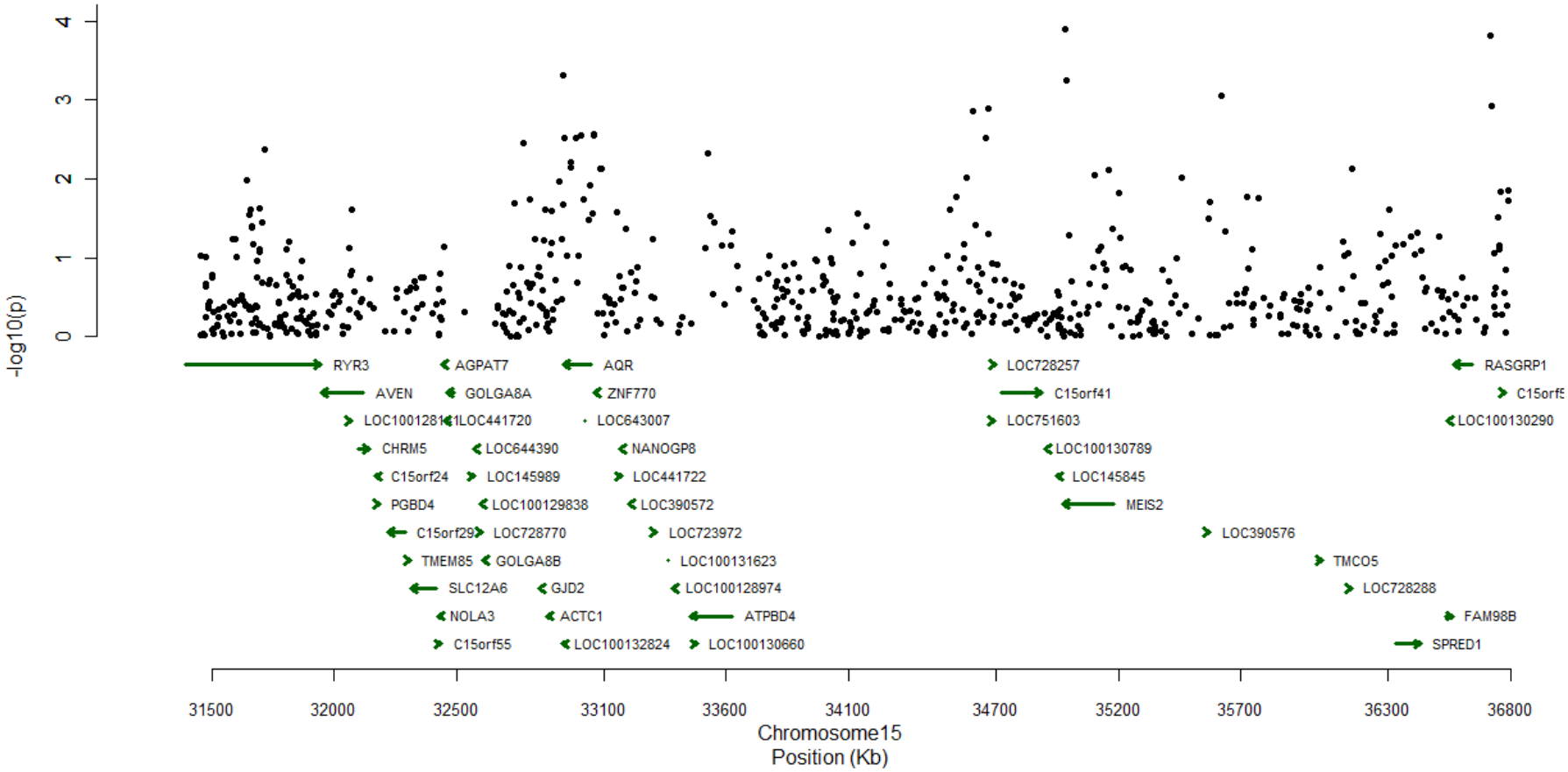


Annexe 3 : Associations avec la MTEV observées *in silico* dans une étude d'association pangénomique cas-témoins (453 cas de MTEV de moins de 50 ans et 1327 témoins)

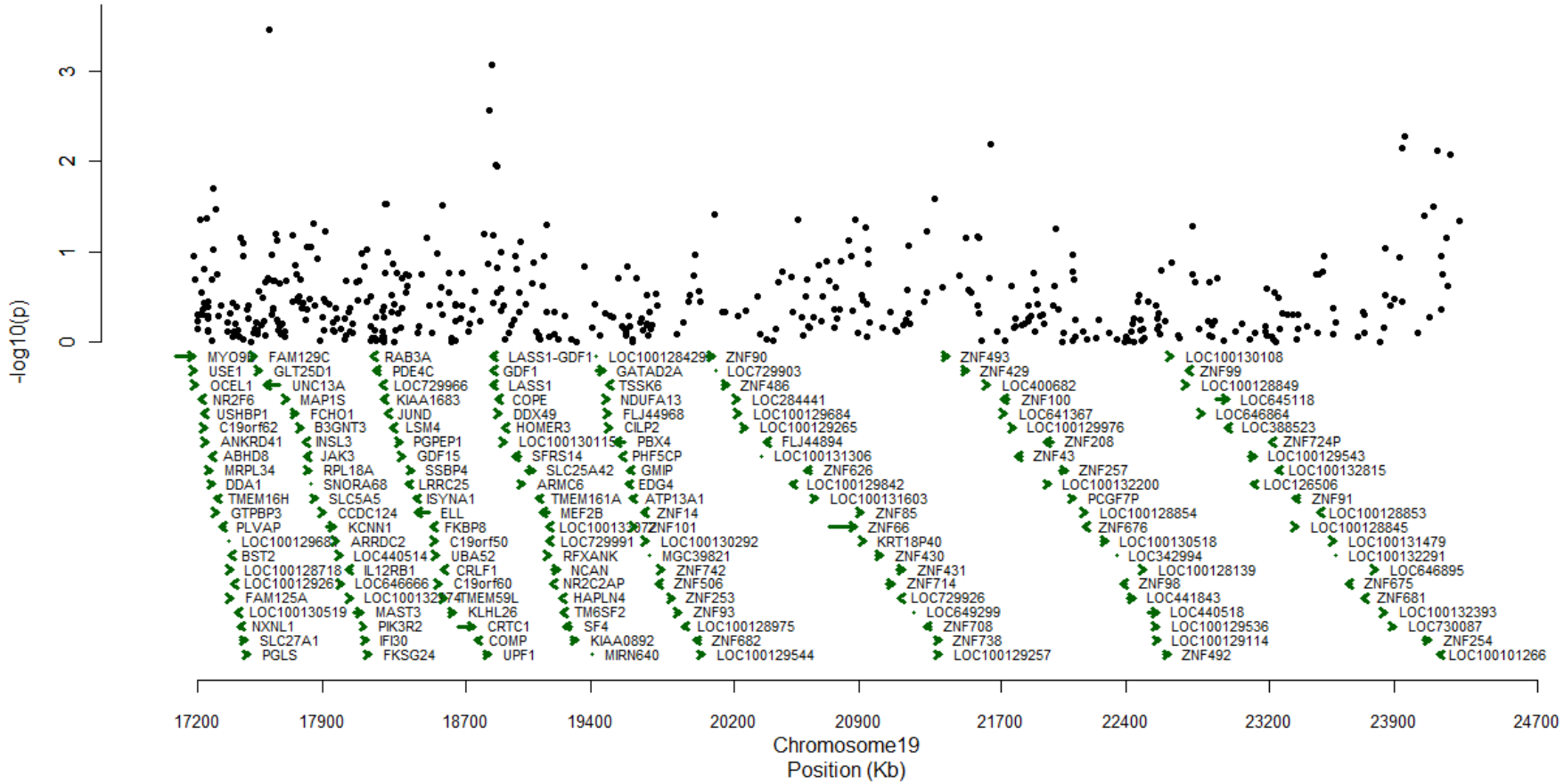
GWAS *in silico* : résultats obtenus dans la région 2q33



GWAS *in silico* : résultats obtenus dans la région 15q14



GWAS *in silico* : résultats obtenus dans la région 19p13



Annexe 4. Calcul de la vraisemblance des paramètres de l'analyse d'association génétique, fondée sur une méthode Décomposition de la Variance

$$L\left(\mu,\sigma_g^2,\sigma_e^2,\beta/\underline{Y},\underline{X}\right)=\left(\frac{1}{\sqrt{2\pi}}\right)^N \frac{1}{\sqrt{|\underline{\underline{\Omega}}|}}\exp\left(-\frac{1}{2}\underline{\Delta}^t\times\underline{\underline{\Omega}}^{-1}\times\underline{\Delta}\right) \qquad (1)$$

Avec μ , la moyenne phénotypique

σ_g^2 , la variance phénotypique due à des caractères génétiques

σ_e^2 , la variance phénotypique due aux expositions environnementales

β , l'effet de l'augmentation d'une unité d'une covariable x sur le phénotype

$\underline{Y}$, le vecteur des phénotypes, de taille N

$\underline{X}$, le vecteur de la covariable x , de taille N

N , le nombre de sujets dans l'échantillon

$\underline{\underline{\Omega}}$, la matrice de covariance phénotypique de taille $N * N$.

Elle se compose de deux covariances, l'une expliquée par le terrain génétique et l'autre par l'exposition environnementale :

$$\underline{\underline{\Omega}} = 2\underline{\underline{\Theta}}\sigma_g^2 + \underline{\underline{I}}\sigma_e^2$$

$\underline{\underline{\Theta}}$, la matrice de taille $N * N$ contenant les coefficients de parentés entre les sujets de l'étude

$\underline{\underline{I}}$, matrice identité de taille $N * N$

$\underline{\Delta}$, vecteur des différences entre, d'une part, les phénotypes observés, et d'autre part, leurs valeurs attendues sur la droite de régression $y = \mu + \beta x$

$$\Delta = \begin{bmatrix} y_1 - \mu - \beta x_1 \\ y_2 - \mu - \beta x_2 \\ \dots \\ y_N - \mu - \beta x_N \end{bmatrix}$$

On peut donc maintenant estimer $\mu,\sigma_g^2,\sigma_e^2$ et β par maximisation de la vraisemblance. Il existe classiquement deux méthodes de maximisation. La première, fondée un calcul exact, consiste à annuler la fonction dérivée de la fonction de vraisemblance. La seconde, empirique, envisage toutes les combinaisons possibles des paramètres, et sélectionne celle qui donne la plus grande vraisemblance. Le logiciel SOLAR (Sequential Oligogenic Linkage Analysis Routines) dans lequel est implémentée la méthode par VC, et que nous avons utilisé pour ce travail, utilise une méthode empirique de maximisation de la vraisemblance. Il s'appuie sur un programme d'optimisation du choix des paramètres afin d'accélérer la progression vers le maximum.

Remarque : Si on suppose que toutes les observations sont indépendantes les unes des autres, $\underline{\underline{\Omega}}$ est une matrice diagonale du type $\sigma^2 \underline{\underline{I}}$ (σ^2 étant la variance phénotypique). On retrouve la formule de la vraisemblance pour des observations issues d'une loi normale. En effet,

$$\left|\underline{\underline{\Omega}}\right| = \left(\sigma^2\right)^N$$

et

$$\underline{\Delta}^t \times \underline{\underline{\Omega}}^{-1} \times \underline{\Delta} = \underline{\Delta}^t \times \frac{I}{\sigma^2} \times \underline{\Delta}$$

$$= \frac{1}{\sigma^2} \times \underline{\Delta}^t \times \underline{\Delta}$$

$$= \sum_{i=1}^{N} \frac{(y_i - \mu - \beta x_i)^2}{\sigma^2}$$

(1) s'écrit alors :

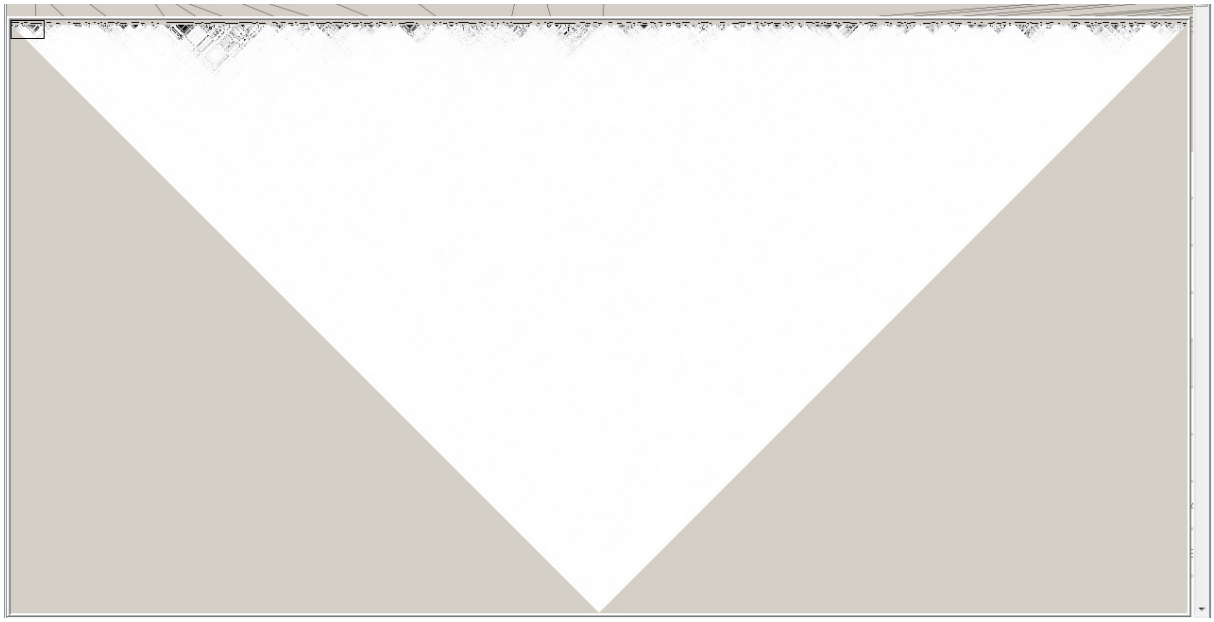
$$L\left(\mu, \sigma_g^2, \sigma_e^2, \beta / \underline{Y}, \underline{X}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(\sum_{i=1}^{N} - \frac{(y_i - \mu - \beta x_i)^2}{2\sigma^2}\right)$$

$$L\left(\mu, \sigma_g^2, \sigma_e^2, \beta / \underline{Y}, \underline{X}\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu - \beta x_i)^2}{2\sigma^2}\right)$$

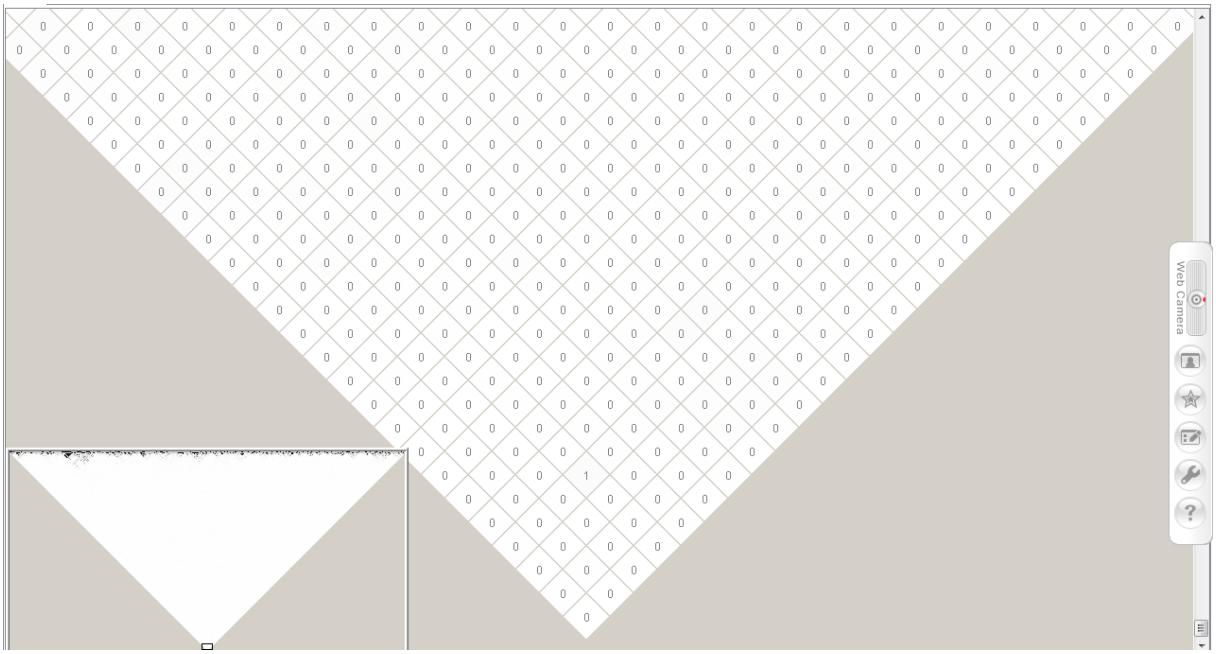
Annexe 5. Comparaison du déséquilibre de liaison des échantillons MARTHA et Familles FVL observé au sein des régions 9q34 et 12q23

Matrice des valeurs r^2 observées dans MARTHA08 en 9q34 entre 132 600 et 137 000 kb

A



B

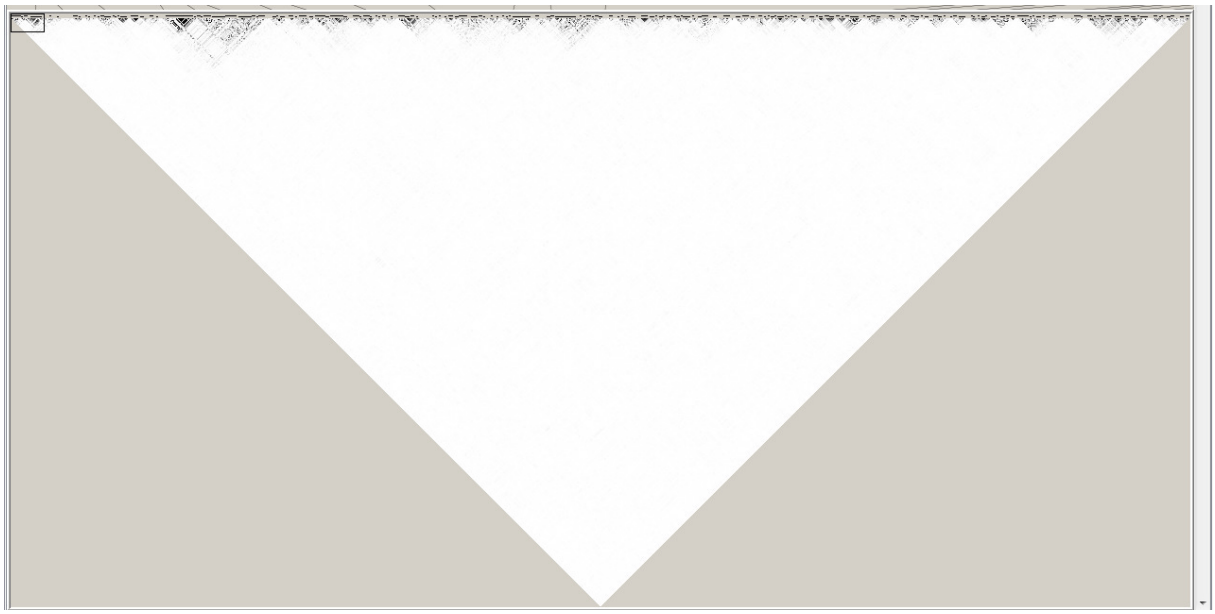


A : ensemble de la région (moyenne de r^2=0,09%)

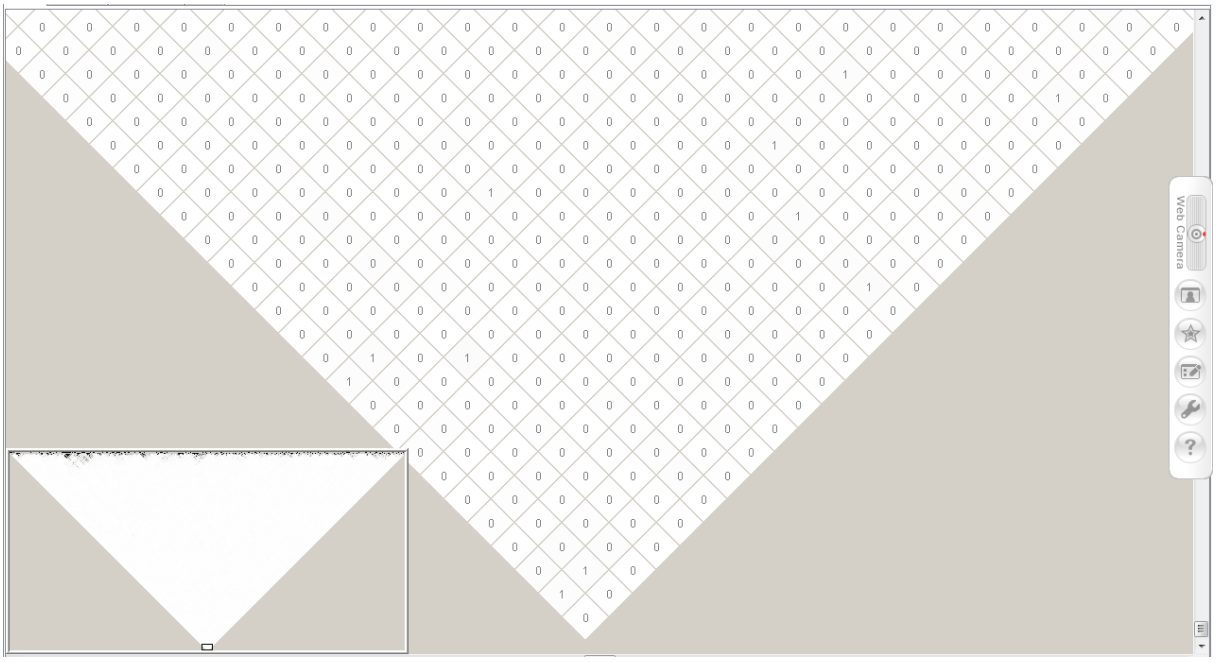
B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs r^2 observées dans MARTHA10 en 9q34 entre 132 600 et 137 000 kb

A



B

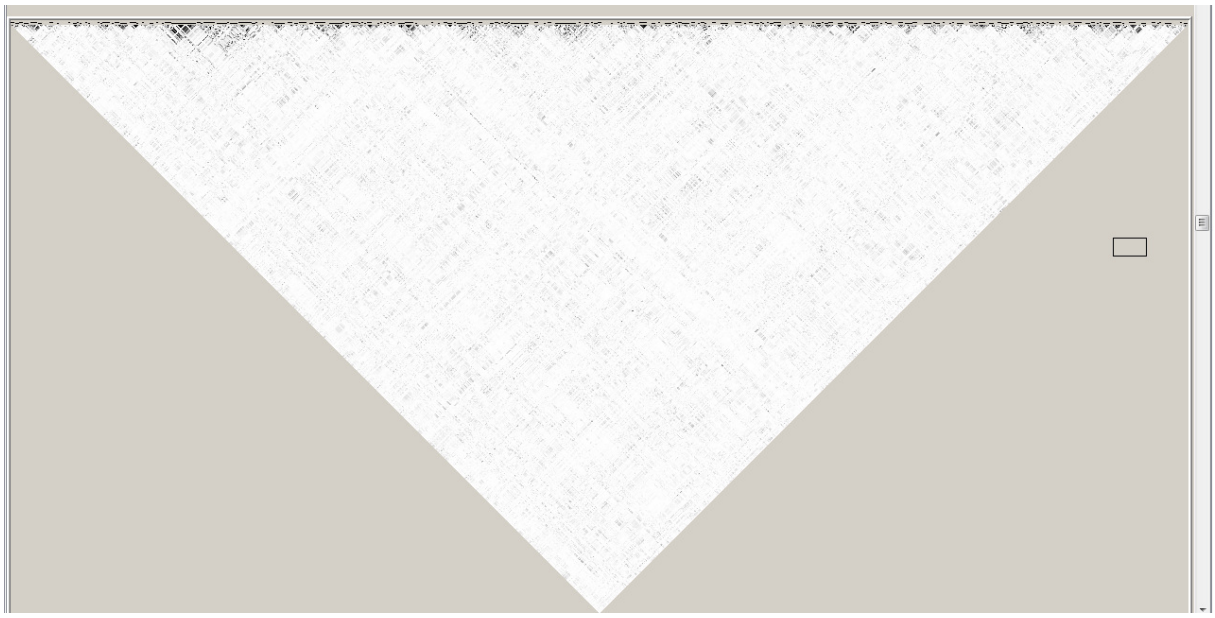


A : ensemble de la région (moyenne de r^2=0,09%)

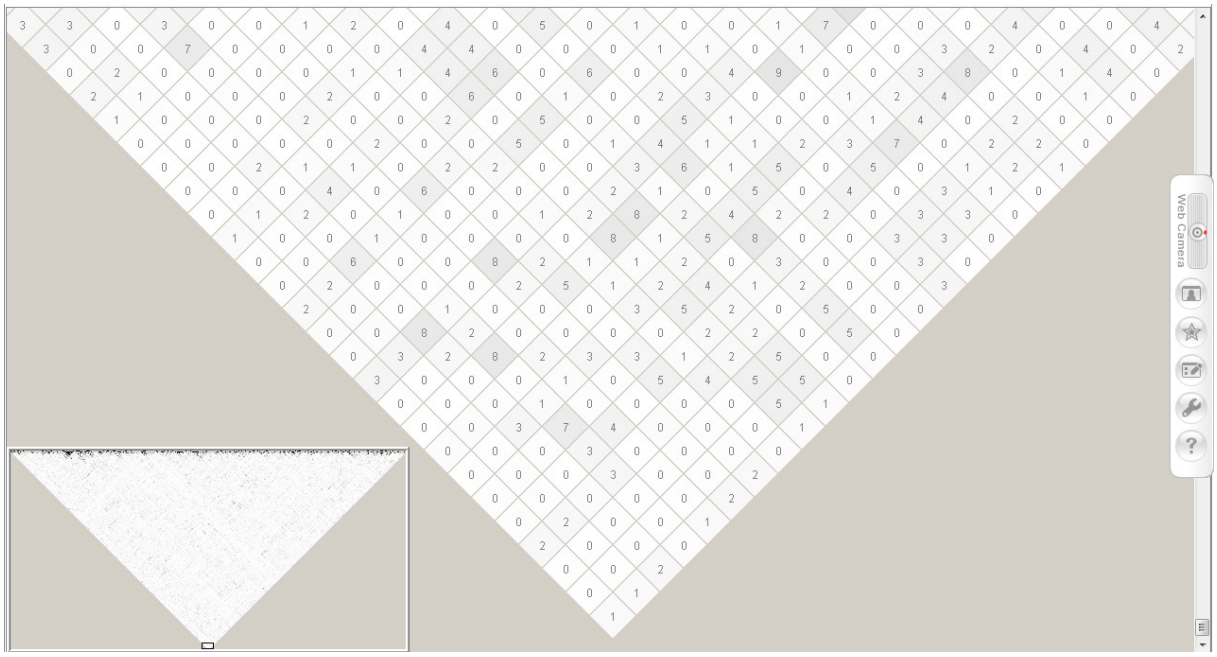
B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs r^2 observées dans Famille-FVL en 9q34 entre 132 600 et 137 000 kb

A



B



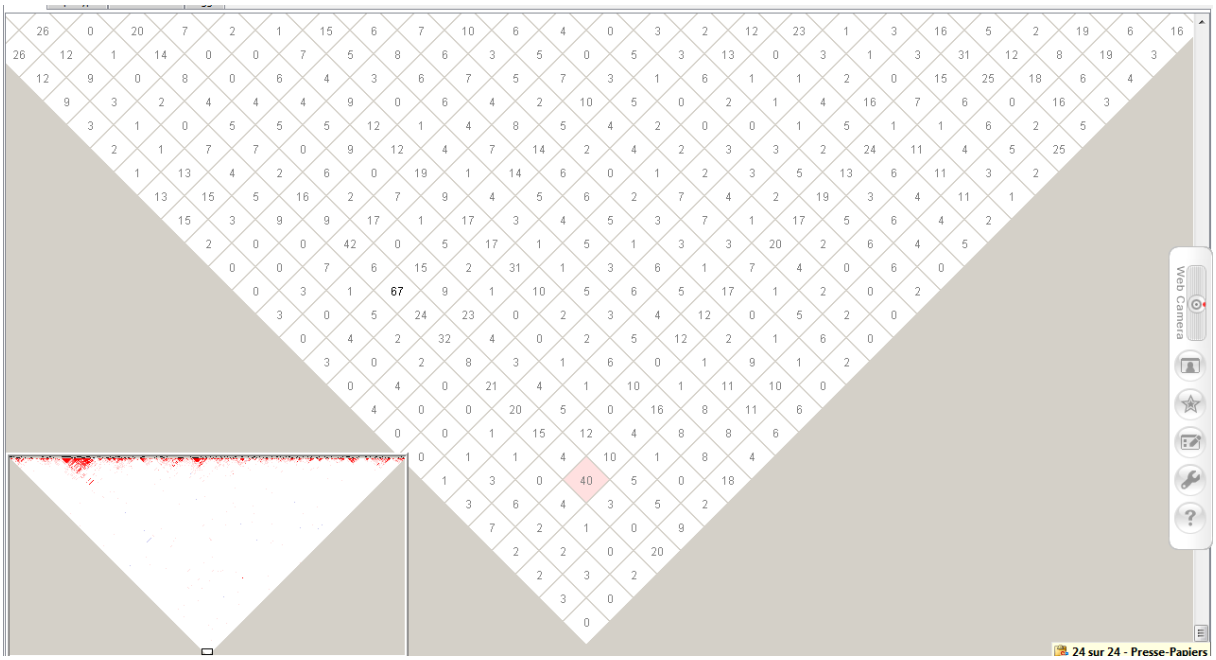
A : ensemble de la région (moyenne de r^2=1,15%)

B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs absolues de D' observées dans MARTHA08 en 9q34 entre 132 600 et 137 000 kb

A

B



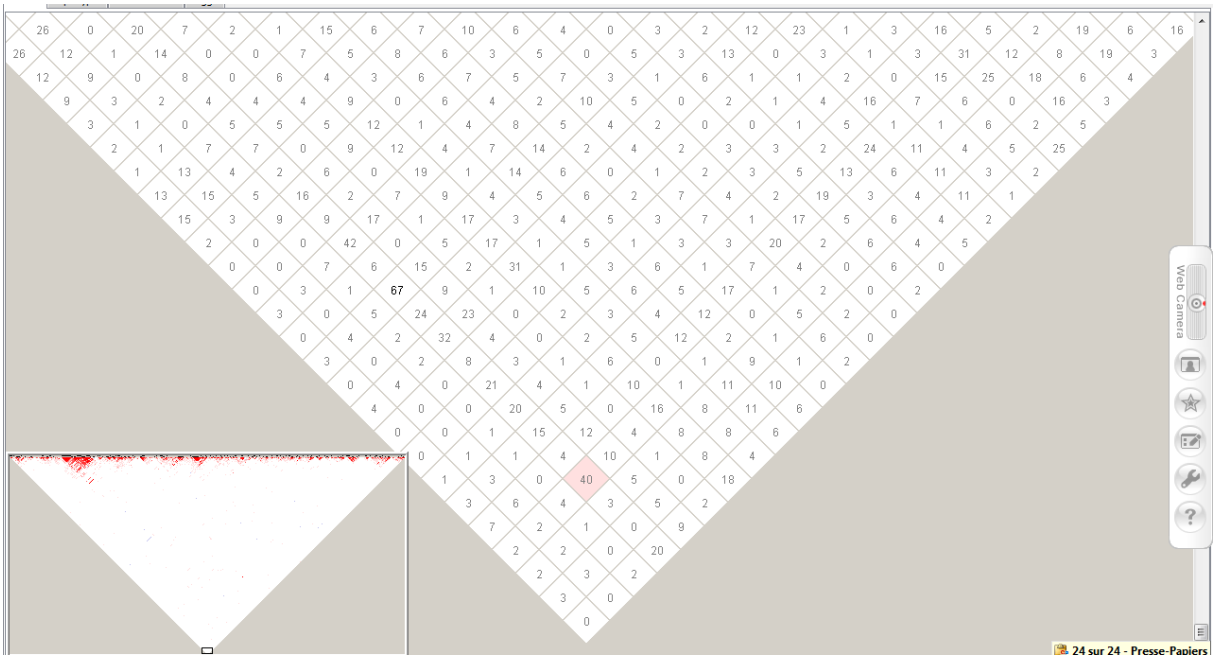
A : ensemble de la région (moyenne des valeurs absolues de D'=0,05)

B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs absolues de D' observées dans MARTHA10 en 9q34 entre 132 600 et 137 000 kb

A

B

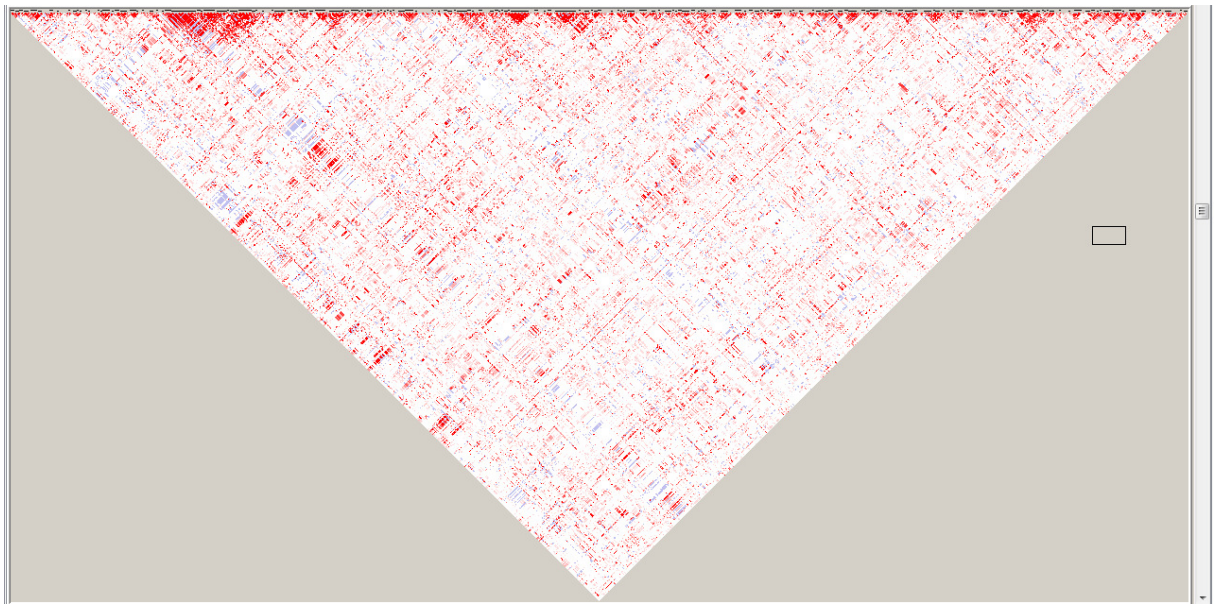


A : ensemble de la région (moyenne des valeurs absolues de D'=0,05)

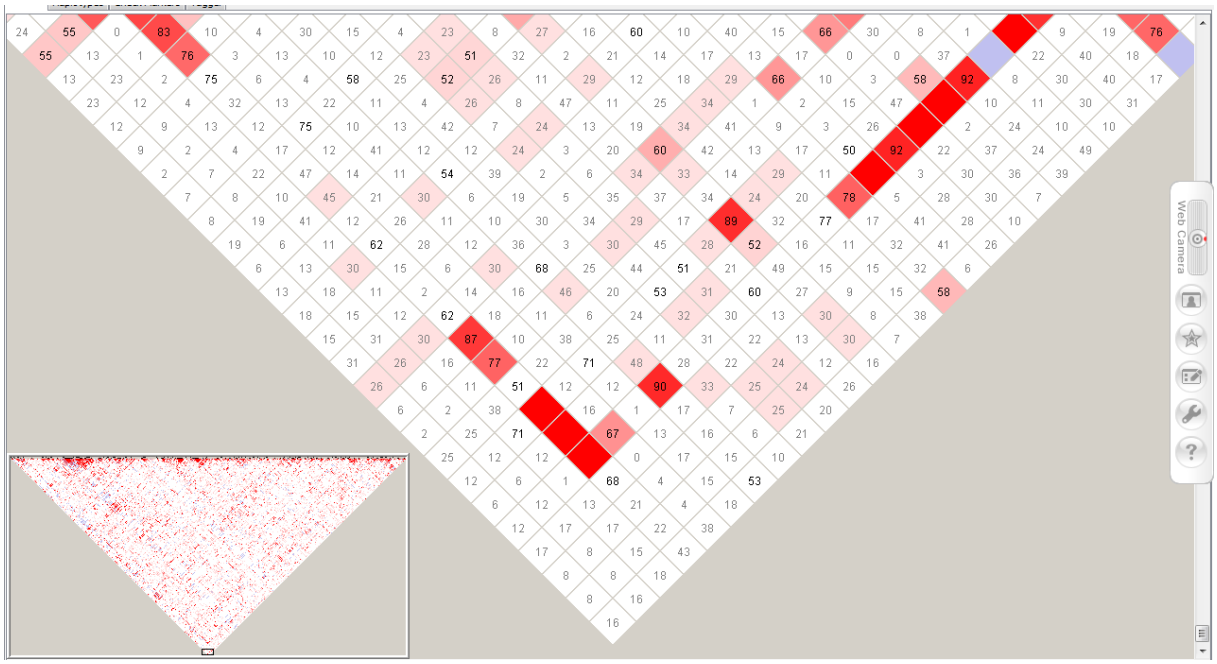
B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs absolues de D' observées dans Familles-FVL en 9q34 entre 132 600 et 137 000 kb

A



B

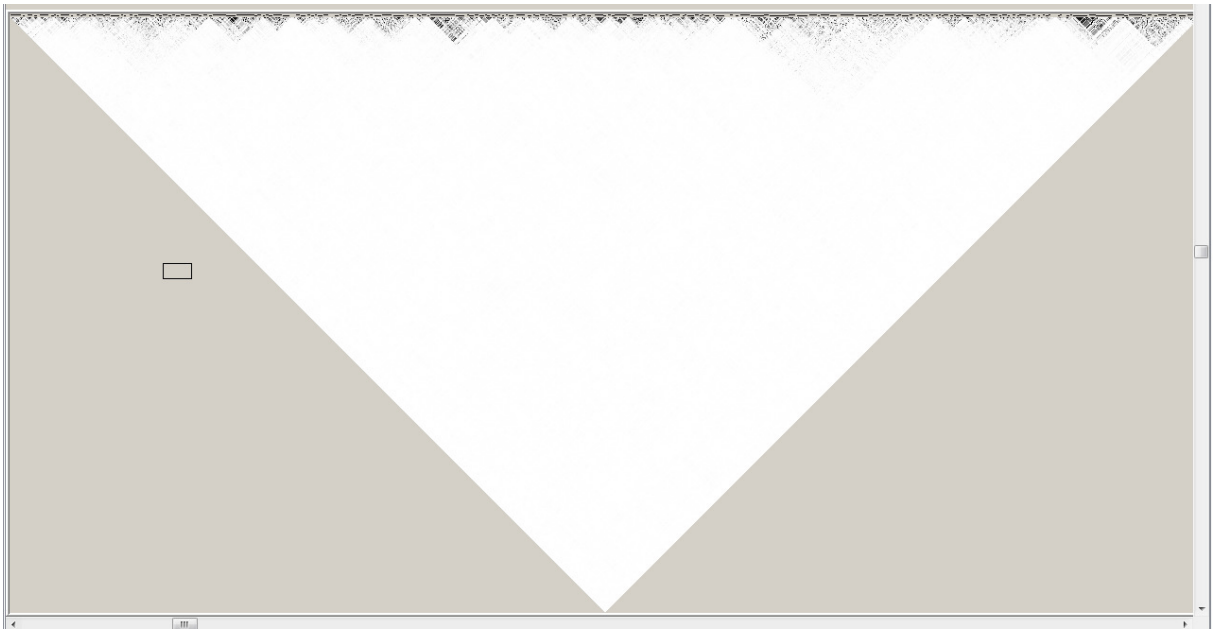


A : ensemble de la région (moyenne des valeurs absolues de D'=0,22)

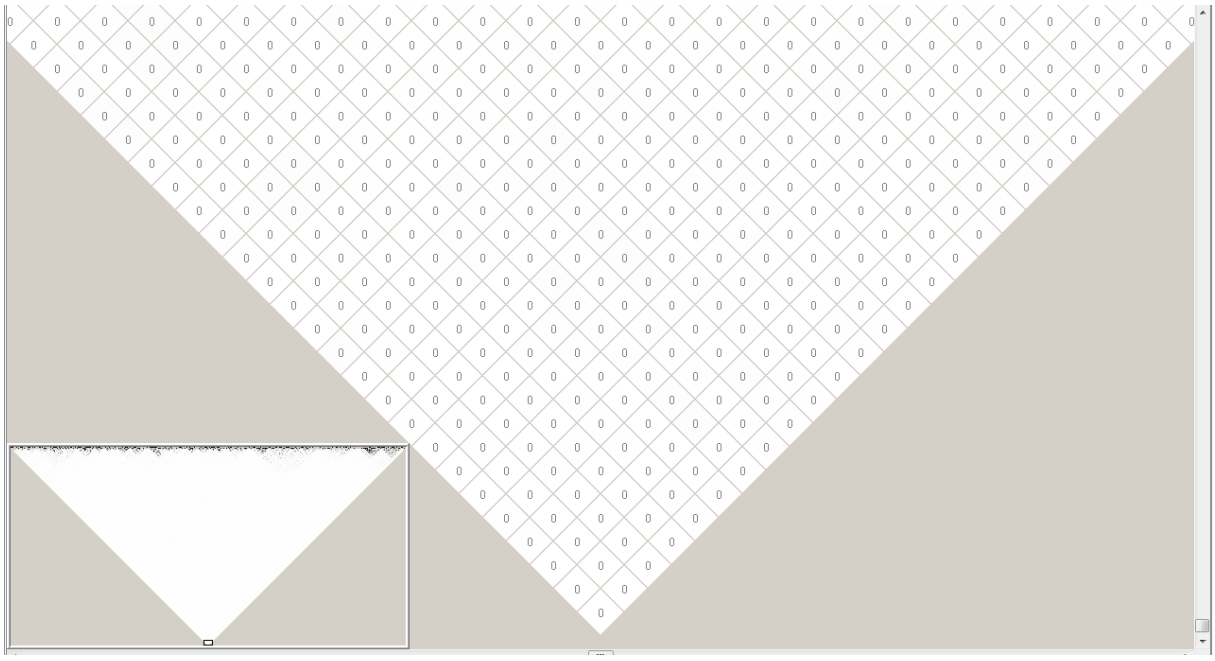
B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs r^2 observées dans MARTHA08 en 12q23 entre 99 000 et 104 300 kb

A



B

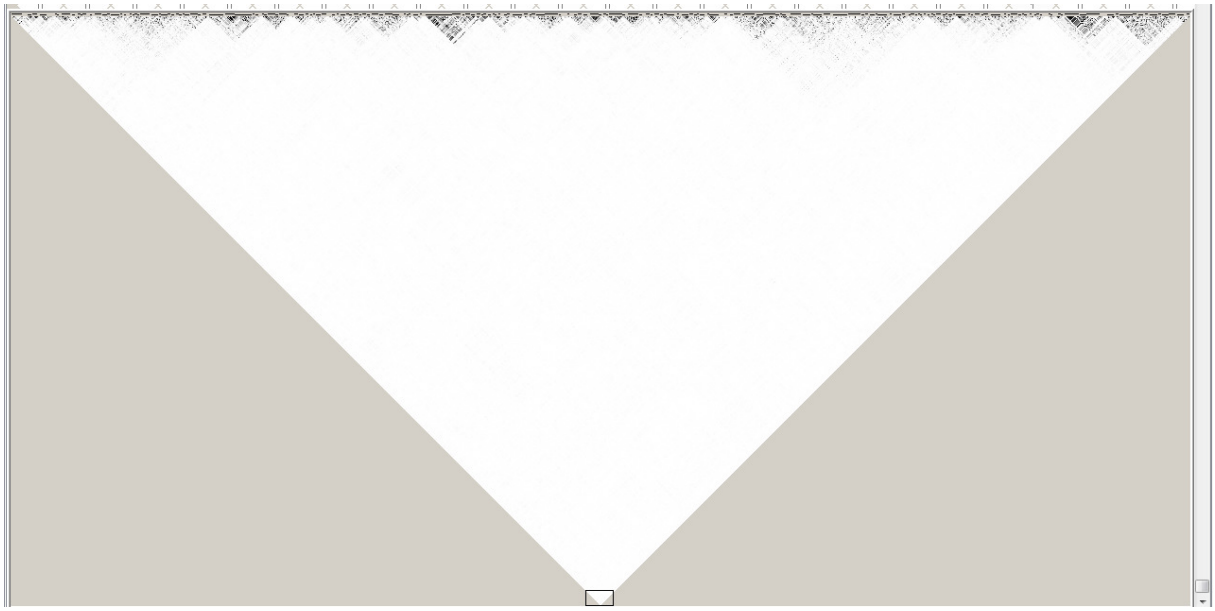


A : ensemble de la région (moyenne de r^2=0,10%)

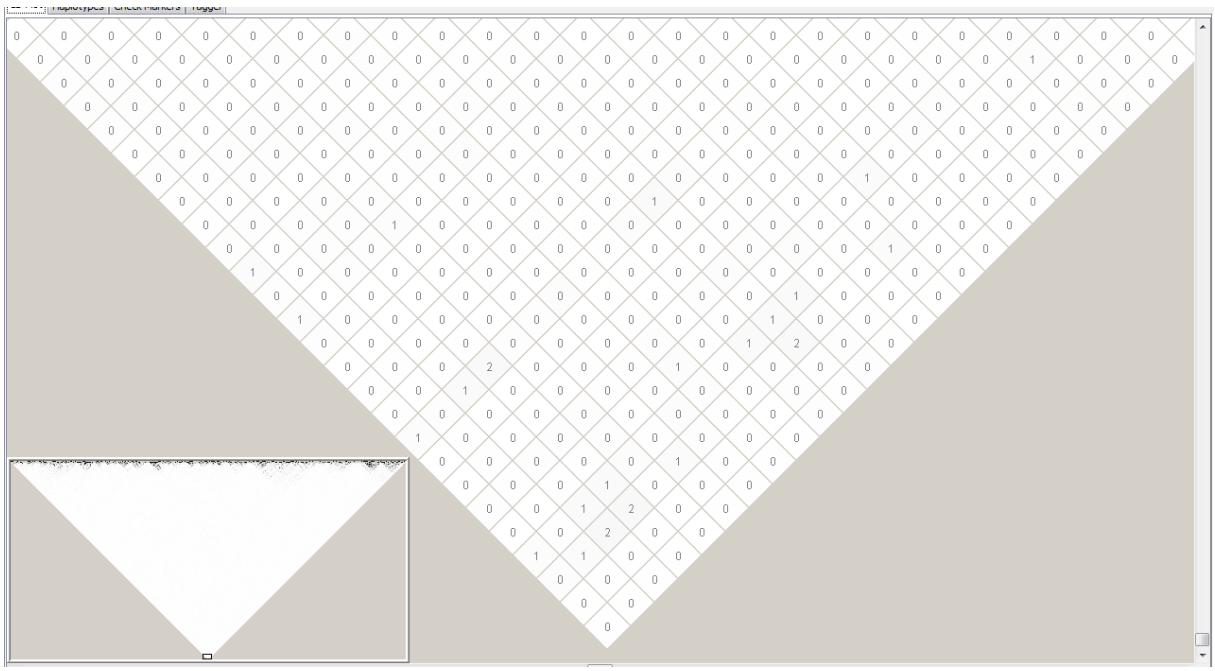
B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs r^2 observées dans MARTHA10 en 12q23 entre 99 000 et 104 300 kb

A



B

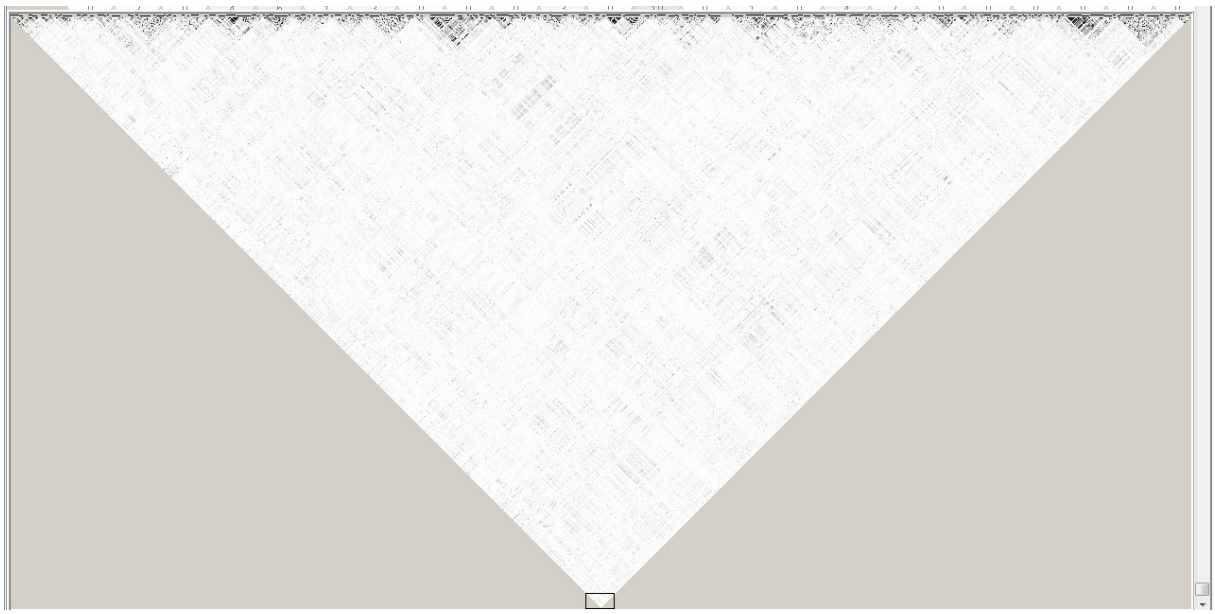


A : ensemble de la région (moyenne de r^2=0,08%)

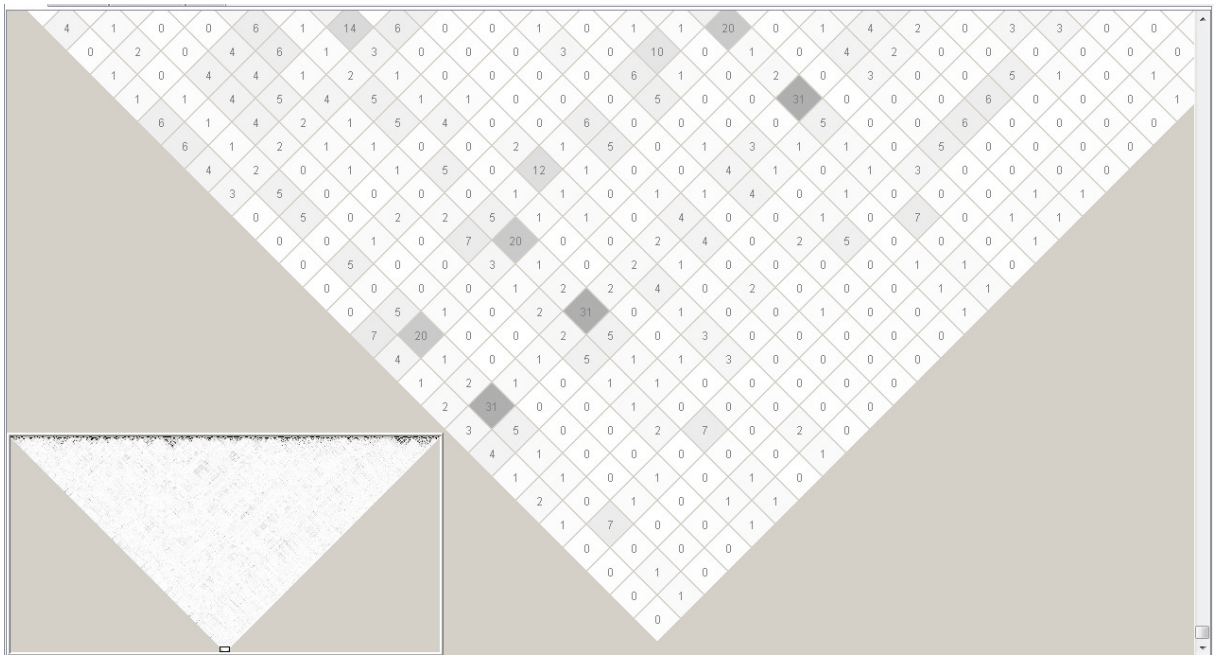
B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs r^2 observées dans Familles-FVL en 12q23 entre 99 000 et 104 300 kb

A



B

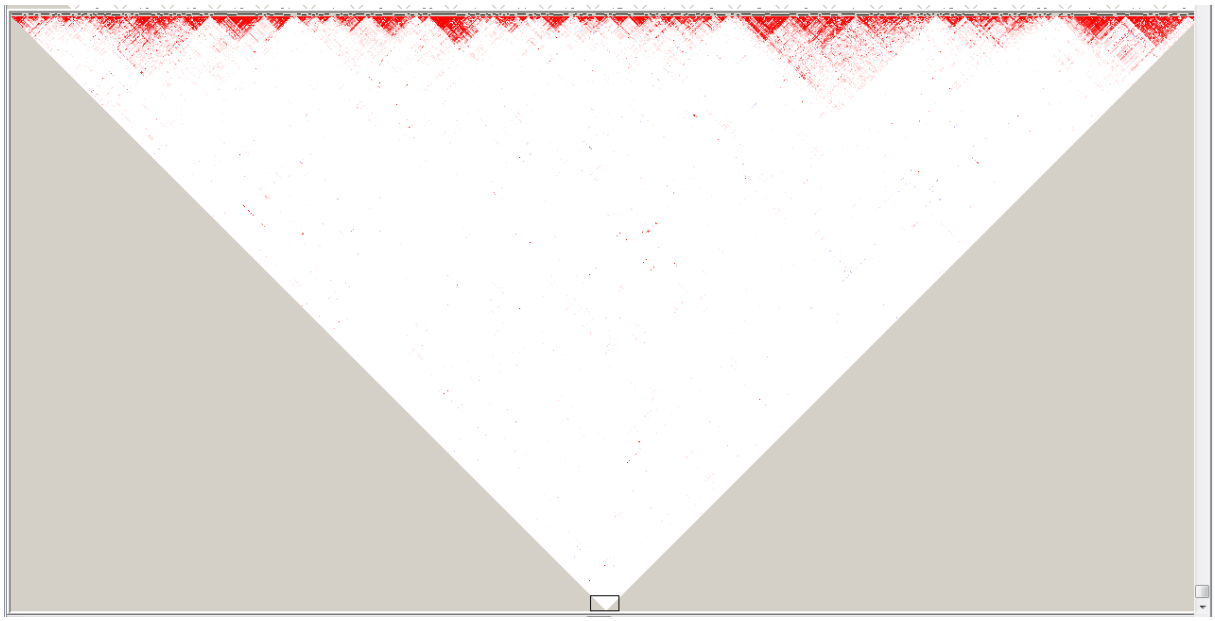


A : ensemble de la région (moyenne de r^2=1,6%)

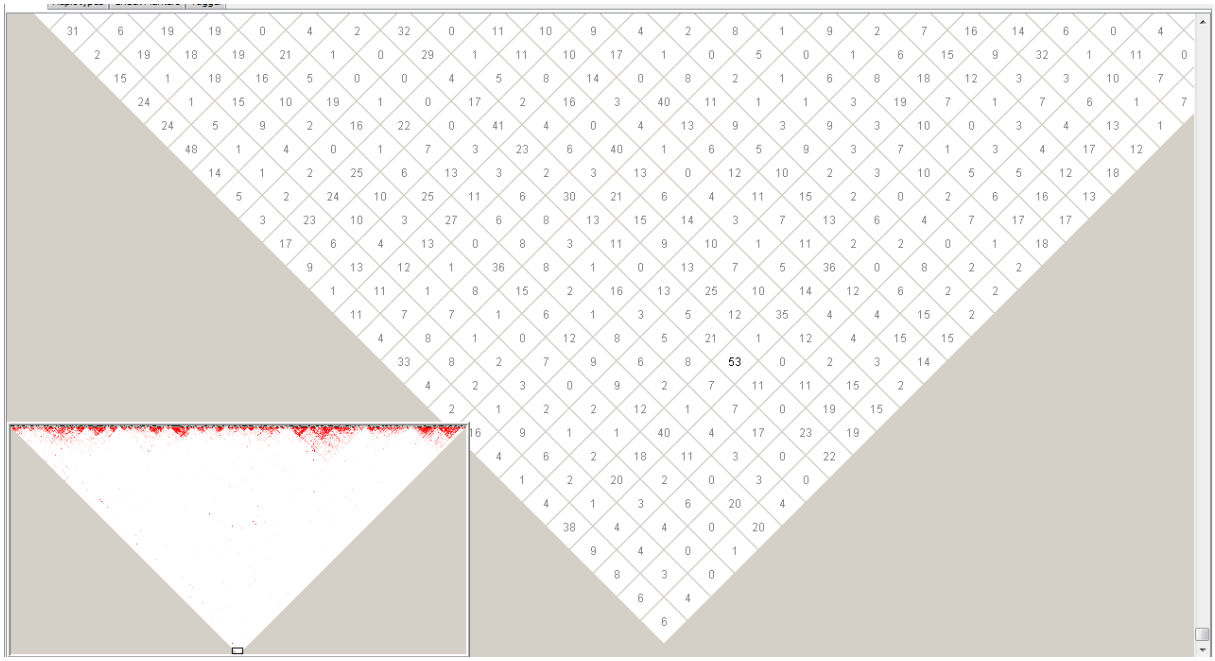
B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs absolues de D' observées dans MARTHA08 en 12q23 entre 99 000 et 104 300 kb

A



B



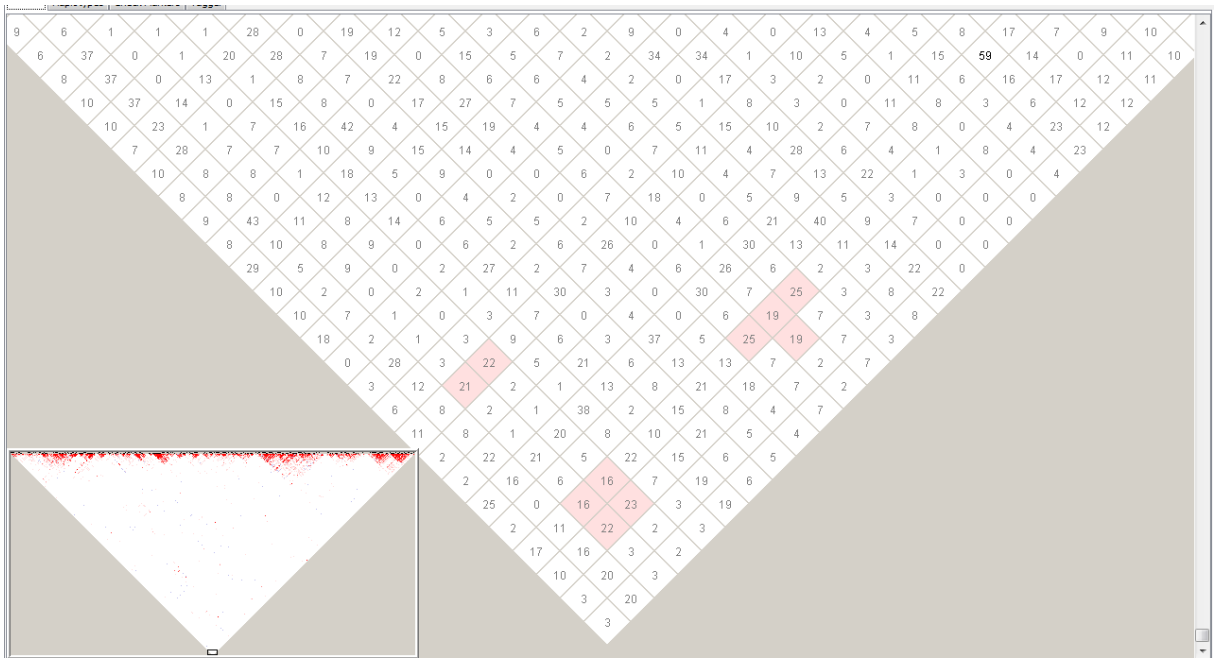
A : ensemble de la région (moyenne des valeurs absolues de D'=0,09)

B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs absolues de D' observées dans MARTHA10 en 12q23 entre 99 000 et 104 300 kb

A

B

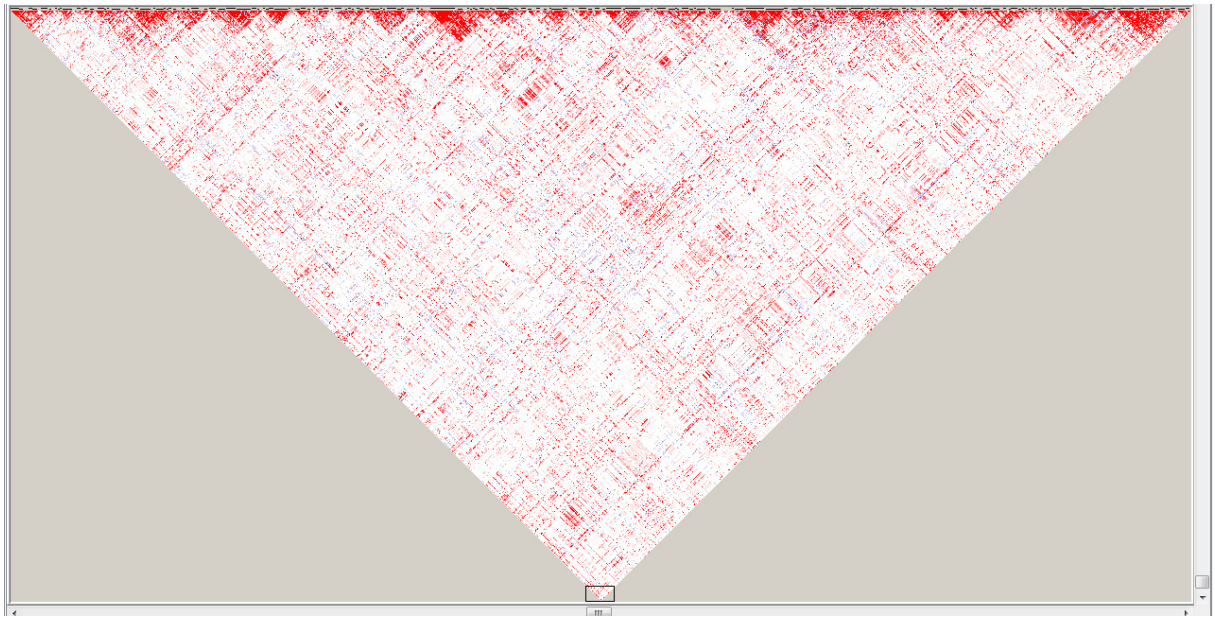


A : ensemble de la région (moyenne des valeurs absolues de D'=0,08)

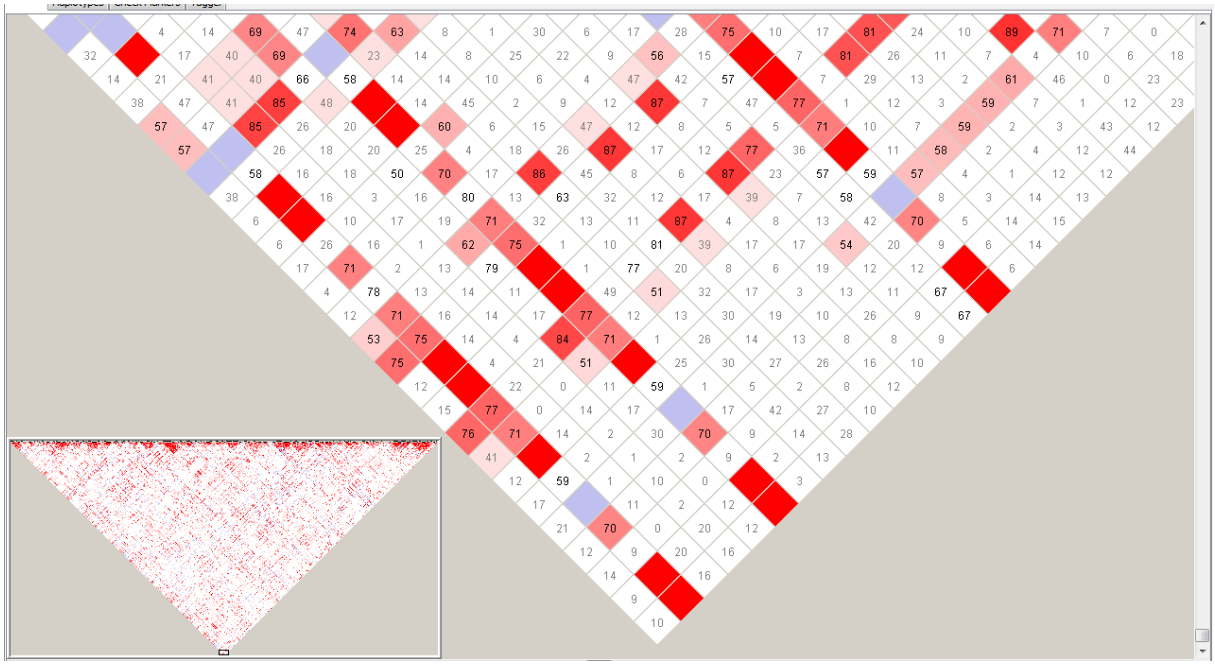
B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Matrice des valeurs absolues de D' observées dans Familles-FVL en 12q23 entre 99 000 et 104 300 kb

A



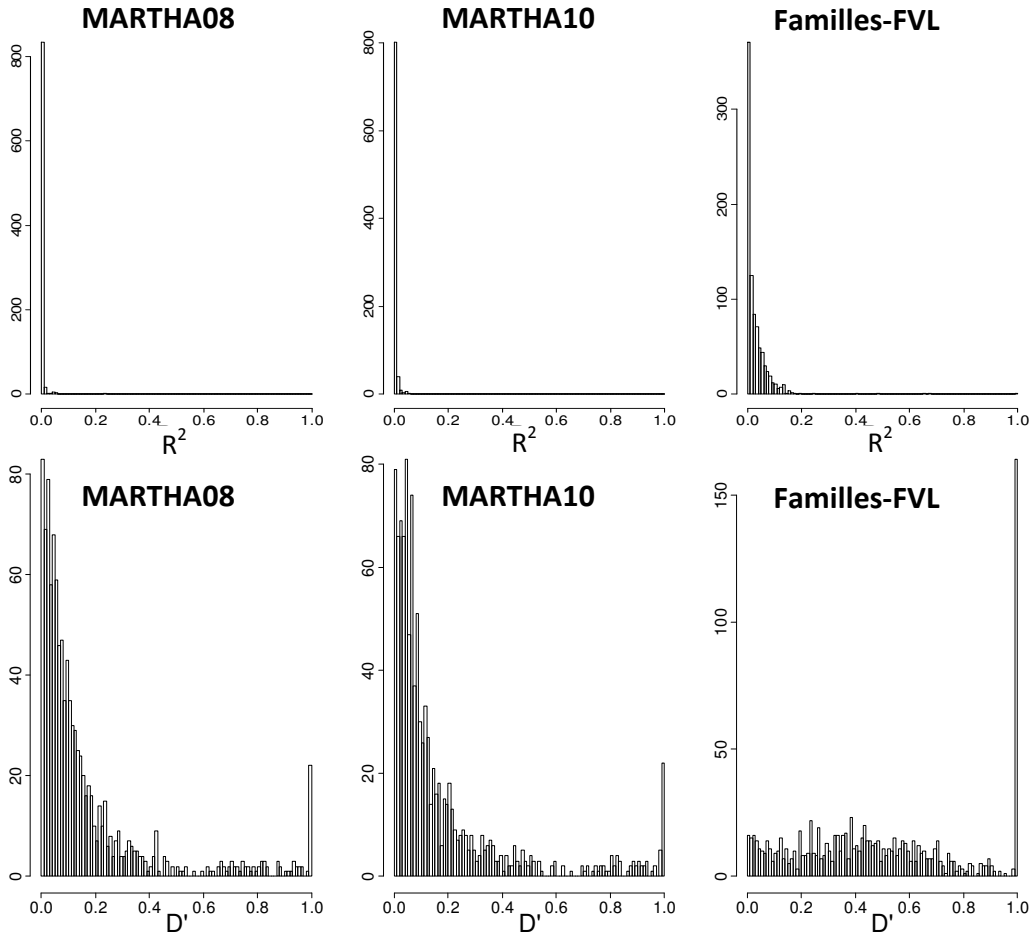
B



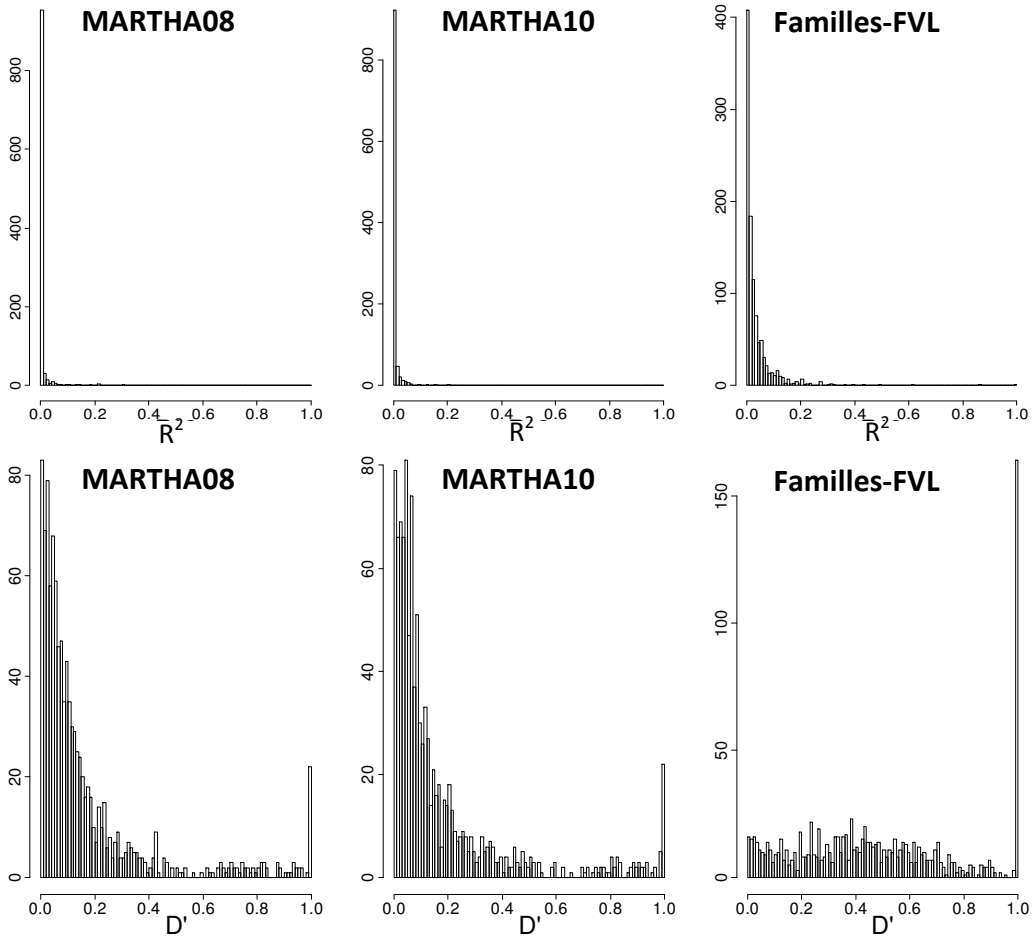
A : ensemble de la région (moyenne des valeurs absolues de D'=0,45)

B : zoom sur la pointe de la figure A, afin d'estimer l'importance de déséquilibre de liaison sur de grandes distances

Distribution de la mesure du DL entre *ABO*-rs505922 et les autres SNPs de la région 9q34



Distribution de la mesure du DL entre TMEM16D-rs7137089 et les autres SNPs de la région 12q23



Annexe 6 . Associations ($p<10^{-5}$) obtenus avec l'échantillon Familles-FVL

Analyse des taux de vWF

SNP	CHR	BP	gene	loc	alleles	MAF	effet	p	qval	p M08	p M10
rs1188414	1	30872516	MATN1	3'UTR	G>A	0.49	-0.47	7.24e-06	0.123	0.985	0.362
rs4494771	2	71574170	DYSF	intron	G>A	0.08	0.76	7.51e-06	0.124	0.921	0.507
rs11901729	2	76801348	LOC647278	3'UTR	G>A	0.37	0.43	8.64e-06	0.127	0.383	0.180
rs26367	5	132917543	FSTL4	intron	A>G	0.19	0.58	7.70e-06	0.125	0.786	0.873
rs33613	5	132925203	FSTL4	intron	G>A	0.19	0.58	7.70e-06	0.125	0.622	0.794
rs381879	5	132961027	FSTL4	intron	A>G	0.19	0.59	4.25e-06	0.107	0.470	0.175
rs393092	5	132978698	FSTL4	5'UTR	G>A	0.18	0.59	6.81e-06	0.121	0.123	0.310
rs13293102	9	132785431	FIBCD1	intron	A>G	0.14	0.63	6.96e-06	0.122	0.459	0.978
rs7035540	9	133426383	LOC642515	3'UTR	G>A	0.16	0.67	1.66e-07	**0.016**	0.105	0.662
rs7024039	9	133427549	LOC642515	3'UTR	A>C	0.16	0.67	1.66e-07	**0.016**	0.091	0.596
rs4740283	9	133438117	RAPGEF1	3'UTR	A>G	0.16	0.67	1.66e-07	**0.016**	0.267	0.771
rs943852	9	133441014	RAPGEF1	3'UTR	G>A	0.16	0.67	1.66e-07	**0.016**	0.287	0.771
rs1410562	9	133473626	RAPGEF1	intron	C>A	0.18	0.59	3.24e-06	0.097	0.351	0.112
rs1887787	9	133479584	RAPGEF1	intron	G>A	0.25	0.49	9.66e-06	0.133	NA	NA
rs2282008	9	133482300	RAPGEF1	intron	A>G	0.18	0.59	3.24e-06	0.097	0.267	0.112
rs17553234	9	133575559	RAPGEF1	intron	A>G	0.13	0.79	1.11e-08	**0.005**	0.240	0.059
rs3012740	9	133644997	RAPGEF1	5'UTR	A>G	0.31	0.46	3.96e-06	0.104	0.505	0.266
rs3011272	9	134873822	GFI1B	3'UTR	G>C	0.36	-0.48	7.47e-06	0.124	0.022	0.151
rs3011274	9	134878337	GTF3C5	5'UTR	A>G	0.36	-0.48	7.47e-06	0.124	0.015	0.104
rs3011278	9	134880812	GTF3C5	5'UTR	A>G	0.36	-0.48	7.47e-06	0.124	0.022	0.151
rs505922*	9	135139050	ABO	intron	A>G	0.38	0.5	3.04e-07	**0.021**	1.94e-10	1.57e-09
rs495828**	9	135144688	ABO	5'UTR	C>A	0.41	0.46	2.00e-06	0.078	4.51e-08	1.15e-04
rs642193	9	135616727	VAV2	3'UTR	A>C	0.45	-0.44	3.16e-06	0.096	0.191	0.232
rs2797826	9	135619220	VAV2	intron	A>G	0.43	0.43	1.46e-06	0.066	0.380	0.921
rs12339163	9	136345009	RXRA	5'UTR	A>G	0.22	0.56	1.43e-06	0.065	0.611	0.259
rs7864699	9	136864192	COL5A1	intron	A>G	0.16	0.65	2.43e-07	**0.019**	0.837	0.562
rs7849389	9	137214193	OLFM1	3'UTR	G>A	0.28	0.48	1.88e-06	0.076	0.848	0.136
rs7956222	12	99730699	TMEM16D	intron	A>G	0.15	0.61	8.20e-06	0.126	0.930	0.897
rs7137089	12	99732588	TMEM16D	intron	G>A	0.15	0.62	2.95e-06	0.093	0.498	0.668
rs8078427	17	11471437	DNAH9	intron	G>A	0.35	-0.44	8.21e-06	0.126	0.318	0.909
rs707312	19	57564835	ZNF610	3'UTR	G>A	0.26	0.48	7.75e-06	0.125	0.073	0.153
rs158536	20	52148709	BCAS1	5'UTR	A>G	0.41	0.45	5.38e-06	0.114	0.151	0.250
rs2064863	20	54396179	AURKA	intron	C>A	0.36	-0.45	8.30e-06	0.126	0.553	0.332

* l'allèle fréquent tague le groupe O (O1 et O2 confondus, versus les autres groupes)
** l'allèle rare tague les groupes A1 (majoritairement) et O2 (minoritairement)

Analyse des taux de vWF ajustés sur ABO

SNP	CHR	BP	gene_symbol	location	alleles	MAF	effet	P (Fam-FVL)	qval	p M08	p-M10
rs6691335	1	88189178	LMO4	3'UTR	A>G	0.19	0.53	5.76e-07	**0.048**	0.83	**0.03**
rs1534452	3	84130486	LOC643665	5'UTR	G>A	0.26	-0.4	6.46e-06	0.074	0.52	0.55
rs7623142	3	84152607	LOC643665	5'UTR	G>A	0.26	-0.4	6.46e-06	0.074	0.54	0.54
rs4856241	3	84177508	LOC643665	5'UTR	G>A	0.26	-0.4	6.46e-06	0.074	0.50	0.57
rs956469	4	11155373	HS3ST1	5'UTR	A>C	0.46	0.37	2.31e-06	0.058	0.55	0.45
rs11945500	4	11156353	HS3ST1	5'UTR	G>A	0.46	0.37	2.82e-06	0.059	0.47	0.46
rs1439594	5	124547087	LOC644659	5'UTR	C>A	0.15	0.52	7.59e-06	0.081	0.08	0.58
rs13354264	5	138511464	SIL1	intron	G>A	0.13	0.55	4.89e-06	0.068	0.15	0.23
rs12653458	5	161131389	GABRA6	3'UTR	G>A	0.1	0.67	8.25e-06	0.084	0.88	0.10
rs1286005	6	6991451	LOC728570	3'UTR	G>A	0.32	0.36	8.75e-06	0.087	0.29	0.53
rs9372313	6	112248783	FYN	intron	G>A	0.07	0.75	3.11e-06	0.059	0.22	0.53
rs2031908	9	78380667	LOC392352	5'UTR	A>G	0.28	0.41	2.25e-06	0.058	0.16	0.62
rs12818362	12	92915037	CRADD	3'UTR	G>A	0.09	0.65	4.91e-06	0.068	0.81	0.44
rs7956222	12	99730699	TMEM16D	intron	A>G	0.15	0.6	1.02e-07	**0.023**	0.69	0.75
rs7137089	12	99732588	TMEM16D	intron	G>A	0.15	0.66	1.78e-09	**0.001**	0.23	0.43
rs6538973	12	99738954	TMEM16D	intron	A>G	0.26	0.4	2.69e-06	0.059	0.89	0.57
rs1554074	12	99764481	TMEM16D	intron	G>A	0.22	0.5	9.55e-07	0.053	0.62	0.61
rs2138046	12	99779713	TMEM16D	intron	A>G	0.22	0.5	9.55e-07	0.053	0.42	0.55
rs703704	12	99797499	TMEM16D	intron	A>G	0.22	0.5	9.55e-07	0.053	0.42	0.94
rs10861171	12	103141908	LOC390354	5'UTR	G>A	0.16	0.55	1.41e-06	0.056	0.68	0.93
rs12828801	12	103297120	TXNRD1	3'UTR	G>A	0.11	0.59	5.76e-06	0.072	0.52	0.13
rs7954880	12	103474111	CHST11	intron	C>A	0.21	0.45	4.52e-06	0.066	0.65	0.20
rs17231251	12	103659698	CHST11	intron	A>G	0.13	0.59	3.37e-06	0.059	0.79	0.57
rs10492014	12	111124937	NULL	intron	G>A	0.12	0.6	3.20e-06	0.059	0.53	0.81
rs17822304	12	111373075	PTPN11	intron	C>A	0.12	0.59	5.45e-06	0.07	0.53	0.58
rs1465542	12	112347326	SDSL	intron	G>A	0.08	0.79	4.47e-07	**0.045**	0.08	0.78
rs736122	12	112391009	LHX5	intron	G>A	0.1	0.62	2.56e-06	0.058	0.69	0.60
rs12817205	12	112402354	LHX5	5'UTR	C>A	0.1	0.61	3.51e-06	0.06	0.93	0.46
rs4767079	12	112405815	LHX5	5'UTR	A>G	0.1	0.62	2.56e-06	0.058	0.91	0.47
rs616135	12	112602185	RBM19	3'UTR	G>A	0.1	0.66	1.30e-06	0.055	0.75	0.35
rs1881666	12	112644629	RBM19	3'UTR	C>A	0.09	0.62	5.47e-06	0.07	0.30	0.17
rs16943162	12	112659157	RBM19	3'UTR	A>C	0.09	0.63	3.36e-06	0.059	0.96	0.20
rs7969141	12	115699000	TMEM118	intron	A>G	0.18	0.5	3.11e-06	0.059	0.13	0.68
rs7314211	12	115718907	TMEM118	intron	A>C	0.17	0.51	2.03e-06	0.058	0.13	0.96
rs6490093	12	115725599	TMEM118	intron	G>A	0.16	0.51	6.19e-06	0.073	0.09	0.15
rs1495933	12	115728126	TMEM118	intron	G>A	0.16	0.52	2.18e-06	0.058	0.08	0.44
rs924075	12	115734417	TMEM118	intron	A>G	0.16	0.52	2.18e-06	0.058	0.03	0.50
rs10774890	12	115741081	TMEM118	intron	C>A	0.16	0.52	2.18e-06	0.058	0.04	0.44
rs903772	12	115774984	TMEM118	3UTR	G>A	0.17	0.5	4.76e-06	0.067	0.09	0.90
rs2393108	12	115778024	TMEM118	3'UTR	G>A	0.15	0.52	8.48e-06	0.085	0.11	0.94
rs4767470	12	115813464	HRK	5'UTR	G>A	0.1	0.65	4.79e-06	0.067	0.40	0.59
rs12866745	13	41057423	KIAA0564	intron	G>A	0.18	-0.5	4.18e-06	0.064	0.62	0.73
rs4942351	13	44201711	LOC144817	3'UTR	G>A	0.14	0.53	8.56e-06	0.086	0.48	**0.04**
rs11841892	13	58199835	LOC341689	3'UTR	G>A	0.35	0.39	1.89e-06	0.057	0.54	0.91
rs9538196	13	58250988	LOC341689	3'UTR	A>G	0.31	0.37	7.14e-06	0.078	0.83	0.96
rs1077829	19	59624545	TTYH1	intron	G>A	0.36	0.39	8.99e-06	0.088	0.61	0.29
rs1541103	21	41535214	BACE2	intron	A>G	0.32	0.41	1.40e-06	0.056	0.35	0.65

Analyse des taux de FVIII

SNP	CHR	BP	gene_symbol	location	alleles	MAF	effet	P	qval	p M08	p-M10
rs2070493	8	37847908	RAB11FIP1	intron	G>A	0.15	0.66	7.33e-06	0.265	0.780	0.894
rs4733677	8	128781003	NULL	5'UTR	G>A	0.17	0.6	4.16e-06	0.265	0.553	0.781
rs7035540	9	133426383	LOC642515	3'UTR	G>A	0.16	0.57	9.83e-06	0.265	0.580	0.568
rs7024039	9	133427549	LOC642515	3'UTR	A>C	0.16	0.57	9.83e-06	0.265	0.555	0.517
rs4740283	9	133438117	RAPGEF1	3'UTR	A>G	0.16	0.57	9.83e-06	0.265	0.518	0.716
rs943852	9	133441014	RAPGEF1	3'UTR	G>A	0.16	0.57	9.83e-06	0.265	0.451	0.716
rs505922	9	135139050	ABO	intron	A>G	0.38	0.46	3.41e-06	0.265	1.18e-06	3.69e-08
rs7864699	9	136864192	COL5A1	intron	A>G	0.16	0.62	8.60e-07	0.265	0.201	0.808
rs11594791	10	3549601	KLF6	3'UTR	A>G	0.04	-1.02	8.44e-06	0.265	0.348	0.958
rs11825217	11	127908214	ETS1	5'UTR	G>A	0.3	-0.5	1.85e-06	0.265	0.316	0.519
rs2001625	11	130103423	C11orf44	3'UTR	G>A	0.46	-0.45	1.96e-06	0.265	0.445	0.103
rs7956222	12	99730699	TMEM16D	intron	A>G	0.15	0.63	4.92e-06	0.265	0.288	0.540
rs7137089	12	99732588	TMEM16D	intron	G>A	0.15	0.62	4.54e-06	0.265	0.620	0.322
rs2141834	18	66329898	GTSCR1	3'UTR	G>A	0.28	-0.5	4.58e-06	0.265	0.124	0.091

Analyse des taux de FVIII ajustés sur les taux de vWF

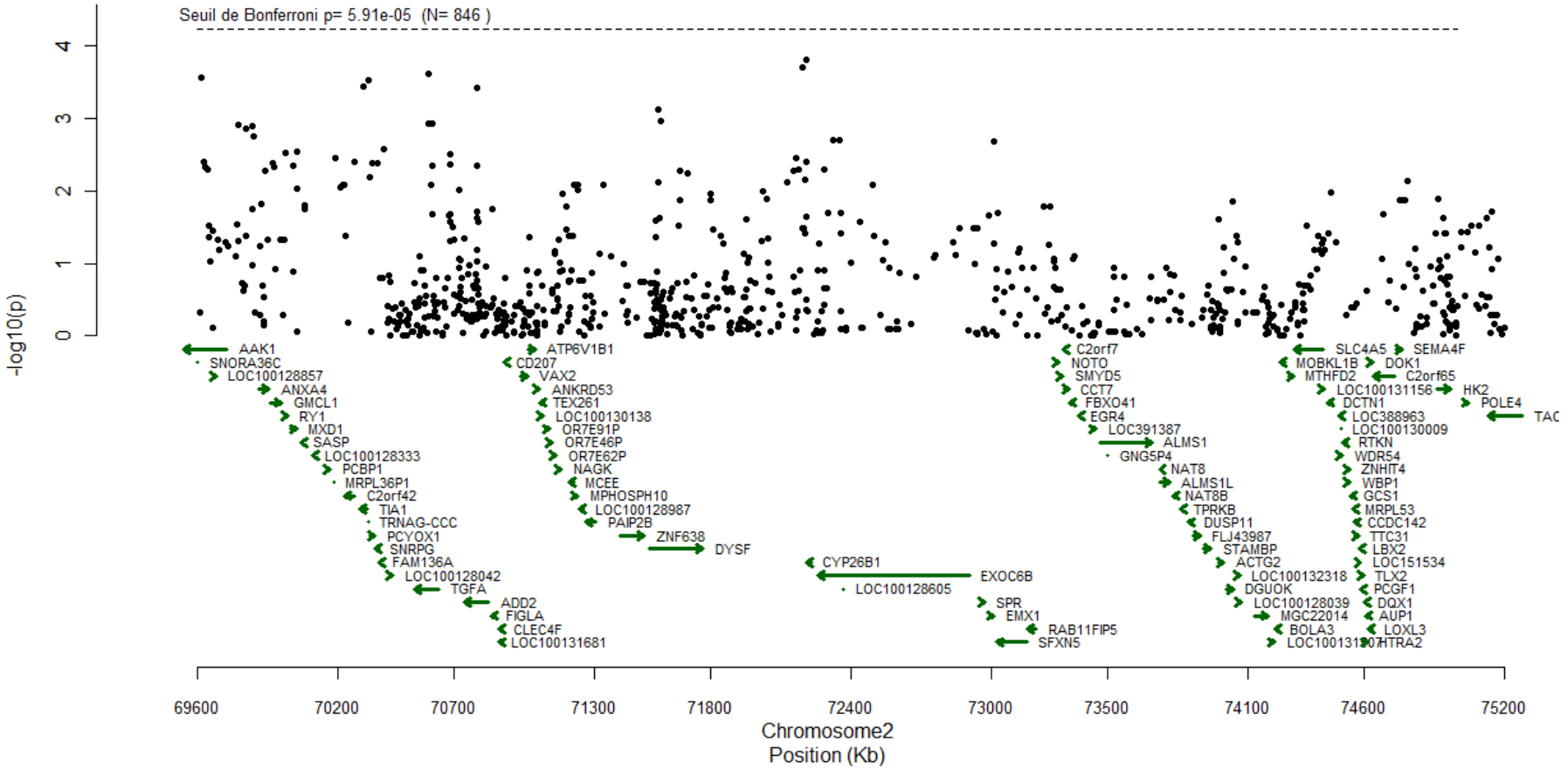
SNP	CHR	BP	gene_symbol	location	alleles	MAF	effet	P	qval	p M08	p-M10
rs4733677	8	128781003	NULL	5'UTR	G>A	0.17	0.62	1.81e-06	0.433	0.263	0.086
rs11594791	10	3549601	KLF6	3'UTR	A>G	0.04	-1.11	1.04e-06	0.433	0.724	0.079
rs11813909	10	78879196	LOC729187	3'UTR	G>A	0.08	0.83	8.97e-06	0.554	0.600	0.280
rs7314453	12	118016813	KIAA1853	intron	A>G	0.07	0.88	5.10e-06	0.514	0.533	0.769
rs11609210	12	118019384	KIAA1853	intron	G>A	0.07	0.88	5.41e-06	0.517	0.533	0.769
rs11611821	12	118020092	KIAA1853	intron	A>G	0.07	0.88	6.59e-06	0.526	0.533	0.769
rs11612089	12	118041732	KIAA1853	intron	A>C	0.07	0.88	5.10e-06	0.514	0.921	0.715

Analyse des taux de FVIII ajustés sur ABO,

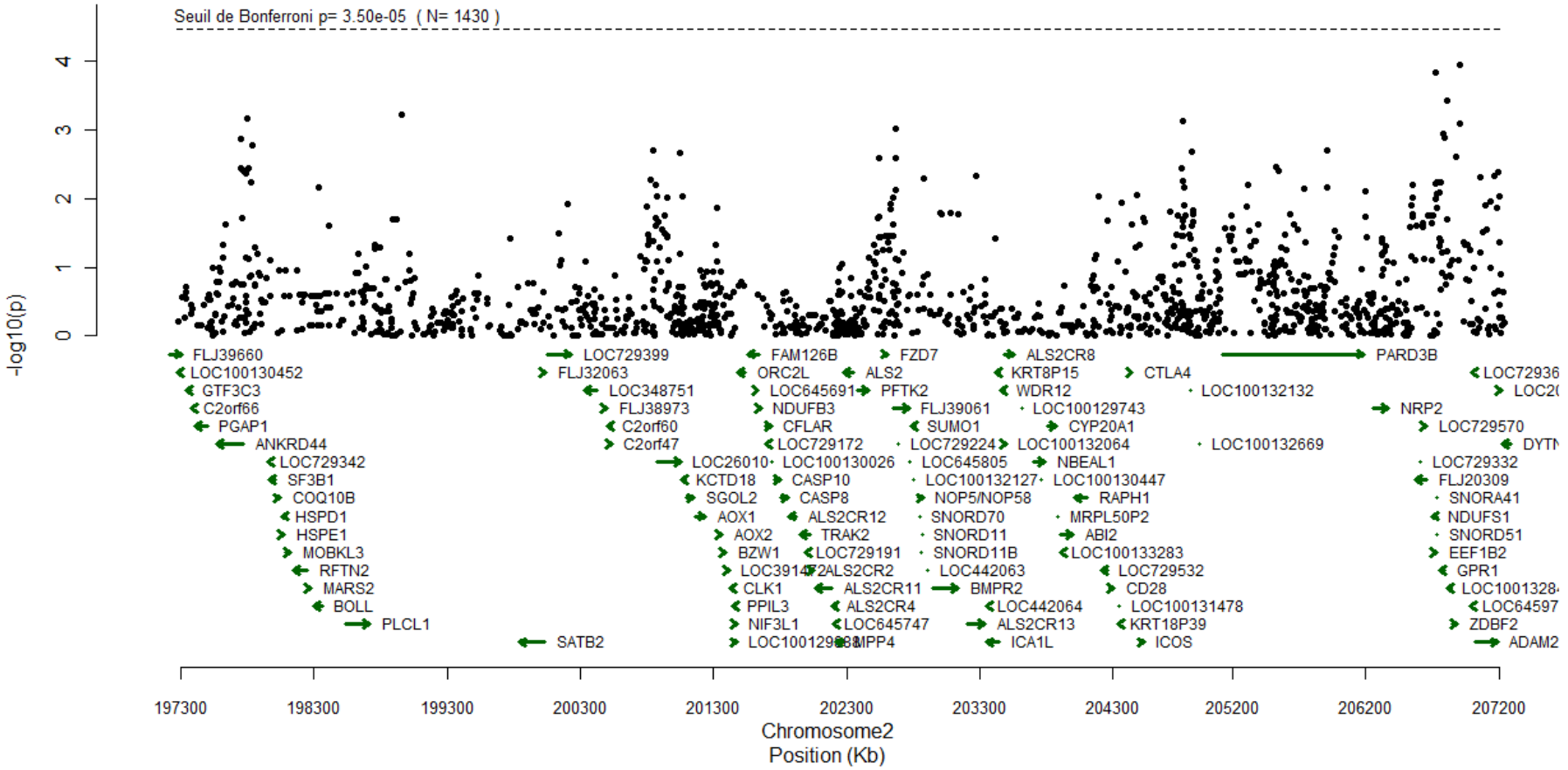
SNP	CHR	BP	gene_symbol	location	alleles	MAF	effet	P	qval	P M08	P M10
rs3827733	1	114140112	RSBN1	intron	A>G	0.2	-0.49	3.52e-06	0.14	0.79	0.94
rs4659867	1	236593267	LOC339535	3'UTR	A>C	0.12	-0.59	3.38e-06	0.138	0.61	0.64
rs1451197	2	2720453	LOC648230	5'UTR	A>G	0.46	0.37	8.12e-06	0.174	0.79	0.27
rs10174217	2	2721657	LOC648230	5'UTR	G>A	0.46	0.37	8.01e-06	0.174	0.70	0.27
rs4733677	8	128781003	NULL	5'UTR	G>A	0.17	0.54	7.52e-07	0.063	0.71	0.69
rs1570500	9	76408875	RORB	intron	A>C	0.25	0.45	6.15e-06	0.165	0.65	0.31
rs11594791	10	3549601	KLF6	3'UTR	A>G	0.04	-0.88	5.98e-06	0.164	0.82	0.57
rs7956222	12	99730699	TMEM16D	intron	A>G	0.15	0.61	1.51e-07	**0.024**	0.63	0.20
rs7137089	12	99732588	TMEM16D	intron	G>A	0.15	0.64	1.63e-08	**0.008**	0.83	0.07
rs6538973	12	99738954	TMEM16D	intron	A>G	0.26	0.42	1.85e-06	0.107	0.43	0.08
rs1554074	12	99764481	TMEM16D	intron	G>A	0.22	0.46	7.97e-06	0.174	0.40	0.39
rs2138046	12	99779713	TMEM16D	intron	A>G	0.22	0.46	7.97e-06	0.174	0.50	0.39
rs703704	12	99797499	TMEM16D	intron	A>G	0.22	0.46	7.97e-06	0.174	0.68	0.48
rs12819674	12	100053011	TMEM16D	3'UTR	G>A	0.1	0.68	6.61e-06	0.167	0.15	0.50
rs2173815	12	100352181	ARL1	5'UTR	G>A	0.13	0.56	5.97e-06	0.163	0.08	0.69
rs7963127	12	100904698	CCDC53	3'UTR	G>A	0.1	0.63	9.42e-06	0.179	0.14	0.69
rs10861272	12	103634668	CHST11	intron	A>G	0.24	0.45	9.96e-06	0.181	0.16	0.12
rs7297415	12	111145487	NULL	intron	G>A	0.14	0.63	8.65e-07	0.068	0.40	0.42
rs11615047	12	111485068	PTPN11	3'UTR	A>C	0.14	0.65	9.90e-08	0.021	0.63	0.59
rs1544656	12	111492716	PTPN11	3'UTR	A>G	0.23	0.44	9.23e-06	0.178	0.81	0.04
rs11612310	12	111510284	PTPN11	3'UTR	A>G	0.13	0.64	2.71e-07	**0.032**	0.81	0.69
rs4239153	17	66882690	hCG_1644301	5'UTR	A>C	0.47	-0.37	9.55e-06	0.179	0.34	0.35
rs7247001	19	40695802	DMKN	coding	A>G	0.11	0.63	2.22e-06	0.117	0.18	0.59
rs1783016	21	26201909	APP	intron	G>A	0.14	-0.59	3.08e-06	0.134	0.21	0.59
rs6003807	22	22264030	IGLL1	5'UTR	C>A	0.06	0.76	5.80e-06	0.162	0.94	0.12

Annexe 7. Association entre les taux de vWF ou de FVIII et les polymorphismes situés dans des signaux de liaison (échantillon Familles-FVL)

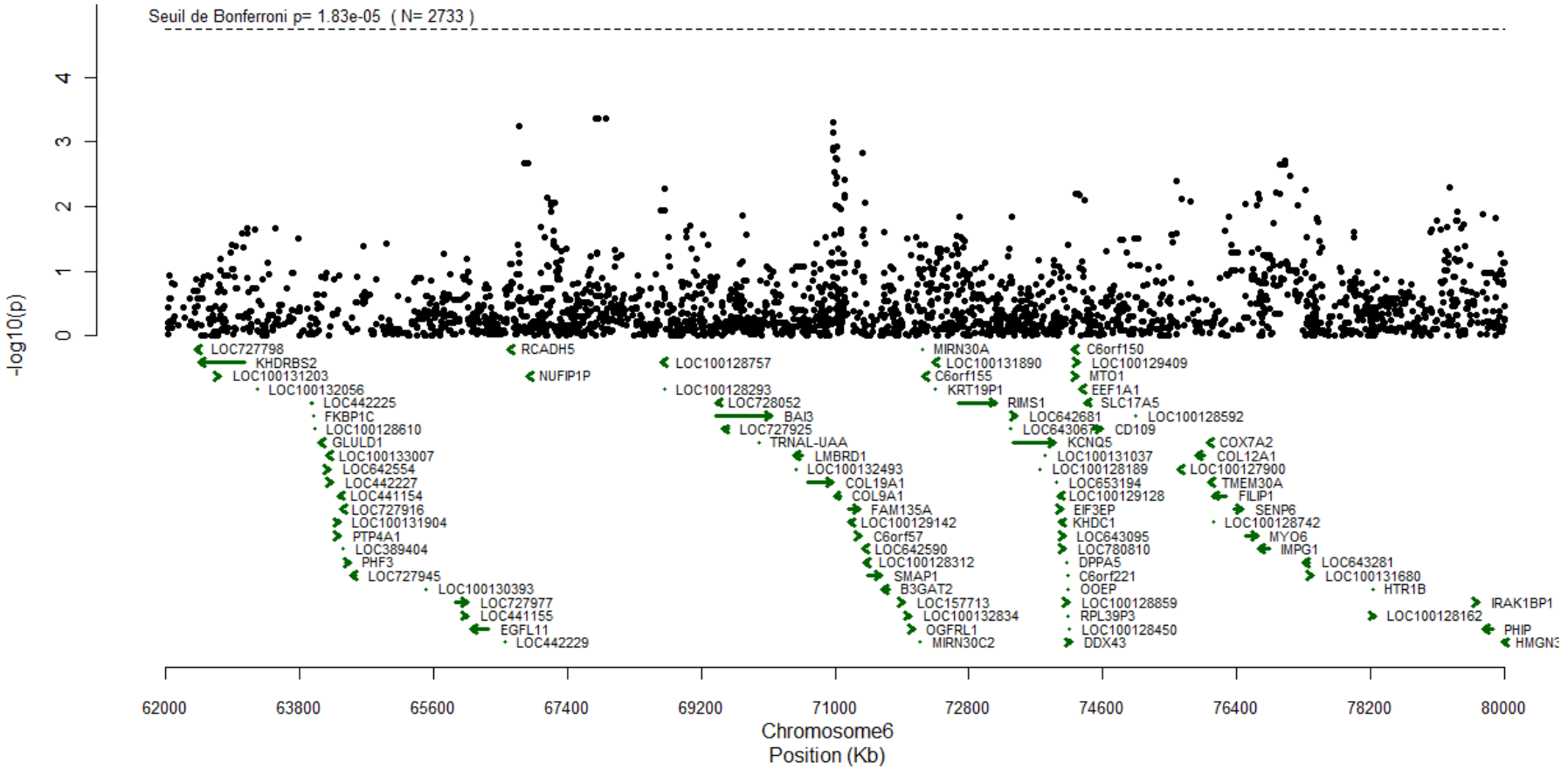
Association entre les polymorphismes du locus 2p12-13 et les taux de FVIII (échantillon Familles-FVL)



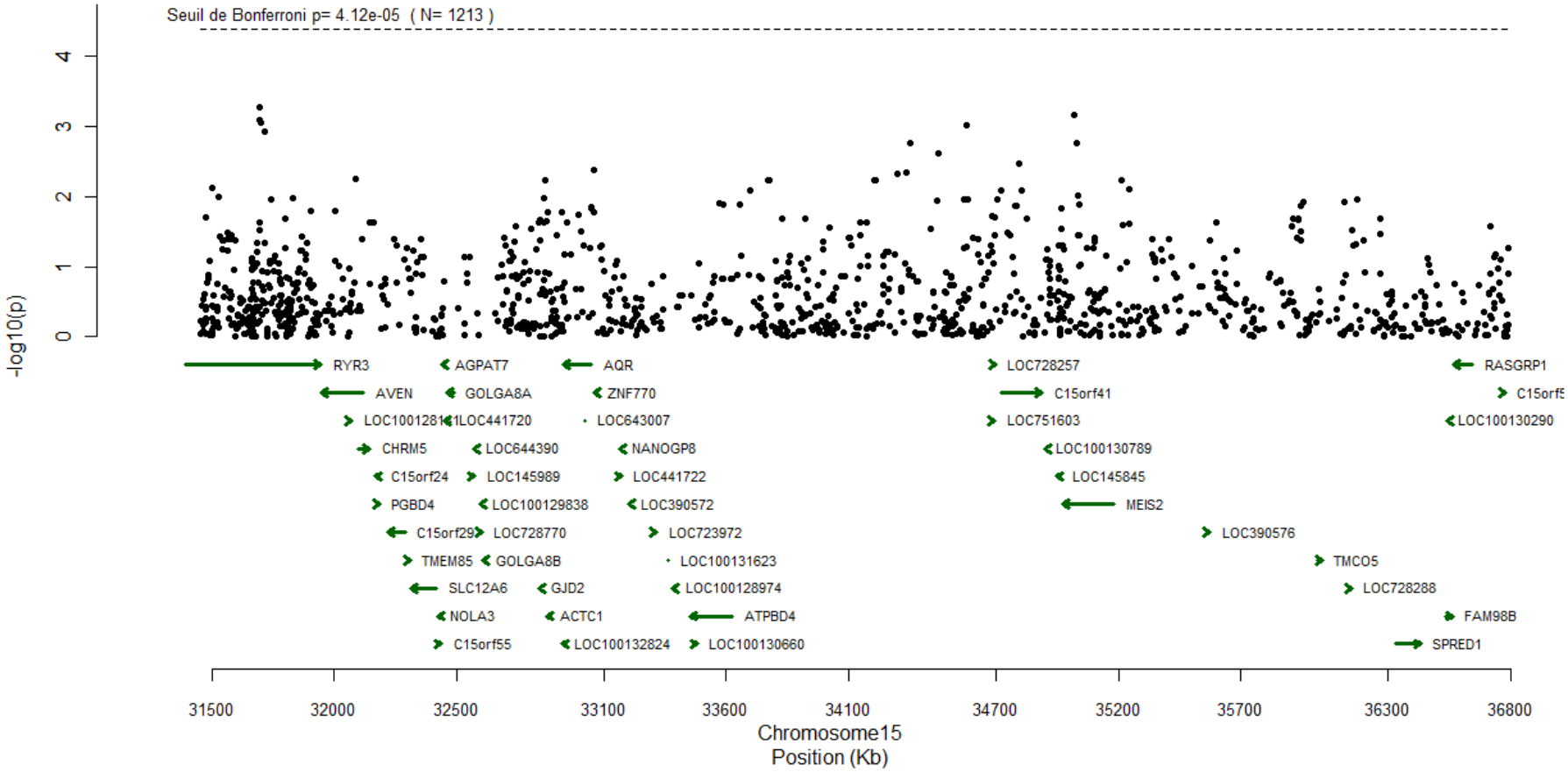
Association entre les polymorphismes du locus 2q33 et les taux de vWF (échantillon Familles-FVL)



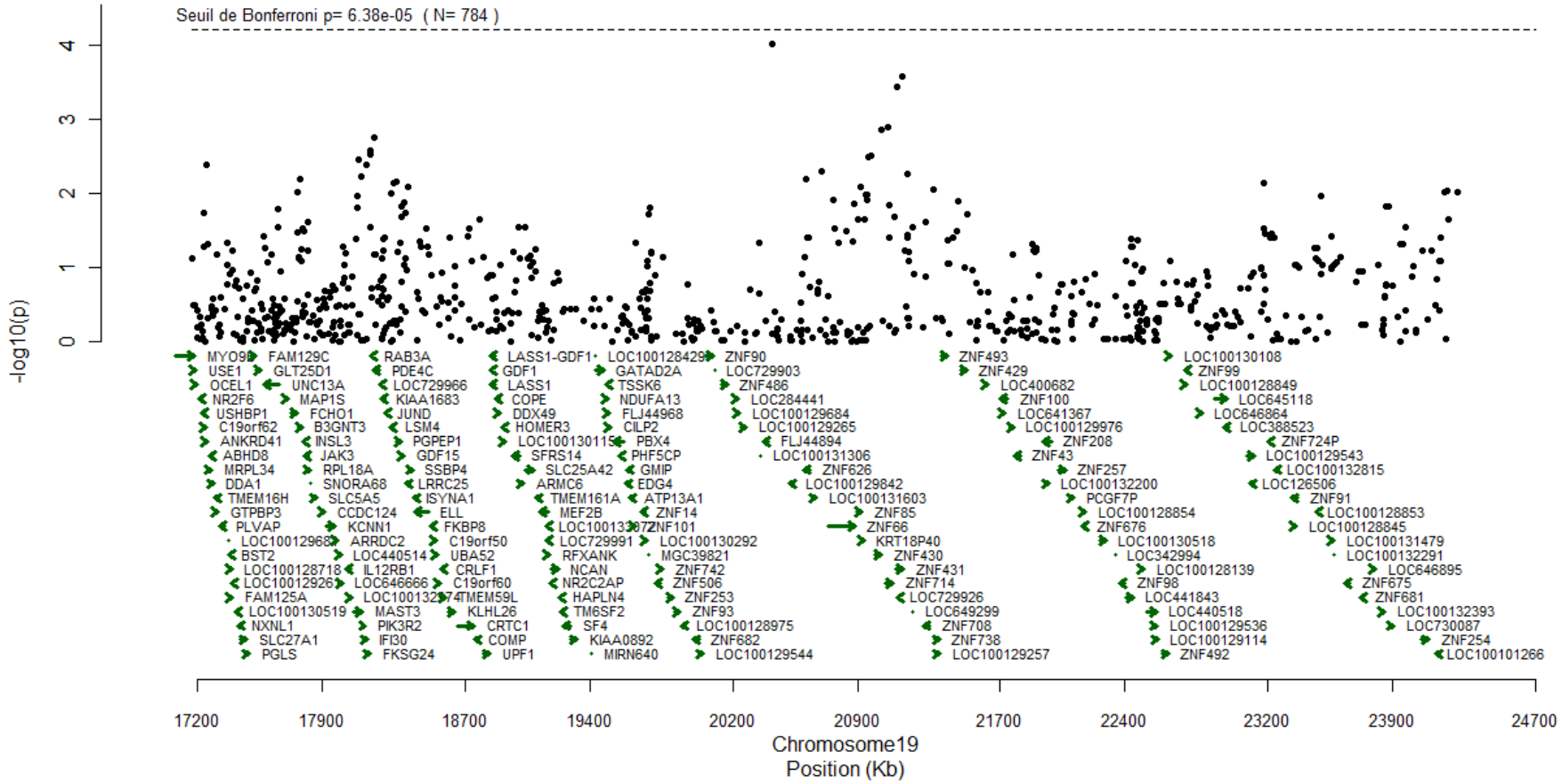
Association entre les polymorphismes du locus 6q13-14 et les taux de FVIII ajustés sur les taux de vWF (échantillon Familles-FVL)



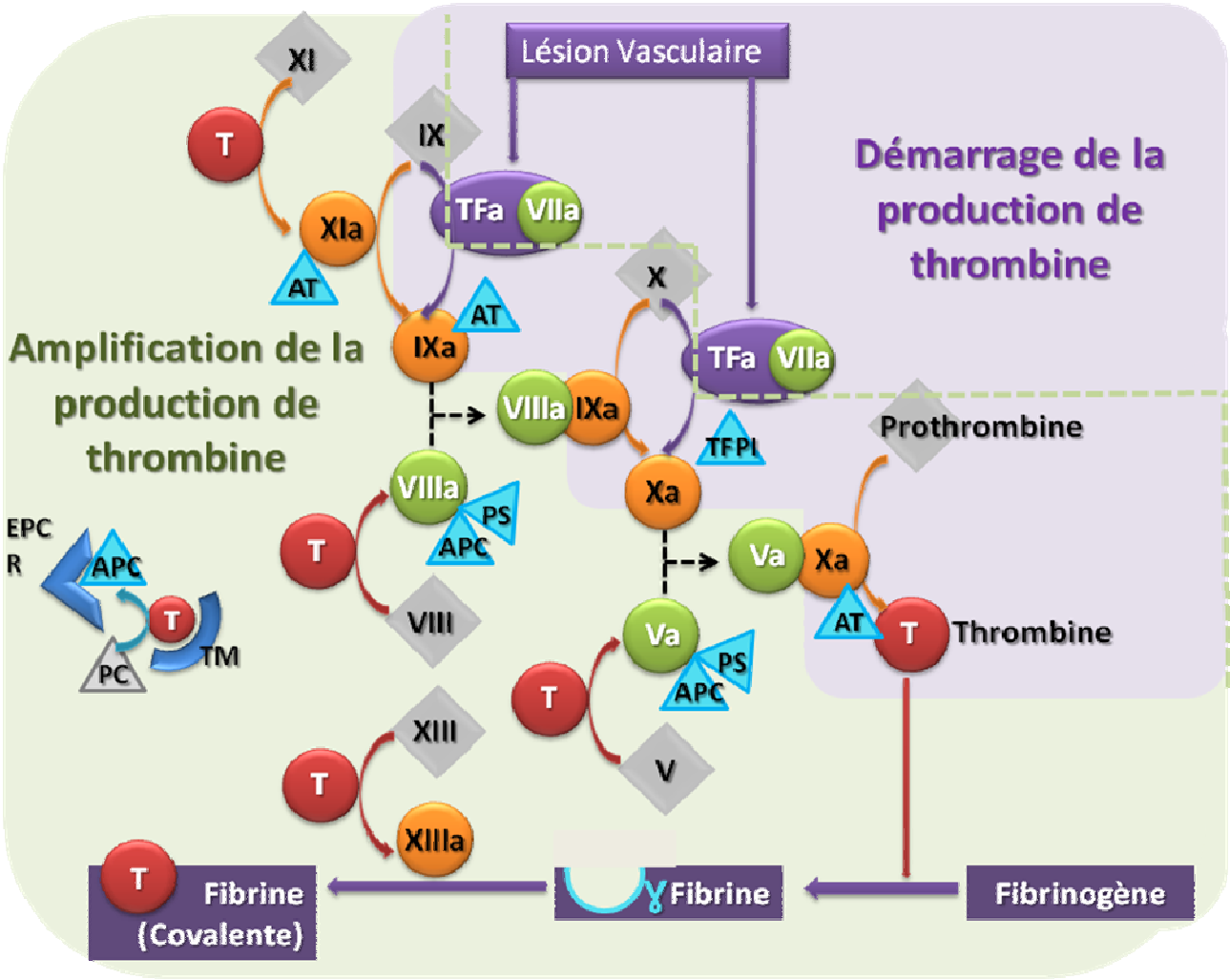
Association entre les polymorphismes du locus 15q14 et les taux de vWF ajusté sur *ABO* (échantillon Familles-FVL)



Association entre les polymorphismes du locus 19p13 et les taux de vWF ajustés sur *ABO* (échantillon Familles-FVL)

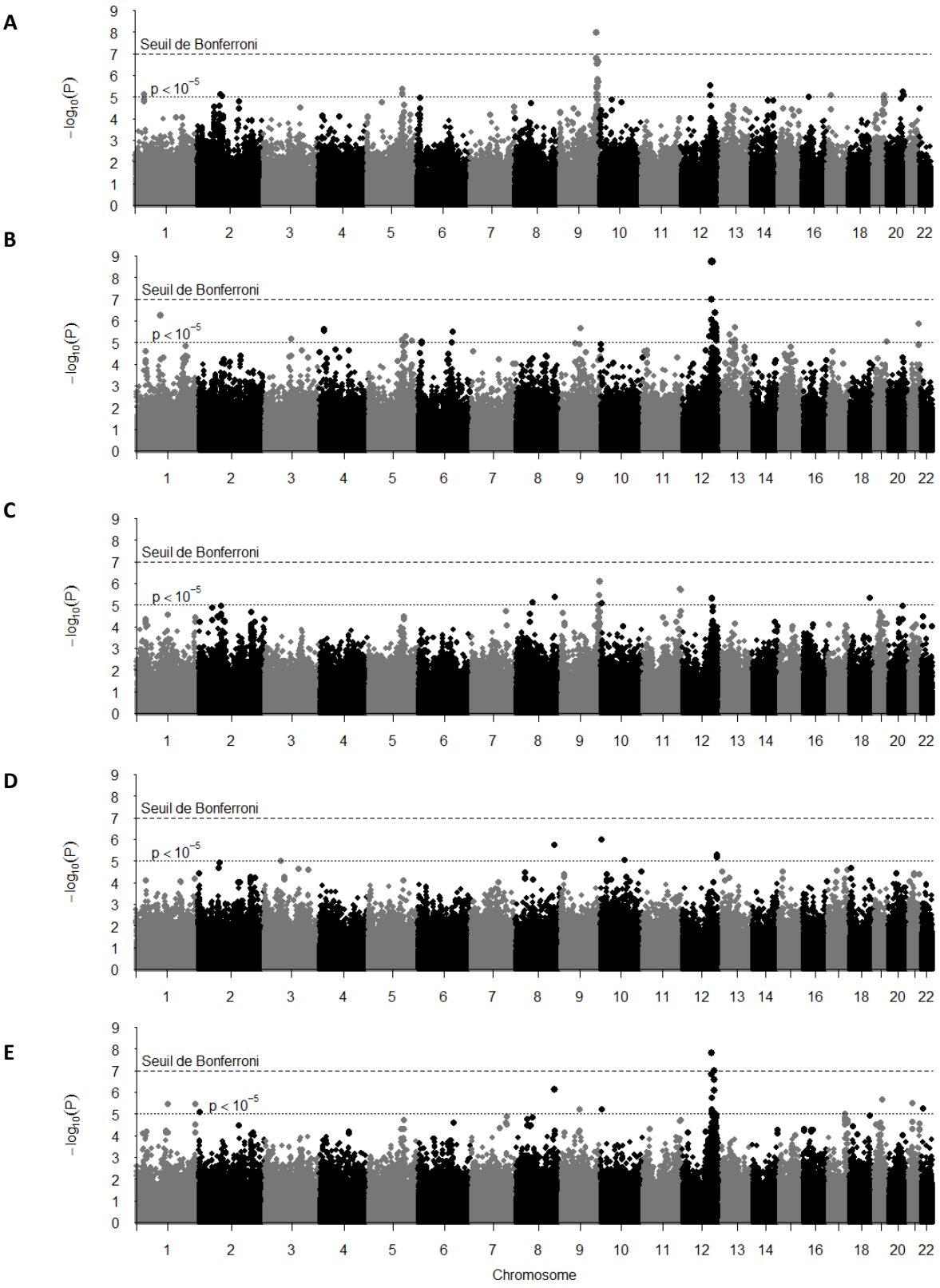


Cascade de la coagulation



La cascade de la coagulation démarre suite au contact du Facteur Tissulaire (TF) avec le sang, induit par une brèche vasculaire. Dans un premier temps (phase de démarrage ou voie extrinsèque, sur fond mauve), le complexe FVII-TF active FIX et FX. La coagulation est maintenue dans un second temps par des réactions initiées par le facteur IXa (phase d'amplification ou voie intrinsèque, sur fond vert). Les deux voies convergent vers une même voie dans laquelle la Prothrombine est convertie en Thrombine. Cette dernière modifie le Fibrinogène en Fibrine. Celle-ci est stabilisée par FXIIIa

Analyses pangénomiques réalisées dans les Familles-FVL vWF (A), vWF ajusté sur *ABO* (B), FVIII (C), FVIII ajusté sur vWF (D), FVIII ajusté sur *ABO*



A : vWF **; B :** vWF ajusté sur *ABO* **; C :** FVIII **; D : ; E :FVIII ajusté sur *ABO***

La p-value du test d'association de chaque SNP est représentée en fonction de ce dernier. Le seuil de Bonferoni est égal à 1.02 10^{-7}

Articles publiés, en rapport avec le travail de cette thèse.

Journal of Thrombosis and Haemostasis, **8**: 2671–2679 DOI: 10.1111/j.1538-7836.2010.04092.x

ORIGINAL ARTICLE

A multi-stage multi-design strategy provides strong evidence that the *BAI3* locus is associated with early-onset venous thromboembolism

G. ANTONI,*† P.-E. MORANGE,‡§ Y. LUO,† N. SAUT,‡§ G. BURGOS,‡§ S. HEATH,¶ M. GERMAIN,* C. BIRON-ANDREANI,** J.-F. SCHVED,** G. PERNOD,†† P. GALAN,‡‡§§ D. ZELENIKA,¶ M.-C. ALESSI,‡§ L. DROUET,* S. VISVIKIS-SIEST,¶¶ P. S. WELLS,*** M. LATHROP,¶ J. EMMERICH,††† D.-A. TREGOUET* and F. GAGNON†

**INSERM UMRS 937, Université Pierre et Marie Curie, Paris, France; †Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, ON, Canada; ‡INSERM, UMR_S 626; §Université de la Méditerranée, Marseille; ¶Commissariat à l'Energie Atomique, Institut de Génomique, Centre National de Génotypage, Evry; **Laboratoire d'Hématologie, CHU, Montpellier; ††Service de Médecine Vasculaire, CHU, Grenoble; ‡‡INSERM, UMR_S U872; §§Département de Santé Publique et d'Informatique Médicale, Faculté de Médecine René Descartes, Paris; ¶¶EA4373 'Génétique Cardiovasculaire', Université Henri Poincaré Nancy I, Nancy, France; ***Ottawa Health Research Institute, Civic Campus, Ottawa, ON, Canada; and †††INSERM U765, Médecine Vasculaire, HTA, Hôpital Européen Georges-Pompidou, Université Paris-Descartes, Paris, France*

To cite this article: Antoni G, Morange PE, Luo Y, Saut N, Burgos G, Heath S, Germain M, Biron-Andreani C, Schved JF, Pernod G, Galan P, Zelenika D, Alessi MC, Drouet L, Visvikis-Siest S, Wells PS, Lathrop M, Emmerich J, Tregouet DA, Gagnon F. A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism. *J Thromb Haemost* 2010; **8**: 2671–9.

Summary. *Background*: Factor VIII (FVIII) and von Willebrand factor (VWF) are two known quantitative risk factors for venous thromboembolism (VTE). *Objectives*: To identify new loci that could contribute to VTE susceptibility and to modulating FVIII and/or VWF levels. *Patients/Methods:* A pedigree linkage analysis was first performed in five extended French-Canadian families, including 253 individuals, to identify genomic regions linked to FVIII or VWF levels. Identified regions were further explored using '*in silico*' genome-wide association studies (GWAS) data on VTE (419 patients and 1228 controls), and two independent case-control studies (MARTHA and FARIVE) for VTE, gathering 1166 early-onset patients and 1408 healthy individuals. Single nucleotide polymorphisms (SNPs) associated with VTE risk were further investigated in relation to plasma levels of FVIII and VWF in a cohort of 108 healthy nuclear families. *Results:* Four main linkage regions were identified, among which the well-characterized *ABO* locus, the recently identified *STAB 2* gene, and a third one, on chromosome 6q13-14, harbouring four non-redundant SNPs, associated with VTE at $P < 10^{-4}$ in the GWAS dataset. The association of one of these SNPs, rs9363864, with VTE was further replicated in the MARTHA and FARIVE studies. The rs9363864-AA genotype was associated with a lower risk for VTE (OR = 0.58 [0.42–0.80], $P = 0.0005$) but mainly in non-carriers of the FV Leiden mutation. This genotype was further found to be associated with the lowest levels of FVIII ($P = 0.006$) and VWF ($P = 0.001$). *Conclusions:* The *BAI3* locus where the rs9363864 maps is a new candidate for VTE risk.

Keywords: factor VIII, genome-wide association studies, linkage, polymorphisms, venous thromboembolism, von Willebrand factor.

Correspondence: France Gagnon, University of Toronto, Dalla Lana School of Public Health, 155 College Street, Toronto, ON, M5T 3M7, Canada.
Tel.: +1 416 978 0130; fax: +1 416 978 8299.
E-mail: france.gagnon@utoronto.ca

Received 22 March 2010, accepted 23 September 2010

Introduction

Genome-wide association studies (GWAS) have brought great hopes for identifying new susceptibility loci in human complex diseases. To date, only one GWAS has been performed in the field of venous thromboembolism (VTE) [1]. Because of the moderate sample size, this GWAS was underpowered to detect, at a genome-wide significance level of $\sim 10^{-7}$, any single nucleotide polymorphism (SNP) whose effect would be smaller than those of the well-known *F5* and *ABO* loci. However, this GWAS detected the marginal associations ($P < 0.05$) [1] of two SNPs located within *GP6* and *CYP4V2/F11* loci that had been previously found associated with VTE in a large-scale analysis centered around non-synonymous SNPs [2]. This emphasizes the need to further explore the heap of SNP *P*-values generated from the VTE GWAS using, for example,

less stringent statistical thresholds, a strategy that led to the identification of *HIVEP1* as a new susceptibility locus for VTE [3]. An alternative strategy is to focus on SNPs or genes involved in the variability of quantitative traits known to be risk factors for VTE, as recently illustrated by a work on Protein S-related phenotypes [4].

In this work, we wish to apply a similar strategy to coagulation factor VIII (FVIII) and its binding protein von Willebrand factor (VWF). These two phenotypes are known quantitative traits presenting good predictive value of VTE occurrence [5–7] and showing evidence for a strong genetic component [8–10]. Up until last year, the only gene reported to be consistently associated with these two traits is the *ABO* blood locus where the non-O blood group is associated with higher VWF and FVIII plasma levels [11,12]. Other genes, such as the structural VWF [13] and *F8* genes [14], the lipoprotein receptor-related protein (*LRP*) gene coding the only known FVIII receptor [15,16] and the *HABP2* gene coding the FVII activating protein [17], have also been suggested to influence FVIII and VWF plasma levels but these associations have not been robustly confirmed. More robust are the associations of several SNPs within the *CLEC4M*, *SCARA5*, *STAB 2*, *STXBP5* and *TC2N* genes with plasma levels of FVIII and VWF, very recently identified by means of a GWAS approach followed by consistent replication [18]. Together with *ABO* blood group, these SNPs explained about 35% of the variability of these two traits and an in-depth investigation of the role of these loci in relation to VTE risk is ongoing (D.-A. Tregouet, personal observation).

In order to identify additional genes that could modulate the variability of FVIII and/or VWF plasma levels and contribute to VTE susceptibility, a four-stage multi-design strategy carried out in five independent samples of French origin was applied here. We first conducted a genome-wide linkage analysis of FVIII and VWF plasma levels in extended pedigrees ascertained through single probands with VTE who were carriers of the *FV* Leiden (FVL) mutation. Second, we analyzed *in silico* data from the VTE GWAS [1], restricting our investigation to linked regions identified in stage 1. Then, we conducted a validation analysis for the associations identified in stage 2 using two independent VTE case-control samples. Finally, VTE-associated SNPs that were located in FVIII and/or VWF linked regions were further analysed and tested for association with FVIII and/or VWF in a cohort of healthy nuclear families.

Methods

The schematic and sequential details of the samples analyzed, with the associated characteristics, are presented in Table 1.

Subjects

Each individual study was approved by its institutional ethics committee and informed written consent was obtained in accordance with the Declaration of Helsinki. All subjects were of European origin.

French-Canadian sample We used five extended French-Canadian pedigrees ascertained through single probands with objectively diagnosed VTE and carrying the FVL mutation. As we sought idiopathic VTE, potential probands with acquired forms of VTE (e.g. probands treated for cancer) and/or rare forms of inherited VTE (e.g. protein S deficiency) were excluded. The proband of each family was recruited through the Thrombosis Clinic of the Ottawa Hospital. Once the proband was identified, the size of the family was the sole criterion for recruitment. That is, the five largest families were recruited. Phenotypic and genotypic data were collected on a total of 253 individuals. The main sample characteristics are shown in Table 1.

'In silico' *GWAS data* For this analysis, we used the results of a previously published GWAS ('*in silico*') comparing 419 early age of onset (< 50 years) VTE cases with 1228 healthy controls [1]. The latter were French subjects selected from the SUVIMAX population [19] while the former were patients diagnosed as having VTE (including deep vein thrombosis and pulmonary embolism) before the age of 50 years. These patients were recruited from four different French medical centers (Grenoble, Marseille, Montpellier and Paris) between 1999 and 2006 and, in order to be eligible, should not have any acquired risk factors nor known genetic risk factors (antithrombin, protein C or protein S deficiencies, and homozygosity for factor V Leiden or FII G20210A).

MARTHA and FARIVE samples MARTHA is a case-control study including 1150 VTE cases and 801 controls, and FARIVE is a multicentric case-control study composed of 607 VTE cases and 607 controls, both studies being thoroughly described elsewhere [1]. Unlike the GWAS patients, no selection on the age of onset of VTE was initially made. As MARTHA was initially designed to study interactions between FVL and prothrombin (PT) G20210A mutations and other VTE risk factors, the controls were a random sample of healthy French individuals and a subsample of healthy heterozygote carriers of the FVL or PT G20210A mutations.

Stanislas A subsample of 108 families was selected from the Stanislas cohort [20]. These families were composed of both parents ($n = 216$) and at least two offspring ($n = 237$), and were volunteering for a free health examination at the Centre for Preventive Medicine in Vandoeuvre-lès-Nancy, France, between 1994 and 1995. All individuals were free from acute or chronic disease. A detailed description of these families is provided elsewhere [15].

Biological parameters

FVIII and VWF phenotypes were measured in the sample of French-Canadian pedigrees and in the Stanislas cohort. Plasma levels of FVIII activity were evaluated by a clotting assay on the BCS instrument (Siemens Diagnostics, Marburg, Germany) in the French-Canadian sample, while plasma FVIII

Table 1 Main design and sample characteristics of the four stages of the multi-stage strategy

	Stage 1 FVL French-Canadian pedigrees	Stage 2 VTE GWAS	Stage 3a MARTHA	Stage 3b FARIVE	Stage 4 Stanislas
Design characteristics					
Main goal	Initial mapping	Linkage region exploration	Validation	Validation	Validation
Epidemiological design	Extended pedigrees	GWAS	Case-control	Case-control	Cohort of healthy nuclear families
Outcome type	Quantitative	Binary	Binary	Binary	Quantitative
Main analytic strategy	Genome-wide linkage	*In silico* GWAS	Candidate SNPs	Candidate SNPs	Candidate SNPs
Main analytic tool	Bayesian MCMC-based joint linkage and segregation	Cochran-Armitage trend test	Logistic regression Cochran-Armitage trend test	Logistic regression Cochran-Armitage trend test	Generalized estimating equations
'Baseline' covariate adjustment	Age, sex	NA	Age, sex, smoking, ABO blood group	Age, sex, smoking, ABO blood group	Age, sex, smoking, Tanner stage, ABO blood group
Sample characteristics		Control/cases	Control/cases	Control/cases	Parents/offspring
Mean age (range)	40.4 (14–93)	50.2/36.2	47.7 (18–74)/32.4 (18–49)	51.5 (18–91)/32.9 (17–49)	45.8 (36–65)/17.7 (8–30)
FVIII (IU/dL)	118.6 (38.51)	NA	NA	NA	99.0 (34.29)/89.5 (29.03)
VWF (IU/dL)	130.3 (53.24)	NA	NA	NA	117.8 (43.24)/112.0 (41.79)
Sex (% female)	50.6	69.9/44.8	52.2/69.9	57.3/82.0	50.0/48.1
History of VTE (%)	5.95	NA	NA	NA	0
Smoking (%)	24.4	NA	27.4/35.4	44.5/48.8	24.1/15.4
PT G20210A carriers (%)	0.40	NA	18.2/13.7	3.2/11.0	NA
FVL carriers (%)	24.9	NA	22.5/30.2	4.6/15.0	NA
ABO blood group (%)					
O	40.6	NA	42.6/24.6	45.2/21.7	42.1/37.1
A	57.8		44.2/59.7	42.1/58.8	47.2/48.9
B	1.6		9.0/10.1	8.7/12.2	8.4/7.6
AB	–		4.2/5.6	4.2/7.2	2.3/6.3

Mean (SD) or percentages are shown when available. NA, not available.

coagulant activity (FVIII:C) was assayed in an automated coagulometer (STA-R; Diagnostica Stago, Asnières, France) in the Stanislas sample [15]. VWF antigen (VWF:Ag) was measured with a commercially available ELISA kit from Diagnostica Stago in both samples. The interassay coefficients of variation for VWF:Ag were 6.1% and 10.8% in the French-Canadian dataset and in the Stanislas sample, respectively. The interassay coefficients of variation for FVIII levels were smaller than 4% and 15% in normal and abnormal samples, respectively, in the French-Canadian dataset, and 10.5% for FVIII:C in the Stanislas sample.

Genotyping and quality control measures

French-Canadian pedigree members were genotyped for a panel of 1079 microsatellite markers with ~3.4 cM average marker density and outsourced to DeCODE (http://www.decode.com/services/microsatellite-genotyping-genome-wide-scans.php). Genotyping errors and family structure misspecifications were identified using Relpair [21] and PREST [22].

Individuals from the GWAS were genotyped with the Illumina Sentrix HumanHap300 Beadchip (Illumina Inc., San Diego, CA, USA) with quality control criteria as previously described [1]. In the MARTHA, FARIVE and Stanislas samples, genotyping was performed using TaqMan technology (Applied Biosystems; Life Technologies, Foster City, CA, USA).

Hardy-Weinberg equilibrium (HWE) was tested by a Pearson chi-squared test with one degree of freedom in controls of the case-control samples. In the Stanislas group, this was carried out in parents only. No deviation from HWE at $P < 0.05$ was observed. The lowest genotype success rates were 98%, 96% and 98% in the MARTHA, FARIVE and Stanislas samples, respectively.

Statistical methods

Genome-wide joint linkage and segregation analyses of plasma levels of FVIII activity and VWF:Ag were performed in the French-Canadian pedigrees using the Loki software, which is based on Bayesian Markov Chain Monte-Carlo methods [23]. The quantification of the linkage signals was carried out using the Bayes Factor (BF) [24], which is the ratio of the posterior to the prior odds of linkage. Based on

interpretation guidelines proposed by Kass and Raftery (1995) [24], we defined BF between 10 and 100 as 'strong evidence' for linkage, and BF > 100 as 'decisive evidence'. In addition, the evidence for linkage on the strongest signals was further assessed through permutations (n = 3000 permutated datasets) to provide an empirical estimate of P-values as per Igo *et al.* [25]. Correlation between FVIII and VWF:Ag was estimated using the generalized estimating equation (GEE) method. This method was also used to test for association between biological parameters and clinical variables such as VTE status and FVL. The GEE method is an efficient statistical technique dealing with the non-independence of family members [26]. We used the geeglm function of the geepack R package.

For the *in silico* GWAS analysis, association of genotypes with VTE risk was assessed using the Cochran-Armitage trend test [27]. In the MARTHA and FARIVE studies, allele frequencies were compared between cases and controls by use of the Fisher test. Logistic regression analyses were carried out to estimate overall genetic odds ratio (OR) adjusted for age, sex, smoking and ABO blood group, separately in the MARTHA and FARIVE samples. Subgroup analyses were performed separately in carriers and non-carriers of the FVL mutation. We tested for homogeneity across studies and across subgroups using the Mantel-Haenzel statistics [28]. As our design strategy was to follow-up the *in silico* GWAS results, which were performed on a sample with early VTE age of onset (< 50 years), we restricted our analyses in the MARTHA and FARIVE samples to the same age group. The total number of cases satisfying these criteria was 916 in MARTHA and 250 in FARIVE. As the MARTHA and FARIVE studies were used to replicate the *in silico* GWAS findings, one-sided tests were performed to specifically investigate whether genetic effects observed in the MARTHA and FARIVE samples were in the same direction as those observed in the *in silico* GWAS.

In the Stanislas cohort, which was composed of family data, we tested for association between the candidate SNPs and the quantitative traits (FVIII and VWF) using the GEE technique [29]. These association analyses were adjusted for age, sex, smoking and Tanner stage, and further on ABO blood group. Again, one-sided P-values were reported as we were specifically interested in testing whether SNPs associated with decreased VTE risk in MARTHA and FARIVE studies showed association with decreased plasma levels of FVIII and VWF, as would be expected from the known biology of these two proteins in relation to VTE.

Results

The characteristics of individual samples used for each stage are summarized in Table 1.

Stage 1: Quantitative traits genome-wide linkage scan (GWLS)

As expected, FVIII activity and VWF:Ag plasma levels were strongly correlated (ρ = 0.78, $P < 10^{-4}$) in the French-Canadian pedigrees. Both trait values were increased in subjects with a personal history of VTE ($P < 10^{-4}$) (Fig. S1) but were not different according to FVL status (Fig. S1). The pedigree linkage analyses identified four regions linked to FVIII, VWF:Ag and FVIII adjusted for VWF:Ag levels (Fig. 1). The strongest signal was observed on chromosome 9q34 between microsatellites D9S1863 and D9S1826. This signal was common to FVIII (BF = 240, $P < 10^{-3}$) and VWF:Ag (BF = 295, $P < 10^{-3}$) levels, and harbours the ABO locus well known for its role in the regulation of plasma levels of FVIII and VWF:Ag. The other three regions were located on 2p12-13 between D2S2152 and D2S2116, on 6q13-14 between D6S257 and D6S1644, and on 12q23 between D12S1727 and D12S1636. The 2p12-13 region was linked to FVIII levels

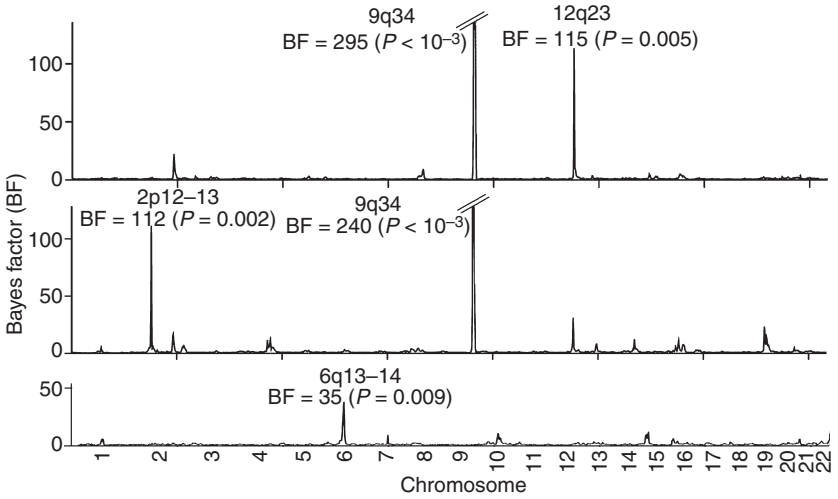


Fig. 1. Genome-wide linkage scan results in the French-Canadian pedigrees. Genome-wide linkage scan on VWF (top), FVIII (middle) and FVIII adjusted for VWF levels (bottom).

(BF = 112, $P = 0.002$), the 6q13-14 to FVIII adjusted for VWF:Ag (BF = 35, $P = 0.009$) and the 12q23 region to VWF:Ag levels (BF = 115, $P = 0.005$). Further adjustment on anticoagulant treatment, FVL and ABO blood groups did not modify the observed linkage signals, except on chromosome 9q34 where signals were strongly reduced after adjusting for ABO blood group.

Stage 2: VTE GWAS

Based on the GWLS results, the 2p12-13, 6q13-14 and 12q23 regions were further investigated by *in silico* association analysis of a previously published VTE GWAS [1]. Our aim was to refine the loci, by narrowing down the regions. For this, we decided to check for SNPs showing evidence of association with VTE at $P < 10^{-4}$ in this GWAS. The threshold of $P < 10^{-4}$ was used in order to limit the number of false positives while maintaining a good sensitivity. This value corresponds to a false discovery rate (FDR) of ~20% at the genome-wide scale and a FDR of 7% based on the number of SNPs (i.e. 3468) within the three investigated linkage regions. There was no evidence of association of any SNP within the 2p12-p13 region with VTE at $P < 10^{-4}$ (Fig. S2), and only one SNP within the 12q23 region (rs1593812) was significantly associated with VTE at this threshold (Fig. S3). This SNP is located within the 5′ region of the *STAB 2* gene recently found to be associated with FVIII and VWF plasma levels in a GWAS [18]. The most interesting and novel results are those observed in the 6q13-q14 region, where four SNPs were found significantly associated with VTE at $P < 10^{-4}$ (Fig. S4). One SNP, rs9363864, was located ~500 kb downstream from the *BAI3* gene, whereas three other SNPs, rs1889879, rs3798992 and rs10485430, were located in the *BAI3* gene. Other SNPs in *BAI3* also showed suggesion of association with VTE at $P < 10^{-2}$ (Table S1). Linkage disequilibrium (LD) (Table S2) and in-depth haplotype analysis searching for the most informative and parsimonious haplotype configuration derived from these four SNPs in terms of association with disease risk [30] led to the conclusion that three SNPs (rs9363864, rs1889879 and rs10485430) characterized the *in silico* association signal. These three SNPs were then genotyped in the MARTHA and FARIVE samples to replicate their association with VTE risk and to confirm that the minor rs9363864-A and rs1889879-C alleles are associated with decreased risk and the minor rs10485430-A allele with increased risk.

Stage 3: VTE case-control analyses

Results of the association analyses are reported in Table 2. While the rs1889879 did not show any trend for association in any of the two studies, two marginal associations consistent with the pattern observed in the GWAS dataset were detected for the other SNPs. In the FARIVE study the rs9363864-A allele was found less frequent in cases than in controls (0.43 vs. 0.48, $P = 0.02$) and in the MARTHA study the rs10485430-A

Table 2 Association of rs9363864, rs1889879 and rs10485430 with early age of onset of VTE in MARTHA and FARIVE

	MARTHA					FARIVE				
rs9363864	GG	GA	AA	G/A		GG	GA	AA	G/A	
Controls	231 (29.1%)	389 (48.9%)	175 (22.0%)	0.54/0.46	$P^* = 0.81$	168 (28.5%)	274 (46.5%)	147 (25.0%)	0.52/0.48	$P = 0.02$
Cases	235 (26.5%)	453 (51.1%)	199 (22.4%)	0.52/0.48		82 (33.6%)	115 (47.1%)	47 (19.3%)	0.57/0.43	
rs1889879	AA	AC	CC	A/C		AA	AC	CC	A/C	
Controls	327 (41.0%)	373 (46.8%)	97 (12.2%)	0.64/0.36	$P = 0.61$	223 (38.0%)	289 (49.2%)	75 (12.8%)	0.63/0.37	$P = 0.28$
Cases	363 (40.6%)	419 (46.8%)	113 (12.6%	0.64/0.36		96 (39.8%)	117 (48.5%)	28 (11.6%)	0.64/0.36	
rs10485430	GG	GA	AA	G/A		GG	GA	AA	G/A	
Controls	560 (70.3%)	222 (27.9%)	14 (1.8%)	0.84/0.16	$P = 0.02$	396 (67.1%)	183 (31.0%)	11 (1.9%)	0.83/0.17	$P = 0.34$
Cases	603 (67.9%)	244 (27.5%)	41 (4.6%)	0.82/0.18		163 (66.0%)	78 (31.6%)	6 (2.4%)	0.82/0.18	

Genotype distributions (with corresponding percentage) are shown in addition to allele frequencies. *One-sided Fisher test *P*-value comparing allele frequencies between cases and controls.

allele was more frequent in cases than in controls (0.18 vs. 0.16, $P = 0.02$). By design, the MARTHA study is enriched with carriers of FVL and PT G20210A mutations compared with the FARIVE study. In the MARTHA study, the percentages of FVL and/or PT G20210A carriers were 44% and 41% in cases and controls, respectively, while these proportions were 24% and 8% in the FARIVE study, as well as in the GWAS sample. To investigate whether the discrepancy in the SNP associations observed in the MARTHA and FARIVE samples was due to a different distribution of FVL and/or PT G20210A mutations in the two samples, we re-analyzed the data by carrier status. Because the Akaike criterion identified the recessive model as the most compatible model for the rs9363864-A allele in the FARIVE sample, Table 3 provides the distribution of the rs9363864-AA genotype in cases and controls, according to the non-carrier/carrier status for FVL and/or PT G20210A mutations. Interestingly, in non-carriers of FVL and/or PT G20210A mutations, homozygotes for the rs9363864-A allele had a lower risk for VTE both in the MARTHA (OR = 0.63 [0.42–0.95], $P = 0.013$) and FARIVE (OR = 0.52 [0.31–0.86], $P = 0.006$) (Table 3) samples. These two ORs were not significantly different ($P = 0.54$) across samples, and the pooled OR estimate was then 0.58 (0.42–0.80) ($P = 0.0005$). This association remained significant at a statistical level of $8\ 10^{-3}$ ($\sim$0.05/6), which corresponds to the Bonferroni threshold correcting for the number of tested SNPs (i.e. 3) and subgroup analysis (i.e. 2). In contrast, the AA genotype did not show any evidence of association with VTE risk (OR = 1.11 [0.68–1.80], $P = 0.67$) in carriers of FVL and/or PT G20210A mutations. The test for homogeneity of the effect of rs9363864 according to carrier status of FVL and/or PT G20210A mutations was significant ($P = 0.03$).

Because of the low allele frequency and small sample size, it was unfortunately not possible to investigate such putative heterogeneity for rs10485430.

Stage 4: Quantitative traits cohort analyses

Finally, based on the results observed in stage 3, both rs9363864 and rs10485430 were further investigated, but now in relation to plasma levels of FVIII:C and VWF:Ag in a cohort of healthy families. According to stage 2 and stage 3 results, we were particularly interested in testing whether the rs9363864-AA genotype was associated with decreased plasma levels and the rs10485430-A allele with increased levels.

While no association was observed between rs10485430 and either phenotypes, the rs9363864-AA genotype was associated with decreased levels of FVIII:C ($P = 0.006$) and VWF:Ag ($P = 0.0013$), which is consistent with the results we observed for VTE risk. These effects were homogeneously observed in parents and offspring (Table 4). The percentage of variability explained by this SNP, after adjusting for age, sex, smoking, Tanner stage and ABO blood group, was relatively weak, with 1.22% and 1.60% for FVIII:C and VWF:Ag, respectively. The rs9363864-AA genotype was no longer associated with FVIII:C after adjusting for VWF:Ag ($P = 0.25$), but its association with VWF:Ag remained marginally significant after adjusting for FVIII:C ($P = 0.036$).

Table 3 Association between rs9363864 and early age of onset of VTE separately in carriers and non-carriers of FVL/PT G20210A mutations

	Non-carriers			Carriers		
	GG/GA	AA	OR*	GG/GA	AA	OR*
MARTHA						
Controls	361 (76.3%)	112 (23.7%)	0.63 (0.42–0.95)	259 (80.4%)	63 (19.6%)	1.03 (0.62–1.73)
Cases	394 (78.5%)	108 (21.5%)	$P = 0.013$	294 (76.4%)	91 (23.6%)	$P = 0.55$
FARIVE						
Controls	401 (74.3%)	139 (25.7%)	0.52 (0.31–0.86)	38 (84.4%)	7 (15.6%)	1.90 (0.47–7.68)
Cases	151 (82.5%)	32 (17.5%)	$P = 0.006$	44 (75.9%)	14 (24.1%)	$P = 0.82$

*Odds ratio adjusted for age, sex, ABO blood group and smoking with corresponding one-tailed *P*-values.

Table 4 Association of rs9363864 and rs10485430 with FVIII:C and VWF:Ag levels in the Stanislas cohort

	Parents				Offspring			
rs9363864	GG	GA	AA		GG	GA	AA	
	$N = 51$	$N = 104$	$N = 55$		$N = 56$	$N = 121$	$N = 54$	
FVIII:C	99.9 (4.9)	102.3 (3.5)	91.8 (4.0)	$P^* = 0.11$	96.8 (4.2)	89.7 (2.5)	82.4 (4.0)	$P = 0.05$
VWF:Ag	118.0 (6.9)	122.7 (4.3)	108.4 (4.9)	$P = 0.04$	115.3 (5.9)	115.9 (4.0)	99.2 (4.4)	$P = 0.01$
rs10485430	GG	GA	AA		GG	GA	AA	
	$N = 155$	$N = 53$	$N = 5$		$N = 173$	$N = 56$	$N = 5$	
FVIII:C	98.3 (2.9)	102.9 (4.0)	86.4 (6.9)	$P = 0.89$	88.7 (2.2)	92.9 (4.1)	85.2 (17.4)	$P = 0.38$
VWF:Ag	116.5 (3.7)	123.4 (4.8)	104.0 (16.4)	$P = 0.89$	110.8 (3.3)	114.2 (4.9)	105.5 (18.3)	$P = 0.60$

Mean (SD). *One-sided association test *P*-value adjusted for age, sex, ABO blood group, smoking and Tanner stage in offspring. We specifically tested whether the rs9363864-AA genotype was associated with decreased plasma levels and the rs10485430-AA and AG genotypes with increased levels.

Discussion

Using a multi-stage multi-design strategy relying on genome-wide and individual SNP approaches, we were able to identify one SNP, rs9363864, strongly associated with early onset VTE. In the combined dataset of two independent case-control studies, the rs9363864-AA genotype was indeed associated with decreased risk of VTE in individuals not carrying FVL and PT G20210A mutations (0.58 [0.42–0.80], $P = 0.0005$). The rs936384 lies in the proximity of the *BAI3* promoter region and was the SNP showing the highest significance at this locus in our *in silico* GWAS study. Other SNPs within *BAI3* and in weak LD with rs9363864 showed a strong association with VTE in the GWAS but their association was not replicated in the two case-control samples. These observations suggest that the studied SNPs are more likely in LD with the functional variant(s) not genotyped with the Illumina chip used in the GWAS. *BAI3* is a large gene spanning 752 kb and, according to public databases, harbours more than 1000 SNPs with reported allele frequencies. A thorough examination of the *BAI3* locus in different samples, and if possible with different LD patterns, could help narrow down the list of potential functional SNPs.

BAI3 was investigated in this work because it was located under a linkage peak for FVIII activity adjusted for VWF, and because it showed promising association ($P < 10^{-4}$) with VTE in a previously published GWAS [1]. When studied in two additional independent VTE case-control samples, the association observed with the lead SNP was confirmed. Interestingly, the genotype associated with lower risk of VTE was also associated with lower plasma levels of FVIII and VWF in an independent cohort of healthy individuals. This observation is quite consistent with what is known about the underlying biology [31] of FVIII and VWF, as individuals with the highest levels of these proteins are at higher risk of VTE [5–7]. Even though the effect on VTE risk of the rs9363864 is of similar amplitude to that of the O and A2 blood groups [1], it only explains about 2% of the variability of the FVIII and VWF compared with the 25% explained by the ABO blood group [15,18]. This suggests that the mechanisms implicating the *BAI3* locus to VTE would not be mediated by FVIII or VWF, unlike the hypothesis underlying the research strategy adopted in our work. One could rather speculate that this locus could exert pleiotropic effects on other risk factors associated with VTE risk. The effect of rs9363864 on VTE risk being restricted to patients not carrying FVL and PT G20210A mutations, and for which we do not yet have any explanation, could support this pleiotropy hypothesis. Finally, rs9363864 was genotyped in the French-Canadian pedigrees to investigate whether it contributes to the identified linkage signal. Adjusting for this SNP did not modify the linkage intensity (data not shown), a phenomenon observed in linkage/association studies in the presence of pleiotropic effects and/or multiple rare variants.

BAI3, for brain-specific angiogenesis inhibitor 3, belongs to a family of transmembrane receptors and shares strong similarities with BAI1 and BAI2 [32,33]. Little is known about the role of this protein in human biology. Shown to be expressed in the brain, it has been suspected to play a role in ischemia-induced brain angiogenesis [33] and, very recently, to be associated with schizophrenia-related clinical phenotypes [34]. However, we do not yet have a hypothesis on the specific mechanisms implicating BAI3 in VTE pathophysiology. Further investigations are needed to identify in which other cell types is *BAI3* expressed, which could provide clues to the underlying biology of this relationship. Nevertheless, an example already exists of a gene expressed in brain tissue that is related to VWF. This gene is *ADAMTS13* [35–37], the VWF-cleavage protease associated with brain ischemic injury and cardiovascular outcomes, especially in young individuals. It would be interesting to test the hypothesis of a putative link between *BAI3* and *ADAMTS13*.

In addition to the well-characterized *ABO* locus and the 6q13-14 region harbouring *BAI3*, our linkage scans identified two other putative regions linked to FVIII and/or VWF, on 2p12-13 and 12q23. Because the 2p12-13 region did not show any SNP associated with VTE at $P < 10^{-4}$ in the *in silico* GWAS, we did not investigate this locus further. However, we cannot exclude that this region harbours susceptibility loci to FVIII/VWF and VTE risk, as our use of a stringent statistical threshold of 10^{-4} could have prevented its detection. Conversely, one SNP, rs1593812, lying in the 12q23 region showed promising association with VTE at $P < 10^{-4}$. This SNP lies in the *STAB2* gene, which was recently identified by the CHARGE Consortium as a new locus influencing FVIII and VWF plasma levels using a GWAS approach in unrelated individuals [18]. Because this region was already known, we chose to focus on the new 6q13-14/*BAI3* locus instead. The *BAI3* locus was not identified in the CHARGE GWAS. The weaker signal we observed at the *BAI3* locus compared with that of the *STAB2* locus (BF = 35 vs. BF = 115) could explain why the former was not detected in this recently published GWAS. Conversely, the CHARGE GWAS identified four other loci (*CLEC4M*, *SCARA5*, *STXBP5* and *TC2N*) influencing the two traits [18] and it is not completely surprising that these loci were not detected by our linkage scans. It is well known that the information provided by linkage analyses is complementary to but different from that provided by association analyses, and that linkage analyses may not be optimal for the detection of loci with very modest effects such as those generally observed in GWAS of unrelated individuals.

The ascertainment strategy for the French-Canadian pedigrees used in our GWLS aimed at limiting genetic heterogeneity. With this aim in mind, probands were not only identified by their disease status but also by their FVL carrier status. Therefore, these pedigrees are not representative of the general population, which could also contribute to differences observed between our work and that of the CHARGE consortium.

Despite the unsolved questions raised in this work, we provide strong statistical arguments for *BAI3* being implicated in the occurrence of VTE through a mechanism that could contribute to modulating plasma levels of VWF and FVIII activity. We also show that the perspective provided by

comparing results from samples of complementary designs, along with the use of quantitative intermediate phenotypes, is an appealing strategy not only to increase the power of detecting genetic risk factors, but also to help disentangle the pathophysiology of a complex disease.

Addendum

G. Antoni carried out the main statistical analyses and wrote the manuscript. M. Germain, S. Heath and Y. Luo participated in statistical analyses and data storing. Genetic laboratory works were performed by N. Saut and G. Burgos. F. Gagnon designed and led the French-Canadian FVL pedigree study; P. S. Wells identified the probands and their relatives for the French-Canadian FVL pedigree study. C. Biron-Andreani, D. Zelenika, G. Pernod, J.-F. Schved, L. Drouet and P. Galan collected the GWAS data. Case-control studies for VTE were coordinated by J. Emmerich, M.-C. Alessi, M. Lathrop and P.-E. Morange while the Stanislas cohort was under the coordination of S. Visvikis-Siest. D.-A. Tregouet, F. Gagnon and P.-E. Morange designed this research study and drafted the manuscript.

Acknowledgements

The French-Canadian FVL pedigree study was supported by grants from the Canadian Institutes of Health Research (MOP86466) and by the Heart and Stroke Foundation of Canada (T6484). The FARIVE study was supported by grants from the Fondation pour la Recherche Médicale, the Program Hospitalier de recherche Clinique, the Fondation de France, and the Leducq Foundation. The MARTHA study was supported by a grant from the Program Hospitalier de la Recherche Clinique. G. Antoni holds an 'INSERM Poste d'accueil' position and M. Germain is supported by a grant funded by the Agence Nationale pour la Recherche (Project ANR-07-MRAR-021). F. Gagnon and P. S. Wells hold Canada Research Chairs. A France-Canada Research Fund 2008 provided opportunities for face-to-face meetings of principal investigators. The authors are grateful to the individuals who contributed to the data collection in the field for each of these studies and to all study participants.

Disclosure of Conflict of Interests

The authors state that they have no conflict of interest.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Characteristics of the French-Canadian pedigrees.
Fig. S2. Allelic association (−log P) between SNPs and VTE within the 69 568 693–76 502 912 region on chromosome 2 in the GWAS (each circle corresponds to one SNP).
Fig. S3. Allelic association (−log P) between SNPs and VTE within the 98 052 454–106 162 994 region on chromosome 12 in the GWAS (each circle corresponds to one SNP).
Fig. S4. Allelic association (−log P) between SNPs and VTE within the 56 026 396–80 407 960 region on chromosome 6 in the GWAS (each circle corresponds to one SNP).
Table S1. Minor allele frequencies of the *BAI3* SNPs in the VTE GWAS.
Table S2. Linkage disequilibrium between the four *BAI3* SNPs identified from the *in silico* VTE GWAS.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

1 Tregouet DA, Heath S, Saut N, Biron-Andreani C, Schved JF, Pernod G, Galan P, Drouet L, Zelenika D, Juhan-Vague I, Alessi MC, Tiret L, Lathrop M, Emmerich J, Morange PE. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood* 2009; **113**: 5298–303.
2 Bezemer ID, Bare LA, Doggen CJ, Arellano AR, Tong C, Rowland CM, Catanese J, Young BA, Reitsma PH, Devlin JJ, Rosendaal FR. Gene variants associated with deep vein thrombosis. *JAMA* 2008; **299**: 1306–14.
3 Morange PE, Bezemer I, Saut N, Bare L, Burgos G, Brocheton J, Durand H, Biron-Andreani C, Schved JF, Pernod G, Galan P, Drouet L, Zelenika D, Germain M, Nicaud V, Heath S, Ninio E, Delluc A, Munzel T, Zeller T, *et al.* A follow-up study of a genome-wide association scan identifies a susceptibility locus for venous thrombosis on chromosome 6p24.1. *Am J Hum Genet* 2010; **86**: 592–5.
4 Buil A, Tregouet DA, Souto JC, Saut N, Germain M, Rotival M, Tiret L, Cambien F, Lathrop M, Zeller T, Alessi MC, Rodriguez de Cordoba S, Munzel T, Wild P, Fontcuberta J, Gagnon F, Emmerich J, Almasy L, Blankenberg S, Soria JM, *et al.* C4BPB/C4BPA is a new susceptibility locus for venous thrombosis with unknown protein S-independent mechanism: results from genome-wide association and gene expression analyses followed by case-control studies. *Blood* 2010; **115**: 4644–50.
5 Koster T, Blann AD, Briet E, Vandenbroucke JP, Rosendaal FR. Role of clotting factor VIII in effect of von Willebrand factor on occurrence of deep-vein thrombosis. *Lancet* 1995; **345**: 152–5.
6 Kraaijenhagen RA, in't Anker PS, Koopman MM, Reitsma PH, Prins MH, van den Ende A, Buller HR. High plasma concentration of factor VIIIc is a major risk factor for venous thromboembolism. *Thromb Haemost* 2000; **83**: 5–9.
7 Tsai AW, Cushman M, Rosamond WD, Heckbert SR, Tracy RP, Aleksic N, Folsom AR. Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE). *Am J Med* 2002; **113**: 636–42.
8 Souto JC, Almasy L, Borrell M, Gari M, Martinez E, Mateo J, Stone WH, Blangero J, Fontcuberta J. Genetic determinants of hemostasis phenotypes in Spanish families. *Circulation* 2000; **101**: 1546–51.
9 de Lange M, Snieder H, Ariens RA, Spector TD, Grant PJ. The genetics of haemostasis: a twin study. *Lancet* 2001; **357**: 101–5.
10 Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM, Coll I, Felices R, Stone W, Fontcuberta J, Blangero J. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. *Am J Hum Genet* 2000; **67**: 1452–9.

11 Souto JC, Almasy L, Muniz-Diaz E, Soria JM, Borrell M, Bayen L, Mateo J, Madoz P, Stone W, Blangero J, Fontcuberta J. Functional effects of the ABO locus polymorphism on plasma levels of von Willebrand factor, factor VIII, and activated partial thromboplastin time. *Arterioscler Thromb Vasc Biol* 2000; **20**: 2024–8.

12 Orstavik KH, Magnus P, Reisner H, Berg K, Graham JB, Nance W. Factor VIII and factor IX in a twin population. Evidence for a major effect of ABO locus on factor VIII level. *Am J Hum Genet* 1985; **37**: 89–101.

13 Keightley AM, Lam YM, Brady JN, Cameron CL, Lillicrap D. Variation at the von Willebrand factor (vWF) gene locus is associated with plasma vWF:Ag levels: identification of three novel single nucleotide polymorphisms in the vWF gene promoter. *Blood* 1999; **93**: 4277–83.

14 Viel KR, Machiah DK, Warren DM, Khachidze M, Buil A, Fernstrom K, Souto JC, Peralta JM, Smith T, Blangero J, Porter S, Warren ST, Fontcuberta J, Soria JM, Flanders WD, Almasy L, Howard TE. A sequence variation scan of the coagulation factor VIII (FVIII) structural gene and associations with plasma FVIII activity levels. *Blood* 2007; **109**: 3713–24.

15 Morange PE, Tregouet DA, Frere C, Saut N, Pellegrina L, Alessi MC, Visvikis S, Tiret L, Juhan-Vague I. Biological and genetic factors influencing plasma factor VIII levels in a healthy family population: results from the Stanislas cohort. *Br J Haematol* 2005; **128**: 91–9.

16 Vormittag R, Bencur P, Ay C, Tengler T, Vukovich T, Quehenberger P, Mannhalter C, Pabinger I. Low-density lipoprotein receptor-related protein 1 polymorphism 663 C > T affects clotting factor VIII activity and increases the risk of venous thromboembolism. *J Thromb Haemost* 2007; **5**: 497–502.

17 Reiner AP, Lange LA, Smith NL, Zakai NA, Cushman M, Folsom AR. Common hemostasis and inflammation gene variants and venous thrombosis in older adults from the Cardiovascular Health Study. *J Thromb Haemost* 2009; **7**: 1499–505.

18 Smith NL, Chen MH, Dehghan A, Strachan DP, Basu S, Soranzo N, Hayward C, Rudan I, Sabater-Lleal M, Bis JC, de Maat MP, Rumley A, Kong X, Yang Q, Williams FMK, Vitart V, Campbell H, Malarstig A, Wiggins KL, Van Duijn CM *et al*. Novel associations of multiple genetic loci with plasma levels of Factor VII, Factor VIII and von Willebrand Factor. The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation* 2010; **121**: 1382–92.

19 Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D, Roussel AM, Favier A, Briancon S. The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Arch Intern Med* 2004; **164**: 2335–42.

20 Visvikis-Siest S, Siest G. The STANISLAS Cohort: a 10-year follow-up of supposed healthy families. Gene-environment interactions, reference values and evaluation of biomarkers in prevention of cardiovascular diseases. *Clin Chem Lab Med* 2008; **46**: 733–47.

21 Epstein MP, Duren WL, Boehnke M. Improved inference of relationship for pairs of individuals. *Am J Hum Genet* 2000; **67**: 1219–31.

22 Sun L, Wilder K, McPeek MS. Enhanced pedigree error detection. *Hum Hered* 2002; **54**: 99–110.

23 Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 1997; **61**: 748–60.

24 Kass RE, Raftery AE. Bayes Factors. *J Am Stat Assoc* 1995; **90**: 773–95.

25 Igo RP Jr, Wijsman EM. Empirical significance values for linkage analysis: trait simulation using posterior model distributions from MCMC oligogenic segregation analysis. *Genet Epidemiol* 2008; **32**: 119–31.

26 Tregouet DA, Tiret L. Applications of the estimating equations theory to genetic epidemiology: a review. *Ann Hum Genet* 2000; **64**: 1–14.

27 Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997; **53**: 1253–61.

28 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22**: 719–48.

29 Tregouet DA, Ducimetiere P, Tiret L. Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. *Am J Hum Genet* 1997; **61**: 189–99.

30 Tregouet DA, Konig IR, Erdmann J, Munteanu A, Braund PS, Hall AS, Grosshennig A, Linsel-Nitschke P, Perret C, DeSuremain M, Meitinger T, Wright BJ, Preuss M, Balmforth AJ, Ball SG, Meisinger C, Germain C, Evans A, Arveiler D, *et al*. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat Genet* 2009; **41**: 283–5.

31 Vlot AJ, Koppelman SJ, Bouma BN, Sixma JJ. Factor VIII and von Willebrand factor. *Thromb Haemost* 1998; **79**: 456–65.

32 Ito J, Ito M, Nambu H, Fujikawa T, Tanaka K, Iwaasa H, Tokita S. Anatomical and histological profiling of orphan G-protein-coupled receptor expression in gastrointestinal tract of C57BL/6J mice. *Cell Tissue Res* 2009; **338**: 257–69.

33 Kee HJ, Ahn KY, Choi KC, Won Song J, Heo T, Jung S, Kim JK, Bae CS, Kim KK. Expression of brain-specific angiogenesis inhibitor 3 (BAI3) in normal brain and implications for BAI3 in ischemia-induced brain angiogenesis and malignant glioma. *FEBS Lett* 2004; **569**: 307–16.

34 DeRosse P, Lencz T, Burdick KE, Siris SG, Kane JM, Malhotra AK. The genetics of symptom-based phenotypes: toward a molecular classification of schizophrenia. *Schizophr Bull* 2008; **34**: 1047–53.

35 Fujioka M, Hayakawa K, Mishima K, Kunizawa A, Irie K, Higuchi S, Nakano T, Muroi C, Fukushima H, Sugimoto M, Banno F, Kokame K, Miyata T, Fujiwara M, Okuchi K, Nishio K. ADAMTS13 gene deletion aggravates ischemic brain damage: a possible neuroprotective role of ADAMTS13 by ameliorating postischemic hypoperfusion. *Blood* 2010; **115**: 1650–3.

36 Zhao BQ, Chauhan AK, Canault M, Patten IS, Yang JJ, Dockal M, Scheiflinger F, Wagner DD. von Willebrand factor-cleaving protease ADAMTS13 reduces ischemic brain injury in experimental stroke. *Blood* 2009; **114**: 3329–34.

37 Bongers TN, de Bruijne EL, Dippel DW, de Jong AJ, Deckers JW, Poldermans D, de Maat MP, Leebeek FW. Lower levels of ADAMTS13 are associated with cardiovascular disease in young patients. *Atherosclerosis* 2009; **207**: 250–4.

BMC Medical Genetics

RESEARCH ARTICLE Open Access

Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels

Guillemette Antoni[1,2,3†], Tiphaine Oudot-Mellakh[2†], Apostolos Dimitromanolakis[3], Marine Germain[1,2], William Cohen[4,5], Philip Wells[6], Mark Lathrop[7], France Gagnon[3], Pierre-Emmanuel Morange[4,5] and David-Alexandre Tregouet[1,2*]

Abstract

Background: Elevated levels of factor VIII (FVIII) and von Willebrand Factor (vWF) are well-established risk factors for cardiovascular diseases, in particular venous thrombosis. Although high, the heritability of these traits is poorly explained by the genetic factors known so far. The aim of this work was to identify novel single nucleotide polymorphisms (SNPs) that could influence the variability of these traits.

Methods: Three independent genome-wide association studies for vWF plasma levels and FVIII activity were conducted and their results were combined into a meta-analysis totalling 1,624 subjects.

Results: No single nucleotide polymorphism (SNP) reached the study-wide significance level of 1.12×10^{-7} that corresponds to the Bonferroni correction for the number of tested SNPs. Nevertheless, the recently discovered association of *STXBP5*, *STX2*, *TC2N* and *CLEC4M* genes with vWF levels and that of *SCARA5* and STAB2 genes with FVIII levels were confirmed in this meta-analysis. Besides, among the fifteen novel SNPs showing promising association at $p < 10^{-5}$ with either vWF or FVIII levels in the meta-analysis, one located in *ACCN1* gene also showed weak association ($P = 0.0056$) with venous thrombosis in a sample of 1,946 cases and 1,228 controls.

Conclusions: This study has generated new knowledge on genomic regions deserving further investigations in the search for genetic factors influencing vWF and FVIII plasma levels, some potentially implicated in VT, as well as providing some supporting evidence of previously identified genes.

Background

Elevated plasma levels of factor VIII (FVIII) and von Willebrand factor (vWF), two key molecules of the coagulation cascade, are well-established risk factors for venous thrombosis (VT) [1-3]. More recent evidence shows that these plasma hemostatic proteins are also risk factors for other cardiovascular diseases (CVD) [4-8]. The broader role of FVIII and vWF is further supported by studies showing that genetic factors modulating the variability of these proteins are also associated with CVD. These include single nucleotide polymorphisms (SNPs) at the *BAI3* [9], *LDLR* [5,10], *VWF* [4] and *ABO* [11] genes, the latter being associated with other quantitative risk factors for CVD [12,13].

The estimated heritability of FVIII and vWF levels range between 40% and 60% [14,15] among which about 20% is attributable to the *ABO* locus. A genome wide association study (GWAS) within the CHARGE consortium [16] has recently identified five new genes, apart from their structural genes and *ABO*, consistently influencing vWF and/or FVIII plasma levels. These include *CLEC4M*, *SCARA5*, *STX2*, *STXBP5* and *TC2N*, collectively explaining ~10% of the variability of each two traits. These observations suggest that there are additional genetic factors remaining to be identified and contributing to the hidden heritability of these quantitative traits.

* Correspondence: david.tregouet@upmc.fr
† Contributed equally
[1]UMR_S 937, INSERM, Boulevard de l'Hopital, Paris, 75013, France
Full list of author information is available at the end of the article

The increased power of selected samples has long been recognized in family-based studies but more recently the putative advantages of carefully selected samples for quantitative trait analysis of unrelated subjects has also been highlighted [17]. Therefore, we undertook the combined analysis of individual data from three GWAS performed in samples of VT patients and in extended families ascertained on VT and Factor V Leiden (FVL) to identify novel genetic factors implicated in the variation of plasma levels of FVIII and vWF.

Methods

Overall strategy

To achieve our primary goal of identifying new genetic factors that could influence vWF and/or FVIII plasma levels, we used data from three carefully selected independent GWAS. Great attention was drawn to the homogeneity across samples in terms of - ethnic background (most individuals were of French origin), - exclusion criteria with respect to rare forms of inherited thrombophilia, - objectively diagnosed VT, - studied intermediate phenotypes (although some adjustments were done) and similar genotyping technologies (Illumina platform).

In the context of quantitative trait GWAS, individual genetic effect sizes are known to be small [18] and it is expected that a number of real associations do not reach genome-wide significance. Therefore, as part of our analytic strategy, we first tested for association in the individual studies, and results observed across samples were combined into a meta-analysis. We then focused on the consistency of associations across studies as our hypothesis was that real associations would more likely be consistently observed across studies given that each study samples were quite homogeneous with respect to the above-mentioned characteristics. Previously reported associations were also investigated using the above strategy.

As genetic variants associated to plasma levels of FVIII and vWF could be risk factors for VT, our secondary goal was to test the identified SNPs with VT using an *in silico* GWAS [19]. Analytic approaches and samples characteristics of the FVIII and vWF GWAS are described below.

FVL-families sample

Five extended French-Canadian families were ascertained through single probands with idiopathic VT diagnosed at the Thrombosis Clinic of the Ottawa Hospital, and carrying the FVL mutation. VT cases secondary to cancer as well as rare forms of inherited VT (protein S, protein C, AntiThrombin deficiencies) were excluded. A pedigree was drawn from interviews with each potential probands. The largest families were invited to participate in the study - the family size and willingness to participate being the only criteria for the selection of the families (see Additional File 1, File S1 for the used questionnaire). The total number of family members was 255. Description of the extended families has been published elsewhere [9].

MARTHA samples

The MARseille THrombosis Association (MARTHA) project is composed of two independent samples of VT patients, named MARTHA08 (N = 1,006) and MARTHA10 (N = 586). MARTHA subjects are unrelated caucasians consecutively recruited at the Thrombophilia center of La Timone hospital (Marseille, France) between January 1994 and October 2005. All patients had a documented history of VT and free of well characterized genetic risk factors including AT, PC, or PS deficiency, homozygosity for FV Leiden or FII 20210A, and lupus anticoagulant. They were interviewed by a physician on their medical history, which emphasized manifestations of deep vein thrombosis and pulmonary embolism using a standardized questionnaire (see Additional file 2, File S2). The thrombotic events were confirmed by venography, Doppler ultrasound, spiral computed tomographic scanning angiography, and/or ventilation/perfusion lung scan. All the subjects were of European origin, with the majority being of French descent.

The main characteristics of the three samples are shown in Table 1.

In silico GWAS study on VT

In a previously published GWAS on VT [19], 419 early age of onset and the idiopathic character of VT (ie without environemental risk factors) (< 50 years) VT cases were compared to 1,228 healthy controls at 291,872

Table 1 Main Characteristics of the Studied Samples

	FVL Families N = 253	MARTHA08 N = 972	MARTHA10 N = 570
Age (SD)	40.4 (17.9)	45.7 (14.9)	49.2 (15.7)
Sex (% female)	50.6%	70.8%	58.2%
Smoking (%)	24.4%	24.9%	22.71%
History of VT (%)	5.95%	100%	100%
PT G20210A carriers	0.40%	15.9%	10.6%
FV Leiden carriers	24.9%	26.6%	14.1%
ABO blood group (%)			
O	40.6%	22.9%	22.4%
A	57.8%	61.8%	59.3%
B	1.6%	10.3%	14.4%
AB	-	5%	3.9%
FVIII (SD) IU/dL	118.6 (38.51)	138.70 (55.34)	130.2 (46.35)
vWF (SD) IU/dL	130.3 (53.24)	152.33 (68.23)	152.9 (63.93)

SNPs. Cases were patients from four different French medical centers (Grenoble, Marseille, Montpellier, Paris) selected according to the same criteria as the MARTHA samples, except with the restriction on age of onset. Controls were French subjects selected from the SUVIMAX population [20].

Measurements

In the French-Canadian (FVL) sample, plasma levels of FVIII activity were measured by a clotting assay on the BCS instrument (Siemens Diagnostics, Marburg Germany) and vWF antigen was measured with a commercially available ELISA kit from Diagnostica Stago. The interassay coefficients of variation for FVIII were ~ 1% and 6.1% for vWF.

In MARTHA subjects, plasma coagulant activity and vWF antigen were assayed in an automated coagulometer (STA-R; Diagnostica Stago, Asnières, France). The interassay coefficients of variation for FVIII and vWF were 6.96% and 2.27% respectively.

Genotyping

The French-Canadian sample was genotyped with the Illumina 660W-Quad Beadchip. The raw datafile contained data for 547,886 autosomal SNPs genotyped on 255 individuals. From these SNPs, 490,083 passed the quality control (QC) criteria of genotyping rate > 90% and more than 20 observations of the minor allele among all individuals. After removing the 88,390 SNPs that failed QC, the overall genotyping rate was 99.88%. The maximum missing rate per sample for all the 255 samples was 3.9%, with an average missing rate of 0.13%. The family structures had previously been checked using 1079 microsatellite markers and RELPAIR [9]. To further verify the correctness of the family structure, we used PREST [21] and computed IBD estimates for all the sample pairs, within and across pedigrees. PREST reported 14,949 Mendelian errors, which is equivalent to a very low Mendelian error rate of 0.012% among all genotypes. Genotypes showing Mendelian inconsistencies were excluded from the analysis. Finally, phenotypic and genotypic data were available on a total of 253 individuals.

The MARTHA08 study sample was typed in 2008 with the Illumina Human610-Quad Beadchip containing 567,589 autosomal SNPs while the MARTHA10 sample was recently typed (beginning of 2010) with the same Illumina Human660W-Quad Beadchip as in the FVL study sample. SNPs showing significant ($P < 10^{-5}$) deviation from Hardy-Weinberg equilibrium, with minor allele frequency (MAF) less than 1% or genotyping call rate < 99%, in each study were filtered out. Individuals with genotyping success rates less than 95% were excluded from the analyses, as well as individuals demonstrating close relatedness as detected by pairwise clustering of identity by state distances (IBS) and multi-dimensional scaling (MDS) implemented in PLINK software [22]. Non-European ancestry was also investigated using the Eigenstrat program [23] leading to the final selection of 972 and 570 patients left for analysis in MARTHA08 and MARTH10, respectively. Plasma vWF levels were available in 834 and 537 MARTHA08 and MARTHA10 patients, respectively; corresponding numbers were 541 and 548 for plasma FVIII levels. A total of 442,728 SNPs were common to the three GWAS datasets (see Additional file 3, Figure S1).

Statistical analysis

In the FVL families, association of SNPs with vWF and FVIII levels was tested by means of measured genotype linear association analysis as implemented in the SOLAR (version 4.0, http://solar.txbiomedgenetics.org/download.html) program. In MARTHA subjects, association was tested using linear model as implemented in the PLINK program [22].

In order to handle differences in phenotype distributions across studies (Figure 1), and any possible deviation from normality, plasma levels of vWF and FVIII were first normalized before any statistical analysis using the normal quantile transformation [24], separately in the French-Canadian sample, MARTHA08 and MARTHA10. This transformation assigns to each observed measurement the quantile value of the standard normal distribution that corresponds to the rank of this measurement in the original untransformed distribution. Transformed variables are then normally distributed making linear models applicable, and linear regression coefficients comparable across studies. Association analyses were then carried out on the transformed variables assuming additive allele effects (0,1, 2 coding according to the number of minor alleles), and adjusting for age, sex and *ABO* blood group as tagged by the *ABO* rs8176746, rs8176704 and rs505922 [19].

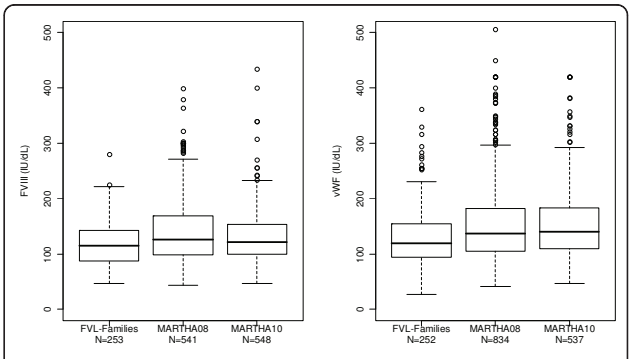


Figure 1 Box Plot Distribution of FVIII (left) and vWF (right) Plasma Levels in the Three GWAS Datasets.

When appropriate, haplotype association analyses were carried out in MARTHA samples using THESIAS software [25] to handle the correlation between SNPs, that is linkage disequilibrium (LD). This widely used software implements a stochastic-EM algorithm that simultaneously estimates the frequencies and the effect on the studied phenotype of each inferred haplotype. Haplotype - phenotype associations are then assessed by means of likelihood ratio tests.

Results obtained in each GWAS datasets were combined in a meta-analysis using the GWAMA program [26]http://www.sph.umich.edu/csg/abecasis/metal. Both fixed-effect and random-effect models- based analyses were conducted. Regression coefficients characterizing the minor allele effect of each SNP were then combined (after having checked that the minor allele was the same in the different populations) using the inverse-variance method to provide an overall allelic estimate. All reported P values were 2-sided.

Results

A total of 442,728 QC-validated SNPs were common to the three GWAS and were tested through a meta-analysis for association with vWF and FVIII plasma levels. Quantile-quantile plots did not reveal any inflation from what was expected under the null hypothesis of no association (Figure 2), and no SNP reached the study-wide significance level of 1.12×10^{-7} that corresponds to the Bonferroni correction for the number of tested SNPs. Applying the less stringent Sidak correction corresponding to a significant threshold of $p = 1.16 \times 10^{-7}$ would not have modified this conclusion. We then further focused on genetic effects that were consistent across studies and with combined p-value of less than 10^{-5}. As fixed-effect and random-effect analyses provided similar results for most of the main associations (Tables 2 &3), the following discussion is based on results obtained from the fixed-effect model analysis.

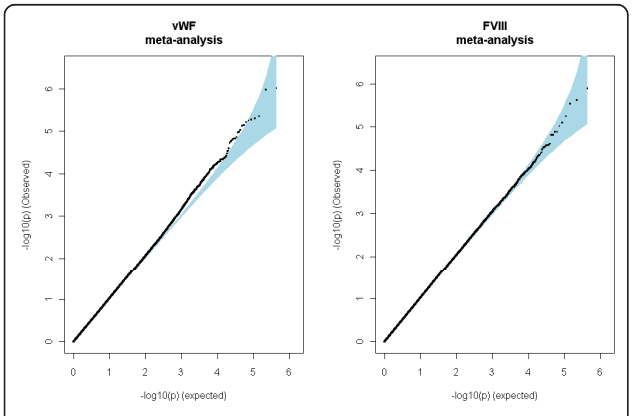


Figure 2 Quantile-Quantile Plots of the Association Results from the Meta-Analysis of the Three GWAS Datasets.

Ten SNPs covering seven different genes (Figure 3 - Table 2) were associated with plasma vWF levels at $p < 10^{-5}$ with no strong evidence for heterogeneity across GWAS as the lowest Mantel-Haenszel observed p-value, p = 0.036, for the ANKDR6 rs645764 would not pass multiple testing correction for testing ten SNPs. The strongest association was observed for rs379440 ($P = 9.82\ 10^{-6}$) mapping the *EPB41L4A* gene (Table 2). Another SNP at this locus was also associated with vWF, rs13361927 ($P = 4.51\ 10^{-6}$), but its association was due to its complete LD with rs379440, with pairwise r^2 of 0.78, 0.69 and 0.62 in FVL, MARTHA08 and MARTHA10, respectively. Other vWF-associated SNPs included the *SAFB2* rs732505 ($P = 9.38\ 10^{-6}$), *VPS8* rs4686760 ($P = 1.08\ 10^{-6}$) and the *KRT18P24* rs1757948 ($P = 7.37\ 10^{-6}$). The last three SNPs, rs1438993, rs10745527, rs2579103 (with P~ $6\ 10^{-6}$), were located at the 12q21.33 locus with no known mapped gene and were in nearly complete association. Altogether, the independent signals derived from the rs4686760, rs379440, rs1757948, rs10745527 and rs732505 explained up to 5.7% and 3.8% of the variability of plasma vWF levels in MARTHA08 and MARTHA10, respectively, and 5.3% in the pooled MARTHA samples.

None of the ten vWF-associated SNPs were associated with plasma FVIII levels (all p > 0.05). However, six additional SNPs were specifically associated to FVIII levels with homogeneous effects (Mantel-Haenszel p-value > 0.05) across studies (Figure 4 - Table 3). The strongest effect ($P = 2.95\ 10^{-6}$) was observed for rs7306642, a non synonymous Pro2039Thr variant within the *S*TAB2 gene, which was one of the recently identified genes by the CHARGE consortium. However, our hit rs7306642 was not in LD with any of the two *S*TAB2 SNPs recently identified, rs4981022 ($r^2 < 0.01$ in the three studies) and rs4981021 that served as a proxy for rs12229292 ($r^2 < 0.07$ in the three studies). Other FVIII-associated SNPs included the rs6708166 ($P = 1.30\ 10^{-6}$) in the proximity of *LBH*, the rs1321761 ~ 300 kb apart from *FAM46A* ($P = 9.54\ 10^{-6}$) and the intronic *VAV2* rs12344583 ($P = 7.92\ 10^{-6}$) (Table 3). Lastly, two SNPs within the *ACCN1* gene, rs1354492 and rs12941510, were found modulating FVIII plasma levels, the A allele of the former being associated with increased FVIII levels ($\beta = +0.16$, $P = 2.42\ 10^{-6}$) and the A allele of the latter being associated with decreased levels ($\beta = -0.17$, $P = 5.67\ 10^{-6}$). These two SNPs were in complete negative LD generating three haplotypes, the sole carrying the rs1354492-A allele being associated with highest levels (see Additional file 4, Table S1). Altogether, these five SNPs (i.e. rs6708166, rs1321761, rs12344583, rs7306642, rs1354492) explained 8.2% and

Table 2 Ten SNPs Showing Association with vWF levels Across the Three GWAS Datasets With Combined Significance P-value < 10^{-5}

Gene	SNP	Alleles*		MAF+	β (SE)	p	I^2	p_{het}	Random Effect		Fixed Effect	
									β (SE)	p	β (SE)	p
			FVL	0.47	-0.16 (0.08)	0.044						
VPS8	rs4686760	A/G	Martha08	0.46	-0.18 (0.04)	$4.11\ 10^{-5}$	0	0.549	0.15 (0.03)	$1.10\ 10^{-6}$	-0.15 (0.03)	$1.08\ 10^{-6}$
			Martha10	0.45	-0.11 (0.05)	0.047						
			FVL	0.15	0.44 (0.11)	$3.08\ 10^{-4}$						
	rs13361927	*G/A*	Martha08	0.06	0.28 (0.09)	0.003	0.53	0.119	-0.28 (0.09)	0.002	0.28 (0.06)	$4.51\ 10^{-6}$
EPB41L4A			Martha10	0.05	0.11 (0.11)	0.316						
			FVL	0.12	0.46 (0.12)	$8.35\ 10^{-4}$						
	rs379440	A/G	Martha08	0.04	0.31 (0.11)	0.004	0	0.502	-0.34 (0.07)	$9.99\ 10^{-7}$	0.34 (0.07)	$9.82\ 10^{-7}$
			Martha10	0.03	0.25 (0.14)	0.071						
			FVL	0.04	-0.01 (0.21)	0.977						
ANKRD6	rs6454764	*C/T*	Martha08	0.06	0.24 (0.09)	0.007	0.70	0.036	-0.29 (0.14)	0.035	0.31 (0.07)	$5.12\ 10^{-6}$
			Martha10	0.05	0.54 (0.12)	$8.97\ 10^{-6}$						
			FVL	0.27	0.34 (0.09)	$2.82\ 10^{-4}$						
KRT18P24	rs1757948	*T/G*	Martha08	0.27	0.1 (0.05)	0.030	0.62	0.071	-0.18 (0.06)	0.003	0.15 (0.03)	$7.37\ 10^{-6}$
			Martha10	0.30	0.15 (0.06)	0.009						
			FVL	0.19	0.15 (0.1)	0.127						
	rs1438993	G/A	Martha08	0.28	0.18 (0.05)	$1.11\ 10^{-4}$	0	0.666	-0.16 (0.03)	$6.34\ 10^{-6}$	0.16 (0.03)	$6.25\ 10^{-6}$
			Martha10	0.27	0.12 (0.06)	0.052						
			FVL	0.20	0.19 (0.1)	0.062						
desert	rs10745527	T/G	Martha08	0.28	0.18 (0.05)	$1.63\ 10^{-4}$	0	0.663	-0.16 (0.03)	$5.51\ 10^{-6}$	0.16 (0.03)	$5.43\ 10^{-6}$
			Martha10	0.27	0.11 (0.06)	0.056						
			FVL	0.18	0.17 (0.11)	0.098						
	rs2579103	T/G	Martha08	0.26	0.19 (0.05)	$8.24\ 10^{-5}$	0	0.533	-0.16 (0.04)	$7.72\ 10^{-6}$	0.16 (0.04)	$7.61\ 10^{-6}$
			Martha10	0.25	0.1 (0.06)	0.090						
			FVL	0.04	-0.02 (0.19)	0.905						
CDH2	rs2298574	A/G	Martha08	0.08	-0.34 (0.08)	$2.77\ 10^{-5}$	0.19	0.290	0.26 (0.07)	$1.81\ 10^{-4}$	-0.27 (0.06)	$5.67\ 10^{-6}$
			Martha10	0.07	-0.24 (0.1)	0.022						
			FVL	0.05	0.32 (0.18)	0.080						
SAFB2	rs732505	G/A	Martha08	0.09	0.24 (0.08)	0.001	0	0.929	-0.25 (0.06)	$9.50\ 10^{-6}$	0.25 (0.06)	$9.38\ 10^{-6}$
			Martha10	0.08	0.25 (0.1)	0.013						

*Common/rare alleles

+ Allele frequency of the minor allele

4.6% of the variability of FVIII levels in MARTHA08 and MARTHA10, respectively, and 6.3% in the combined MARTHA samples.

We then used our GWAS datasets to investigate SNPs that had previously been reported associated with vWF and/or FVIII [4,5,9,16]. As shown in Supplementary Table two, marginal associations (P < 0.05) with vWF levels at *STXBP5*, *VWF*, *STX2*, *TC2N* and *CLEC4M* were also observed in our study, the strongest (P = $1.3\ 10^{-4}$) being for SNP rs216335 at the structural *VWF* gene. All these associations were consistent (i.e the same allele was associated with a genetic effect in the same direction on the studied phenotype) with those previously reported. Together, these associations explained an additional 1.4% and 3.2% of the variance of plasma levels of vWF in MARTHA08 and MARTHA10, respectively. We did not observe any evidence for an effect of *S*TAB2 rs4981022 or *BAI3* rs9363864, while the effect of *SCARA5* rs2726953 was heterogeneous across the studies. For FVIII levels, we observed marginal associations of *SCARA5* rs9644133 (P = 0.009) and *VWF* rs1063856 (P = 0.020) that were consistent with those previously reported (Table 4), these two SNPs explaining 0.7% and 0.2% of FVIII variability in MARTHA08 and MARTHA10, respectively. No trend for association was observed for the previously reported associations with *STXBP5*, *S*TAB2 nor *LDLR* SNPs (Table 5).

We have recently observed that, among the newly identified vWF and/or FVIII genes by the CHARGE consortium, *TC2N* could also be associated with VT risk [27]. Therefore

Table 3 Six SNPs Showing Association with FVIII Activity Across the Three GWAS Datasets With Combined Significance P-value < 10^{-5}

Gene	SNP	Alleles*		MAF+	β (SE)	p	I^2	p_{het}	Random Effect		Fixed Effect	
									β (SE)	p	β (SE)	p
LBH	rs6708166	G/A	FVL	0.41	-0.12 (0.09)	0.156	0	0.478	-0.17 (0.04)	1.32 10^{-6}	-0.17 (0.04)	1.30 10^{-6}
			Martha08	0.40	-0.23 (0.06)	8.98e-05						
			Martha10	0.42	-0.15 (0.05)	0.007						
FAM46A	rs1321761	T/C	FVL	0.42	-0.20 (0.08)	0.014	0	0.451	-0.15 (0.04)	9.67 10^{-6}	-0.15 (0.04)	9.54 10^{-6}
			Martha08	0.45	-0.10 (0.06)	0.074						
			Martha10	0.47	-0.19 (0.05)	5.93e-04						
VAV2	rs12344583	A/G	FVL	0.17	0.28 (0.11)	0.012	0	0.716	0.20 (0.04)	8.03 10^{-6}	0.20 (0.04)	7.92 10^{-6}
			Martha08	0.20	0.19 (0.07)	0.006						
			Martha10	0.18	0.17 (0.07)	0.012						
STAB2	rs7306642	*C/A*	FVL	0.16	0.52 (0.12)	1.36e-05	0.59	0.086	0.31 (0.10)	0.002	0.30 (0.06)	2.95 10^{-6}
			Martha08	0.07	0.22 (0.11)	0.057						
			Martha10	0.07	0.20 (0.1)	0.052						
ACCN1	rs1354492	*G/A*	FVL	0.53	0.09 (0.08)	0.293	0.39	0.192	0.16 (0.04)	5.47 10^{-6}	0.16 (0.03)	2.41 10^{-6}
			Martha08	0.49	0.23 (0.05)	1.20e-05						
			Martha10	0.47	0.12 (0.05)	0.027						
	rs12941510	G/A	FVL	0.22	-0.29 (0.1)	0.004	0.12	0.321	-0.17 (0.04)	2.18 10^{-5}	-0.17 (0.04)	5.67 10^{-6}
			Martha08	0.31	-0.17 (0.06)	0.002						
			Martha10	0.33	-0.12 (0.06)	0.029						

*Common/rare alleles

+ Allele frequency of the minor allele

we investigated the effect of the SNPs identified in our meta-analysis on the risk of VT. Our working hypothesis was that SNPs associated with increased (decreased, resp.) plasma levels of these two molecules could be associated with increased (decreased, resp.) risk of disease. For this, we used the results of our previously published GWAS based on 419 VT patients and 1228 healthy subjects (*in silico* association) [19]. As indicated in Table 6, only two SNPs, *VPS8* rs4686760 and *ACCN1* rs12941510, showed some trend of association consistent with our hypothesis. The rs4686760-G allele found associated with decreased vWF levels was slightly less frequent in VT patients than in controls (0.441 vs 0.475, $P = 0.101$) and the rs12941510-A allele, associated with decreased FVIII levels, was also less frequent in cases than in controls (0.310 vs 0.350, $P = 0.046$). These associations can only be considered as suggestive as they would not pass correction for multiple testing. Nevertheless, the observed homogeneity of the allele frequencies of these two SNPs across all genotyped patients is noteworthy. Combining all the VT patients (n = 1946), and comparing to the healthy controls of the *in silico* GWAS, the association of rs4686760 with VT remained (0.454 vs 0.475, $P = 0.108$), and that of rs12941510 was strengthened (0.314 vs 0.348, $P = 0.0056$) (Table 7).

Discussion

Theoretically, a sample size of 1,624 unrelated individuals should have a power of 95% to detect, at the significant level of 1.12 10^{-7}, the additive allele effect of a SNP explaining at least 3% if the variability of a quantitative trait [28]. This power would decrease to 86% and 66% for a SNP explaining 2.5% and 2%, respectively. Our meta-analysis of 1,624 carefully selected samples did not reveal any genome-wide significant association suggesting that the additional common SNPs tagged by current GWAS array and influencing vWF and FVIII plasma levels left to be identified would, if any, individually explain less than 2% of the variability of these two traits.

By lowering the statistical stringency to $p < 10^{-5}$ but focusing on the homogeneity of the effects observed in three independent samples, we identified several novel candidate genes that could contribute to modulate the variability of vWF and FVIII, and that deserve to be further studied. The novel candidate genes for vWF are *VPS8*, *EBP41L4A*, *KRT18P24*, *SAFB2* and a region on 12q21.3 where no known gene maps. Unfortunately, little is known about the biology of the associated proteins and their role in cardiovascular diseases. Among these, *VPS8* stands out. The rs4686760-G allele of the *VPS8* gene, which was associated with decreased vWF levels, was also observed less frequently in VT cases than in healthy controls (0.45 vs 0.48) in the *in silico* GWAS, although this observation did not reach significance ($P = 0.10$). The vacuolar protein sorting 8 homolog gene (*VPS8*) is involved in protein traffic between the golgic

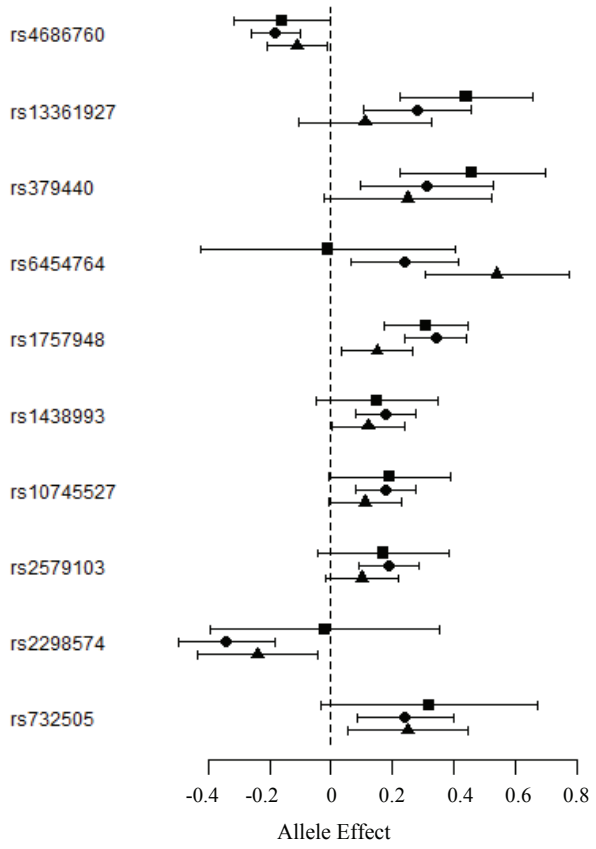


Figure 3 Forest plot representation of the ten SNPs that associated the most with vWF levels in the Three GWAS Datasets. Results observed in the FVL families, MARTHA08 and MARTHA10 studies are depicted by square, circle and triangle, respectively.

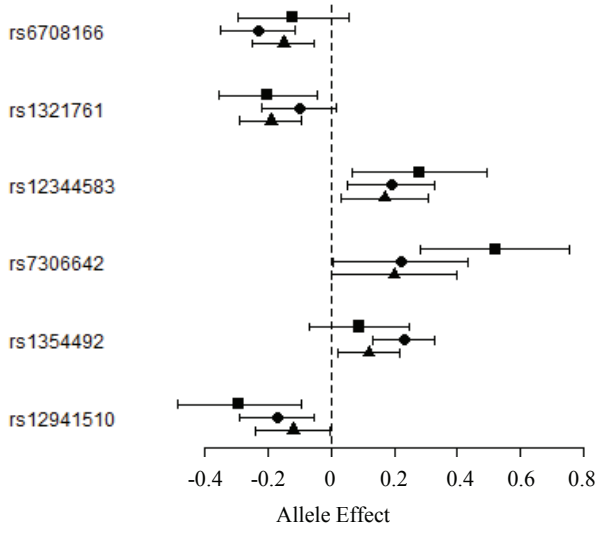


Figure 4 Forest plot representation of the six SNPs that associated the most with FVIII levels in the Three GWAS Datasets. Results observed in the FVL families, MARTHA08 and MARTHA10 studies are depicted by square, circle and triangle, respectively.

appartus and the vacuaole [29] and could participate to the regulation of urokinase-type plasminogen activator [30], the latter known to be involved in thrombosis.

For FVIII levels, the candidate genes identified in our study were *LBH*, *FAM46A*, *VAV2*, *S*TAB2 and *ACCN1*. Both *LBH* and *VAV2* genes are thought to be involved in angiogenesis. The transcriptional cofactor limb-bud-and-heart (Lbh) was discovered as a small acidic nuclear protein highly conserved among species [31]. It has been demonstrated a dramatic suppression of VEGF mRNAs in cells that overexpress Lbh [32]. Vav2 is a guanine nucleotide exchange factor for Rho family proteins. The expression of a dominant negative form of Vav2 suppress the Vascular Endothelial-Protein Tyrosine Phosphatise (VE-PTP)-induced changes in endothelial cell morphology, such changes being implicated in regulation of angiogenesis [33].

Interestingly, we had previously shown that *S*TAB2 was located within a linkage peak for vWF levels in our FVL extended families [9] while almost concomitantly *S*TAB2 SNPs were found associated with both FVIII and vWF in the CHARGE consortium GWAS [16]. However, the non-synonymous rs7306642 (Pro2039Thr) found associated here with FVIII levels did not show a homogeneous effect on vWF levels across the three GWAS datasets (data not shown), and was in very low LD with others *S*TAB2 SNPs found associated with these plasma levels. The substitution of a Proline by a Threonine at position 2039 is predicted to be damaging according to web resources http://genetics.bwh.harvard.edu/pph/index.html; http://www.rostlab.org/services/SNAP. Investigating the effect of this substitution on VT risk would have been relevant but the corresponding SNP did not pass quality control in our *in silico* GWAS. These observations nevertheless suggest that an in-depth haplotype analysis of the *S*TAB2 gene are required to gain better insight into which SNPs more likely influence plasma levels of FVIII and/or vWF.

ACCN1, encoding an amiloride-sensitive cation channel implicated in cell growth and migration [34], is another gene that deserves greater attention as its genetic variability was found here associated with both FVIII levels and VT risk. However, the SNP that seemed to modulate FVIII levels the most, rs1354492, was not the one that showed association with the disease. This could suggest that either different SNPs distinctly influence plasma levels and VT risk, or that the identified SNPs are in LD with unmeasured variant(s) that could simultaneously influence both phenotypes.

Table 4 Association of Previously Identified SNPs with vWF Levels in the three GWAS Datasets

Gene	SNP	Alleles*		MAF+	β (SE)	p	I^2	p_{het}	Random Effect		Fixed Effect	
									β (SE)	p	β (SE)	p
			FVL	0.42	0.04 (0.08)	0.618						
BAI3	rs9363864	A/G	Martha08	0.52	0.03 (0.04)	0.421	0	0.838	0.02 (0.03)	0.461	0.02 (0.03)	0.461
			Martha10	0.49	-0.002 (0.05)	0.973						
			FVL	0.43	-0.08 (0.08)	0.366						
STXBP5	rs9390459	G/A	Martha08	0.42	-0.06 (0.04)	0.197	0	0.545	-0.09 (0.03)	0.005	-0.09 (0.03)	0.005
			Martha10	0.43	-0.13 (0.05)	0.011						
			FVL	0.20	-0.08 (0.10)	0.446						
SCARA5	rs10866867(1)	G/T	Martha08	0.25	0.17 (0.05)	4.88e-04	0.71	0.03	0.05 (0.07)	0.466	0.09 (0.04)	0.015
			Martha10	0.25	0.01 (0.06)	0.830						
			FVL	0.06	-0.28 (0.19)	0.141						
	rs216335(2)	G/A	Martha08	0.08	-0.23 (0.08)	0.003	0	0.945	-0.23 (0.06)	$1.31\ 10^{-4}$	-0.23 (0.06)	$1.30\ 10^{-4}$
			Martha10	0.06	-0.21 (0.11)	0.059						
			FVL	0.45	0.07 (0.08)	0.371						
VWF	rs1063856(3)	A/G	Martha08	0.37	0.08 (0.05)	0.094	0	0.889	0.09 (0.03)	0.006	0.09 (0.03)	0.006
			Martha10	0.38	0.11 (0.05)	0.041						
			FVL	0.48	-0.04 (0.08)	0.612						
	rs7306706	A/G	Martha08	0.45	0.02 (0.04)	0.634	0	0.754	0.01 (0.03)	0.664	0.01 (0.03)	0.664
			Martha10	0.46	0.03 (0.05)	0.604						
			FVL	0.30	-0.05 (0.09)	0.601						
STAB2	rs4981022	T/C	Martha08	0.30	0.02 (0.05)	0.652	0	0.541	-0.01 (0.03)	0.664	-0.01 (0.03)	0.664
			Martha10	0.28	-0.06 (0.06)	0.333						
			FVL	0.33	0.01 (0.09)	0.863						
STX2	rs4334059(4)	C/T	Martha08	0.37	0.08 (0.04)	0.067	0.01	0.363	0.1 (0.03)	0.004	0.1 (0.03)	0.003
			Martha10	0.36	0.15 (0.06)	0.008						
			FVL	0.52	0.05 (0.08)	0.548						
TC2N	rs2402074(5)	G/A	Martha08	0.48	0.04 (0.04)	0.382	0	0.509	0.07 (0.03)	0.033	0.07 (0.03)	0.033
			Martha10	0.47	0.12 (0.05)	0.030						
			FVL	0.22	-0.07 (0.1)	0.515						
CLEC4M	rs868875	A/G	Martha08	0.32	-0.10 (0.05)	0.036	0	0.762	-0.08 (0.03)	0.026	-0.08 (0.03)	0.026
			Martha10	0.35	-0.05 (0.06)	0.424						

* Common/rare alleles

+ Allele frequency of the minor allele

(1) rs10866867 serves as proxy for rs2726953 (r^2 = 0.92); (2) rs216335 serves as proxy for rs216318 (r^2 = 1)

(3) rs1063856 serves as proxy for Rs1063857 (r^2 = 1); (4) rs4334059 serves as proxy for rs7978987 (r^2 = 1.0

(5) rs2402074 serves as proxy for rs10133762 (r^2 = 0.96); No good proxy with r^2 > 0.5 was available for the *VWF* rs4764478

Our meta-analysis was also able to replicate several of the previously reported associations between SNPs and vWF/FVIII levels. Replicated associations include vWF-associated SNPs at *STXBP5*, *VWF*, *STX2*, *TC2N* and *CLEC4M* genes, and FVIII-associated SNPs within *SCARA5* and *VWF* genes. Other previously reported associations were not replicated, such as those involving *LDLR*, *BAI3*, and *S*TAB2 SNPs [5,9,16]. In addition to a lack of power, as previously discussed, this could be due to differential effects of SNP in normal range of plasma levels compared to the higher levels observed in VT patients. This could apply to the association of *BAI3* with vWF levels observed in healthy nuclear families [9] where plasma levels were lower than those observed in our VT samples. Conversely, this explanation does not completely hold for the *LDLR* SNPs that were found associated with FVIII activity in a population [5] where FVIII activity in healthy individuals were at higher levels than those observed in our VT patients. Besides, in these two studies, different methods from those we have used here were employed to measure vWF and FVIII activity, and this could also contribute to the discrepancies observed in our study.

Conclusions

In conclusion, a carefully planned meta-analysis of three independent samples gathering 1,624 individuals genotyped for more than 400,000 SNPs all over the genome

Table 5 Association of Previously Identified SNPs with FVIII Activity in the three GWAS Datasets

Gene	SNP	Alleles*		MAF+	β (SE)	p	I^2	p_{het}	Random Effect		Fixed Effect	
									β (SE)	p	β (SE)	p
			FVL	0.43	0.15 (0.08)	0.083						
STXBP5	rs9390459	*G/A*	Martha08	0.42	-0.08 (0.06)	0.158	0.65	0.059	-0.02 (0.06)	0.795	-0.04 (0.03)	0.310
			Martha10	0.43	-0.07 (0.05)	0.199						
			FVL	0.24	-0.08 (0.1)	0.433						
SCARA5	rs9644133	C/T	Martha08	0.17	-0.16 (0.07)	0.029	0	0.753	-0.12 (0.05)	0.009	-0.12 (0.05)	0.009
			Martha10	0.18	-0.10 (0.07)	0.152						
			FVL	0.45	0.11 (0.08)	0.170						
VWF	rs1063856	A/G	Martha08	0.37	0.09 (0.06)	0.114	0	0.843	0.08 (0.03)	0.020	0.08 (0.03)	0.020
			Martha10	0.38	0.06 (0.05)	0.249						
			FVL	0.27	-0.13 (0.09)	0.146						
STAB2	rs4981021[(1)]	G/A	Martha08	0.32	-0.02 (0.06)	0.737	0	0.389	-0.02 (0.04)	0.521	-0.02 (0.04)	0.521
			Martha10	0.29	0.02 (0.06)	0.782						
			FVL	0.14	-0.03 (0.11)	0.816						
	rs2228671	C/T	Martha08	0.11	0.11 (0.09)	0.193	0.46	0.157	-0.01 (0.07)	0.890	-0.01 (0.05)	0.894
LDLR			Martha10	0.10	-0.13 (0.09)	0.161						
			FVL	0.38	-0.25 (0.09)	0.005						
	rs688	C/T	Martha08	0.45	0.06 (0.05)	0.235	0.79	0.010	-0.05 (0.08)	0.531	-0.02 (0.03)	0.652
			Martha10	0.45	-0.007 (0.05)	0.901						

* Common/rare alleles

+ Allele frequency of the minor allele

(1) rs4981021 serves as proxy for rs12229292 (r^2 = 0.88)

Table 6 In Silico Association With Venous Thrombosis of the Identified vWF- and FVIII Associated SNPs

		Alleles*	Minor Allele Frequency		Cochran Armitage *P*-value
			Cases	Controls	
	vWF associated SNPs				
VPS8	rs4686760	A/G	0.441	0.475	*P* = 0.101
EPB41L4A	rs13361927	G/A	0.065	0.062	*P* = 0.797
KRT18P24	rs1634352†	G/A	0.284	0.318	*P* = 0.055
12q21.33	rs1438933	G/A	0.256	0.294	*P* = 0.051
CDH2	rs2298574	A/G	0.084	0.093	*P* = 0.444
SAFB2	rs732505	G/A	0.061	0.064	*P* = 0.713
	FVIII associated SNPs				
VAV2	rs12344583	A/G	0.217	0.193	*P* = 0.133
ACCN1	rs1354492	G/A	0.476	0.469	*P* = 0.740
ACCN1	rs12941510	G/A	0.310	0.350	*P* = 0.046

*Common/minor alleles

† serves as proxy for rs1757948 (r^2 = 1).

No good proxy with r^2 > 0.80 was available for rs6708166 (*LBH*), rs1321761 (*FAM46A*) and rs7306642 (STAB2)

Table 7 Genotype Distributions of rs4686760 and rs12941510 Across VT Samples

	rs4686760			
	AA	AG	GG	MAF [2]
MARTHA08	271	502	198	0.462
MARTHA10	173	281	115	0.449
GWAS patients	129	196	81	0.441
All VT patients	573	979	394	0.454
GWAS controls	354	581	292	0.475
	Test of association $P = 0.108$[1]			
	rs12941510			
	AA	AG	GG	MAF
MARTHA08	93	409	469	0.306
MARTHA10	67	243	259	0.331
GWAS patients	45	161	199	0.310
All VT patients	205	813	927	0.314
GWAS controls	139	576	512	0.348
	Test of association $P = 0.0056$			

[1] Cochran Armitage trend test

[2] Minor Allele Frequency

replicated very recent findings but did not reveal any new genetic factors that could individually explain at least 2% of the plasma variability of vWF and FVIII levels.

Additional material

Additional file 1: FVL Family Questionnaire.

Additional file 2: MARTHA questionnaire. Excel file illustrating the questionnaire used for selecting MARTHA VT patients.

Additional file 3: Figure S1. Genotype filtering strategy applied to the three GWAS datasets. [1] A genotype calling rate of > 0.90 was used in the FVL families and a threshold of 0.99 was used for the MARTHA patients. [2] SNPs with minor allele frequency less than 0.04 and 0.01 in FVL families and MARTHA patients, respectively, were excluded from the analysis. [3] SNPs demonstrating deviation from Hardy-Weinberg equilibrium at $p < 10^{-5}$ were excluded. 217 SNPs failed the genotype calling criterion simultaneously in the three study samples and this number was 19,111 for the minor allele frequency criterion. 19 SNPs failed the Hardy-Weinberg criterion in MARTHA08 and MARTHA10.

Additional file 4: Table S1. Haplotype Association Analysis of *ACCN1* rs1354492 and rs12941510 With Plasma FVIII levels in MARTHA08 and MARTHA10 Studies. [1] Haplotypic effect associated with each haplotype by comparison to the most frequent AG haplotype under the assumption of haplotype additive effects. Analyses were adjusted for age, sex and *ABO* blood group.

Acknowledgements
The French-Canadian FVL family study was supported by grants from the Canadian Institutes of Health Research (MOP86466) and by the Heart and Stroke Foundation of Canada (T6484). The MARTHA studies were supported by a grant from the Program Hospitalier de la Recherche Clinique. G.A hold an "INSERM Poste d'accueil" position and T.O.M was supported by a grant from the Fondation pour la Recherche Médicale. F.G and P.W. hold Canada Research Chairs. A France-Canada Research Fund 2008 provided opportunities for face-to-face meetings of lead collaborators.

Author details
[1]UMR_S 937, INSERM, Boulevard de l'Hopital, Paris, 75013, France. [2]UMR_S 937, ICAN Institute, Université Pierre et Marie Curie, Boulevard de l'Hopital, 75013, Paris, France. [3]Dalla Lana School of Public Health, University of Toronto, College Street, Toronto, M5T 3M7, Ontario, Canada. [4]UMR_S 626, INSERM, rue Saint-Pierre, Marseille, 13385, France. [5]UMR_S 626, Université de la Méditerranée, rue Saint-Pierre, Marseille, 13385 France. [6]Department of Medicine, Ottawa Hopital Research Institute, Carling Avenue, Ottawa, K1Y 4E9, Ontario, Canada. [7]Institut de Génomique, Centre National de Génotypage, Commissariat à l'Energie Atomique, rue Gaston Crémieux, Evry, 91057, France.

Authors' contributions
PEM, ML, FG and DAT designed the study and directed its implementation. GA, TOM and AD carried out statistical analyses. MG and WC were responsible for data collection and database management. GA drafted the article that was further reviewed by PEM, FG and DAT.
All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 7 May 2011 Accepted: 2 August 2011
Published: 2 August 2011

References

1. Koster T, Blann AD, Briet E, Vandenbroucke JP, Rosendaal FR: **Role of clotting factor VIII in effect of von Willebrand factor on occurrence of deep-vein thrombosis.** *Lancet* 1995, **345**:152-155.
2. Kraaijenhagen RA, in't Anker PS, Koopman MM, Reitsma PH, Prins MH, van den Ende A, *et al*: **High plasma concentration of factor VIIIc is a major risk factor for venous thromboembolism.** *Thromb Haemost* 2000, **83**:5-9.
3. Tsai AW, Cushman M, Rosamond WD, Heckbert SR, Tracy RP, Aleksic N, *et al*: **Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE).** *Am J Med* 2002, **113**:636-642.
4. van Schie MC, de Maat MP, Isaacs A, van Duin CM, Deckers JW, Dippel DW, *et al*: **Variation in the von Willebrand Factor gene is associated with VWF levels and with the risk of cardiovascular disease.** *Blood* 2011, **117**:1393-1399.
5. Martinelli N, Girelli D, Lunghi B, Pinotti M, Marchetti G, Malerba G, *et al*: **Polymorphisms at LDLR locus may be associated with coronary artery disease through modulation of coagulation factor VIII activity and independently from lipid profile.** *Blood* 2010, **116**:5688-5697.
6. Whincup PH, Danesh J, Walker M, Lennon L, Thomson A, Appleby P, *et al*: **von Willebrand factor and coronary heart disease: prospective study and meta-analysis.** *Eur Heart J* 2002, **23**:1764-1770.
7. Folsom AR, Rosamond WD, Shahar E, Cooper LS, Aleksic N, Nieto FJ, *et al*: **Prospective study of markers of hemostatic function with risk of ischemic stroke. The Atherosclerosis Risk in Communities (ARIC) Study Investigators.** *Circulation* 1999, **100**:736-742.
8. Cambronero F, Vilchez JA, Garcia-Honrubia A, Ruiz-Espejo F, Moreno V, Hernandez-Romero D, *et al*: **Plasma levels of von Willebrand factor are increased in patients with hypertrophic cardiomyopathy.** *Thromb Res* 2010, **126**:e46-50.
9. Antoni G, Morange PE, Luo Y, Saut N, Burgos G, Heath S, *et al*: **A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism.** *J Thromb Haemost* 2010, **8**:2671-2679.
10. Vormittag R, Bencur P, Ay C, Tengler T, Vukovich T, Quehenberger P, *et al*: **Low-density lipoprotein receptor-related protein 1 polymorphism 663 C > T affects clotting factor VIII activity and increases the risk of venous thromboembolism.** *J Thromb Haemost* 2007, **5**:497-4502.
11. Carpeggiani C, Coceani M, Landi P, Michelassi C, L'Abbate A: **ABO blood group alleles: A risk factor for coronary artery disease. An angiographic study.** *Atherosclerosis* 2010, **211**:461-466.
12. Teupser D, Baber R, Ceglarek U, Scholz M, Illig T, Gieger C, *et al*: **Genetic regulation of serum phytosterol levels and risk of coronary artery disease.** *Circ Cardiovasc Genet* 2010, **3**:331-339.

13. Barbalic M, Dupuis J, Dehghan A, Bis JC, Hoogeveen RC, Schnabel RB, *et al*: **Large-scale genomic studies reveal central role of ABO in sP-selectin and sICAM-1 levels.** *Hum Mol Genet* 2010, **19**:1863-1872.
14. Souto JC, Almasy L, Borrell M, Gari M, Martinez E, *et al*: **Genetic determinants of hemostasis phenotypes in Spanish families.** *Circulation* 2000, **101**:1546-1551.
15. Morange PE, Tregouet DA, Frere C, Saut N, Pellegrina L, Alessi MC, *et al*: **Biological and genetic factors influencing plasma factor VIII levels in a healthy family population: results from the Stanislas cohort.** *Br J Haematol* 2005, **128**:91-99.
16. Smith NL, Chen M-H, Dehghan A, Strachan DP, Basu S, Soranzo N, *et al*: **Novel associations of multiple genetic loci with plasma levels of Factor VII, Factor VIII and von Willebrand Factor. The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium.** *Circulation* 2010, **121**:1392-1392.
17. Abecasis GR, Cookson WO, Cardon LR: **The power to detect linkage disequilibrium with quantitative traits in selected samples.** *Am J Hum Genet* 2001, **68**:1463-1474.
18. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, *et al*: **Biological, clinical and population relevance of 95 loci for blood lipids.** *Nature* 2010, **466**:707-713.
19. Tregouet DA, Heath S, Saut N, Biron-Andreani C, Scheved JF, Pernod G, *et al*: **Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach.** *Blood* 2009, **113**:5298-5303.
20. Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, *et al*: **The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals.** *Arch Intern Med* 2004, **164**:2335-2342.
21. Sun L, Wilder K, McPeek MS: **Enhanced pedigree error detection.** *Hum Hered* 2002, **54**:99-110.
22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559-575.
23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**:904-909.
24. Peng B, Yu RK, Dehoff KL, Amos CI: **Normalizing a large number of quantitative traits using empirical normal quantile transformation.** *BMC Proc* 2007, **1(Suppl 1)**:S156.
25. Tregouet DA, Garelle V: **A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies.** *Bioinformatics* 2007, **23**:1038-1039.
26. Magi R, Morris AP: **GWAMA: software for genome-wide association meta-analysis.** *BMC Bioinformatics* 2010, **11**:288.
27. Morange PE, Saut N, Antoni G, Emmerich J, Tregouet DA: **Impact on venous thrombosis risk of newly discovered gene variants associated with FVIII and VWF plasma levels.** *J Thromb Haemost* 2011, **9**:229-231.
28. Gauderman WJ, Morrison JM: **Quanto 1.1: a computer program for power and sample sizee calculations for genetic-epidemiology studies.** 2006 [http://hydra.usc.edu/gxe].
29. Chen YJ, Stevens TH: **The VPS8 gene is required for localization and trafficking of the CPY sorting receptor in Saccharomyces cerevisiae.** *Eur J Cell Biol* 1996, **70**:289-297.
30. Agaphonov M, Romanova N, Sokolov S, Iline A, Kalebina T, *et al*: **Defect of vacuolar protein sorting stimulates proteolytic processing of human urokinase-type plasminogen activator in the yeast Hansenula polymorpha.** *FEMS Yeast Res* 2005, **5**:1029-1035.
31. Briegel KJ, Baldwin HS, Epstein JA, Joyner AL: **Congenital heart disease reminiscent of partial trisomy 2p syndrome in mice transgenic for the transcription factor Lbh.** *Development* 2005, **132**:3305-3316.
32. Conen KL, Nishimori S, Provot S, Kronenberg HM: **The transcriptional cofactor Lbh regulates angiogenesis and endochondral bone formation during fetal bone development.** *Dev Biol* 2009, **333**:348-358.
33. Mori M, Murata Y, Kotani T, Kusakari S, Ohnishi H, Saito Y: **Promotion of cell spreading and migration by vascular endothelial-protein tyrosine phosphatase (VE-PTP) in cooperation with integrins.** *J Cell Physiol* 2010, **224**:195-204.
34. Vila-Carriles WH, Kovacs GG, Jovov B, Zhou ZH, Pahwa AK, Colby G, *et al*: **Surface expression of ASIC2 inhibits the amiloride-sensitive current and migration of glioma cells.** *J Biol Chem* 2006, **281**:19220-19232.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1471-2350/12/102/prepub

doi:10.1186/1471-2350-12-102
Cite this article as: Antoni *et al.*: Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. *BMC Medical Genetics* 2011 **12**:102.

Impact on venous thrombosis risk of newly discovered gene variants associated with FVIII and VWF plasma levels

P.-E. MORANGE,* N. SAUT,* G. ANTONI,† J. EMMERICH‡ and D.-A. TRÉGOUËT†

**INSERM, UMR_S 626, F-13385, Marseille, France; Université de la Méditerranée, Marseille; †INSERM UMR_S 937, Université Pierre et Marie Curie (UPMC, Paris 6), Paris; ‡INSERM U765, Médecine vasculaire – HTA, Hôpital Européen Georges-Pompidou, Université Paris-Descartes, France*

To cite this article: Morange P-E, Saut N, Antoni G, Emmerich J, Trégouët D-A. Impact on venous thrombosis risk of newly discovered gene variants associated with FVIII and VWF plasma levels. *J Thromb Haemost* 2011; **9**: 229–31.

Through a genome-wide association study (GWAS), seven new single nucleotide polymorphisms (SNPs) were recently found to be associated with plasma levels of factor (F)VIII and von Willebrand Factor (VWF) by the CHARGE consortium [1]. These SNPs were located in five genes that have not previously been suspected to be associated with these protein levels. More specifically, VWF plasma levels were found to be influenced by SNPs located in *TC2N* (rs10133762), *CLEC4M* (rs868875), *SCARA5* (rs2726953), *STAB 2* (rs4981022) and *STXBP5* (rs9390459). The last three genes were also found to be associated with FVIII plasma levels, either through the same SNP (rs9390459 for *STXBP5*) or via different SNPs (rs9644133 for *SCARA5* and rs12229292 for *STAB 2*). Because high plasma levels of these two proteins are considered risk factors for venous thrombosis (VT) [2–4], we investigated the influence of these SNPs on VT risk using the results of our previously published GWAS on VT (*in silico* GWAS) and two candidate gene case–control studies, MARTHA and FARIVE, all three based on independent French populations [5].

Two of these seven SNPs, rs9390459 and rs9644133, were available in the DNA chip array used in our previous GWAS, but none of them showed any trend of an association with VT risk (Table 1). The five other SNPs identified in CHARGE were not typed in our GWAS but according to the HapMap database, three of them, rs12229292, rs2726953 and rs10133762, were in strong linkage disequilibrium with SNPs available in our GWAS. Their best proxies were rs4981021 (pairwise $r^2 = 0.879$), rs4276643 ($r^2 = 0.884$) and rs1884841 ($r^2 = 0.961$), respectively (Table 1). Among those, only the rs1884841 showed a promising significant association with VT (Table 1) as the rs1884841-T allele was more frequent in cases than in controls (0.498 vs. 0.432, $P = 6.45 \times 10^{-4}$). The latest two hit SNPs identified in CHARGE, rs4981022 and rs868875, were not in strong LD with any of the GWAS typed SNPs, with best proxies being rs608773 ($r^2 = 0.124$) and rs4964629 ($r^2 = 0.266$), respectively.

As a consequence, we decided to genotype the rs1884841 in the MARTHA study composed of 1150 VT patients and 801 healthy individuals [5] for further support of association with VT as well as both SNPs that did not have good proxies in our GWAS, rs4981022 and rs868875. While the rs4981022 did not show any evidence of an association with VT risk in MARTHA, the rs868875-G allele was found to be less frequent in cases than in controls (0.292 vs. 0.338, $P = 0.0025$) and the rs1884841-T more frequent in cases than in controls (0.487 vs. 0.437, $P = 0.0024$) (Table 1). Consequently, these last two SNPs were further examined for support of an association with VT in a sample of 594 VT patients and 588 controls part of the FARIVE study [5]. We were not able to confirm the association of the rs868875 with VT risk as, in FARIVE, its G allele was slightly more frequent in cases than controls, a result that was in the opposite direction from that observed in MARTHA. However, even although the association did not reach significance, which is likely as a result of underpowered sample size, the rs1884841-T allele was found to be more frequent in FARIVE cases than in FARIVE controls (0.472 vs. 0.453) as observed in MARTHA. We then used logistic regression analyzes to assess the association of the rs1884841 with VT risk after adjusting for age, gender, ABO blood group, FV and FII Leiden mutations, separately in MARTHA and FARIVE. The Mantel–Haenszel method was then used to test for the homogeneity across studies. Compared with the most frequent rs1884841-CC genotype, the CT genotype was associated with an adjusted odds ratio (ORs) [95% confidence interval (CI)] for VT of 1.17 [0.92–1.50] and 1.27 [0.94–1.72] in MARTHA and FARIVE, respectively. These two ORs were not statistically different ($P = 0.696$) and led to a combined OR of 1.21 [1.00–1.47] ($P = 0.049$). Similarly, as the adjusted ORs for the TT genotype were not significantly different ($P = 0.387$) across MARTHA (1.51 [1.12–2.03]) and FARIVE (1.22 [0.84–1.78]), the combined estimate was 1.39 [1.10–1.76] ($P = 0.005$). The ORs associated with the CT and TT genotypes being of similar amplitude and not significantly different from each other ($P = 0.360$), these results showed that carrying the rs1884841-T allele, either at the heterozygous or homozygous state (i.e. under a dominant model), was associated with an adjusted

Correspondence: Professor Pierre-Emmanuel Morange, Laboratory of Haematology, CHU Timone, 246, rue Saint-Pierre, 13385 Marseille Cedex 05, France.
Tel.: +33 4 91 38 60 49; fax: +33 4 91 94 23 32.
E-mail: pierre.morange@ap-hm.fr

DOI: 10.1111/j.1538-7836.2010.04082.x

Received 9 August 2010, accepted 23 September 2010

Table 1 Association of factor (F)VIII/von Willebrand factor (VWF) associated polymorphisms with venous thrombosis (VT)

				In silico GWAS		MARTHA		FARIVE	
Gene	SNP	Allele	CHARGE association*	Controls (n = 1228)	Cases (n = 419)	Controls (n = 801)	Cases (n = 1150)	Controls (n = 588)	Cases (n = 594)
CLEC4M	rs868875	A/G	↓ vWF	No proxy available		0.338	0.292	0.301	0.315
						OR = 0.81; P = 0.0025		OR = 1.07; P = 0.434	
SCARA5	rs9644133	C/T	↓ FVIII	0.178	0.195	Not investigated		Not investigated	
				OR = 1.12; P = 0.272					
	rs4276643†	T/C	↑ vWF	0.302	0.297	Not investigated		Not investigated	
				OR = 0.98; P = 0.800					
STAB 2	rs4981021‡	C/T	↑ FVIII	0.284	0.304	Not investigated		Not investigated	
				OR = 1.10; P = 0 .251					
	rs4981022	T/C	↓ vWF	No proxy available		0.315	0.301	Not investigated	
						OR = 0.92; P = 0.374			
STXBP5	rs9390459	G/A	↓ vWF	0.435	0.418	Not investigated		Not investigated	
				OR = 0.93; P = 0.372					
TC2N	rs1884841§	C/T	↑ vWF	0.432	0.498	0.437	0.487	0.453	0.472
				OR = 1.30; $P = 6.45 \times 10^{-4}$		OR = 1.22; P = 0.0024		OR = 1.08; P = 0.341	

Frequencies of the underlined minor alleles are shown in this table, separately in controls and cases, with the corresponding allelic odds ratios (OR). Reported P-values were derived from the Cochran-Armitage trend test. All genotyped single nucleotide polymorphisms (SNPs) followed Hardy–Weinberg equilibrium in each study, separately in cases and controls.

*Direction of association of the underlined minor alleles observed (or expected in case of proxy) in the CHARGE GWAS on FVIII/VWF levels. †rs4276643 serves as a proxy for the rs2726953 (r^2 = 0.884). ‡rs4981021 serves as a proxy for the rs12229292 (r^2 = 0.879). §rs1884841 serves as a proxy for the rs10133762 (r^2 = 0.961).

increased risk for VT of 1.28 [1.11–1.48] ($P = 9.9 \times 10^{-4}$) in the combined MARTHA and FARIVE studies. No heterogeneity was observed according to age of onset, gender, smoking, ABO blood group nor FV/FII Leiden mutation (data not shown).

Using a multi-stage strategy, we identified one SNP, rs1884841, which was associated with VT risk in three independent French case–control samples. This SNP was studied because it is in nearly complete association with the rs10133762 found modulating plasma levels of VWF in the CHARGE consortium [1]. Interestingly, according to the HapMap database, the rs10133762-G allele found to be associated with increased VWF levels [1] corresponds to the rs1884841-T allele associated with increased risk of VT, this observation being consistent with the well-known association between VWF levels and VT risk. It is worth noting that the rs10133762-G allele was associated with increased VWF plasma levels but no association was reported with FVIII [1]. In the *TC2N* gene where these two SNPs map has three different isoforms, one of them has an additional exon in its 5′ sequence. While the rs10133762 is intronic whatever the isoform, the rs1884841 lies either in the first intron or in the promoter region according to the referred isoform. One could feel inclined to claim that the effect of rs1884841-T (or of any other SNP in strong linkage disequilibrium with it) is mediated through an effect on VWF levels. Unfortunately, we were not able to definitively assess this (likely) hypothesis as plasma levels of VWF were not available in MARTHA nor FARIVE.

Other SNPs were assessed for their association with VT in this report because they were also found to influence plasma levels of FVIII and/or VWF, most of them having an even stronger effect on VWF than the *TC2N* rs10133762 (see Table 2 of [1]). However, no strong evidence for association was obtained for any of them. This does not mean that the genes they are lying in are not susceptibility genes to VT. This is especially true for *STAB 2* and *STXBP5* where the direction of associations with VT of the studied SNPs (rs4981021, rs4981022 and rs9390459) was consistent with the direction of their associations with FVIII/vWF observed in CHARGE [1]. However, the moderate sample size of the *in silico* GWAS we used at the first step of our multi-stage analysis may have limited the power to select SNPs for further genotyping in MARTHA. Based on the allele frequencies and allelic odds ratio observed in the *in silico* GWAS, the power of the MARTHA study would have anyway been very low to detect the association with VT, if any, of rs9644133 (28%), rs4276643 (6%), rs4981021 (27%) and rs9390459 (18%). As these power values only increase up to 42%, 7%, 41% and 26%, respectively, with MARTHA and FARIVE combined, this illustrates the need for collecting larger studies in order to support the relationship of these SNPs to VT risk. In addition, we could not exclude the possibility of heterogeneity across MARTHA and FARIVE that would have prevented us from confirming in FARIVE the association in MARTHA at the *CLEC4M* rs868875. MARTHA cases are patients referred to thrombophilia centers generally younger and more often smokers than FARIVE cases that were recruited from the general population [5]. In addition, MARTHA population was enriched with Factor V Leiden or FII G20210A mutations, 50% in cases and 41% in controls, compared to 18% and 8% in FARIVE, respectively. However, stratified analysis of this SNP according to age of onset, gender, smoking, ABO blood

group or FV/FII mutations did not reveal such possible heterogeneity (data not shown).

In conclusion, we observed strong arguments in favor of the *TC2N* gene (also referred to as *MTAC2D1*) as a new candidate gene for VT whose functional variant(s) remains to be identified and biological mechanisms relating it to VT physiopathology to be elucidated.

Acknowledgements

This work was supported by a grant from the Assistance Publique des Hopitaux de Marseille (AORC 2009).

Disclosure of Conflict of Interests

The authors state that they have no conflict of interest.

References

1 Smith NL, Chen MH, Dehghan A, Strachan DP, Basu S, Soranzo N, Hayward C, Rudan I, Sabater-Lleal M, Bis JC, de Maat MP, Rumley A, Kong X, Yang Q, Williams FM, Vitart V, Campbell H, Malarstig A, Wiggins KL, van Duijn CM *et al*. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation* 2010; **121**: 1382–92.

2 Kraaijenhagen RA, in't Anker PS, Koopman MM, Reitsma PH, Prins MH, van den Ende A, Buller HR. High plasma concentration of factor VIIIc is a major risk factor for venous thromboembolism. *Thromb Haemost* 2000; **83**: 5–9.

3 Koster T, Blann AD, Briet E, Vandenbroucke JP, Rosendaal FR. Role of clotting factor VIII in effect of von Willebrand factor on occurrence of deep vein thrombosis. *Lancet* 1995; **345**: 152–5.

4 Tsai AW, Cushman M, Rosamond WD, Heckbert SR, Tracy RP, Aleksic N, Folsom AR. Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE). *Am J Med* 2002; **113**: 636–42.

5 Tregouet DA, Heath S, Saut N, Biron-Andreani C, Schved JF, Pernod G, Galan P, Drouet L, Zelenika D, Juhan-Vague I, Alessi MC, Tiret L, Lathrop M, Emmerich J, Morange PE. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood* 2009; **113**: 5298–303.

Effect of provision of the NHS NPSA oral anticoagulant therapy patient information pack upon patients' knowledge and anticoagulant control

L. FAIRBAIRN-SMITH,* W. COPE,† B. ROBINSON,† F. KAMALI* and H. WYNNE†

**Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne; and †Newcastle upon Tyne Hospitals NHS Foundation Trust, Freeman Hospital, Newcastle upon Tyne, UK*

To cite this article: Fairbairn-Smith L, Cope W, Robinson B, Kamali F, Wynne H. Effect of provision of the NHS NPSA oral anticoagulant therapy patient information pack upon patients' knowledge and anticoagulant control. *J Thromb Haemost* 2011; **9**: 231–3.

Good control of anticoagulation within the target prothrombin time range lowers patients' risk of thromboembolic events and major bleeding [1]. Insufficient knowledge about oral anticoagulant therapy is a major contributor to bleeding risk in older patients [2]. Providing written information about warfarin therapy has been shown to improve stability of anticoagulant control [3], but is inadequately delivered [4]. In March 2007, the NHS National Patient Safety Agency (NPSA) [5] issued a patient safety alert setting out actions that can make anticoagulant therapy safer, recommending that patients should receive appropriate verbal and written information both at the start and throughout their course of therapy. It introduced a patient information booklet to contribute to this. The object of the present study was to investigate the effect of introducing this booklet upon knowledge about therapy and stability of control. Ethical approval was obtained from the Newcastle and North Tyneside Research Ethics Committee.

Thirty-five consecutive patients attending an anticoagulant monitoring clinic of Newcastle upon Tyne Hospitals NHS Foundation Trust for at least 7 months were approached and asked to participate in assessing the acceptability and the effect upon their knowledge of the NPSA oral anticoagulant therapy information. Twenty-four consenting participants completed a validated questionnaire [6] asking about warfarin, at baseline

Correspondence: Hilary Wynne, Newcastle upon Tyne Hospitals NHS Foundation Trust, Freeman Hospital, Newcastle upon Tyne NE7 7DN, UK.
Tel.: +44 191 2231683; fax: +44 191 2231249.
E-mail: hilary.wynne@nuth.nhs.uk

DOI: 10.1111/j.1538-7836.2010.04079.x

Received 8 September 2010, accepted 17 September 2010

OPEN ACCESS Freely available online

PLoS one

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

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

1 INSERM UMR_S 937; ICAN Institute, Université Pierre et Marie Curie, Paris 6; Paris, France, **2** INSERM, UMR_S 626, Marseille, France; Université de la Méditerranée, Marseille, France, **3** INSERM UMR_S 915; CNRS ERL3147; Institut du Thorax; Nantes, France, **4** INSERM U744, Lille, France; Institut Pasteur de Lille, Lille, France Université de Lille Nord de France, Lille, France, **5** Commissariat à l'Energie Atomique, Institut de Génomique, Centre National de Génotypage, Evry, France, **6** INSERM UMR_S 708, Université Pierre et Marie Curie (UPMC, Paris 6), Paris, France, **7** INSERM U888, Hôpital La Colombière, Montpellier, France, **8** INSERM, U897, Bordeaux, France; Université Victor Segalen, Bordeaux, France, **9** INSERM U765, médecine vasculaire - HTA, hôpital européen Georges-Pompidou, Université Paris-Descartes, Paris, France, **10** CHRU de Lille, Lille, France

Abstract

Background: Venous Thrombosis (VT) is a common multifactorial disease associated with a major public health burden. Genetics factors are known to contribute to the susceptibility of the disease but how many genes are involved and their contribution to VT risk still remain obscure. We aimed to identify genetic variants associated with VT risk.

Methodology/Principal Findings: We conducted a genome-wide association study (GWAS) based on 551,141 SNPs genotyped in 1,542 cases and 1,110 controls. Twelve SNPs reached the genome-wide significance level of 2.0×10^{-8} and encompassed four known VT-associated loci, *ABO*, *F5*, *F11* and *FGG*. By means of haplotype analyses, we also provided novel arguments in favor of a role of *HIVEP1*, *PROCR* and *STAB2*, three loci recently hypothesized to participate in the susceptibility to VT. However, no novel VT-associated loci came out of our GWAS. Using a recently proposed statistical methodology, we also showed that common variants could explain about 35% of the genetic variance underlying VT susceptibility among which 3% could be attributable to the main identified VT loci. This analysis additionally suggested that the common variants left to be identified are not uniformly distributed across the genome and that chromosome 20, itself, could contribute to ~7% of the total genetic variance.

Conclusions/Significance: This study might also provide a valuable source of information to expand our understanding of biological mechanisms regulating quantitative biomarkers for VT.

Citation: Germain M, Saut N, Greliche N, Dina C, Lambert J-C, et al. (2011) Genetics of Venous Thrombosis: Insights from a New Genome Wide Association Study. PLoS ONE 6(9): e25581. doi:10.1371/journal.pone.0025581

Editor: Heribert Schunkert, Universitätsklinikum Schleswig-Holstein - Campus Luebeck, Germany

Received June 7, 2011; **Accepted** September 6, 2011; **Published** September 27, 2011

Funding: M.G. was supported by grants funded by the Agence Nationale pour la Recherche (Project ANR-07-MRAR-021) and the Program Hospitalier de recherche Clinique (PHRC2009 RENOVA-TV). T.O.M. was supported by a grant from the Fondation pour la Recherche Médicale. The MARTHA project was supported by a grant from the Program Hospitalier de Recherche Clinique and the FARIVE study by grants from the Fondation pour la Recherche Médicale, the Program Hospitalier de recherche Clinique (PHRC 20002; PHRC2009 RENOVA-TV), the Fondation de France, and the Leducq Foundation. Statistical analyses benefit from the C2BIG computing centre funded by the Fondation pour la Recherche Médicale and La Région Ile de France. The 3C Study is conducted under a partnership agreement between Inserm, the Victor Segalen –Bordeaux II University and Sanofi-Synthélabo. The Fondation pour la Recherche Médicale funded the preparation and first phase of the study. The 3C-Study is also supported by the Caisse Nationale Maladie des Travailleurs Salariés, Direction Générale de la Santé, Mutuelle Generale de l'Education Nationale, the Institut de la Longévité, Agence Française de Sécurité Sanitaire des Produits de Santé, the Regional Governments of Aquitaine, Bourgogne and Languedoc-Roussillon and, the Fondation de France, the Ministry of Research-Inserm Programme 'Cohorts and collection of biological material'. The Lille Génopôle received an unconditional grant from Eisai. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have read the journal's policy and have the following conflicts. Funding was received from Sanofi-Synthélabo. The Lille Génopôle received an unconditional grant from Eisai. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: david.tregouet@upmc.fr (D-AT); pierre.morange@ap-hm.fr (P-EM)

Introduction

Venous thrombosis (VT) is a common multifactorial disease affecting two individuals out of one thousand a year and associated with a mortality rate of 10% [1]. Recurrence risk of VT is about 6% a year, and post-thrombotic disease occurs within the next 5 years following a VT event in about 25% of patients [2]. It has been reported that 25,000 individuals die from the consequences of VT each year in England [3] and that the disease has a substantial economic costs [4,5]. Despite these striking elements, venous thrombosis can be considered as the Cinderella of genetic research on thrombotic disorders compared to arterial and

cerebral thrombosis. Even though genetic factors are estimated to explain up to 60% of the VT heritability [6,7], VT genetics has not benefit a lot from the genome wide association study (GWAS) revolution. While several GWAS and meta-analysis of GWAS have been conducted for arterial and cerebral thrombosis on thousands of individuals [8–14], only one GWAS on VT has been reported so far [15], and on a rather small sample of 419 cases and 1,228 controls. Before this GWAS was carried out, well-established susceptibility genes for VT were *SERPINC1*, *PROC*, *PROS1*, *FII*, *FGG*, *FV* and *ABO* [16]. The latter two loci were the only genomic regions that reached genome-wide statistical significance in the VT GWAS. Nevertheless, using additional strategies to assess the most promising associations generated by this GWAS, other VT-associated loci were robustly identified, *HIVEP1* [17], *C4BPA* [18], and *TC2N* [19]. Two additional VT-associated loci, *GP6* and *F11* [20], were also robustly identified through another large-scale association study, focusing mainly on non-synonymous polymorphisms.

In our quest to identify novel susceptibility genes for VT beyond those already known (Figure 1), we report the results of a second GWAS based on a larger sample size (1,542 cases and 1,110 controls) and exploring a larger number of single nucleotide polymorphisms (SNPs) (551,141 vs 317,139 in the previous GWAS [15]). The overall sequential procedure of this work was summarized in Figure 2. A standard GWAS comparing VT patients participating in the MARTHA project [21] to healthy individuals from the Three-City Study (3C) [22] was first performed to identify genome-wide significant associations of SNPs with VT risk (stage I). Second, results from this GWAS were combined to those of our previously published GWAS on VT (referred to as "in silico GWAS" in the rest of the document) [15] to detect novel associations that would not have been declared significant at stage I. At this stage (stage II), both raw and imputed genotyped data analyses were carried out. In addition to this standard GWAS, we performed a candidate gene association analysis using less stringent statistical thresholds. SNPs demonstrating suggestive evidence of association with the disease were then planned to be further tested for replication in independent case-controls studies. A new estimate of the genetic variance associated with VT susceptibility was also derived using a novel methodology for GWAS data.

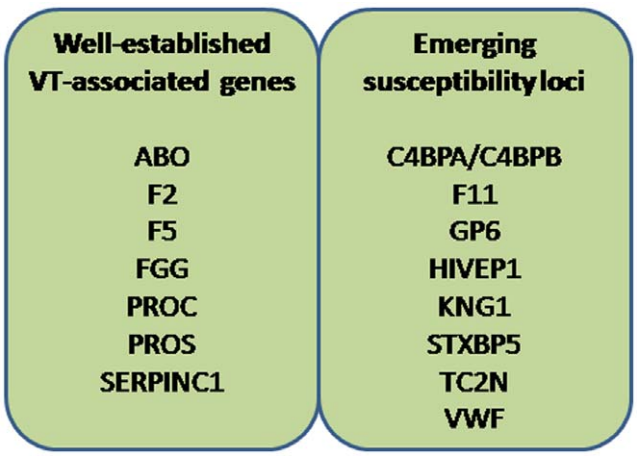


Figure 1. Known candidate loci for VT.
doi:10.1371/journal.pone.0025581.g001

Results

Genome Wide Association analysis

Stage I. A quantile-quantile (Q-Q) plot representation of the whole set of association results was compatible with what was expected under the assumption of no genetic association (Figure 3) and the corresponding genomic control (GC) value was 1.04. Among the 491,258 tested SNPs at this stage, twelve were significant at the fixed genome-wide threshold of $p<2.0\times10^{-8}$ (Table 1). These SNPs were located within *ABO*, *FGG*, *F5*, and *F11* loci, four well-established VT-associated genes. The *F5* and *ABO* hit SNPs included those already identified in our previous GWAS [15] while the *FGG* VT-associated SNP was the rs2066865, located in the 3′UTR region of the gene and known to influence both fibrinogen γ′ levels and VT risk [15,23,24]. The *F11* hit SNP was the rs10029715. According to the SNAP software [25] based on HapMap 3 (release 2), this SNP is in modest linkage disequilibrium (LD) with two *F11* SNPs recently found to independently affect VT risk [20,26], rs2036914 and rs2289252 ($r^2=0.094$, $D'=+0.66$, and $r^2=0.076$; $D'=+0.75$, respectively). However, these two SNPs were not available in our genotyping array. Several other *F11* SNPs showed suggestive statistical associations with VT at $p<10^{-4}$ (Table S1) and their haplotype analysis showed that the *F11* association signal was driven by two common yin/yang haplotypes (Table 2). We then used the HapMap data to infer the *F11* haplotypic structure derived from the two rs2036914 and rs2289252 SNPs and those found associated with VT in our sample. It is interesting to note that the yin-yang pattern described above is still present when the rs2036914 and rs2289252 are included in the analysis (Table S2).

Stage II. The results of the new GWAS were combined to those obtained on the previous *in silico* GWAS [15] through a meta-analysis totaling 1,961 VT cases and 2,238 controls. Using either the 253,355 genotyped SNPs common to both GWAS studies or 2,475,305 observed or imputed SNPs, no novel association was detected (Figure 4). In the imputation analysis, only 99 SNPs reached genome-wide significance and they were all located within the *ABO*, *F5*, *FGG*, or *F11* loci (Table S3). Detailed regional association plots for these four loci are shown in Figure 5.

Candidate Gene association analysis

We then further explored the association results obtained in the discovery GWAS by focusing on SNPs located within candidate genes as it is now well-admitted that genuine association may be hidden in the heap of non genome-wide significant associations. Forty-nine genes were selected as candidates because they were either known to participate to the coagulation/fibrinolysis cascade, were already shown to be associated with VT , or had recently been identified through GWAS as modulating the variability of quantitative traits known/hypothesized to be associated with VT risk (Table S1).

In addition to the genome-wide significant loci discussed above, four candidate loci were found to harbor SNPs showing suggesting evidence of association with VT at $p<10^{-3}$ (Table S1). These SNPs were rs169715 and rs2228220 in *HIVEP1*, rs6060278 and rs6088735 in *PROCR*, rs4981021 in *STAB2* and rs8074026 in *SERPINF2*.

The two *HIVEP1* SNPs, rs169715 and rs2228220, were in modest LD with each other ($r^2=0.09$, $D'=+0.45$). Their haplotype analysis suggested that the observed effects would be additive (Table 3), the rs169715-G allele being associated with an adjusted OR of 1.57 [1.17–2.09] ($p=2.60\ 10^{-3}$) and the rs2228220-G allele with an OR of 1.35 [1.10–1.66] ($p=3.98\ 10^{-3}$). In addition, this haplotype analysis suggested that the effect

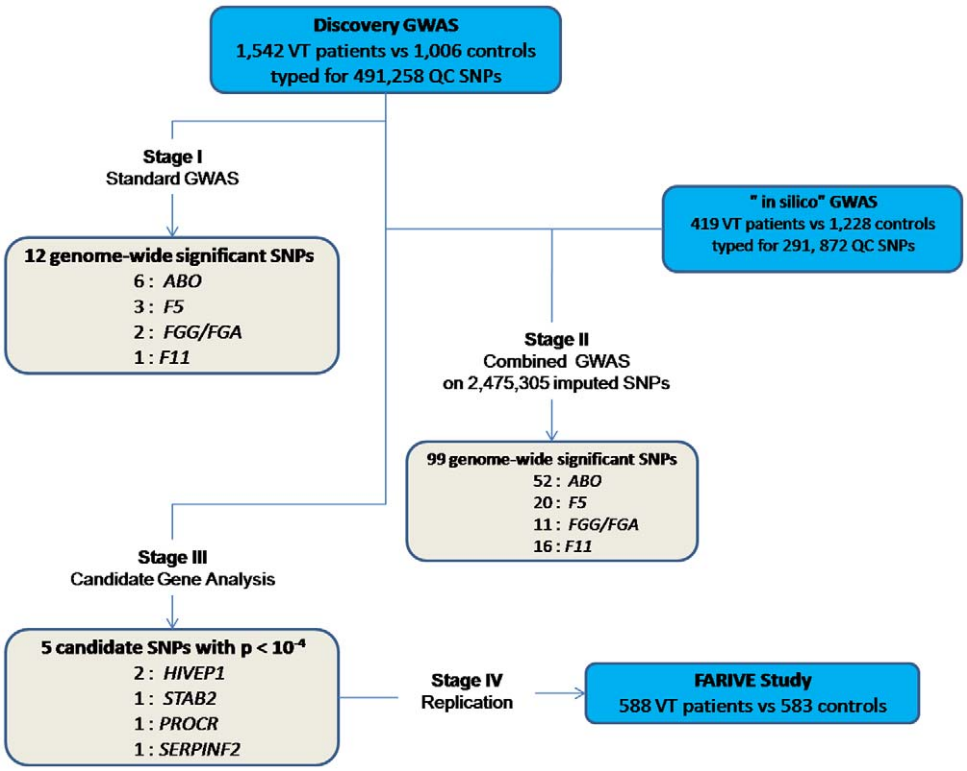


Figure 2. Main outlines of the adopted sequential GWAS strategy.
doi:10.1371/journal.pone.0025581.g002

on VT risk of the previously identified *HIVEP1* rs169713 [17] was due to its LD with the rs169713 ($r^2 = 0.01$, $D' = 1$) and rs2228220 ($r^2 = 0.06$, $D' = -0.50$) *HIVEP1* hit SNPs (Table 3).

The two *PROCR* SNPs, rs6060278 and rs6088735, were in complete association ($r^2 = 1$) and were also in strong negative LD ($r^2 = 0.03$, $D' = -1$) with the rs867186 variant, also known as Ser219Gly. The role of the latter in VT risk is still a matter of debate [15] and its association with VT was borderline in our discovery GWAS (Table S1). Nevertheless, a haplotype analysis of the rs6088735 and rs867186 suggested that both SNPs could act additively on the risk of disease (Table 4). When adjusted for rs6088735, the rs867186-G allele was associated with an OR for VT of 1.33 [1.11–1.60] ($p = 2.34\ 10^{-3}$) whereas , adjusted for rs867816, the rs6088735-T allele was associated with an OR of 1.35 [1.19–1.54] ($p = 5.47\ 10^{-6}$) (Table 4). Interestingly, according to the SNAP database [25], the rs6088735 is in complete association ($r^2 = 1$) with the *EDEM2* rs6120849 that was recently found associated with protein C levels in the ARIC study [27]. The rs6088735-T allele associated with increased risk of VT corresponds to the rs6120849-T allele that was associated with decreased protein C levels, an observation consistent with the known association of decreased PROC levels and VT risk [28].

The *STAB2* rs4981021-T allele was associated with increased risk of VT 1.29 [1.14–1.46] ($p = 3.17\ 10^{-4}$) but this association did not reach significance (OR = 1.10, p = 0.251) in the "in silico GWAS". Conversely, the most significant *STAB2* SNP in the latter study was the rs1593812 already discussed in [29]. Haplotype analysis of these two SNPs suggested that they define at least one common at-risk haplotype for VT in both GWAS datasets (OR = 2.18 [1.66–2.87] ($p = 2.16\ 10^{-8}$) (Table 5). Note that the rs4981021-T is a good proxy ($r^2 = 0.88$) for the rs12229292-T not typed in our GWAS array but recently found associated with increased Factor VIII levels [30].

The association observed at *SERPINF2* was novel, the rs8074026-T allele being more frequent in MARTHA cases than in 3C healthy controls (0.29 vs 0.24, $p = 6.87\ 10^{-4}$). This SNP was not typed in our previous GWAS in patients with early age of onset of VT but a similar trend was observed using imputation data (0.22 vs 0.19, p = 0.289). This SNP was therefore further explored in the FARIVE study, but the association was not confirmed as the rs8074026-T allele tended, conversely, to be less frequent in cases than in controls (0.23 vs 0.25, p = 0.520). This polymorphism was then not further studied.

Genetic variance analysis

The next step of our analysis consisted in getting an overall estimate of the genetic variance (h^2) of VT as well as an estimate of the contribution of the main susceptibility loci discussed above. The application of the GCTA software [31] to the discovery GWAS led to an estimate of 0.357 ± 0.049. This estimate was obtained using an assumed prevalence of 0.001 for the disease according to [1]. This estimate was strongly dependent on the value of the assumed prevalence (Table 6). When the GCTA analysis was applied to the first GWAS data set with an assumed prevalence of 0.001, the genetic variance estimate was 0.223 with a larger standard error, 0.108, an estimate that was nevertheless consistent with that observed in the discovery GWAS.

The relative contribution of each chromosome is summarized in Figure 6 for a prevalence of 0.001. While *F5* and *ABO* explained ~1%, each, of the genetic variance, and the *FGG* and *F11* together only 1.2%, we note that the chromosome most contributing to the estimate ($6.9\% \pm 1.2$) was chromosome 20. This observation hold whatever the assumed value for the prevalence of VT (Table 6). We then turned back to the original GWAS results and focused on chromosome 20 SNPs. Only eleven chromosome 20 SNPs

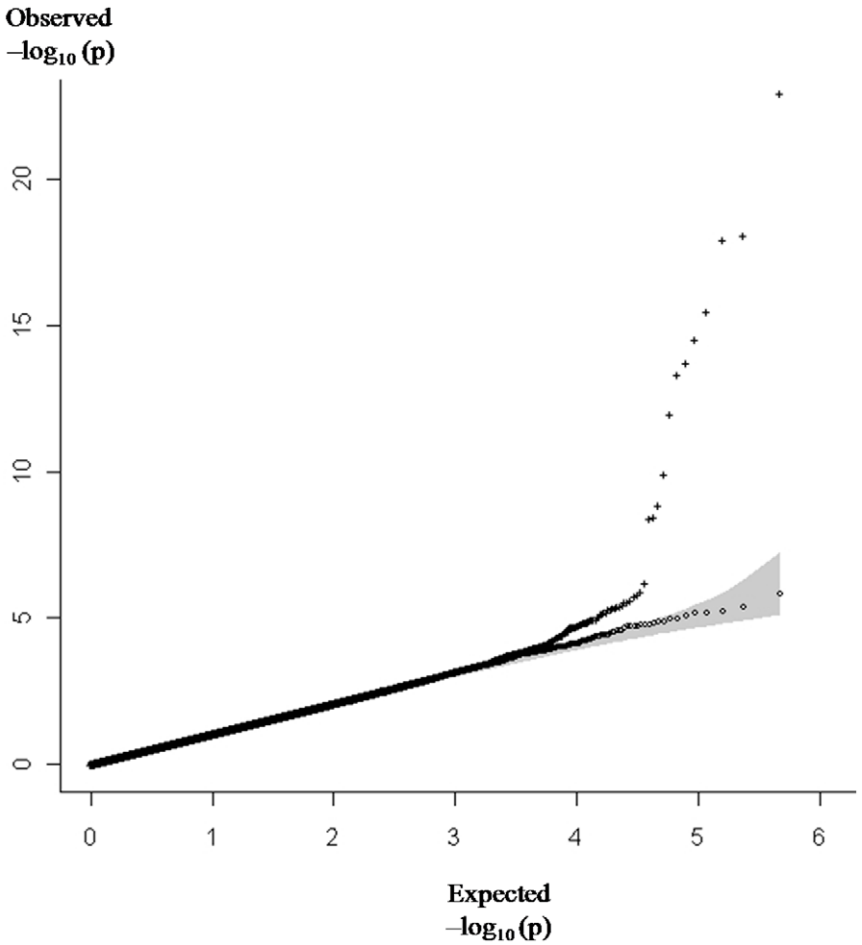


Figure 3. Quantile-Quantile plot representation of the GWAS results obtained from 491,258 studied SNPs. Q-Q plot derived from all SNP p-values is illustrated by + . The exclusion of 878 SNPs located within ±500 kb of the *ABO*, *F5*, *FGG* and *FXI* loci, the four main well-established VT-associated loci, lead to the Q-Q plot symbolized by ○ with its 95% confidence interval in shaded area.
doi:10.1371/journal.pone.0025581.g003

Table 1. Stage I - Minor allele frequencies distribution of SNPs demonstrating association with VT at $p_{EIGENSTRAT} < 2.0\ 10^{-8}$ in a GWAS sample of 1,542 VT cases and 1,110 controls.

CHR	Position	Gene	SNP	Alleles[(1)]	Cases	Controls	P[(2)]	P[(3)]
1	167401751	*NME7*	rs16861990	C/A	0.134	0.058	5.53×10^{-20}	2.75×10^{-15}
1	167695568	*SLC19A2*	rs1208134	C/T	0.133	0.056	4.06×10^{-21}	3.29×10^{-16}
1	167758179	*F5*	rs2420371	G/A	0.151	0.066	3.24×10^{-23}	8.44×10^{-19}
4	155744726	*FGG*	rs2066865	A/G	0.280	0.209	3.44×10^{-9}	1.17×10^{-10}
4	155720638	*FGA*	rs6825454	C/T	0.299	0.228	1.39×10^{-8}	1.32×10^{-9}
4	187459594	*F11*	rs10029715	C/T	0.115	0.172	1.09×10^{-9}	3.20×10^{-9}
9	135126961	*ABO*	rs2073828	A/G	0.321	0.406	1.91×10^{-10}	3.57×10^{-9}
9	135129086	*ABO*	rs657152	C/A	0.494	0.383	5.55×10^{-20}	1.10×10^{-18}
9	135138468	*ABO*	rs500498	T/C	0.332	0.432	3.54×10^{-14}	1.03×10^{-12}
9	135139050	*ABO*	rs505922	C/T	0.489	0.350	1.53×10^{-25}	1.06×10^{-23}
9	135139543	*ABO*	rs630014	A/G	0.381	0.485	9.26×10^{-15}	4.40×10^{-14}
9	135144688	*ABO*	rs495828	T/G	0.357	0.264	8.82×10^{-13}	1.78×10^{-14}

[(1)]Common/minor alleles.
[(2)]P-value of the Cochran-Armitage Trend test.
[(3)]Association test p-value corrected for principal components (EIGENSTRAT program).
doi:10.1371/journal.pone.0025581.t001

Table 2. Haplotype association analysis of *F11* hit SNPs with VT risk in a sample of 1,542 VT cases and 1,110 controls.

Polymorphisms				Haplotype Frequencies	
rs925451	rs10029715	rs1008728	rs13133050	Controls	Cases
				n = 1110	n = 1542
A	T	T	C	0.338	0.403
A	T	C	A	0.033	0.027
A	C	C	C	0.016	0.011
G	T	T	C	0.257	0.248
G	T	C	A	0.196	0.201
G	C	C	C	0.043	0.041
G	C	C	A	0.106	0.051

F11 haplotypes were more strongly associated with VT ($p = 1.05\ 10^{-12}$) than single SNP alone (best p-value = $1.09\ 10^{-9}$) and the association was likely due to two common haplotypes, ATTC and GCCA, differing at all studied sites ("yin-yang" haplotypes), the former being associated with increased risk of VT, the latter with decreased risk. Compared to the GTTC haplotype, the ATTC haplotype was associated with an increased risk of VT (OR = 1.218 [1.048–1.416], p = 0.0099 while the GCCA haplotype was associated with a decreased risk of the disease (OR = 0.493 [0.391–0.623], $p = 3.39\ 10^{-9}$).
doi:10.1371/journal.pone.0025581.t002

corresponding to four different loci *RSPO4*, *C20orf23/SNRPB2*, *MYLK2* and *PREX1*, showed association at $p<10^{-4}$ with VT (Table 7). We further investigated whether these four loci, in addition to the *PROCR* locus mentioned above that was also located on chromosome 20, could substantially contribute to explain the genetic variance contribution of chromosome 20. After discarding the genetic influence of these five loci, the remaining estimate associated with chromosome 20 SNPs was 5.5%±1.2. Further analyses also suggested that ~80% of the chromosome 20's contribution came from SNPs located in its 20p arm (Table 6).

Discussion

In this work we reported the results of a second GWAS on VT that, when added to the previous one, gathered a total sample of 1,961 cases and 2,338 controls, all of French origin. With such a sample size, our study had a power of 80% at the genome-wide significance level of $\sim 2\times10^{-8}$ to detect the allelic effect of any SNP associated with an OR of 1.40 provided that its minor allele frequency is greater than 0.20 [32]. These values perfectly matched those observed at the four loci (*F5*, *FGG*, *F11* and *ABO*) that reached genome-wide significant in this report (Table S3), these four loci being now well-established susceptibility genes to VT [15,16,26]. The contribution of *F5* and *ABO* in VT susceptibility has been extensively discussed and the functional role of *FGG* rs2066865 has already been established [3,23,24]. Conversely, the identification of the functional *F11* variant(s) hypothesized to be tagged by the ying/yang haplotype structure discussed above deserves additional work.

In order to achieve a 80% power for detecting ORs of magnitude 1.30, 1.25 and 1.20, the statistical stringency would have to be lessened to 3×10^{-5}, 6×10^{-4}, and 9×10^{-2}, respectively, at the risk of increasing false positive rates. In an attempt to increase the power of our analysis while limiting for false positives, we therefore focused on all SNPs located within ~fifty candidate genes and demonstrating suggestive statistical evidence ($p<10^{-4}$) for association with VT using raw genotype data. Suggestive associations were observed for the *HIVEP1*, *PROCR*, *STAB2* and *SERPINF2* loci (Table S1). Associations with VT have already been reported for the first three genes [15,17,29] and disentangling their exact genetic contribution to VT susceptibility would warrant additional extensive works. The latest suggestive association was observed for the *SERPINF2* rs8074026. *SERPINF2* is an obvious candidate for VT as it codes for a serpine protease inhibitor that acts as a inhibitor of plasmin. However, no trend for association was observed in the replication study.

Following the findings of a GWAS on aPTT levels [33], a candidate biomarker for VT, we have recently suggested that the

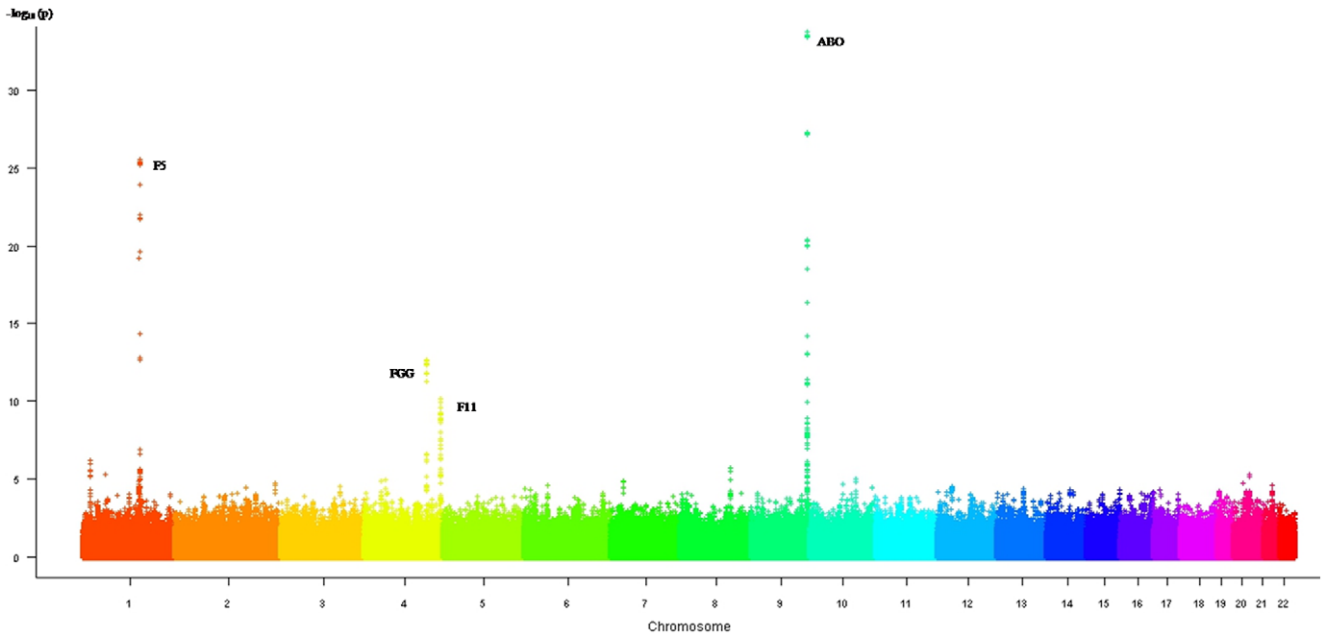


Figure 4. Manhattan plot of the association results from the combined analysis of two imputed GWAS data sets for 2,475,305 SNPs.
doi:10.1371/journal.pone.0025581.g004

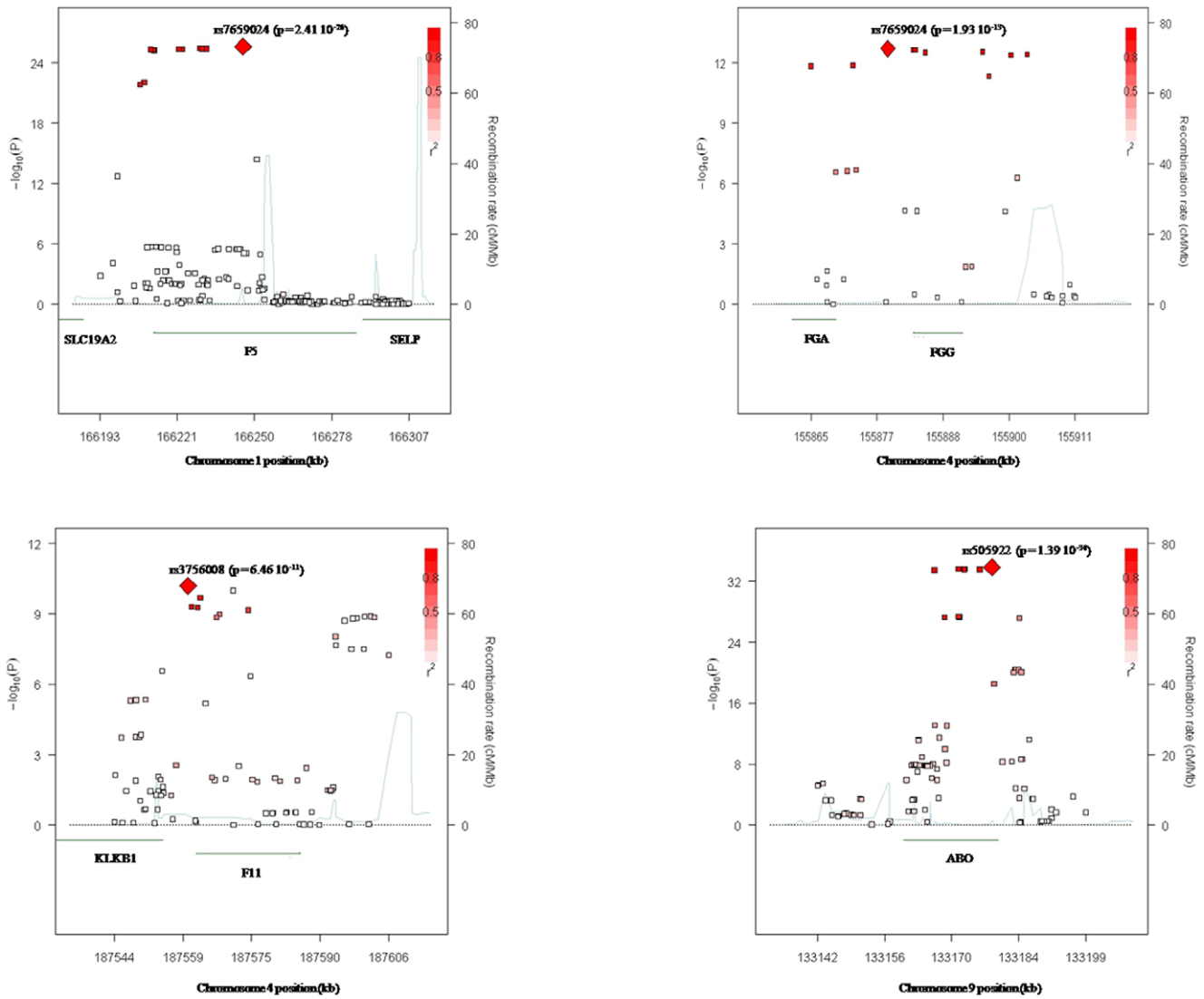


Figure 5. Regional association plots at the four genome-wide significant loci using imputed SNPs. Four genome-wide significant loci were *F5* (top left), *FGG* (top right), *F11* (bottom left) and *ABO* (bottom right). These plots were drawn from the SNAP software [25].
doi:10.1371/journal.pone.0025581.g005

KNG1 Ile581Thr variant (rs710446) could also be a risk factor for VT using data from our discovery GWAS and the FARIVE studies. This variant only reached a significance of $p = 1.17\ 10^{-3}$ in the discovery GWAS (Table S1) highlighting the need for exploring in more details the list of less significant p-values, in particular by use of external information on candidate quantitative risk factors. Two other SNPs have recently been suggested to influence VT-risk, *STXBP5* rs1039084 and *VWF* rs1063856 [34]. These were not available in the "in silico" GWAS, but using QC imputed data in the whole set of 1,961 cases and 2,338 controls, the rare allele of the *VWF* rs1063856 was marginally associated with the risk of VT (OR = 1.10 [1.00–1.21], p = 0.042) , an association consistent with that previously reported [34]. Conversely, we did not observe any trend of association for the *STXBP5* rs1039084 rare allele (OR = 0.97 [0.89–1.06]; p = 0.55), even if this OR was of similar amplitude with that observed in the MEGA study (OR = 0.91 [0.86–0.97]) [34]. These two associations were previously observed in a meta-analysis of studies gathering about 5,000 cases and 5,000 controls, underlying the low power of our study to detect modest genetic effect as already discussed above. Large GWAS samples gathering at least ~20,000 patients would be required in order to detect genome-wide significant ORs of ~1.10 and, for the moment, we are far from reaching such sample size by contrast to international consortia on coronary artery disease [35] . Another limitation of this work could be related to the selection of the GWAS subjects. Controls were part of a national GWAS sample of French healthy individuals that were not matched to VT cases, in particular for gender and sex. Nevertheless, all known or suspected VT-associated loci were identified in our work suggesting a rather modest influence of imperfect matching between cases and controls. Conversely, VT patients homozygous for the FV Leiden or FII 20210A mutation or with anti-thrombin, protein C or protein S deficiencies were not included in this work. It is very unlikely that the selection on FV Leiden homozygosity had affected our results as the *F5* gene is among the four loci that reached genome-wide significance in our study. Note that the FII 20210 mutation (rs1799963) was not available in the imputed reference datasets. However, one cannot exclude that the other exclusion criteria may have affect our power to identify novel VT-associated variants, in particular through a

Table 3. Haplotype analysis of *HIVEP1* haplotypes derived from rs169713, rs169715 and rs2228220 in a sample of 1,542 cases and 1,110 controls.

Polymorphisms			Haplotype Frequencies	
rs169713	rs169715	rs2228220	Controls	Cases
			n = 1,110	n = 1,542
C	A	A	0.198	0.206
C	A	G	0.042	0.054
T	A	A	0.698	0.643
T	A	G	0.026	0.037
T	G	A	0.018	0.027
T	G	G	0.017	0.031

HIVEP1 haplotypes were strongly associated with VT risk ($\chi^2 = 30.22$ with 5df, $p = 1.33\ 10^{-5}$).
All haplotypes carrying the rs169715-G or the rs2228220-G alleles tended to be more frequent in cases than in controls, suggesting that both alleles could act additively to influence VT risk. This hypothesis were then tested and was not rejected ($\chi^2 = 0.75$ with 2 df, $p = 0.686$). After adjusting for rs2228220, the OR associated with the rs169715-G allele was 1.57 [1.17–2.09] ($p = 2.60\ 10^{-3}$) and the OR associated with rs2228220-G adjusted for rs169715 was 1.35 [1.10–1.66] ($p = 3.98\ 10^{-3}$). After adjusting for these two SNPs, the rs169713-C allele was not significant (OR = 1.11 [0.97–1.26], p = 0.129).
doi:10.1371/journal.pone.0025581.t003

modulation of anti-thrombiin, protein C or protein S levels. It is nevertheless worthy of note that the *PROCR* locus that was found influencing the most protein C levels in the ARIC GWAS [27], was among the top 8 most significant VT-associated loci in our GWAS.

The second original aspect of our work is the application of a novel statistical methodology to get an estimate of the genetic variance of VT. This approach requires several assumptions including a fixed value for the disease prevalence, additive genetic effects and the existence of an underlying liability characterized by a threshold above which the disease status is called. Using the latest known estimate of the VT prevalence [1], we showed that the genetic variance could be ~35%, an estimate slightly lower than those obtained from families studies [6,7]. While the four main VT-linked loci, *FV*, *ABO*, *FGG* and *F11*, altogether contributed to about ~3% of the total genetic variance it was striking to observe that chromosome 20 was the chromosome contributing the most to the total genetic variance with about ~7% of the total genetic variance. Further analyses including chromosome-wide haplotype and homozygosity mapping analyses are ongoing to further investigate the chromosome 20 genetic architecture in relation to VT risk.

Table 4. Haplotype analysis of *PROCR* haplotypes derived from rs6088735 and rs867186 in a sample of 1,542 cases and 1,110 controls.

Polymorphisms		Haplotype Frequencies	
rs6088735	rs867186	Controls	Cases
		n = 1,110	n = 1,542
C	A	0.673	0.603
C	G	0.095	0.115
T	A	0.232	0.282

PROCR haplotypes were strongly associated with VT risk ($\chi^2 = 26.51$ with 2 df, $p = 1.75\ 10^{-6}$). Compared to the most frequent CA haplotype, the CG and TA haplotypes were associated with an increased OR for VT of 1.33 [1.11–1.60] ($p = 2.34\ 10^{-3}$) and 1.35 [1.19–1.54] ($p = 5.47\ 10^{-6}$).
doi:10.1371/journal.pone.0025581.t004

Table 5. Haplotype analysis of *STAB2* haplotypes derived from rs1593812 and rs4981021 in two GWAS data sets.

Polymorphisms		Haplotype Frequencies			
		"in silico GWAS"		Discovery GWAS	
rs1593812	rs4981021	Controls	Cases	Control	Cases
		n = 1,228	n = 419	n = 1,110	n = 1,542
A	C	0.629	0.573	0.646	0.587
A	T	0.248	0.240	0.219	0.246
G	C	0.091	0.114	0.104	0.111
G	T	0.032	0.073	0.031	0.056

Compared to the most frequent AC haplotype, the GT haplotype was associated with an increased risk of 2.43 [1.60–3.71] ($p = 3.0\ 10^{-5}$) and 2.01 [1.40–2.88] ($p = 1.48\ 10^{-4}$) in the "In silico" and discovery GWAS respectively. The combined Mantel-Haenszel OR associated with the GT haplotype compared to the AC haplotype was then 2.18 [1.66–2.87] ($p = 2.16\ 10^{-8}$).
doi:10.1371/journal.pone.0025581.t005

In conclusion, this work provided new information about the genetic susceptibility to VT and strongly suggested that chromosome 20 genes warrant specific attentions. It generated a wealth of valuable genetic information to those showing interest in disentangling the genetic architecture of VT.

Materials and Methods

Ethics Statement

Each individual study was approved by its institutional ethics committee and informed written consent was obtained in accordance with the Declaration of Helsinki. All subjects were of European origin. All subjects were of European origin.

Ethics approval were obtained :

- for MARTHA, from the "Departement santé de la direction générale de la recherche et de l'innovation du ministère" (Projects DC: 2008-880 & 09.576).
- for FARIVE, from the "Comité consultatif de protection des personnes dans la recherche biomedicale" (Project n° 2002-034)
- for the 3C study, from the institutional ethics committees of the Kremlin-Bicetre Hospital.

Studies

Stage I - Discovery GWAS. MARTHA patients (n = 1,592) are unrelated VT patients, mainly of French origin, consecutively recruited at the Thrombophilia center of La Timone hospital (Marseille, France) between January 1994 and October 2005. All patients had a history of a first VT event documented by venography, Doppler ultrasound, angiography and/or ventilation/perfusion lung scan. They were all free of any chronic conditions and free of any well characterized genetic risk factors including anti-thrombin, protein C or protein S deficiency, homozygosity for FV Leiden or FII 20210A, and lupus anticoagulant. A more detailed description of these patients can be found in [21]. These VT patients were compared to healthy individuals from the 3C study.

Table 6. Relative contribution of each chromosome on the total genetic variance of VT according to the assumed prevalence of the disease.

	prevalence			
	0.001	**0.005**	**0.01**	**0.05**
Total Genetic	**0.357**	**0.480**	**0.561**	**0.860**
Variance ± SE	**±0.047**	**±0.067**	**±0.078**	**±0.120**
chromosome				
1	0.016	0.033	0.038	0.036
2	0.015	0.020	0.023	0.036
3	0.019	0.026	0.031	0.045
4	0.041	0.052	0.060	0.094
5	0.013	0.017	0.020	0.029
6	0.006	0.012	0.014	0.016
7	0.011	0.012	0.014	0.027
8	0.004	0.021	0.025	0.010
9	0.039	0.049	0.058	0.089
10	0.040	0.042	0.049	0.094
11	0.007	0.000	0.002	0.013
12	0.020	0.018	0.022	0.047
13	0.008	0.009	0.010	0.019
14	0.002	0.017	0.019	0.003
15	0.017	0.021	0.024	0.039
16	0.007	0.010	0.012	0.016
17	0.006	0.004	0.005	0.015
18	0.008	0.008	0.010	0.017
19	0.001	0.017	0.020	0.037
20	0.069[a]	0.081	0.094	0.159
21	0.008	0.009	0.011	0.019
22	0.000	0.000	0.000	0.000

Estimates were obtained from the discovery GWAS data and adjusted for gender and principal components.

[a]When chromosome 20 SNP data were split into two parts, one including 6,769 SNPs on the shortest 20p arm and the other 7,170 SNPs on the longest 20q arm, their relative contribution on the genetic variance were 0.056±0.013 and 0.013±0.007.

doi:10.1371/journal.pone.0025581.t006

The 3C Study is a population-based, prospective (4-years follow-up) study, initially set-up to investigate the relationship between vascular factors and dementia. It has been carried out in three French cities: Bordeaux (southwest France), Montpellier (southeast France) and Dijon (central eastern France). A sample of non-institutionalised subjects aged over 65 was randomly selected from the electoral rolls of each city. Between January 1999 and March 2001, 9,686 subjects meeting the inclusion criteria agreed to participate. Following recruitment, 392 subjects withdrew from the study. Thus, 9,294 subjects were finally included in the study (2,104 in Bordeaux, 4,931 in Dijon and 2,259 in Montpellier). At the baseline clinical examination, blood samples were obtained from 8,707 individuals. For the present study, a random sample of 1,140 subjects free of any chronic diseases was selected to serve as controls.

Stage II - *In silico* GWAS study. In a previously published GWAS on VT [15], 419 early age of onset (<50 years) VT cases were compared to 1,228 healthy controls at 291,872 SNPs. Cases were patients from four different French medical centers (Grenoble, Marseille, Montpellier, Paris) selected according to the same criteria as the MARTHA patients, except with the restriction on age of onset. Controls were French subjects selected from the SUVIMAX population [36].

Stage III - Replication studies. For the replication of the GWAS findings, the FARIVE study [15], a multicenter case-control study for first episode of VT composed of 607 cases and 607 healthy individuals, all of French origin, was used.

Genotyping and Quality control

Stage I - Discovery GWAS. A subsample of 1011 VT patients were typed with the Illumina Human 610-Quad Beadchip while the remaining 586 VT patients were typed with the Illumina Human660W-Quad Beadchip. Individuals from the 3C study were also typed with Illumina Human 610-Quad Beadchip. A set of 551,141 SNPs including 537,883 autosomal SNPs and 13,258 sex-linked SNPs was common to the three samples.

Individuals with genotyping success lower than 95% (n = 18) were excluded from the analyses as were individuals demonstrating close relatedness (n = 67). This latter was assessed by pairwise clustering of identity by state distance (IBS) and multi-dimensional scaling (MDS) using the PLINK software [37]. The Eigenstrat program [38] was further used to detect individuals of non-European ancestry. SNPs showing significant ($p<10^{-5}$) deviation from Hardy-Weinberg Equilibrium (HWE) in controls, with minor allele frequency (MAF) less than 1% in the combined cases/controls samples or genotyping call rate <99% were filtered out. This lead to the final analysis of 481,002 autosomal and 10,256 sex-linked SNPs in a sample of 1,542 VT patients and 1110 healthy individuals.

Stage II - *In silico* GWAS study. Individuals participating in this previous GWAS were genotyped for 317,139 SNPs using the Illumina Sentrix HumanHap300 Beadchip among which 291,872 satisfied the quality control criteria previously described [15]. Individuals of non European ancestry had been also excluded from this analysis [15].

Stage III - Replication studies. In FARIVE, the rs8074026 was genotyped by allele-specific PCR (also referred to as ARMS i.e amplification refractory mutation system) with success rate of 97.5%.

Statistical Analysis

Genome-wide association study. Genome-wide association analysis of autosomal SNPs was conducted using the Eigenstrat program that correct for any uncontrolled population stratification [38]. The genomic control (GC) inflation factor was also computed according to the median test statistic [39]. X-linked SNPs association was tested using the PLINK software [37] while adjusting for first four principal components.

Haplotype analysis. To handle the linkage disequilibrium (LD) between SNPs of interest at specific loci, haplotype analysis was performed by use of the THESIAS program [40].

Imputation. In both GWAS datasets, imputation of 2,557,252 autosomal SNPs was conducted using the MACH (v1.0.16a) software (http://www.sph.umich.edu/csg/abecasis/mach/) according to the CEU HapMap 2 release 21 (build 35) reference dataset. A logistic regression analysis was then conducted to evaluate the association of each SNP with VT risk in an additive genetic model, in which allele dosage (0 to 2 copies of the minor allele) of imputed SNPs was analyzed. Analyses were adjusted for the first four principal components and were performed using the mach2dat (v 1.08.18) software (http://genome.sph.umich.edu/wki/Minimac).

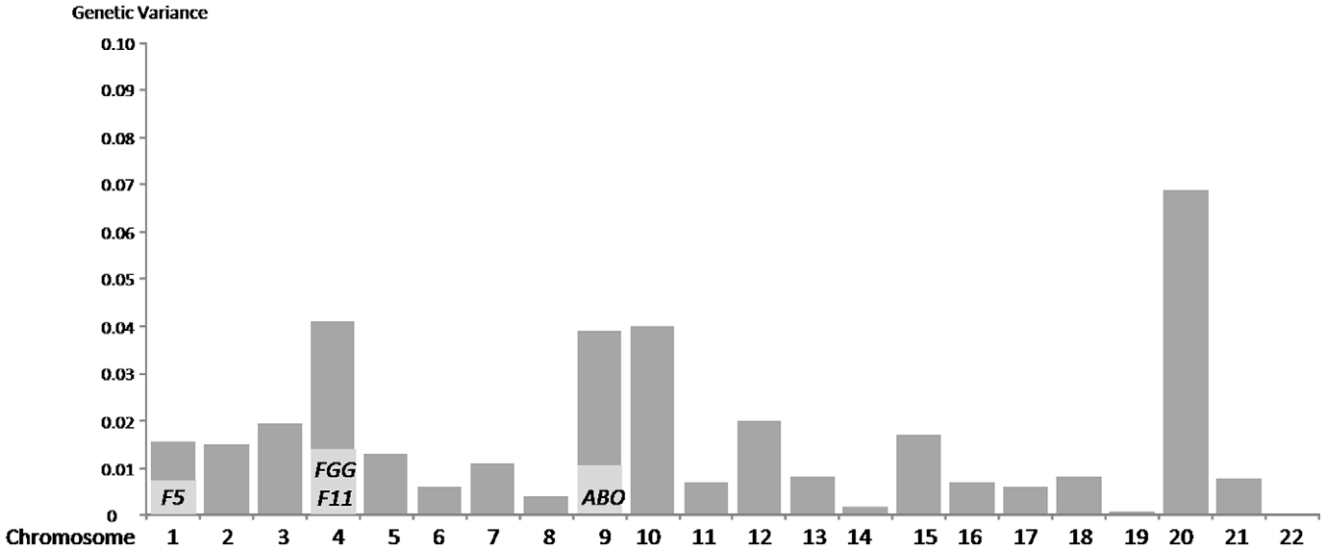


Figure 6. Distribution of VT genetic variance across chromosomes. In light grey is shown the relative contribution of specific loci.
doi:10.1371/journal.pone.0025581.g006

Meta-Analysis. All SNPs with acceptable imputation quality (r2≥0.3) in both imputed GWAS datasets were entered into a meta-analysis, leading to 2,475,305 SNPs left for statistical association analysis. For the meta-analysis, a fixed-effect model relying on the inverse-variance weighting was used as implemented in the METAL software (http://www.sph.umich.edu/csg/abecasis/metal). Homogeneity of associations across the two GWAS studies was tested using the Mantel-Haenszel method [41].

For all these GWAS analyses, a statistical threshold of 2.0×10^{-8} was used to declare genome-wide significance. This value corresponds to the family-wise error rate of 0.05 corrected for the number of studied SNPs (2,475,305) according to Bonferroni correction.

Replication. Association of SNPs tested for replication with VT was assessed by use of the Cochran-Armitage trend test [42]. Logistic regression analysis was further used to estimate genetic effects, expressed in terms of Odds Ratio (OR), adjusted for age, gender, FV leiden and ABO blood group.

Genetic Variance Estimation. The recently proposed GCTA methodology was used to investigate the genetic variance of VT [31,43]. Briefly, this method consists in estimating the genetic relationship between unrelated individuals from genome-wide SNPs information and in incorporating it into a regression model to provide an estimate of the genetic variance of a given phenotype. For a binary phenotype such as VT, it assumes the existence of an underlying normally distributed liability variable, with individuals being affected if their liability exceeds a threshold which may depends on covariates such as gender. We computed the genetic relationships separately from all SNPs of a given chromosome and assessed the contribution of each chromosome on the genetic variance. A similar approach was applied to estimate the contribution of specific loci of using all SNPs within ±5 Mb of each locus. All analyses were adjusted for gender and

Table 7. Minor allele frequencies distribution of chromosome 20 SNPs demonstrating association with VT at $p_{EIGENSTRAT}<1.00\times10^{-4}$ in a GWAS sample of 1,542 VT cases and 1,110 controls.

CHR	Position	Gene	SNP	Alleles[1]	Cases	Controls	P[2]
20	960026	*RSPO4*	rs11696364	C/A	0.059	0.102	1.53×10^{-6}
20	16273269	*C20orf23*	rs4814475	G/A	0.181	0.228	3.98×10^{-5}
20	16312593	*C20orf23*	rs6034465	C/T	0.158	0.204	6.73×10^{-6}
20	16574082	*C20orf23*	rs964216	C/T	0.116	0.084	4.55×10^{-5}
20	16575913	*C20orf23*	rs13038362	A/C	0.114	0.084	7.41×10^{-5}
20	16580409	*SNRPB2*	rs6135823	T/C	0.115	0.083	5.38×10^{-5}
20	29886239	*MYLK2*	rs17340555	T/C	0.110	0.084	4.53×10^{-5}
20	46604745	*PREX1*	rs1883888	G/A	0.350	0.286	1.11×10^{-5}
20	46613060	*PREX1*	rs878198	A/C	0.354	0.292	3.67×10^{-5}
20	46622891	*PREX1*	rs4810820	T/C	0.364	0.306	9.95×10^{-5}
20	46624742	*PREX1*	rs6012481	C/T	0.345	0.287	2.71×10^{-5}

[1]Common/minor alleles.
[2]Association test p-value corrected for principal components (EIGENSTRAT program).
doi:10.1371/journal.pone.0025581.t007

principal components as indicated in the GCTA documentation [31].

Supporting Information

Table S1 Allele frequencies of candidate gene SNPs in the discovery GWAS sample of 1,542 VT cases and 1,110 controls. [(1)] Common/minor alleles. [(2)]Association test p-value corrected for principal components (EIGENSTRAT program). [(3)] P-value of the Cochran-Armitage Trend test corrected for the genomic control factor. Genes were selected as candidates for VT because either: - (A): they belong to the coagulation cascade (Blood 2000; 95:1517–1532;Blood 2008; 112: 19–27). - (B): or they belong to the fibrinolytic cascade (Blood 2000; 95:1517–1532; Semin Thromb Hemost. 2009;35:468–77). - (C): or they harbours SNPs that have been associated with VT risk (Blood 2010; 115:4644–4650; JAMA 2008; 299:1306–1314; Am J Hum Genet 2010; 86:592–595; J Thromb Haemost 2010; 8:2671–2679). - (D): or they mapped loci found through recent GWAS associated with quantitative biomarkers of VT such as D1: vWF & FVIII (Circulation 2010; 121:1382–1392). D2: Platelet volume (Am J Hum Genet 2009; 84:66–71). D3: Protein C levels (Blood 2010; 116:5032–5036). D4: aPTT (Am J Hum Genet 2010; 86:626–631). D5: PAI-1 levels (Blood 2010; 116:2160–2163).
(DOC)

Table S2 Haplotype structure derived from VT-associated *F11* SNPs in the HapMap database. Haplotype frequencies were estimated using the Haploview software from the HapMap 3 (release 2) data. SNPs identified in the LETS study (Li Y et al. J Thromb Haemost 2009;7:1802–1808) are shown in bold, others were those identified in the current MARTHA project.
(DOC)

Table S3 Stage II - Genome-wide significant ($p<2.01\ 10^{-8}$) SNP imputed associations with VT in the combined discovery and *in silico* GWASes of 1,961 cases and 2,338 controls. [(1)] Common/rare alleles. [(2)] Minor allele frequency. [(3)] Odds Ratio associated with the minor allele estimated from the Mach2dat imputation software, after adjusting for principal components. [(4)] Combined p-values computed using the inverse-variance model as implemented in METAL software. All shown imputed SNPs satisfied the imputation quality criteria (r2.hat>0.3).
(DOC)

Acknowledgments

We wish to kindly thank Professor Frits Rosendaal for his fruitful discussions all through the process of this work.

Author Contributions

Conceived and designed the experiments: D-AT P-EM PA ML. Performed the experiments: NS CP DZ. Analyzed the data: MG NG TO-M GA D-AT CD J-CL FC LT. Contributed reagents/materials/analysis tools: P-EM PA ML LL A-MD MB JE NS M-CA WC. Wrote the paper: MG D-AT P-EM.

References

1. White RH (2003) The epidemiology of venous thromboembolism. Circulation 107: I4–8.
2. Prandoni P, Bernardi E, Marchiori A, Lensing AW, Prins MH, et al. (2004) The long term clinical course of acute deep vein thrombosis of the arm: prospective cohort study. Bmj 329: 484–485.
3. Coombes R (2005) Venous thromboembolism caused 25,000 deaths a year, say MPs. Bmj 330: 559.
4. MacDougall DA, Feliu AL, Boccuzzi SJ, Lin J (2006) Economic burden of deep-vein thrombosis, pulmonary embolism, and post-thrombotic syndrome. Am J Health Syst Pharm 63: S5–15.
5. Beckman MG, Hooper WC, Critchley SE, Ortel TL (2010) Venous thromboembolism: a public health concern. Am J Prev Med 38: S495–501.
6. Larsen TB, Sorensen HT, Skytthe A, Johnsen SP, Vaupel JW, et al. (2003) Major genetic susceptibility for venous thromboembolism in men: a study of Danish twins. Epidemiology 14: 328–332.
7. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, et al. (2000) Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. Am J Hum Genet 67: 1452–1459.
8. Ikram MA, Seshadri S, Bis JC, Fornage M, DeStefano AL, et al. (2009) Genomewide association studies of stroke. N Engl J Med 360: 1718–1728.
9. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, et al. (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 41: 334–341.
10. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, et al. (2007) A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316: 1491–1493.
11. Erdmann J, Willenborg C, Nahrstaedt J, Preuss M, Konig IR, et al. (2010) Genome-wide association study identifies a new locus for coronary artery disease on chromosome 10p11.23. Eur Heart J 32: 158–168.
12. Erdmann J, Grosshennig A, Braund PS, Konig IR, Hengstenberg C, et al. (2009) New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat Genet 41: 280–282.
13. Tregouet DA, Konig IR, Erdmann J, Munteanu A, Braund PS, et al. (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat Genet 41: 283–285.
14. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, et al. (2007) Genomewide association analysis of coronary artery disease. N Engl J Med 357: 443–453.
15. Tregouet DA, Heath S, Saut N, Biron-Andreani C, Schved JF, et al. (2009) Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. Blood 113: 5298–5303.
16. Rosendaal FR, Reitsma PH (2009) Genetics of venous thrombosis. J Thromb Haemost 7 Suppl 1: 301–304.
17. Morange PE, Bezemer I, Saut N, Bare L, Burgos G, et al. (2010) A follow-up study of a genome-wide association scan identifies a susceptibility locus for venous thrombosis on chromosome 6p24.1. Am J Hum Genet 86: 592–595.
18. Buil A, Tregouet DA, Souto JC, Saut N, Germain M, et al. (2010) C4BPB/C4BPA is a new susceptibility locus for venous thrombosis with unknown protein S-independent mechanism: results from genome-wide association and gene expression analyses followed by case-control studies. Blood 115: 4644–4650.
19. Morange PE, Saut N, Antoni G, Emmerich J, Tregouet DA (2011) Impact on venous thrombosis risk of newly discovered gene variants associated with FVIII and VWF plasma levels. J Thromb Haemost 9: 229–231.
20. Bezemer ID, Bare LA, Doggen CJ, Arellano AR, Tong C, et al. (2008) Gene variants associated with deep vein thrombosis. Jama 299: 1306–1314.
21. Morange PE, Oudot-Mellakh T, Cohen W, Germain M, Saut N, et al. (2011) KNG1 Ile581Thr and susceptibility to venous thrombosis. Blood 117: 3692–3694.
22. 3C Study Group (2003) Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. Neuroepidemiology 22: 316–325.
23. Uitte de Willige S, de Visser MC, Houwing-Duistermaat JJ, Rosendaal FR, Vos HL, et al. (2005) Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen gamma' levels. Blood 106: 4176–4183.
24. Uitte de Willige S, Pyle ME, Vos HL, de Visser MC, Lally C, et al. (2009) Fibrinogen gamma gene 3′-end polymorphisms and risk of venous thromboembolism in the African-American and Caucasian population. Thromb Haemost 101: 1078–1084.
25. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938–2939.
26. Li Y, Bezemer ID, Rowland CM, Tong CH, Arellano AR, et al. (2009) Genetic variants associated with deep vein thrombosis: the F11 locus. J Thromb Haemost 7: 1802–1808.
27. Tang W, Basu S, Kong X, Pankow JS, Aleksic N, et al. (2010) Genome-wide association study identifies novel loci for plasma levels of protein C: the ARIC study. Blood 116: 5032–5036.
28. Folsom AR, Aleksic N, Wang L, Cushman M, Wu KK, et al. (2002) Protein C, antithrombin, and venous thromboembolism incidence: a prospective population-based study. Arterioscler Thromb Vasc Biol 22: 1018–1022.
29. Antoni G, Morange PE, Luo Y, Saut N, Burgos G, et al. (2010) A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism. J Thromb Haemost 8: 2671–2679.

30. Smith NL, Chen MH, Dehghan A, Strachan DP, Basu S, et al. (2010) Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. Circulation 121: 1382–1392.
31. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76–82.
32. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38: 209–213.
33. Houlihan LM, Davies G, Tenesa A, Harris SE, Luciano M, et al. (2010) Common variants of large effect in F12, KNG1, and HRG are associated with activated partial thromboplastin time. Am J Hum Genet 86: 626–631.
34. Smith NL, Rice KM, Bovill EG, Cushman M, Bis JC, et al. (2011) Genetic variation associated with plasma von Willebrand factor levels and the risk of incident venous thrombosis. Blood 117: 6007–6011.
35. Coronary Artery Disease (C4D) Genetics Consortium (2011) A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat Genet 43: 339–344.
36. Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, et al. (2004) The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. Arch Intern Med 164: 2335–2342.
37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
39. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
40. Tregouet DA, Garelle V (2007) A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies. Bioinformatics 23: 1038–1039.
41. Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22: 719–748.
42. Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53: 1253–1261.
43. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569.