Genetic and Epigenetic Determinants of Thrombin Generation Potential : an epidemiological approach

Maria-Ares Rocanin-Arjo

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE 420 :
SANTÉ PUBLIQUE PARIS SUD 11, PARIS DESCARTES

Laboratoire : Equip 1 de Unité INSERM UMR_S1166
Genomics & Pathophysiology of Cardiovascular Diseases

THÈSE DE DOCTORAT

SANTÉ PUBLIQUE - GÉNÉTIQUE STATISTIQUE

par

Ares ROCAÑIN ARJO

Genetic and Epigenetic Determinants of Thrombin Generation Potential: an epidemiological approach.

Date de soutenance : 20/11/2014

Composition du jury :

Directeur de thèse :	David Alexandre TREGOUET	DR, INSERM U1166, Université Paris 6, Jussieu
Rapporteurs :	Guy MEYER	PU_PH, Service de pneumologie. Hôpital européen Georges Pompidou
	Richard REDON	DR, Institut thorax, UMR 1087 / CNRS 6291 , Université de Nantes
UMR Examinateurs :	Laurent ABEL	DR, INSERM U980, Institut Imagine
	Marie Aline CHARLES	DR, INSERM U1018, CESP

Al meu pare
(*to my father*
/à mon père)

Your genetics load the gun.
Your lifestyle pulls the trigger.
Mehmet Oz

Reserve your right to think,
for even to think wrongly
is better than not to think at all.
Hypatia

If you want to know about thrombin- measure thrombin"
HC Hemker

Aknowledgements / Remerciemets / Agraïments

C'est incroyable, mais malgré ce que je craignais au début, ces trois ans ont passé très vite ! Il y a trois ans, déjà, j'ai commencé ma première journée dans l'unité U937 (maintenant UMR_S1166). Tout juste 5 jours avant, j'arrivais avec mes deux valises. Je m'en souviens encore : Moi, attendant le bus 351 à l'extérieur du terminal de CDG. Une dame m'avait alors demandé comment se rendre au centre de Paris. Puis elle m'avait dit: "Je viens pour visiter Paris, vous aussi ?". Et je lui avais répondu : "Non, je viens pour y vivre, je vais faire un doctorat.". Deux heures plus tard et une bonne route à travers les banlieues parisiennes, je débarquais sur la place Gallieni pour retrouver celui qui serait mon colocataire et que je ne connaissais pas encore. Quels premiers jours ! Que d'émotions ! La veille de commencer ma thèse, à l'occasion d'Halloween, j'étais allée au cinéma Le Champo, où j'avais passé la nuit devant 3 films d'horreur. En sortant au petit matin, avec mon premier croissant et mes nouvelles connaissances, nous avions décidé de rentrer à pied à travers les rues de Paris. Quelle belle promenade. Il n'y avait personne et en arrivant à la Seine nous avions découvert Notre Dame illuminée par le soleil levant et toute pour nous. Que Paris est belle !

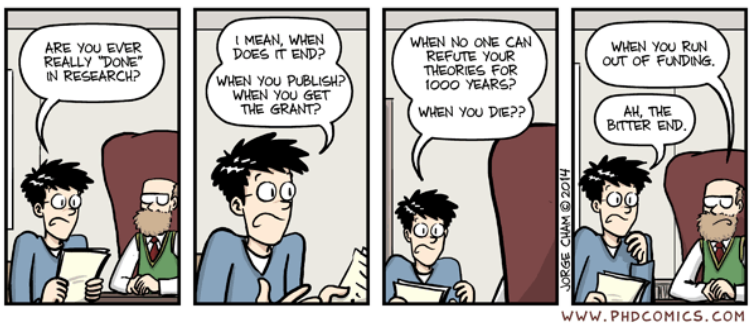
J'ai aussi des souvenirs de mes premiers jours au labo et notamment des 3-4 alertes incendie que nous avions eues en trois jours. Je me souviens de la première fois que j'ai entendu la sirène des pompiers, en pensant, "C'est grave ? Qu'est-ce qu'il faut faire ? C'est une alerte ? Que font les autres ? Rien. Tranquilles... Donc, ok c'est rien ". Je me rappelle aussi de la petite souris qui habitait dans la cuisine. Maintenant, à la fin de ma thèse, nombre d'expériences et de moments vécus, qui m'ont appris aussi bien professionnellement qu'humainement, se rappellent à moi.

Je tiens d'abord à remercier François Cambien et Laurence Tiret pour m'avoir accueillie tout au début au sein de l'ancienne unité U937 intitulée "Génomique Cardiovasculaire".

Je voudrais aussi remercier l'ensemble du jury d'avoir accepté de consacrer de leur temps pour évaluer ce travail : mes deux rapporteurs, le Professeur Guy Meyer et Monsieur Richard Redon, pour leur relecture profonde et leurs critiques ; je remercie aussi mon examinateur Monsieur Laurent Abel et mon examinatrice et présidente de jury Madame Marie-Aline Charles.

Je remercie le professeur Pierre Emmanuel Morange pour son important apport dans ce travail et pour avoir mis en place l'étude MARTHA, une étude très puissante sans laquelle je n'aurais pu faire ma thèse. De même, je remercie France Gagnon et les members de l'étude des "Three City" pour m'avoir permis d'utiliser des données issues de leur cohorte.

Je remercie également l'ensemble de l'unité, ceux qui sont là actuellement ainsi que ceux qui y sont passés que ce soit pour un bon moment ou tout simplement de passage. Votre sympathie, votre complicité et votre soutien constant m'ont permis de me sentir intégrée et m'ont donné la sensation d'appartenir à une petite "famille" : Merci aux filles, aux nouvelles Beata et Maguelonne ainsi qu'aux deux Marie-Laure. Merci Carole pour ta joie et ton effort pour que je fasse du sport. Merci Claire pour m'avoir fait découvrir le reblochon et pour m'avoir permis de me maintenir en contact avec la biologie. Vielen danke Ulli (ma belle!!) pour tout: les fêtes, les instants de complicité et pour avoir partagé avec moi tous ces moments. Zahia, pour ta sympathie constante et pour partager avec moi le goût pour le café crème. Grazie mille Veronica pour ton soutien constant. Je vous remercie Ewa, Laurence et Hervé pour nos conversations dans la cuisine. Merci Badreddine pour tes agréables "salut" dans le couloir et "bon appétit" en passant par la cuisine. Merci Nadjim et Lyamine pour être toujours si galants. Merci François (le roux, juif, irlandais,...) pour tes cours de "français des jeunes". Merci Nathalie pour nos discussions politiques et éthiques et pour rendre le monde administratif moins effrayant et plus accessible. Merci Henri, revenu juste pour la fin, pour tes blagues provocatrices et ton ironie. Merci beaucoup à mon compagnon de "fatigue" et bon ami, Dylan (Methylman), qui est toujours dispo pour faire un petit "break" pour se détendre et discuter mais surtout pour m'aider avec les machines. Merci à Maxime même si nous n'avons jamais été en même temps dans l'unité. Merci à Raphaele, Ricardo, Vinhou, Nico, Marie et Guillaume (Boby), pour leur sympathie, leur joie et leur bonne humeur, pour les activités "afterwork chez Jimmy", pour les questions "Nico", pour les "hotdogs chilenos", pour l'hippodrome, pour le cours d'escalade ou le poulet au curry thaï de Vinh! Et aussi merci aux stagiaires (oompa loompas) Lynn, Bathou, Yasmine, Evangelia, Francesca, "Doudou", Pauline, Bastien, pour leurs croissants, Romain qui est passé au coté "compagnon de fatigues" et merci aussi aux canadiens avec qui je garde une amitié précieuse, Martin et Jessica.



Je remercie vivement mes copines de bureau: Marine et Sophie, pour avoir égayé mon quotidien, pour avoir répondu à toutes mes questions, pour votre précieuse aide en relisant ce mémoire, pour m'avoir "supportée" dans les derniers mois difficiles, pour nos moments

comiques en racontant nos expériences personnelles mais aussi gourmands en parlant de recettes. Vous allez me manquer. Ce sera bizarre de ne plus vous voir après ces 3 ans !

Enfin je veux remercier plus particulièrement mon directeur de thèse, David-Alexandre Trégouët pour sa confiance en moi, surtout au début quand il ne me connaissait pas, en m'aidant pour ma première inscription à la faculté, pour la recherche d'appartement et pour les démarches administratives. Merci beaucoup d'avoir accordé ta confiance à une biologiste sans expérience pour réaliser ce projet. Merci pour ton très bon encadrement, pour ton aide en tout moment au cours de ces trois ans et pendant l'écriture de ma thèse. Merci pour avoir été patient avec ma façon de faire (lui très organisé et moi j'ai une tendance au désordre), et pour ton exigence, qui m'a certes valu certains moments difficiles à endurer, mais m'a permis de me surpasser, de m'organiser et d'avoir le travail que je présente ici.

Je vais désormais retourner à mes origines et continuer dans ma langue natale, le Catalan, pour mieux exprimer mes remerciements à mes amis et à ma famille: *Sou moltes persones que també heu aportat el vostre granet de sorra i m'heu permès arribar fins aquí.*

Tots els meus companys de grups anteriors: Magda, Georgios, Marc, Robert, Esther, Mireia, i Pedro, els primers en ensenyar-me i formar-me com a investigadora i amb els que encara conservo una boníssima amistat. Després, les pipetes Raquel, Sonia, Biel i Miquel Soute, pels nostres dinars, esmorçars, cafès, birres i moments a la cabanya i els tecles Angel, Alfonso, Anna i Leonor que em van començar a mostrar en els "lab-meetings" el món dels bioinformàtics i estadístics. Moltíssimes gràcies, José Manuel Soria, per dornar-me l'oportunitat de fer la tesi a Paris, interessant-te i demanant per beques als teus col·laboradors. Encara recordo el moment en que m'ho vas comunicar: em vas trucar perquè m'havia demanat festa per poder saltar en paracaigudes. Dues emocions fortes en un mateix dia. Recordo com vaig estar pensant el què hauria de fer. Vull reiterar les gràcies a Sonia per la temporada que va estar a Paris, en la qual em va ajudar molt al inici de la meva estada que sempre és dura, 6 mesos on vam compartir converses post dinar, moments explosius de desesperació,fins i tot cases. Em van ajudar molt.

Moltes gràcies a tots els companys de la facultat (Roser, Patricia, Cris D, Sergi, Maria, Beth i Cris P, Arnau, Mendoza, Pau, Xavi, Irene, Celia) per les nostres trobades, que no han parat, coincidint com podíeu quan era, o ja més tard érem, a Barcelona. Han arribat a convertir-se com en un grup de teràpia per a doctorants amb anècdotes, converses profundes sobre la ciència, sopars, birretes, braves,...les braves són mooolt importants! Moltes Gràcies Cris D per la correcció d'una part del treball, sobretot per les molèsties en fer-ho ràpid.

Moltes gràcies també a les "nenes del cole": Marta, Mireia, Delia, Laura, Patri, Neus i consorts per interessar-vos sempre per com m'anava i pel vostre suport. Sembla mentida que després de 10 anys, quan vàrem deixar l'escola, i dels diferents camins que hem agafat cadascuna, continuem encara juntes! Moltes gràcies també als "frifris", per donar-me suport amb el seu humor constant.

Moltes gràcies a la meva petita família parisina: alguns presents des del principi acollint-me des dels primers dies i als que ens hem anat trobat de la manera menys probable. Vosaltres feu possible viure a Paris amb menys enyorança; Marina, totes les valencianetes, Jacob, Belen, Chantal, Valentine, Andrea, Yaiza, Jordi i Jerôme (français mais entre catalans, merci pour ton bon humeur), gràcies.

Finalment, els més importants per a mi. Moltes gràcies a la meva família pel seu suport en aquests tres anys, interessant-se sempre i venint-me a visitar quan podien. No puc expressar en paraules l'agraïment profund que sento per ma germana, Anaïs, en Guifré (si, ja ets de la família) i ma mare, Rosamari, els quals han fet gests titànics per ajudar-me en tot aquest procés. Venint-me a veure al preu que fos, o traslladant-se a Paris, deixant la preuada mar per venir a viure amb mi. Gràcies també per corregir moltes parts d'aquest treball i per donar-me forces i consell en tot el camí, per estar-hi sempre. Sense ells no crec que hagués arribat tan lluny. El seu recolzament en tot moment ha estat molt valuós i clau en molts moments. Guardo pel final, una persona molt especial per mi, que malauradament ja no és aquí. Ell sempre va tenir la il·lusió de que jo fes un doctorat i per què negar-ho, la seva voluntat ha estat una part important en la empresa d'aquesta tesi. Ell també, sense voler-ho, em va donar el rumb, fent-me interessar per la trombosi i sé que a ell, amant de tot allò francès, li hagués agradat compartir amb mi aquest procés. Moltes gràcies papa!

I. **Scientific formation and contribution**

- Main publications
  - ② **Ares Rocanin-Arjo,** William Cohen, Laure Carcaillon, Corinne Frère, Noémie Saut, Luc Letenneur, Martine Alhenc-Gelas, Anne-Marie Dupuy, Marion Bertrand, Marie-Christine Alessi, Marine Germain, Philipp S. Wild, Tanja Zeller, Francois Cambien, Alison H. Goodall, Philippe Amouyel, Pierre-Yves Scarabin, David-Alexandre Trégouët, Pierre-Emmanuel Morange and and the CardioGenics Consortium. Blood. 2014;123:777-85.
  - ② **A. Rocanin-Arjo, J.** Dennis, P. Suchon, D. Aïssi, V.Truong, D-A. Trégouët, F. Gagnon, P-E. Morange. *Thrombin Generation Potential and Whole-Blood DNA methylation*. Thrombosis Research.
- Collaborations
  - Aïssi D, Dennis J, Ladouceur M, Truong V, Zwingerman N, **Rocanin-Arjo A**, et al. Genome-Wide Investigation of DNA Methylation Marks Associated with FV Leid**en Mutation. PL**oS One 2014;9:e108087.

- Oral communications
  - ☺ **Rocanin-Arjo A**, Carcaillon L, Cohen W, Saut N, Germain M, Letenneu*r L, Alhenc-Gelas M, Dupuy AM, Bertrand M, Amouyel P, Scarabi*n PY, Trégouët DA, Morange PE. *A meta-analysis of genome-wide association studies identifies a novel locus associated with Thrombin generation potential*. American Society of Human Genetics Conference 2013 Boston October 22nd-26th.
  - ☺ **Rocanin-Arjo A**, Carcaillon L, Cohen W, Saut N, Germain M, Letenneur L, Alhenc-Gelas M, Dupuy AM, Bertrand M, Amouyel P, Scarabin PY, Trégouët DA, Morange PE. *A meta-analysis of genome-wide association studies identifies a novel locus associated with Thrombin generation potential*.CORDDIM.12 septembre.

- Poster communications
  - 🗏 **Rocanin-Arjo A**, Carcaillon L, Cohen W, Saut N, Germain M, Letenneur L, Alhenc-Gelas M, Dupuy AM, Bertrand M, Amouyel P, Scarabin PY, Trégouët DA, Morange PE*. A meta-analysis of genome-wide association studies identifies a novel locus associated with Thrombin generation potential.* European Human Genetics Conference 2013 Paris June 8th-13th

- **Rocanin-Arjo A**, Carcaillon L, Cohen W, Saut N, Germain M, Letenneur L, Alhenc-Gelas M, Dupuy AM, Bertrand M, Amouyel P, Scarabin PY, Trégouët DA, Morange PE*. A meta-analysis of genome-wide association studies identifies a novel locus associated with Thrombin generation potential.* ICAN conferences 2013 Paris December 12th-14th
- **Rocañín-Arjó A,** Pierre Suchon, Noémie Saut, David-Alexandre Trégouët, Pierre-Emmanuel Morange. *Association of Orosomucoid plasma levels with thrombin generation potential and cardiometabolic biomarkers*. ICAN 2014 22th September, Paris, France, (Poster).

- Courses and formation
    - "Leena Peltonen Summer School of Human Genomics", Welcome Trust Conference Centre, 18 – 22 August 2013, Hinxton, Cambridgeshire, United Kingdom.
    - "2nd Non-coding genome", Institute Curie, 10-14 December 2012, Paris, France.

IV. **List of principal abbreviations**

TGP: Thrombin generation potential.

MARTHA: MARseille THrombosis Analysis Study.

3C: Three City Study.

FII: factor II or prothrombin.

FIIa: active factor II or thrombin.

FVIII: factor VIII(FVIIIa active form).

FVL: factor V Leiden (FVLa active form).

VT: Venous Thrombosis.

vWF: von Willebrand factor.

FDR: False Discovery Rate.

GWAS: GenomeWide Association Study.

EWAS: Epigenetic Wide Association Study.

IBD: Identical By Descent.

MAF: Minor Allele Frequency.

MCMC: Monte-Carlo Markov Chain Method.

QTL: Quantitative Trait Locus.

Résumé en français

Contexte du projet et motivation

La thrombine, aussi connue sous le nom "facteur de coagulation II", est synthétisée dans le foie sous forme de zymogène: la prothrombine (ou FII), qui, une fois relâchée, est clivée, en exposant ses domaines et en activant ses fonctions (FIIa). Sa fonction la plus connue et la plus importante est celle de de catalyseur de la formation de fibrine qui constituera le caillot, d'activateur d'autres facteurs pro-coagulants mais également de déclencheur d'enzymes anti-coagulantes[8].

L'activation de la thrombine est le résultat de la cascade de coagulation à laquelle participent de nombreuses protéines (Tableau 1.1, page 6) s'activant les unes les autres (Figure 1.2., page 7). Le début de la coagulation a lieu quand le facteur tissulaire (TF) est exposé au sang, à cause d'une lésion des vaisseaux. Ensuite TF active le facteur VII (FVIIa, a pour activé) devenant le complexe TF-FVIIa qui active alors deux autres facteurs: le facteur IX et le facteur X. Le FXa seul est peu efficace et ne protéolyse qu'une petite quantité de prothrombine[4]. Cette petite, mais suffisante, quantité de thrombine activée (2 nmol / L) commence à interagir avec des cofacteurs et substrats en provoquant le début de réactions en chaîne menant à la coagulation[1,3,4,9,15]. La thrombine protéolyse le fibrinogène en formant des unités de fibrine (Figure 1.3., flèche 1, page 10) qui vont être assemblées en fibres à fin de stabiliser le caillot. En outre, elle protéolyse les facteurs V et VIII (Figure 1.3., flèche 4, page 10) qui vont aider à amplifier l'activation de la thrombine. La thrombine va également activer des facteurs anticoagulants qui vont aider à contrôler et finaliser la coagulation.

La thrombine participe également, au-delà de l'hémostase, à d'autres systèmes physiologiques, comme par exemple les systèmes immunitaire, nerveux, gastro-intestinal, et musculo-squelettique. Elle interagit avec des protéines et des récepteurs, en activant les plaquettes et les cellules endothéliales, et en stimulant l'adhésion, l'angiogénèse, la croissance cellulaire, la différenciation, la prolifération, la vasoconstriction et l'inflammation. Tout ce large éventail de fonctions est modulé grâce à une combinaison complexe de substrats et de cofacteurs[4] (Tableau 1.1. et 1.2., pages 6-7).

Pour toutes les fonctions expliquées ci-dessus, la régulation de la thrombine est fondamentale pour une physiologie normale. Des niveaux déséquilibrés de thrombine se traduisent par différents types d'anomalies. Les plus connues et étudiées sont la thrombose/l'hémophilie, l'inflammation et l'athérosclérose. Il est alors essentiel de pouvoir

mesurer son activité aussi précisément que possible afin d'étudier et de comprendre les conséquences physiopathologiques et moléculaires de ses anomalies. Les tests classiques reflètent seulement 5% de l'ensemble du processus[1] de l'activation et de l'activité de la thrombine.

Le potentiel de génération de thrombine (TGP, en anglais) est un test qui a été mis à point pour mesurer la quantité potentielle de thrombine qui est capable d'être activée au moment de coagulation[45]. Il peut être vu comme un reflet l'ensemble du processus de la coagulation allant de son l'initiation, sa propagation et sa terminaison-amplification[47]. Il est plus sensible aux déficits de facteurs de coagulation: activateurs (FVII, FV, FX) et inhibiteurs (antithrombine, protéine C, protéine S) et à de nombreux troubles de la coagulation associés à une résistance à la protéine C activée (APC, en anglais). En outre, il est sensible à tous les types d'anticoagulants ou d'autres médicaments.

Le TGP est mesuré à partir d'une méthode, communément appelée thrombogramme calibré automatisé (CAT, en anglais), qui consiste en deux mesures simultanées de l'activation de la thrombine avec un substrat fluorochromé pour un même échantillon de plasma. L'une mesure la génération de thrombine (TG) et l'autre sert de calibrage pour corriger le biais entre le signal fluorochrome et l'activation de la thrombine. Dans le test TG, la thrombine est produite dans une réaction de coagulation activée par du facteur tissulaire, des phospholipides et du calcium. La quantité de thrombine générée est alors mesurée en temps réel par la capture du signal de fluorescence qui émet le substrat consommé par la thrombine. Ce signal est capturé et corrigé simultanément pour être affiché sous la forme du courbe (appelée thrombogramme) (Figure 2.1., page 24).

A partir de cette courbe, il est possible d'extraire plusieurs paramètres quantitatifs (Figure 2.2., page 25). Les plus utilisés sont : 1- le temps de latence (*Lagtime* en anglais) qui mesure le temps écoulé depuis le moment où la coagulation est déclenchée jusqu'à le début de formation du caillot; 2 - le potentiel endogène de thrombine (ETP pour *Endogeneous Thrombin Potential* en anglais) qui représente la totalité de thrombine activée (l'aire sous la courbe du thrombogramme) et permettant de représenter plus précisément l'état de coagulation et toute l'activée enzymatique autour la génération de la thrombine; 3- le hauteur du pic (*Peak*, P) qui mesure la quantité maximale de thrombine activée à un moment donné du processus de coagulation.

Le test TGP est sensible aux variations des facteurs FX, FIX, FVII et FVIII, du fibrinogène, des D-dimères (produits de la formation des unités de fibrine à partir de fibrinogène) et de protéines anticoagulantes. Le TGP est également associé à l'indice de masse corporelle (IMC), l'âge et le sexe. En ce qui concerne les facteurs génétiques connus

pour influencer la variabilité interindividuelle du TGP, il en existait deux au moment où je commençais mon projet de thèse: les polymorphismes rs1799963 (20210G> A) et rs3136516 (19911A> G) du gène codant pour la prothrombine (F2).

Des niveaux élevés de TGP, en particulier d' ETP, ont été associés au risque de thrombose veineuse[80–82], aux accidents vasculaires cérébraux (AVC) ischémiques aigus[223] et à l'infarctus du myocarde[35]. En revanche, des niveaux bas d'ETP sont associés à des troubles de la coagulation[46,86] (Figure 2.6.). Le TGP a également été trouvé associé à diverses désordres cardiométaboliques comme l'athérosclérose[87,224], l'obésité[60], le diabète de type 2[88], la néphropathie diabétique[89], et des troubles hépatiques comme la cirrhose[225]. Des troubles inflammatoires ont également été associés aux TGP tels que la maladie de Crohn[91], la septicémie[92] et la drépanocytose[93].

Objectifs:

Comme mentionné ci-dessus, seuls deux polymorphismes génétiques, tous les deux situés dans le gène *F2*, étaient connus pour influencer les taux plasmatiques de TGP. Cependant, ils n'expliquent que 11,3% de la variabilité interindividuelle de ce biomarqueur. Mon projet de thèse avait pour objectifs d'identifier de nouveaux facteurs génétiques et épigénétiques modulant ce phénotype et ses biomarqueurs associés.

Partie I Identification de nouveaux déterminants génétiques contrôlant la variabilité interindividuelle plasmatique du potentiel de génération de thrombine

Dans le cadre de mon projet de thèse, j'ai mené la toute première étude d'association génome-entier (*GWAS* pour Genome Wide Association Study en anglais) sur les biomarqueurs du TGP (à savoir ETP, Peak et Lagtime), afin d'identifier de nouveaux gènes participants à la variabilité du TGP.

Dans une première étape de découverte, j'ai testé l'association entre 6 652 054 de polymorphismes génétiques, plus précisément des polymorphismes de substitution d'un seul nucléotide (*SNP* pour *Single Nucleotide Polymorphism* in anglais), et la variabilité plasmatique des trois biomarqueurs du TGP dans deux échantillons rassemblant 1967 sujets. Outre les deux polymorphismes du gène F2 déjà connus et mentionnés ci-dessus, j'ai identifié un polymorphisme au sein du gène ORM1, rs150611042, montrant une association statistique très forte ($p = 3.36\ 10^{-7}$) avec le temps de latence (Lagtime). J'ai ensuite cherché à répliquer cette association dans deux autres échantillons indépendants rassemblant 1254 sujets. Dans cette deuxième étape de validation, le polymorphisme rs150611042 a été

retrouvé fortement associé ($p = 6.07\ 10^{-9}$) à la variabilité interindividuelle du temps de latence. J'ai ensuite montré que, dans deux cohortes indépendantes, l'une de 745 sujets et l'autre de 1374 individus, l'allèle du polymorphisme rs150611042 qui diminuait le temps de latence était également associé à une diminution de l'expression monocytaire du gène ORM1 ($p = 8.70\ 10^{-10}$ et $p = 5.21\ 10^{-16}$, respectivement). Des expériences fonctionnelles ont été réalisées pour compléter ce travail et ont confirmé *in vitro* l'association entre la molécule codée par le gène ORM1 et la génération de thrombine.

En conclusion, cette partie de mon travail de thèse a permis d'identifier un nouveau gène participant à la régulation de la production de thrombine. Les mécanismes précis de cette régulation restent cependant à élucider.

Partie II Marques de méthylation d'ADN associées au potentiel de génération de thrombine

La méthylation de l'ADN est une marque épigénétique qui consiste en l'addition enzymatique d'un groupement méthyle (CH3, une molécule de carbone et trois d'hydrogène) à la position 5 de carbone de la cytosine principalement dans le contexte de la séquence 5'-cytosine-guanine (communément appelé dinucléotide CpG). Ce groupe méthyle est transféré par une ADN méthyltransférase à partir d'une autre molécule appelée S-adenylmethylcisteine (SAM), qui devient une S-adenylhomocysteine (SAH). Dans le même temps, l'ADN méthyltransférase attire des protéines qui vont se lier l'ADN et réguler l'expression et la structure de la chromatine[161].

Ces dernières années, de plus en plus d'études ont démontré le rôle de la méthylation de l'ADN dans le développement des maladies humaines. Les premiers résultats ont été observés dans le domaine du cancer, mais étendus rapidement à d'autres maladies humaines complexes telles que la sclérose en plaques, le diabète, les maladies inflammatoires et cardiovasculaires[166,175,176,226,227].

Dans la deuxième partie de mon travail de thèse, j'ai utilisé des données de méthylation mesurées dans des échantillons d'ADN sanguin à partir d'une technologie haut-débit, la puce HumanMethylation450K développée par la société Illumina, pour rechercher des marques de méthylation associées à la variabilité plasmatique des mêmes 3 biomarqueurs du TGP que ceux utilisés dans mon projet GWAS (Partie I). Cette puce permet de mesurer les niveaux de méthylation de l'ADN d'environ 480 000 sites CpG répartis tout au long du génome. Son utilisation dans de grandes cohortes épidémiologiques fait l'objet d'un enthousiasme très important en raison des récents succès qu'elle a permis d'obtenir pour détecter des marques

de méthylation associés à des facteurs environnementaux, génétiques et biologiques[185,228]. Dans le cadre de mon projet, j'ai eu accès à deux échantillons, l'un de 238 sujets, l'autre de 187, dans lesquels cette puce de méthylation avait été utilisée à partir de l'ADN issu du sang périphérique des sujets, et pour lesquels les biomarqueurs du TGP avaient aussi été mesurés. J'ai ainsi réalisé la première étude de recherche agnostique d'associations entre des niveaux de méthylation de sites CpG et la variabilité plasmatique du TGP. Ce type de recherche est communément appelée MWAS pour *Methylome-Wide Association Study* en anglais. J'ai suivi une stratégie de recherche assez similaire à celle que j'avais appliquée pour mon étude GWAS. Malheureusement, je n'ai identifié aucune association robuste entre des niveaux de méthylation de l'ADN sanguin et les biomarqueurs de la génération de thrombine.

Conclusions

L'étude GWAS que j'ai menée sur les taux plasmatiques de 3 marqueurs de la génération de thrombine a permis d'identifier un nouveau gène, ORM1, participant à la modulation du temps de latence. Le polymorphisme que j'ai identifié n'est vraisemblablement pas le variant fonctionnel et les mécanismes d'action d'ORM1 sur la génération de thrombine ne sont clairement pas caractérisés. Mon travail n'a donc ouvert qu'une petite porte vers un champ de recherches beaucoup plus vaste qui reste à explorer. D'autres pistes mériteraient d'être également explorées pour mieux disséquer les mécanismes génétiques associés à la génération de thrombine. La recherche que j'ai effectuée ne s'est concentrée que sur des associations univariées entre SNPs et TGP alors qu'il serait également intéressant d'étudier si des phénomènes d'interaction entre SNPs peuvent également contribuer à influencer la génération de thrombine. De plus, alors que je me suis concentrée sur 3 biomarqueurs du TGP, il existe d'autres biomarqueurs qu'il serait également possible d'étudier dans le contexte d'une étude GWAS, ce qui pourrait mener à l'identification d'autres mécanismes de régulation.

L'étude MWAS que j'ai conduite sur les taux plasmatiques de TGP s'est avérée moins fructueuse que l'étude GWAS. Plusieurs explications peuvent être relevées. Tout d'abord, la taille des échantillons que j'ai analysés dans ce projet était relativement modeste par rapport à ceux dont je disposais pour mon étude GWAS. Je ne peux donc pas exclure un manque de puissance pour détecter des effets épigénétiques modérés, et disposer d'autres échantillons mesurés à la fois pour la méthylation de l'ADN et le TGP serait alors le bienvenu. De plus, cette étude MWAS a été réalisée à partir de l'ADN du sang périphérique qui n'est peut-être pas le bon modèle pour étudier des mécanismes de méthylation de l'ADN

de protéines principalement exprimées dans le foie. Il serait plus pertinent de pouvoir disposer de l'information sur la méthylation dans des cellules hépatocytaires mais à l'échelle épidémiologique, cela semble un peu plus compliqué.

Enfin, j'ai mené de manière indépendante mon étude GWAS et mon étude MWAS sans avoir intégré les résultats de l'une pour augmenter la puissance de l'autre. De nombreux travaux indiquent que les niveaux de méthylation d'un site CpG peuvent également être sous le contrôle génétique de polymorphisme(s). Il serait donc tout à fait intéressant de combiner les résultats que j'ai obtenus dans mes 2 projets pour essayer d'identifier des polymorphismes génétiques influençant les taux de TGP via des mécanismes de régulation de la méthylation d'ADN. Affaire à suivre...

TABLE OF CONTENTS

Index of Figures

Index of Tables

Introduction

Thrombin is a crucial protein that is mostly known for its central role in the haemostasis but that also takes part in a wide variety of other physiological functions, such as inflammatory response, cell differentiation and proliferation, vasoconstriction and cell adhesion. Anomalies in its activity, due to either lost or excess of function, can onset different diseases. Thrombosis is one of the most important ones able to trigger myocardial infarction, atherosclerosis, inflammation or stroke. To study any disease, it is highly recommended to have access to a proper, reliable and easy method to measure and detect as many phenotypic and symptomatic shades as possible in order to use it in routine clinical practice. In case of thrombin-related anomalies, the used procedures are composite measurements of coagulation protein concentrations, of platelet counts and of function/activity assays (e.g., time to clot formation). However, the majority of those methods are not specifically measuring thrombin or they get a reduced information of its activity ("scratch the surface") because they do not measure the total amount of thrombin activated in the coagulation reaction.

Recently, a measure called Thrombin Generation Potential (TGP) has been proposed to get a much closer representation of the *in vivo* thrombin activation all along the coagulation response of an individual. This measure brings the opportunity to better investigate the process of thrombin activation in plasma. TGP has already been used for studying thrombin-associated disorders in human and these works have confirmed that this measure is sensible to the already known thrombin-associated environmental and genetic factors. Being such a solid method, it is interesting to further study new factors that might influence thrombin generation and that could be later used in clinical diagnosis or prevention of diseases.

To contextualize my research, I will first start reviewing what is known about the thrombin protein, its functions and implications in human physiology. As it is the protein that is assessed by the TGP method, I want to stress its importance in human (also in animal) physiology and, hence, supporting the interest of the main subject of my PhD project: the identification of new genetic and epigenetic factors influencing the inter-individual plasma variability of TGP. I will also introduce some description about the technical method used for measuring TGP and summarize the main discoveries and works carried out so far on this subject. All along this document, I will refer to TGP as the measurement used for

investigating thrombin generation in a given individual. Hence, I might also define it as a phenotype because it characterizes a trait or state of the physiology of that individual.

After introducing these general descriptions, my work will be divided in two main parts: first, I will focus on the search for new genetic factors associated to TGP I have conducted in the framework of a genome-wide association study; second, I will analyze epigenetic data with the aim of looking for DNA methylation patterns associated to the TGP inter-individual variability in plasma.

The result of these works constitute my PhD project entitled: *Genetic and Epigenetic Determinants of Thrombin Generation Potential: an epidemiological approach*.

MOTIVATIONS AND BACKGROUND

Chapter 1. Thrombin

'The living enzyme of my blood' is how Walter Seegers, one of the pioneers studying thrombin, defined this protein ten years ago[1].

Thrombin, also known as coagulation factor II, is a 36000 Da serine protease composed of two chains: a light one (or A) and a heavy one (or B); unified by a covalent disulfide bound. His precursor, the zymogen Prothrombin (or FII), is synthesized in the liver and, once released, it is cleaved, exposing the domains and activating its functions (FIIa). Like others clotting factors, prothrombin is part of the vitamin K dependent family and is characterized by a γ-carboxyglutamic acid domain (Gla-domain) that depends on the union of the vitamin K and calcium (calcium-binding domain). Both molecules allow prothrombin to easily anchor into the cell membrane where it is functionally more active[2,3]. Other important domains for thrombin activities are the Sodium-binding site, and the exosites I and II[3–5]. The first is essential to modulate the ambivalent main functions. In normal physiological conditions, 60% of thrombin has this site occupied by Na+ making it more efficient to interact with pro-coagulation factors. On the other hand, without Na+, thrombin is more anti-coagulant. The two other thrombin epitopes are the binding sites of cofactors and proteins. In a general manner, the exosite I binds to fibrinogen, factor V, factor VIII, factor FXI, PAR1, thrombomodulin and protein C; while exosite II binds to heparin, glycoproteins and other cofactors that make thrombin more efficient in its functions[4,6,7] (Table 1.1.).

1.1. The functions of thrombin

The most fundamental and studied function of thrombin is his central and pivotal role in haemostasis: catalyzing the formation of Fibrin that will form the thrombus (or fibrin clot), activating other pro-coagulant factors and also triggering anti-coagulation enzymes[8].

Thrombin also participates, beyond the haemostasis, in other physiological mechanisms , such as the immune, nervous, gastrointestinal, and musculoskeletal systems. Thrombin interacts with proteins and receptors, activating platelets and endothelial cells, and stimulating cell adhesion, angiogenesis, cell growth and differentiation, proliferation, vasoconstriction and inflammation. All this wide variety of

functions is modulated by a complex combination of substrates, cofactors and their plasma levels[4] (Table 1.1. and 1.2.).

Table 1.1. Thrombin actions and substrates[3–5].

Thrombin's substrate	Cofactor	Thrombin-Domain	Function	Na^{++}/Ca^{++}
Fibrinogen	-	exosite I	Activation of fibrinogen -> fibrin	+
FVIII	-	exosite I & II	Activation of FVIII -> FVIIIa	+
FV	-	exosite I & II	Activation of FV -> FVa	+
FXIII	fibrin	exosite I	Plus fibrin activation of FXIII -> FXIIIa = stabilization of fibrin	+
FXI	GpIbα	exosite I	Activation of FXI -> FXIa = activation of intrinsic pathway	+
Thrombin activatable fibrinolysis inhibitor (TAFI)	TM		Activation of TAFI=TAFIa-> anti-fibrinolysis	
Glycoprotein V	GpIbα_exositeII	exosite I	Activation of platelets	
Thrombomodulin (TM)		exosite I	Physical inhibition to procoagulant actions	-
Protein C (PC)	TM (exosite I and sometimes II)	exosite I	Activation of PC resulting in: Inactivation of Va ->Vi and VIIIa->VIIIi	-
Heparin cofactor II (HCII)	glycosaminoglycans		Acts like antithrombin	
plasminogen activator inhibitor (PAI-1)			Inhibition of plasmin formation; fibrinolysis	
Glycoprotein GpIb-IX-V platelet receptor complex		exosite II	Cofactor. Helps to cleavage PAR and glycoprotein V by thrombin, FXI,.	
protease activator inhibitor (PAR) receptor present in almost all cell types		exosite I	<u>Healing tissue injuries</u> Activates platelet activation factor (PAF) Cell growth and differentiation Smooth muscle, macrophage and endothelial cell proliferation Angiogenesis <u>Pro-inflammatory</u> Leukocyte adhesion Lymphocytes mitogenesis Activation: IL-6, IL-8 <u>Vasoconstriction and vascular permeability</u>	

	procoagulant		Anticoagulant
	Activation thrombin		Thrombin-PAR functions

Table 1.2. Interactions where thrombin is the substrate.

Enzyme	Cofactor	Thrombin-Domain	Function
FX		-	First activator of thrombin from prothrombin. Exposure active domains
FVa-FXa complex		-	Activation of thrombin from prothrombin. Exposure active domains
Thrombomodulin (TM)		exosite I	Cofactor
Antithrombin (AT)	Non or glycosaminoglycans: heparin, heparan sulfate proteoglycans- (exosite II)	-	Without cofactor: AT inhibits thrombin reversibly. With a cofactor is irreversible. Also Inactivates of Xa -> Xi and IXa->Ixi;
Heparin cofactor II (HCII)	glycosaminoglycans		Acts like AT
Tissue factor pathway inhibitor (TFPI)			Fibrinolysis-anticoagulant

1.1.1. Thrombin in Haemostasis

Haemostasis is the main physiological regulatory mechanism of the blood flow and of its integrity. It includes both the formation of the thrombus (i.e coagulation) and its dissolution (i.e fibrinolysis) [9]. The dynamics of the coagulation can be explained in three phases[10].

1) The initiation of coagulation.
2) The propagation and amplification.
3) The termination.

This last step and the fibrinolysis are both the other pan of the haemostasis balance, dissolving the formed thrombus. Thrombin is key for haemostasis balance due to its paradoxical main functions. First, it triggers the formation of clots when the vessels are

damaged and then helps to stop the process to avoid their progress into the normal vasculature[4].

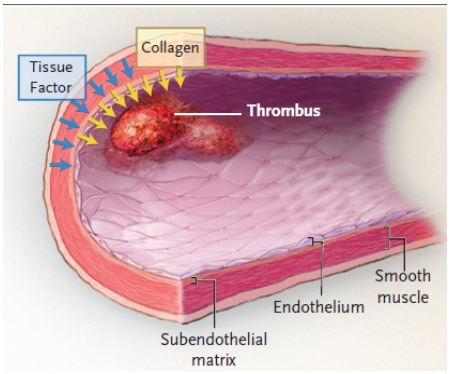


Figure 1.1. Image of a vein section where a thrombus is being formed[9]. The collagen and the tissue factor are components of the vessel wall located under the first layer of cells (endothelium). When the vessel is injured, collagen (yellow arrows) and tissue factor (blue arrows) will get in the blood flow and trigger the coagulation.

1.1.1.1. Blood coagulation: *Initiation. Platelet plug formation*

At the beginning, when a vessel is damaged, collagen and von Willebrand factor become exposed to the blood (Figure 1.1.) and attract the platelets to the injured point starting to build the clot. Both collagen and vWF activate platelets through glycoprotein (GP) membrane receptors such as the GPIIb-V-IX and GP1bα respectively[4,9], and that causes i.e. the releasing of P-selectins in the cell surface to help the cell-cell adhesion[11]. This step is called primary haemostasis. But this first plug of platelets is unstable and needs the fibrin fibbers to consolidate it.

On the surface membrane of these first aggregated platelets takes place the thrombin production and then its activity will contribute to transform fibrinogen into fibrin[4]. That is why the anchoring of thrombin to the cell surface through the Gla-domain and calcium binding domain is very important (page 5). Thrombin activation is the result of the coagulation cascade, a mechanism very efficient where many proteins are involved (Table 1.3.) and ones activate the following others (Figure 1.2.). The cascade can also be described based on two pathways: the intrinsic and the extrinsic. At the beginning, they were considered as equal part of the initiation process of coagulation. Nowadays, it is suggested that the coagulation initiation step would correspond to the extrinsic pathway and the intrinsic one would be part of the amplification and prolongation[12] (Figure 1.2.).

The coagulation cascade is triggered by the tissue factor (TF), which is a membrane receptor. It is normally present in extravascular tissues but when it is exposed to blood, meaning there is some kind of vessel damage, the coagulation starts[4]. Its first action is to activate factor VII (FVIIa, activated) becoming the TF-FVIIa complex that then activates two other factors: factor IX and factor X. The extrinsic and intrinsic pathways both lead to the

activation of FX, reason why FX is established as the connection between the two pathways and thereafter the common coagulation pathway starts[13]. FXa alone is scarcely efficient and only proteolyses a small amount of prothrombin to its active state thrombin, plus other secondary fragments: fragment F1+2 and meizothrombin[4]. Together, the meizothrombin and thrombin join to adrenergic receptors in the nearby smooth muscle cells resulting in a vasoconstriction of the area of clotting helping to stop the bleeding[7].

Table 1.3 List of the principal coagulation factors[14].

Factor	Current synonyms and/or former terminology
I	Fibrinogen
II	Prothrombin
III	Thromboplastin, tissue extract
IV	Calcium
V	Proaccelerin, labile factor, accelerator globulin (AcG)
VI	First used by Owren to describe what is now recognized as activated V (Va), but which was initially also referred to as accelerin. This numeral is now redundant and no longer used
VII	Proconvertin, serum prothrombin conversion accelerator (SPCA), stable factor, autoprothrombin I
VIII	Antihaemophilic factor (AHF), antihaemophilic globulin (AHG), platelet cofactor I, antihaemophilic factor A
IX	Plasma thromboplastin component (PTC), Christmas factor, antihaemophilic factor B, platelet cofactor II, autoprothrombin II
X	Stuart–Prower factor
XI	Plasma thromboplastin antecedent (PTA)
XII	Hageman factor
XIII	Fibrin stabilizing factor (FSF), Laki–Lorand factor (LLF), fibrinase

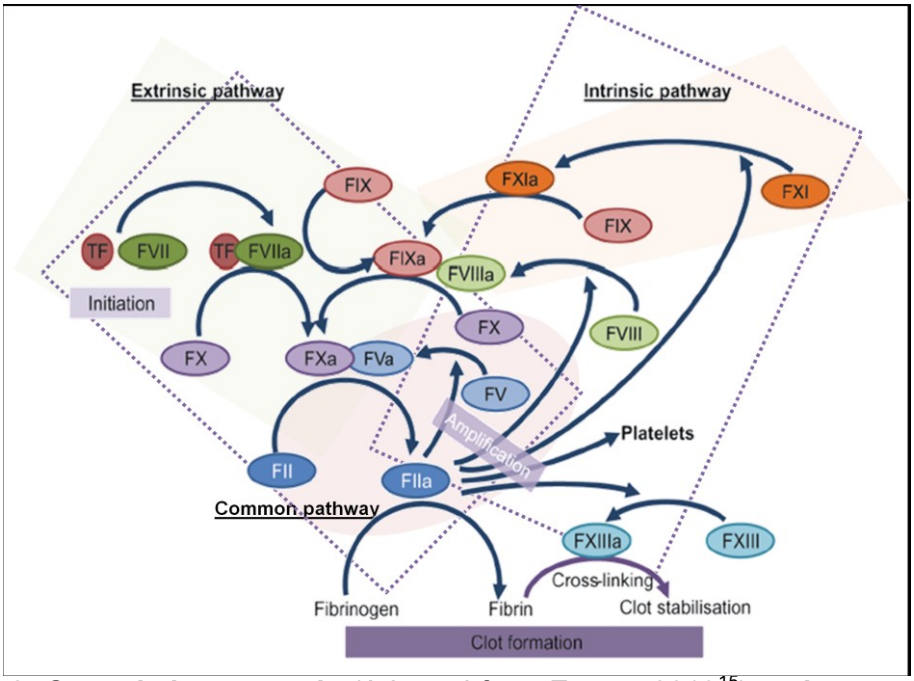


Figure 1.2. Coagulation cascade (Adapted from Esmon 2013[15]) **and comparison of the two nomenclatures**. The two different pathways, extrinsic and intrinsic, allow thrombin activation. Then, with dot-lined square (left) the initiation phase, which starts with the presence of tissue factor in the blood (possibly a vessel damage). Afterwards, with the apparition of the first amount of thrombin, the amplification phase (the right dot-lined square) is activated and helps to amplify its own activation (positive feedback).

1.1.1.2. Blood coagulation: *Propagation and amplification*

Once that small, but sufficient, amount of prothrombin is activated (2nmol/L), thrombin starts to interact with cofactors and substrates causing the burst of coagulation[1,3,4,9,15]. In this part of the process, the presence of sodium ions (Na^{++}) is essential, as it will give to thrombin the suitable allosteric conformation to be bound easily with pro-coagulant substrates. Besides, calcium, also called coagulation factor IV, is important for the activity of thrombin and different proteic complexes.

The most important thrombin's substrate is fibrinogen that is proteolysed to form fibrin units (Figure 1.3., arrow 1). These units are assembled in fibers that tangled the platelets and other necessary cells and molecules, and all together help to fix and stabilize the clot. Factor XIII is also activated (FXIIIa) by thrombin (Figure 1.3., arrow 2) and it acts cross-linking the fibrin fibers and helping to form the net of the thrombus (Figure 1.3., arrow 3).

Furthermore, thrombin proteolyses the factors V and VIII (Figure 1.3., arrows 4) that are both important to ameliorate the efficiency of the factors Xa and IXa, respectively. FVIIIa with FIXa form the IXa-VIIIa complex (tenasa complex) that activates more factor Xa. This conjugation makes FIXa 10^5-10^6 folds more active (Figure 1.3., arrow 5). Then, FXa combined with factor Va, become the Xa-Va complex (the prothrombinase complex). This complex also increases by $3 \cdot 10^5$ folds the thrombin activation[1,16] (Figure 1.3., arrow 6).

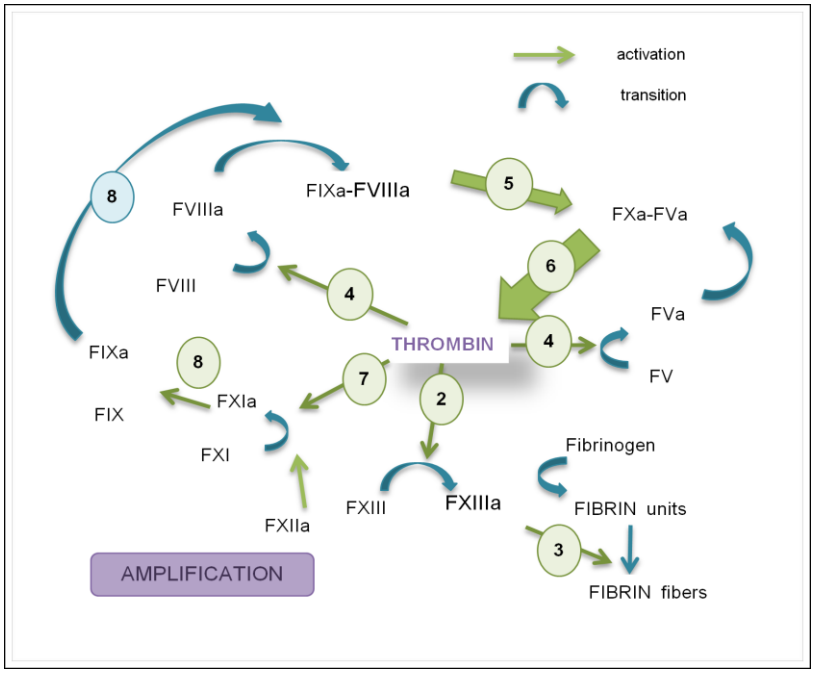


Figure 1.3. Amplification phase of coagulation. Scheme of the main thrombin activations to form fibrin and to increase its own formation (the burst). The 1st, 2nd and 3rd arrows are implicated in the fibrin formation; and the others arrows in the coagulation amplification.

Thrombin also activates factor XI, which becomes FXIa (Figure 1.3., arrow 7) helped by FXIIa whose activation *in vivo* is still controversial[9,17]. This part of the cascade is equivalent to the intrinsic pathway (Figure 1.2.), also known as the contact pathway (or even kallikrein/kinin system) because FXII *in vitro* is activated by the negative charges present in the surface of sampling tubes[8]. Once FXI is activated into FXIa, it triggers the coagulation activating even more FIXa (Figure 1.3., arrow 8). From here starts the common pathway again. FIXa complexes with FVIIIa to boost the activation of FXa and together activate more FXa (Figure 1.2., and 1.3.). This intrinsic pathway is very important to amplify the coagulation and the density of the fibrin. Deficiencies (mild or total) of FXI produce a very variable bleeding disease also called Haemophilia C[18]. Instead, FXII deficiencies are not associated to bleeding disorders, which makes it interesting as a possible anti-coagulation treatment[17].

Concurrently with all the reactions described above, thrombin activates more platelets and endothelial cells allowing increased cell aggregation by P-selectin and mobilization of more receptors to anchor more prothrombin[3].

1.1.1.3. Blood coagulation: *Termination*

A tightly and accurate regulation of coagulation is very important to ensure that it do not pass beyond to a physiologically normal vessel. Antithrombin, tissue factor pathway inhibitor (TFPI) and protein C are the main proteins controlling the amounts of thrombin generated. In this situation, a lack of sodium ions gives thrombin the ability to attract more anti-coagulant factors and cofactors.

There are two kinds of inhibitors: the indirect and the direct (Figure 1.4.).

Indirect inhibitors

The indirect inhibitors are those that cleave o inactivate the necessary proteins for the thrombin activation. These include TFPI, protein Z, ADAMTS13 and thrombomodulin.

TFPI is one of the most important and principal inhibitors of the coagulation[19] with big affinity for TF-FVIIa and FXa (Figure 1.3.). It is widely distributed in healthy arteries, in the endothelial cells, in smooth muscular cells and in macrophages, and is colocalized with TF[8]. It only takes place in early stages of the coagulation and TFPI inactivates the TF-VIIa complex and also the factor Xa (FXa) forming a TF-VIIa-Xa-TFPI complex. The latter blocks the formation of the efficient prothrombinase complex (FXa-FVa). TFPI has also been proposed to inhibit FXa, when it is complexed with protein S[20]. A dysfunction of TFPI

protein in the system is related to risk of thrombosis and disseminated intravascular coagulation[19].

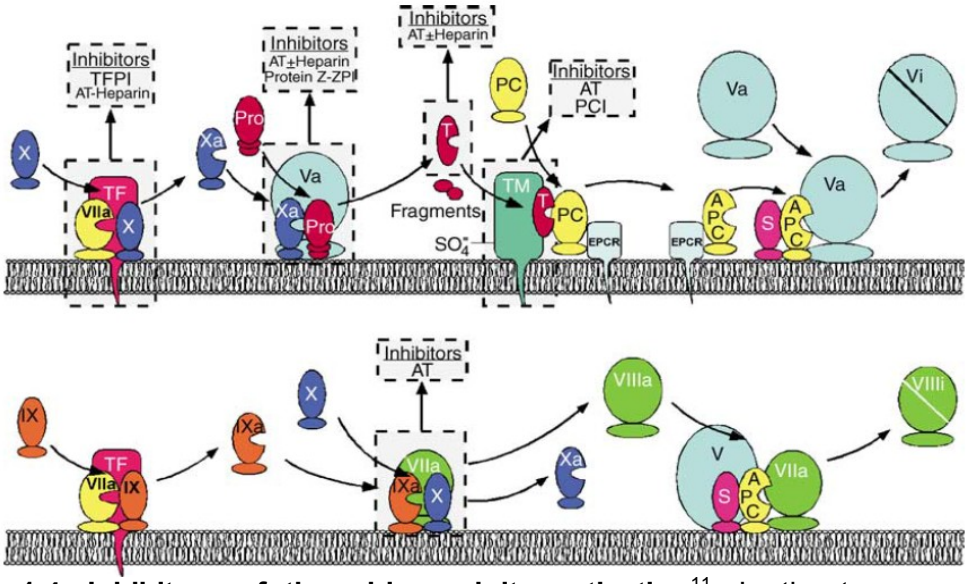


Figure 1.4. Inhibitors of thrombin and its activation[11]. In the two panels are represented some of the coagulation inhibitors: antithrombin (AT), Heparin, tissue factor pathway inhibitor (TFPI), the Z-Z protein inhibitor (Z-ZPI), and protein C (PCI or APC).

As an activator of thrombin, FXa is also the target of other inhibitor proteins such as antithrombin (see below) and serpin PZ-dependent protease inhibitor. PZ corresponds to protein Z, which is a member of the vitamin K-dependent family (FII, VII, IX, X, protein C and protein S) but has no proteolytic activity. However, it acts as a cofactor of the serpin PZ-dependent protein (Z-ZPI, or SERPINA)[11].

ADAMTS13 also acts as an indirect inhibitor in very early stages of coagulation. ADAMTS13 overbreaks von Willebrand factor and makes impossible its interaction with platelets, hence avoiding their aggregation. Mutations in *vWF* or *ADAMTS13* genes that can modify the efficiency of ADAMT13 for easily tagging vWF are responsible for bleeding disorders known as von Willebrand disease (vWD).

Finally, thrombomodulin (TM) is a protein located in the endothelium surface that connects to thrombin as a cofactor by the exosite I and anchors thrombin to the cell membranes. Therefore, it helps and increases the probability of union between thrombin (now T-TM) and the anticoagulant protein C (PC). PC also bound to endothelial cell protein C receptor (EPCR), making it easier the interaction[4]. The complex T-TM activates protein C (APC) and once liberated from its membrane receptor, APC can interact with its cofactor protein S, and together inactivate the factors Va and VIIIa to Vi and VIIIi (i for inactivate).

Direct inhibitors

The direct inhibitors are those who bind "directly" to free activated thrombin and inactivate it. These include thrombomodulin, again, together with antithrombin, heparin cofactor II and alpha 2 macroglobulin.

Thrombomodulin (TM) is additionally considered as direct inhibitor because it interacts with thrombin by the same domain as fibrinogen does, reducing the free thrombin. Hence, it is a physical inhibitor, avoiding the activation of fibrin, platelets and endothelial cells.

Antithrombin (AT) inactivates thrombin by cleaving it. This action can be reversible if AT acts alone but if thrombin is bound to a glycosaminoglycan molecule, for instance heparin, the attraction between AT and thrombin increases and the inactivation is irreversible. AT also inhibits the coagulation factors IXa, Xa and XIa stopping the thrombin generation. From the same AT family, heparin cofactor II (HCII) acts similarly binding to thrombin when there are glycosaminoglycans acting as cofactors.

Finally, alpha2 macroglobulin (a2M) is not really a specific thrombin inhibitor but better known for its relation with Alzheimer disease. It is a blood protease inhibitor that bounds to free thrombin that avoids thrombin functions[21].

1.1.1.4. Thrombin and *(Anti-) Fibrinolysis*

Part of the haemostasis, the fibrinolysis is the process that dissolves the thrombus. It starts few days after the formation of the thrombus. One of the main proteins in this process is plasmin, which breaks the fibrin net connections. Plasmin is activated by tissue plasminogen activator (t-PA) from its precursor, the plasminogen.

As already mentioned, thrombin also activates FXIII helping to stabilize the thrombus. Additionally, when thrombin is bound to thrombomodulin, it activates thrombin activatable fibrinolysis inhibitor (TAFI) that inhibits the t-PA, *ergo* TAFI inhibits the fibrinolysis[3,11,22]. However, when the complex thrombin-TM activates APC, it also promotes the fibrinolysis because APC inhibits the plasminogen activator inhibitor (PAI-1, SERPINE1). Thus it helps to produce plasmin, and that results with the dissolution of fibrin[3].

1.1.2. Thrombin and Inflammation

Inflammation is part of the protective mechanisms of the body as a response of the immunity system to pathogens. The relation between thrombin and inflammation is very

tight, is a cross-talk where they activate each other (Figure 1.5.). Inflammation and sepsis are defence mechanisms against bacterial spreading and lead their destruction.

Thrombin activates platelets by interacting with the protease-activated receptors (PAR). This stimulates the G-protein pathway which can result in the production of chemokines (as monocyte chemoattractant protein-1) that will attract leukocytes on platelets and endothelial cells. At the same time, platelets express protein P-selectin, CD40 ligands and binding receptors favoring the aggregation of those leukocytes. These cell-cell interactions produce the releasing of cytokines such as Interleukin-6 and 8 (IL-6, IL-8), tumour necrosis factor-α (TNFα) and growth factors (as vascular endothelial cell GF) by monocytes and endothelial cells. Thrombin also proteolyses complement components, e.g. activating C3 and C5. All these reactions together result in a inflammatory response[11] by thrombin and normally, it can be stopped by the action of anticoagulant factors such TFPI, APC and AT and also TAFI[11,23].

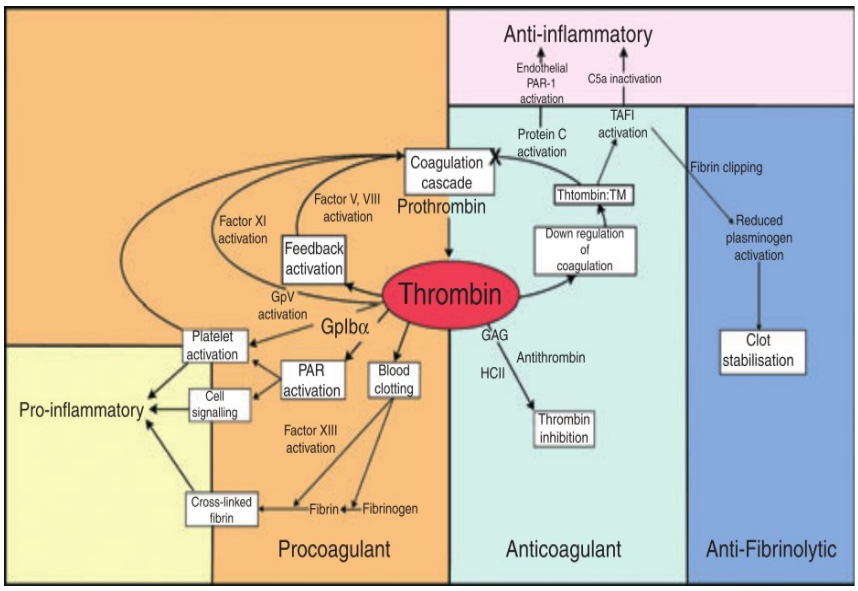


Figure 1.5. The Cross talk between the coagulation, fibrinolysis and inflammation responses[4].

1.1.3. Other thrombin functions

Since the discovery of PAR (PAR 1,3 and 4) in 1991 by Coughlin *et al.*[24], thrombin started to be considered more than a coagulation protein. These receptors are membrane proteins present in many cell types not only from the vascular and immune systems but also from nervous, gastrointestinal, and musculoskeletal systems (Figure 1.6.).

On the vascular cells, the thrombin-PAR (T-PAR) stimulus activates the protein G resulting in angiogenesis, cell growth and differentiation, and smooth muscle cell, macrophage and endothelial cell proliferation. T-PAR stimulates different growth factors

such as connective tissue growth factor (GTGF) and vascular endothelial GF (VEGF). GTGF stimulates the fibroblast mitogenesis and chemotaxis that produce procollagen and fibronectin[3]. VEGF affects the endothelial cells inducing cellular migration, endothelial cell proliferation, and vascular tube formation promoting angiogenesis. These responses permit to heal and repair the damaged tissue of the vessel. Besides, VEGF also promotes the generation of more thrombin.

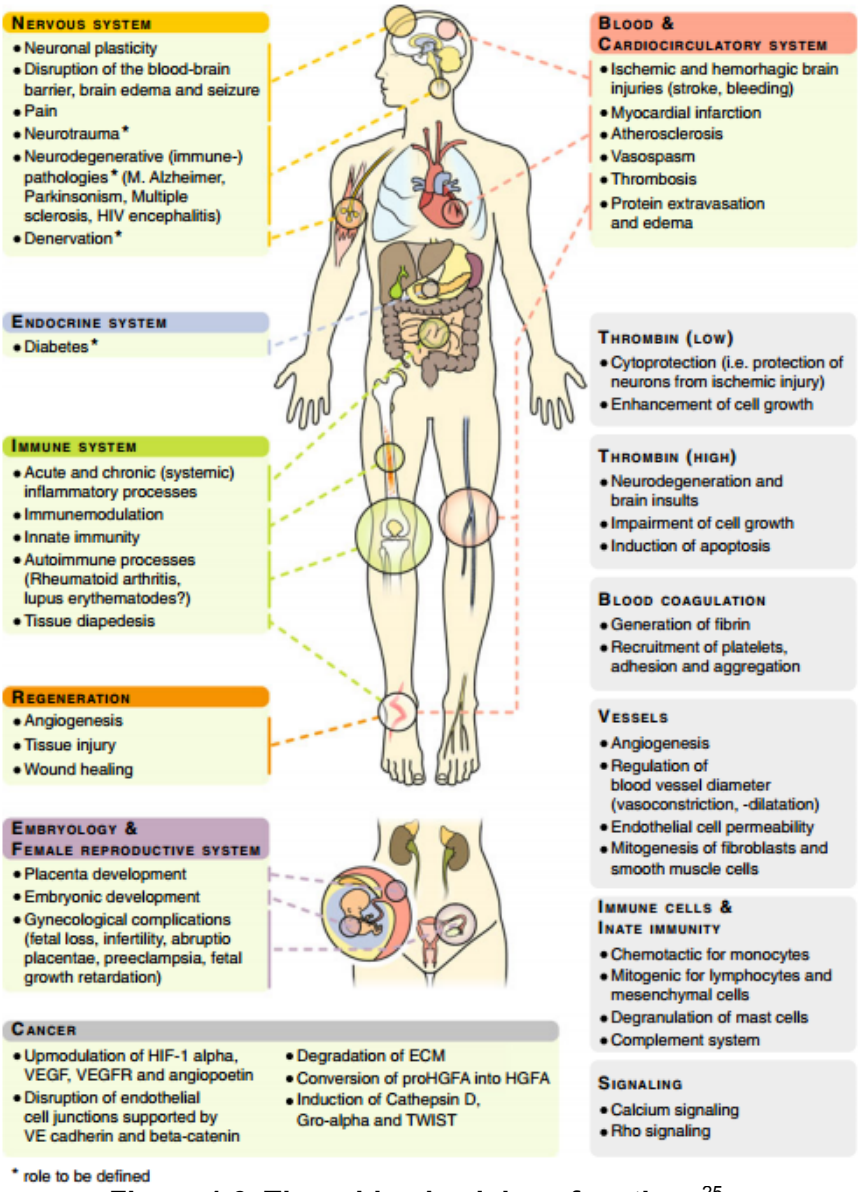


Figure 1.6. Thrombin physiology functions[25].

Additionally, T-PAR interaction stimulates the secretion of nitric oxide (NO). On one hand, NO produces vasoconstriction at the site of the injury contributing to stop the loss of blood. On the other hand, it leads to vasodilation of the normal endothelium to permit the blood flow beyond the injury[3].

In the musculoskeletal system, T-PAR promotes osteoblasts, myoblasts and chondrocytes proliferation, which might be important in cases of bone injuries and fractures. Thrombin also has been found highly expressed in neonatals and embryos having important influences in muscle denervation and during brain development[3].

Neurons, astrocytes and olygodendrocytes express also PAR and release thrombin. Their interaction leads to brain development, synaptic plasticity and neuroprotection[3,26].

Every cell that express PAR, could be stimulated by thrombin, therefore there could be still many thrombin actions to investigate. Although PAR can be activated by other proteases (e.g. FVIIa or FXa), thrombin seems to be the most efficient one[3].

1.2. Related diseases and disorders

For all the functions explained above, a careful thrombin regulation is fundamental for normal physiology and unbalanced levels of thrombin result in different kind of disorders. The most known and studied are: thrombosis, haemophilia, inflammation and atherosclerosis.

1.2.1. Thrombosis and haemophilia

Regarding the main functions of thrombin, two diseases are related to excess or deficiency of thrombin: thrombosis and haemophilia (respectively).

Thrombosis is a very critical disorder that consists in a tendency to form thrombus, increasing the risk of vessel obstruction and interruption of the blood flow. There are two kinds of thrombosis: arterial and venous (Figure 1.7.). The arterial thrombosis is characterized by an excess of platelet activation, generally due to the rupture of an atherosclerotic plaque, whereas venous thrombosis is more related to an excess of activation of the clotting factors and/or decrease of the blood flow[27].

Therefore, all factors that increase the normal amount of thrombin in plasma may increase the risk of arterial and venous thrombosis when mutated[8]. Some of the most well known factors are blood group and the gene mutations: prothrombin G20210A or FV Leiden. However, the risk can be also caused by: up-regulation of FVII, FX, FV, FVIII, FXI, FXIII, TAFI and logically FII; and/or down-regulation of PC, PS, TM, AT, TFPI among others[19,22,28,29]. Total deficiencies in these latter proteins are either incompatible with life or cause bleeding disorders such as haemophilia[30].

Haemophilia is a bleeding disease caused by a decreased capacity to form a thrombus. The most known causes are the decreased level of factors VIII (Haemophilia A) and FIX (Haemophilia B) in plasma. Additionally, deficiency of FXI (Haemophilia C) and von Willebrand disease (vWD) can also induce bleeding disease, however they are less known and also more variable in phenotype[31].

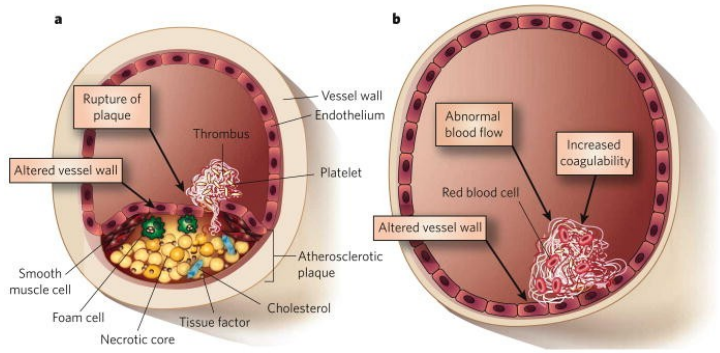


Figure 1.7. Arterial vs. venous thrombosis and their different possible causes[18]. a) is an example of arterial thrombosis and b) a case of venous thrombosis.

1.2.2. Inflammation diseases

As mentioned before, the relation between thrombin and inflammation is very tight. Actually they are mutually reinforcing processes[9]: increased thrombin plasma levels can produce inflammation, and decreased levels act as protective phenotype in inflammation.

In turn, inflammatory cytokines and the TNFα are related to more platelet reactivity enhancing coagulation by stimulating the release of tissue factor. Additionally, TNFα downregulates the action of thrombomodulin and endothelial cell protein C receptor. In acute phase responses, the complement C4b binding protein (C4bBP) levels increase and it is very often found in complex with the protein S. That results in a PS deficiency and a decreased function of APC[32,33].

In some inflammatory diseases or infections (e.g. sepsis) there are symptoms of thrombosis. For instance in *asthma*, characterized by a reversible obstruction of the airflow, chest tightness, dyspnea and cough, coagulation in the airways is induced[34].

1.2.3. Atherosclerosis

Atherosclerosis is a disorder characterized by the formation of cholesterol plaques in the vessels of the arteries. Besides the cholesterol, the plaque is composed by platelets, macrophages (which become foam cells after endocytosis of the deposited lipids), neutrophils, T and B cells, endothelial cells, and smooth muscle cells (SMC). All of them contribute to the development of the plate.

The components sensitive to thrombin are the platelets and macrophages, which express PAR in their membranes. These interactions reinforce the atherosclerotic plates, although the mechanism is unknown[35].

The different components of the atherosclerotic plaques have also an influence on thrombosis[8]. Actually, there is a close interaction between atherosclerosis, inflammation and thrombosis (Figure 1.8.), as thrombosis can stimulate both mechanisms. Inflammation contributes from early to final phases of the plaque: formation and rupture. Leukocytes and platelets express aggregation proteins and cytokines allowing both the accumulation of platelets and the production of tissue factor. The SMC inside the atherosclerosis plaque produce collagen. When plaque rupture occurs; TF and collagen are released into the vessel lumen, triggering thrombosis[36]. Additionally, the plaques in the vessels physically complicate the blood flow, occupying, in the worst case scenario, more than half of the lumen.

Figure 1.8. Cross-talk between atherosclerosis, inflammation and thrombosis[36].

1.2.4. Cancer and other related disorders

Besides the previous diseases, thrombin has been related to some other diseases, for instance cancer or central nervous system (CNS) anomalies.

The capacity of thrombin to stimulate the proliferation, aggregation and mobilisation of cells, makes it a "potential collaborator" for tumours. Thrombin promotes angiogenesis by activating endothelial cells (T-PAR interaction) and stimulating the expression of growth factors. T-PAR also contributes to the expression of integrins (membrane proteins), influencing both the attachment and the mobilization of cells, hence in a cancer environment, to metastasis. For instance thrombin has been related with colon adenocarcinoma and up regulations of PAR with breast, prostate, melanoma and other cancers[3].

In CNS, abnormal levels of thrombin seem to contribute through different mechanisms to stroke by: vascular disruption, inflammatory response, oxidative stress, and cellular toxicity[26]. Moreover, high levels of thrombin and PAR are related to low survival of astrocytes or neurons whereas low levels can be neuroprotective[3].

There are also disorders that contribute to thrombotic stages, e.g., obesity, liver diseases, diabetes and renal disorders. The first one it is also implicated in inflammation responses and probably atherosclerosis. The second one is related to haemostasis because the liver tissue synthesizes the majority of coagulation factors. Thus, in case of cirrhosis, infections or malfunctions, thrombin production and activity, thus hemostase, may be affected. Then, patients of type 2 of diabetes mellitus (T2DM) with insulin resistance tend to have enhanced coagulation activation and hypofibrinolysis. This may be due to a high plasma concentrations of procoagulant proteins and decreased anticoagulants ones in T2DM patients[37]. Finally, renal disorders have been shown to predispose to thrombotic disorders or uremic bleeding although the exact underlying mechanisms remain unknown[38].

1.3. Thrombin tests and measurements

Given the importance of thrombin in coagulation and other physiological processes, it became essential to measure its activity as accurately as possible to study and understand the molecular and physiological consequences of its abnormalities. So far, the routine tests used are:

1.3.1. The prothrombin time (PT)

The prothrombin time is a measurement created by Quick in 1935 to fill the need for quantification of coagulation defects in new-borns and jaundice symptoms[39]. The test consists in the activation *in vitro* of coagulation system in a sample of plasma after incubation with thromboplastine reagent (tissue factor) and $CaCl_2$. The time between the trigger and the clot formation is recorded in seconds. This measure permits to study the deficiency of factor II (prothrombin), V, VII, X and fibrinogen. Normally, this test is performed in combination with activated partial thromboplastine time.

1.3.2. The activated partial thromboplastin time (aPTT)

Langdell designed this coagulation measurement also in 1935, and modified it *a posteriori* in 1961[39]. It consists in the induction of plasma coagulation by adding an activator of factor XII and a platelet substitute. Then the time that takes to get the clot is measured.

This test allows detecting abnormalities in factors XII, XI, IX, VIII, X, V, prothrombin and fibrinogen.

Both aPTT and PT tests are regularly used to monitor the anticoagulant treatment.

1.3.3. The activated clotting time (ACT)

The ACT measurement uses whole blood samples taking into account the importance of platelets and phospholipids in the role of coagulation[39]. The triggering substance to activate the coagulation is a solution containing glass beads, celite or kaolin. All three have large contact surfaces allowing "the contact" pathway activation, thus the FXII activation. Then the time to form the clot is recorded. The test is also very easy and quick, it requires a small blood sample, and can be performed just before extraction by a non-trained person.

1.3.4. Prothrombin fragment

Prothrombin fragment 1+2 (F1+2) is a specific molecular marker of thrombin generation in the body. Circulating levels have previously been reported to be a marker of both acute venous and arterial thrombotic events and also to be a predictor for new thrombotic events and mortality[40].

These conventional coagulation tests just reflect 5% of the total process[16,30,41] of thrombin activation and activity. Despite they are the affordable tests to measure this activity and informative for haemostasis defects, they may not be the best way to study thrombin activity, not detecting some dysfunctions and disorders as some kinds of thrombosis[42,43]. As Professor HC Hemker, one of the most important researchers in thrombin, wrote: "If you want to know about thrombin- measure thrombin"[44]. Fortunately since 2002-2003 there is an optimal model of haemostatic function which permits to measure the thrombin activity being the most closed to *in vivo* coagulation: the thrombin generation potential (TGP).

Chapter 2. Thrombin Generation Potential

Thrombin generation potential (TGP) is a measure of the potential amount of thrombin that is possible to activate[45] in a coagulation reaction, reflecting all the capacity[46] rather than a normal coagulation[30]. Also, it can be expressed as the representation of the coagulation process from initiation, propagation-amplification and termination[47].

Depending on the interest of each study, it can be performed in platelet poor plasma (PPP) or platelet rich plasma (PRP). The first one reflects any coagulation factor deficiency except FXIII, activators (e.g., FIX, FV, FX) and inhibitors (antithrombin, protein C, protein S), thus many kinds of clotting disorders specially APC resistance. Additionally it is sensitive to all kind of anticoagulants or other drugs[40]. The second one adds the influence of platelets in the coagulation process, the vWF platelet-activation function, other thrombotic disorders and it is sensitive not just to anticoagulants but also to anti-platelet aggregation drugs[45,48].

2.1. Short history of the thrombin generation measurements

The thrombin generation test is not a new technique although, it has been widely used in the last decade. The first reported experiments, in which TGP was studied, dated back to 1901 by M. Arthur who tested dog defibrinated blood samples. The technique consisted in the incubation of the samples, then taking aliquots at regular intervals and introducing them into a known amount of fibrin to measure the time to clot formation. The more thrombin was activated, the less time to clot formation. Then, the data was plotted as "concentration of thrombin" versus the time[49] obtaining a curve of thrombin generation.

The first time this technique was applied using human blood was in 1939. However the protocol was still being modified (components and solutions), not finding a simple and reliable test[50]. The modifications continued like in the works of Pitney *et al.* 1953[51], Macfarlane et al 1953[49]. In 1993, Hemker finally proposed a continuous registration of thrombin generation using optical lecture with a chromogenic thrombin substrate to measure the real time coagulation[52]. He proposed a continuous registration of thrombin generation using optical lecture with a chromogenic thrombin substrate to measure the real

time coagulation. Additionally, the substrate had to be slow-reactive to avoid his depletion before the end of the reaction. Before the use of the chromogenic components, the tests were applied in PPP, PRP or whole blood, defibrinated or just diluted. However, this new method reduced the option to PPP defribrinated samples without background noises caused by other bodies or molecules.

Posteriorly, Ramjee proposed in 2000 to substitute the chromogenic substrate for a fluorogenic, permitting to test again all kinds of samples[53]. Besides, it brought the possibility of measure in 96-well plates, thus 96 samples at the same time that notably reduced the time of measure.

Finally, in 2003, HC Hemker improved the protocol and the technique by developing the calibrated automated thrombogram (CAT)[45] based on the fluorochrome substrate. The innovation was the additional calibration test measured with the sample at the same time, which consists in a known amount of thrombin consuming also the fluorogenic substrate. This complementary test permits to correct the bias between the fluorochrome signal and the thrombin activation, which are not constantly proportional as the fluorochrome signal increases. This correction permits to convert the fluorescence into thrombin concentration[45].

2.2. Method of measurement: calibrated automated thrombogram (CAT)

As mentioned before, there are different approaches to assess the generation of thrombin. However, the most recently used, the one also used in this work, is the CAT, also called thrombin generation potential (TGP). The following description of the method is based on different reports: Hemker *et al* 2002[48], Hemker *et al*. 2003[45], Lavigne-Lasside *et al.* 2010[54] and Castoldi *et al* 2011[55]. This method consists in two simultaneous measurements in the same plasma sample. One will be called the measurement or thrombin generation well (TG) and the other the calibration well (CL). The basic necessary solutions and components of this experiment are:

- **Plasma** - As mentioned before, the plasma can be rich or poor in platelets. Generally, blood samples are extracted from antecubital venipuncture. Then a PRP is obtained from the supernatant (first visible liquid phase) of the sample after centrifugation at 265g for 10 min. The platelets are counted to check that the sample has less than 200.000 platelets/µL, the threshold for the method sensitivity. If needed, the sample has to be diluted with the corresponding PPP. For an optimal experiment, this type of plasma has to be used in no more than one hour after extraction to avoid biases

caused by the platelet "aging". For a PPP, the sample is centrifuged twice at 2900-3000g for 10 min. All platelets precipitate in a pellet to the bottom.

- **Buffer base** - The buffer base solution used to dilute and mix the other components. It is formed of: 20 mM Hepes, 140 mM NaCl, 5 g/l bovine serum albumin (BSA), in a pH = 7.35.
- **Trigger buffer** - The triggering buffer is a solution that will activate the coagulation. It is composed by: in case of PRP: just 3 pM recombinant tissue factor (rTF) and buffer base, and in case of PPP:30 pM rTF and 24 µM phospholipid vesicles in the buffer base.
- **FluCa** - It is the solution containing the fluorescent substrate: Z-Gly-Gly-Arg-AMC. Thrombin proteolyses this molecule releasing AMC, which will emit fluorescence at a 390 nm wavelength of excitation and a 460 nm of emission. It is formed by: 2.5 mM of fluorescent substrate, 60 mg/mL of BSA and 0.1 M CaCl2 in buffer.
- **Calibrator solution** - It is the solution used to calibrate and correct the TG measurement. It consists in thrombin complexed with the non-specific inhibitor α2-macroglobulin (α2M-T), which will "protect" thrombin from specific anticoagulants and minimize thrombin activity. It is a mixture of prothrombin, α2-macroglobulin, FX, FV and phospholipids vesicles in Hepes buffer that will react and form the α2M-T. The amount used in the reaction is 600 nM of α2M-T.

The support used is a 96-well specific fluorometer plate with 390/460 filters, excitation and emission wavelength respectively. That permits to test 48 samples in one plate. For each plasma sample, the two mixtures corresponding to TG measure and CL measure (Figure 2.1.).

In the TG test, thrombin is generated in a coagulation reaction activated with the trigger buffer. This thrombin is measured at real time by capturing the fluorescence signal that emits the substrate consumed by thrombin. This signal is captured and displayed in a curve (Figure 2.1.). On the other hand, the calibration test has no coagulation reaction but just the activity of the α2M-thrombin consuming the fluorescent substrate. The resulting curve of the calibrator (from a known amount of thrombin) will correct the bias between the fluorescent curve and the thrombin curve. The dedicated software program used is the Thrombinoscope (Synapse BV, Maastricht, The Netherlands). The program reads the two tests at the same time and automatically corrects the thrombin generation measure. As a result it returns a TG curve of nM of thrombin per min called thrombogram.

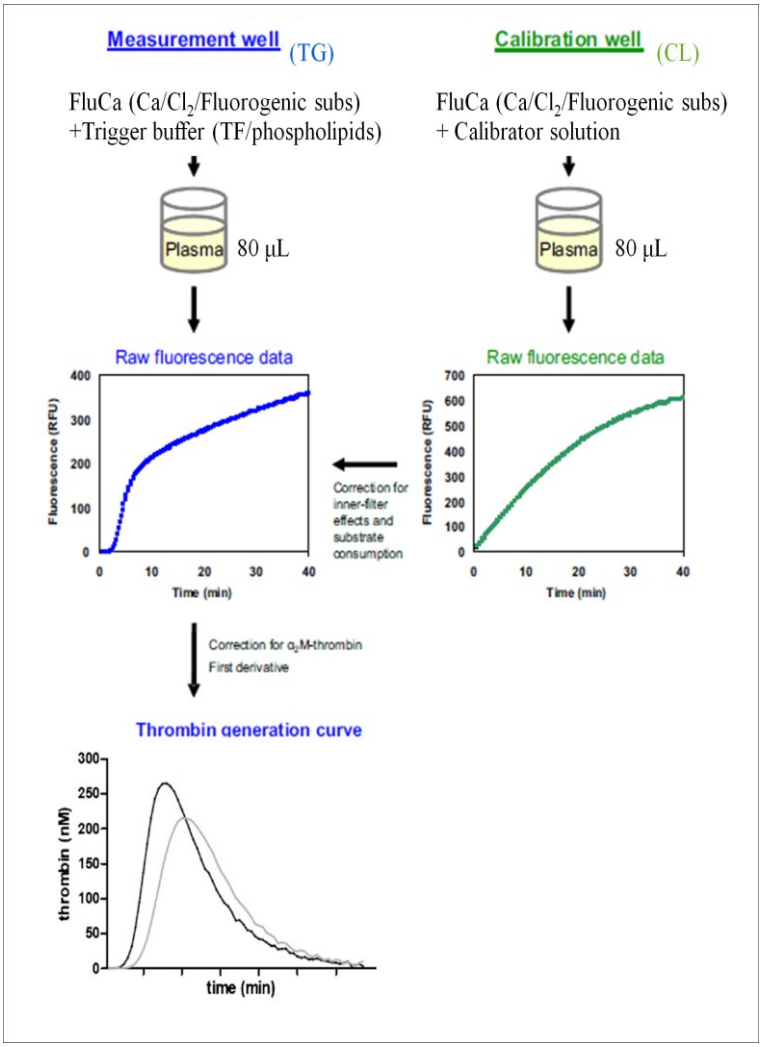


Figure 2.1. Diagram of the CAT method measures for a plasma sample. Adapted from of *Castoldi et al.* 2011[55]. The measurement or thrombin generation (TG) and the calibrator (CL) are simultaneously tested and combined to display the final corrected thrombin generation curve.

Variations of the test

Some variations can be applied changing or testing different concentrations of tissue factor and phospholipids in the reaction. Then, it is also possible to add to the mixture some anticoagulants such as APC with TM and AT to better study their pathways (Figure 2.3). Additionally, it is recommended for some studiesto inhibit the intrinsic pathway by inhibiting FXII with Corn trypsin inhibitor (CTI) and to just measure the intrinsic pathway and their coagulation factors. Finally, it seems that residual platelet can influence the measure; in order to avoid so Chantarangkul *et al.* proposed to use a greater concentration of phospholipids (> 1.5 nM)[56].

2.3. The parameters (biomarkers)

From the thrombogram it is possible to extract several quantitative and biological parameters. In this work they are also referred as biomarkers (Figure 2.2.).

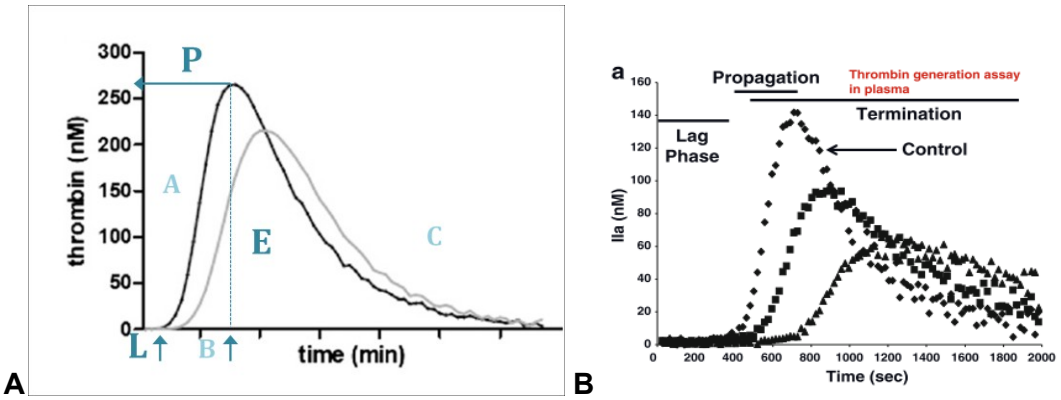


Figure 2.2. Parameters of Thrombin Generation Potential measured by thrombogram. A- (L) is Lagtime (min), (P) is Peak height (nM), (E) is Endogenous thrombin potential (ETP) (nM x min), which are the main parameters used in this project. Then, (A) maximal rising slope (nM/min), (B) time to peak (min) and some also consider (C), start time of tail (min). B- the comparison of the thrombin generation curve with the coagulation cascade[57].

Although the most used are Lagtime, endogenous thrombin potential (ETP) and Peak, there are also other available parameters as: time to peak, velocity and start time of tail.

The Lagtime (L)- It is an equivalent measure of clotting time, as it is the time passed from the moment that the coagulation is triggered until the clot start to form (fibrin). This last point has been arbitrarily proposed as the moment when 10 nM of thrombin are produced[45]. It is measured in minutes (min).

Endogenous thrombin potential (ETP) (E)- It is the area under de curve (AUC) that represents all the potential enzymatic labour of thrombin when is activated. Therefore, it is the parameter that will better represent the state of coagulation. It is measured in units of nanomolar (NM = nmol/L) for the time in minutes (nM·min).

Peak (P)- It is the maximum amount of thrombin activated in all the process. It is highly correlated with ETP. It is measured in units of nanomolar (nM or nmol/L).

Time-to-peak (TTP) (B)- It represents the time that takes to get the peak of thrombin activity from the moment the coagulation started (by the trigger). It corresponds to the initiation phase plus the amplification phase of coagulation. It is also measured in minutes as Lagtime.

Velocity (A)- It is the slope from the first thrombin activity until the peak, representing how fast the reaction reaches the maximum of thrombin activity (nM/min)

Start time of tail (ST)(C)- It is the moment of time (in minutes) when the activation of thrombin decreases reflecting the start of termination phase and the activity of the anticoagulants.

As a general example, a hypercoagulation state is characterized by a small lagtime and large ETP and Peak. Conversely, hypocoagulation is association with delayed lagtime and decreased ETP and Peak.

2.4. Known biological, environmental and genetic determinants of TGP

TGP can also be defined as the collective activity of all components that participate in the coagulation cascade and the influencing pathways. Hence, it is suitable that this phenotype would be determined by different factors: genetic and environmental. Since the publication of this method, due to its close similarity to an *in vivo* coagulation and its medical relevance, numerous works have been conducted to detect its sensitivity to deficiencies, disorders, and mutations or to study its associations with other biological and clinical characteristics.

Besides the levels of thrombin itself, the test is also sensitive to variations of the cascade factors: FX, FIX, FVII, FVIII, fibrinogen andD-dimer; and to anticoagulants: TFPI, AT, PS and PC[20,21,54,58–60] (Figure 2.3.). Any mutation or disorder increasing or decreasing thrombin levels or functions influence the final curve of thrombin generation as an indirect implication.

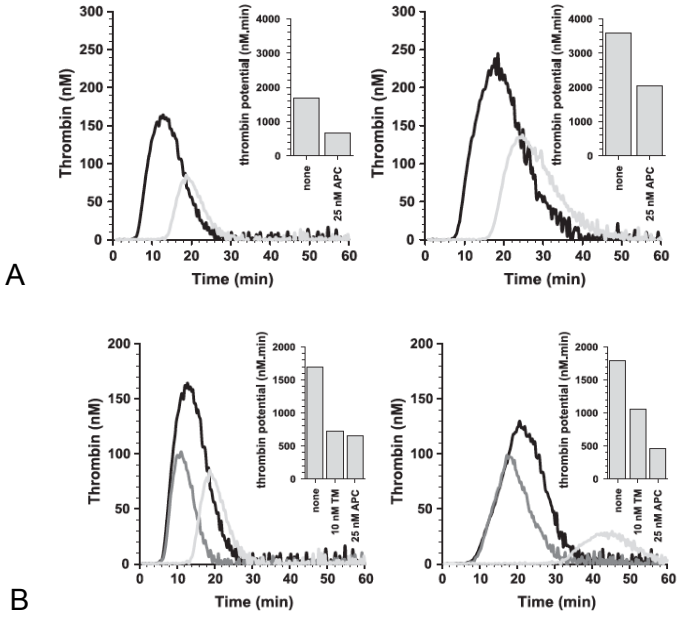


Figure 2.3. Comparison of thrombin generation curves of (left) controls, (right A) an individual with antithrombin deficiency and (right B) an individual with deficiency of protein C.A) In black the basic thrombin generation test, and in light grey, the test adding 25 nM of active PC (APC). The differences between a healthy individual and an individual with deficiency of AT are visible. B) The grey line test with a variation, the addition of 10 nM of TM and the light grey the test with 25 nM of APC. The right curve, in comparison with controls, is a slight prolongation of the peak before decrease, whereas the controls quickly decrease. It displays also the method differences between the addition of TM or APC.

TGP is also associated to body mass index (BMI) presenting longer Lagtimes but also higher ETP[60,61]. Age can also be a determinant, presenting differences between children, young and old adults[58,60–62] (Figure 2.4.).

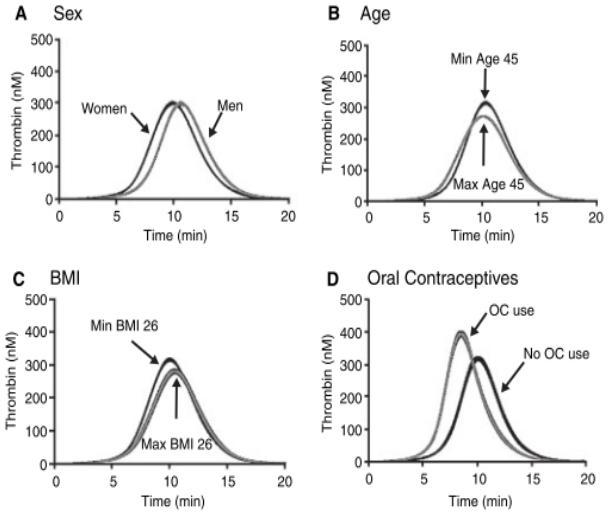


Figure 2.4.: Comparison of thrombin generation curves by sex (A), age (B), BMI (C) and oral contraceptives (D)[61].

Another major determinant of TGP variability observed is the gender[63] (Figure 2.5.). Women are more prone to coagulate: lower Lagtimes and increased ETP and Peak. Also the levels of free PS and TFPI are lower in females, also reflected in higher ETP and Peak in the measurement with the addition of APC [58]. There are also slight evidences of thrombotic changes reflected by TGP variability across the different phases of the menstruation cycle [64]. Besides, women often have more risk to develop thrombopathologies due to the use of oral contraceptives or hormonal postmenopausal treatments[21]; always adding more risk when the treatments are based on estrogens and oral prophylaxis. TGP can be enhanced in women with treatments, and also in tests performed with the addition of APC[45,61,65].

Other studies have been made to search for environmental factors that could also modulate the TGP curve. For example, a small study investigated whether alcohol has any effect on TGP, but no associaiton was observed[66]. Additionally another study presented some preliminary experiments aiming at finding whether diet could influence TGP levels[67].

Last but not least, two genetic factors were known to influence TGP when I started my PhD project: the rs1799963 (20210G>A) and the rs3136516 (19911A>G) from the prothrombin gene (FII)[68–70]. The rs1799963-A is present in 2% of the population and well established as a mutation in the extreme 3'-UTR since 1996[71]. It is located in the exact start point of the poli-A chain, of the gene and related to increased levels of prothrombin[69,70,72]. However the exact mechanism by which this mutation increases prothrombin levels

remains uncertain. It might be due to a better stability of the mRNA, a more efficient transcription or both. This mutation has been also related to risk of venous thrombosis, heterozygous carriers being at 2-3[71,72] and even a 7[69] fold increased risk and homozygous carriers at 11 fold risk[73].

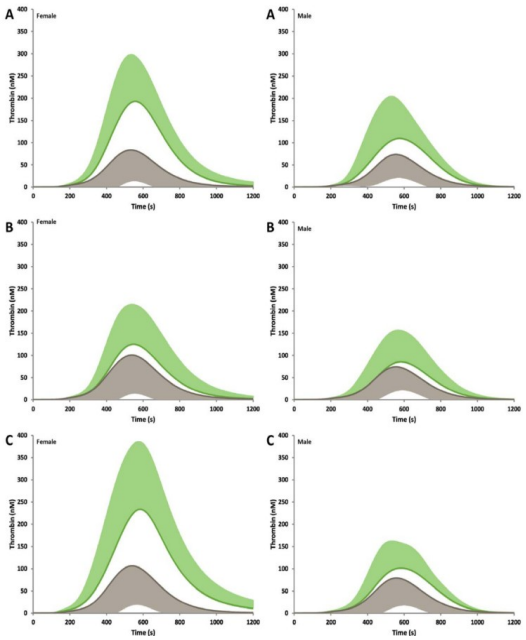


Figure 2.5. Thrombin generation in females vs males[63]. In A thrombin generation curve from a family of individuals with a mutation in protein C (green) versus the non carriers (grey). In B with a prothrombin mutation (with the same colour code). In C individuals with a thrombosis history.

The second F2 polymorphim known to influence TGP, rs3136516, is located in the last intron of the gene (the 13th intron). This polymorphism is very frequent in the population, with minor allele G frequency of almost 50%. This allele is associated with increased prothrombin levels (and then to TGP) in a dose-dependent manner[68]. The nucleotide change, originates an enhancer domain of splicing but has not been proved that is the cause of the increased levels[69]. The association of this polymorphism with the risk of venous thrombosis is still a matter of question. Some studies observed so[74,75] whereas others did not find associations with this polymorphism[76–79].

TGP is sensible to these two genetic variations resulting in increased ETP and Peak[6,54]. Both polymorphisms explain an 11.3% of the variability of TGP (calculated with the populations used in this project, see Chapter 5, page 53).

2.5. TGP related diseases

The levels of TGP have also been related to some diseases or risks, accordingly to the previous section and the thrombin implications in some diseases (page 15).

The levels of TGP, especially increased ETP, have been associated to venous thrombosis[41,80–82], acute ischemic stroke[83] and myocardial infarction[35,84]. It has also been

found very useful in monitoring anticoagulant treatment[85], e.g. after an acute myocardial infarction[40]. In contrast, lower levels of ETP are associated with bleeding disorders[41,46,86] (Figure 2.6.).

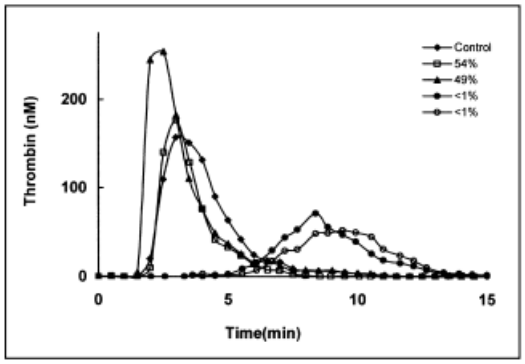


Figure 2.6. Thrombin generation curves from individuals with different levels of FXI deficiency[86]. Each line represents a different grade of heamophilia C indicated in the legend (top right corner).

TGP has been related to other diseases as well including atherosclerosis[87], obesity[60], type 2 diabetes[88], diabetic nephropathy[89], or hepatic disorders as cirrhosis[90]. Inflamatory-related disorders also have been associated to TGP as Crohn's disease[91], sepsis [92]and sickle cell disease[93].

2.6. Main Objectives

The thrombin generation potential has demonstrated so far its capacity to characterize the individual coagulation state and to detect possible disorders or deficiencies compared with other routine tests. TGP demonstrates a strong inter-individual variability that is mainly explained by environmental factors an, the main genetic factors observed to date are the mutations in the prothrombin gene in chromosome 11. This PhD project was proposed to identify more genetic factors that modulate this phenotype and its biomarkers.

The aim of my PhD project was twofold:

1 - To identify new genes that could influence plasma TGP variability. For this, I conducted a meta-analysis of two French genome-wide association studies (GWAS) on TGP biomarkers.

2 - To investigate whether DNA methylation marks could also contribute to the inter-individual variability of TGP plasma levels. For this, I conducted a meta-analysis of two methylome-wide association studies (MWAS) on TGP biomarkers from DNA methylation measured in peripheral blood DNA.

PARTI

IDENTIFICATION OF NOVEL GENETIC DETERMINANTS CONTROLLING THE INTER-INDIVIDUAL PLASMA VARIABILITY OF THROMBIN GENERATION POTENTIAL

This part contains the achievement of the first objective of my PhD project, the identification of new genetic determinants of the variability of Thrombin Generation Potential (TGP).

The adopted strategy was to perform a genome-wide association study (GWAS) (definition in page 36) of three TGP biomarkers in two independent studies. Then the study-specific results were combined into a meta-analysis to obtain candidate single nucleotide polymorphisms (SNPs) with strong statistical evidence for association with the biomarkers. These candidate SNPs were tested afterwards for replication in two additional independent cohorts.

In the next pages, I will first describe the main principles and procedures underlying the application of a GWAS strategy and the different cohorts I have used for the main analysis. Then, I will describe the results I have obtained using this GWAS approach that are summarized in my first publication (Rocanin-Arjo *et al.* 2014[94]) attached at the end of this chapter.

Chapter 3. Genome-Wide Association Studies: concepts

The main objective of this project has been the study of inherited determinants that may influence thrombin generation potential in an epidemiological context. That means to study the trait (or disease) in human populations and find their genetic patterns and causes, occurrence and distribution to better understand and "control" the health of the population[95]. These traits can be quantitative such as: protein levels, weight or amount of adipose tissue; or discrete like: sick / healthy. For decades, there have been two main research strategies in the study of genetic factors: linkage and association analysis.

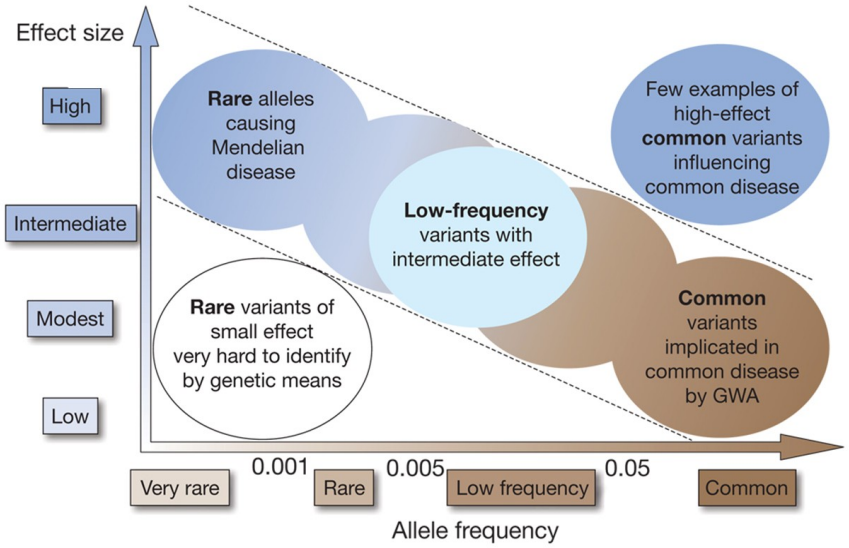


Figure 3.1. Relation between the frequency of the alleles and their effect (or causal relation) upon the trait (adapted from of Manolio *et al.* 2009[96]).

The first method, linkage, is an approach that permits to identify and map regions of the genome that may contain causal genes, which in turn are transmitted to the offspring. Comparing members of families, it is possible to delimitate blocks of DNA, using known genetic variants (called markers) that have been transmitted together among affected members and where there might be a causal variant. This approach has been and still is very helpful in studies of Mendelian diseases, which are caused by "one" mutation, normally rare, but with a high effect on the phenotype or disease (Figure 3.1.). For that reason it is better to study them in families with affected members where their frequency would be increased compared to the general population (Figure 3.2.).

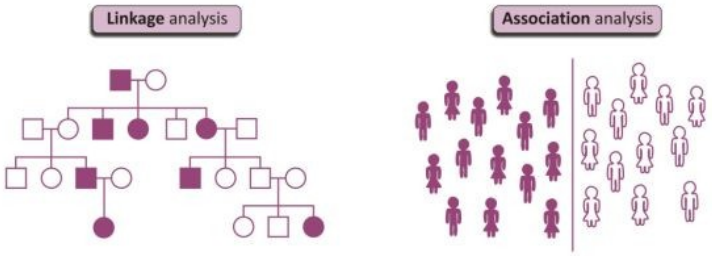


Figure 3.2. Linkage analysis vs. Association analysis[97]. Linkage analyses are performed in Family studies and association analysis are generally assessed in independent samples.

On the other hand, association studies also test known genetic makers, generally SNPs (see in the following chapter 5), in association to a trait of interest and normally, in groups of independent individuals (Figure 3.2.). These kind of studies are based in the linkage disequilibrium, LD (not to confuse with the family linkage). Theoretically, when two polymorphisms, P_A and P_B, are biallelic, A/a and B/b (A and B being the most frequent alleles) and independent and from a panmictic population, they present the following haplotypes: AB, Ab, aB and ab, which are distributed randomly among the individuals. Hence the frequencies for P_A and P_B are:

a) For allele A $= f_A$ and for the allele a: $f_a = 1 - f_A$

b) For allele B $= f_B$ and for the allele a: $f_b = 1 - f_B$

Then the frequencies of their combinations or haplotypes, are the product of their allelic frequencies. For example, AB would be the following in the population $f_{AB} = f_A \cdot f_B$ and similarly for the other combinations.

However, when the allele combinations are not aleatory, and for example the allele b is always transmitted together with allele A, then it exists a disequilibrium where this two alleles are linked. The haplotypes of the population would be: Ab, AB and aB. If, in addition, $f_A = f_B$, the distribution becomes: AB and ab, which represents a case of perfect LD. Therefore the frequencies are not based on the allelic frequencies but influenced by a disequilibrium with a value *D* (following the same example of AB)

- The frequency of AB: $f_{AB} = f_A \cdot f_B + D$
- The frequency of Ab: $f_{Ab} = f_A \cdot f_b - D$
- The frequency of aB: $f_{aB} = f_a \cdot f_B - D$ (in this example, this one would be 0)
- The frequency of ab: $f_{ab} = f_a \cdot f_b + D$

The value *D* is the difference between the observed haplotypic frequencies and the expected ones (in equilibrium). This value depends on the allelic frequencies that are specific of each population, so, it cannot be used to compare.

$$D = f_{AB} - f_A \cdot f_B$$
$$= f_{ab} - f_a \cdot f_b$$
$$= -f_{Ab} + f_A f_b$$
$$= -f_{aB} + f_a f_B$$
$$= f_{AB} \cdot f_{ab} - f_{Ab} \cdot f_{aB}$$

The standardization of the *D* is called *D'*:

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

where D_{max} is *D* in the case of complete LD

$$D_{\max} = \min (f_A \cdot f_B, f_a \cdot f_b) \text{ if } D > 0$$
$$D_{\max} = \min (f_A \cdot f_b, f_a \cdot f_B) \text{if } D < 0$$

D' is equal to 0 when there is equilibrium, and to 1 when there is complete LD where the alleles A and B are associated or -1 when A and b instead.

Another normalized measure of LD is r^2, a correlation between A>a and B>b. It takes values between 0 and 1 where 0 again means equilibrium but 1 doesn't mean complete LD like in D'. The value increases when the linkage between the two polymorphisms is more important.

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

Therefore, the more polymorphisms are used in an association study, the better to increase the power of detection and probabilities to find variants associated to the disease in LD with the markers. If an association is observed between a polymorphism and a phenotype, this could be because: the tested polymorphism is indeed a functional variant directly affecting the phenotype or the polymorphism is in "close proximity" with the underlying functional variant(s). This proximity is generally expressed in terms of linkage disequilibrium (LD).

Even though this strategy is applicable in family studies, it has been mainly applied in designs of independent samples, such as case-control studies. It has also been very successful to investigate multifactorial traits, i.e., traits that are influenced by environmental factors and multiple genetic determinants. Each determinant or factor has modest effect but

when accumulated with others, results into the trait or disease (Figure 3.1.). Up to the 2000s, the common association study approach was the "candidate gene association study" where the tested SNPs were selected according to some *a priori* knowledge about the inquiry trait. Generally, the SNPs were located in a gene whose product was known to participate to the biological mechanisms underlying the studied phenotype. For example, this strategy has been used to show that the two prothrombin (F2) gene variants discussed in Chapter 2 (page 27), rs1799963 and rs3136516, were influencing thrombin generation.

Then the development of high-throughput genotyping technologies revolutionized the genetic research offering the possibility to study simultaneously hundreds of thousands of SNPs all over the genome without any restriction about their location within specific candidate gene or genes. This leads to the emergence of the Genome Wide Association Study (GWAS) era, still relying on the key concept of LD. In the last ten years, this strategy has become a standard method to identify novel susceptibility genes for common, but also rare, diseases and quantitative traits. Thanks to it more than 2000 robust associations within more than 300 complex diseases and traits have been identified in the past 7 years[98], permitting also to create more than 2,000 genetic tests[99].

As part of my PhD project, I conducted the first GWAS on TGP biomarkers, the only one performed so far on such phenotypes, with the aim at identifying novel genes participating to the control of TGP.

Chapter 4. Genome-Wide Association Studies "for dummies": step-by-step analysis

The following chapters aim at exposing the general procedures for a good genome-wide association analysis (GWAS) (Figure 4.1.) I followed in my project, from the genotype determination of the samples up to the final interpretation of the results.

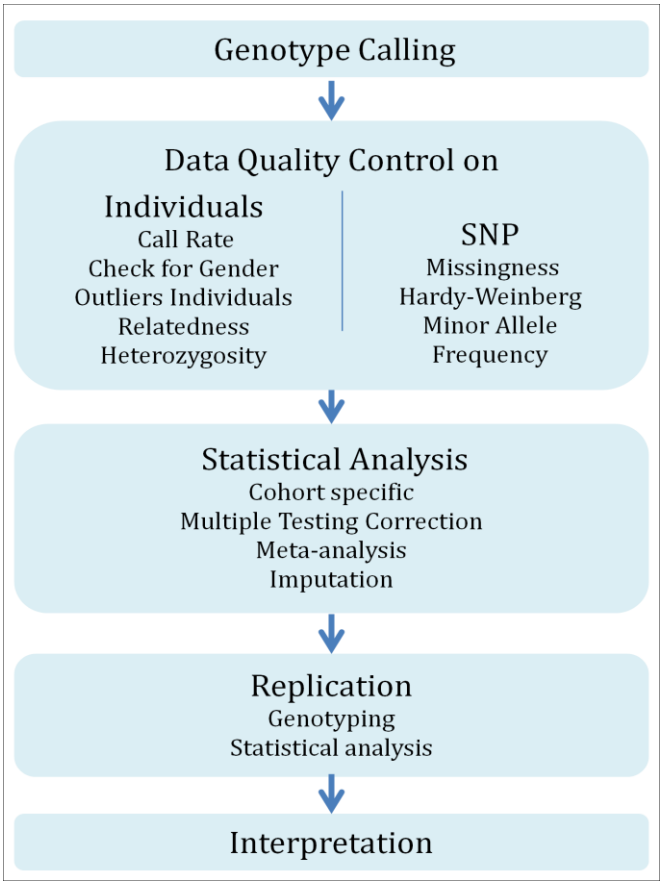


Figure 4.1. Diagram of the steps to follow to perform a GWAS.

4.1. Genotype Calling

One of the essential elements for an association analysis is the genetic data of the samples under study. The genetic data most frequently used are polymorphisms, also called markers, which are known genome variations easy to determine (or to genotype). Among all

known polymorphisms, the mostly used in GWAS are the single nucleotide polymorphisms (SNP).

The application of GWAS has been possible thanks to the development of DNA microarray techniques enabling to simultaneously genotype a huge number of SNPs. The latest generation of such arrays enables now to genotype ~5 Millions of SNPs on a single microarray.

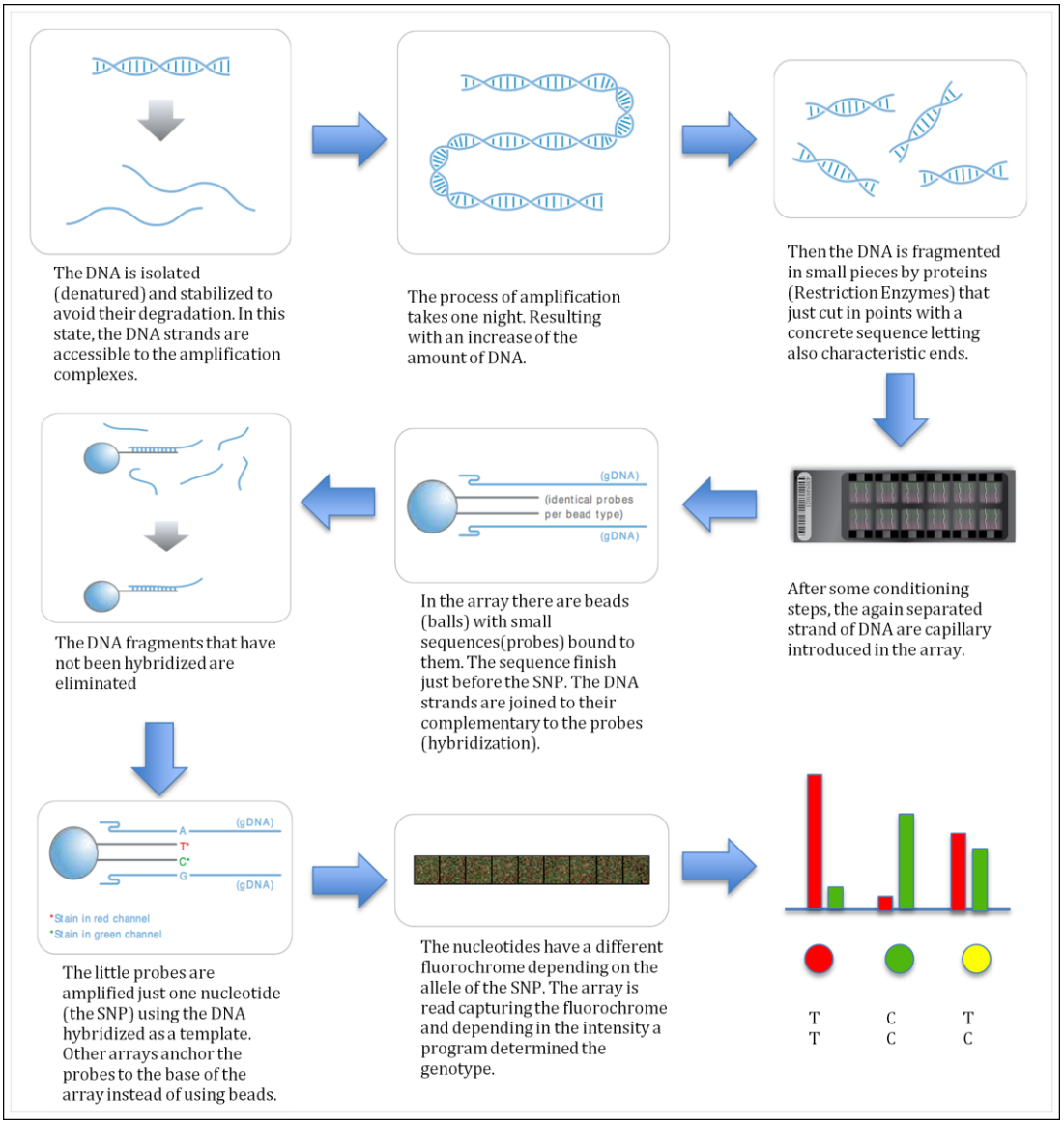


Figure 4.2. Representation of a microarray genotyping using as example Illumina images.

The fundamentals of this technique are explained in Figure 4.2. using images of Illumina®[100] as an example. This method consists in a multiple step procedure where the DNA sample is first amplified to increase the number of copies, and fragmented in small sequences. Then these short sequences are inserted in the array, where they hybridized to

their complementary probes. Those probes finish just before the SNP of interest. Then to determine those SNPs there is an amplification reaction to add just one nucleotide (the SNP itself). The nucleotides are bound to two different fluorochromes (or dyes): one for one allele (A1) and for the other (A2). Finally, the excited fluorochromes are read by a scanner or spectrophotometer and translated into a genotype. The possible results can be:

- A high fluorescence of one of the dyes, and a very small of the other, corresponding to the homozygosis (A1A1, A2A2).
- A combination of the two fluorochromes corresponding to heterozygosis (A1A2).

The fluorochromes intensity can be represented more visually in a 2D graph (Figure 4.3.):

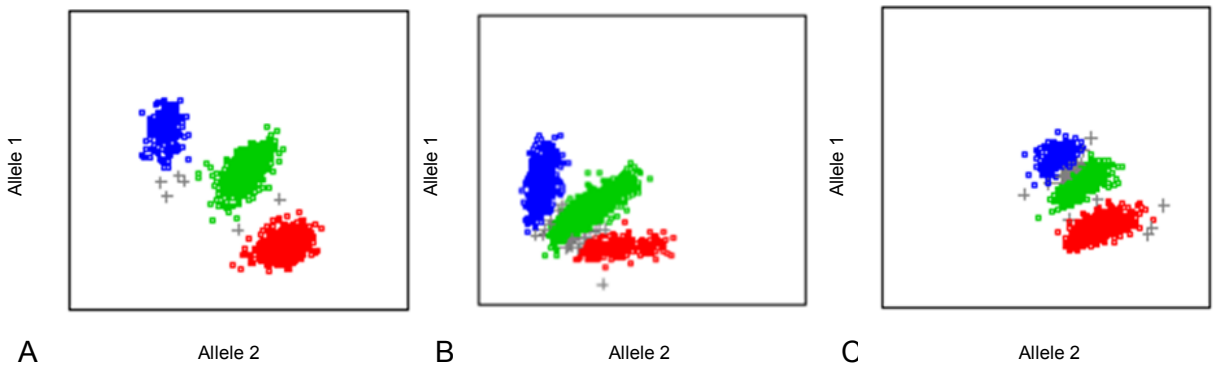


Figure 4.3. Examples of genotype calling results of a SNP. The y axis represents the intensity of one of the fluorochromes marking the allele 1 and in the x axis the intensity of the other fluorochrome for the second allele. A is an example of a good result where it is possible to differentiate well the three genotypes: A1A1 (blue dots), A1A2 (green dots) and A2A2 (red dots) with some indeterminate individuals represented as grey pluses. B is an example of a difficult determination, but it still maintains two groups of dots with a high intensity of one fluorochrome and low for the other and a group of dots in the middle. On the contrary in C, one of the genotypes (the blue) is very ambiguous and has a high intensity for both dyes.

4. 2. Data Quality Control

Before embarking in the statistical analysis of the genetic data there are some important quality controls (QCs) to apply to avoid any bias in the subsequent analysis and to minimize the risk of false positive and false negative results. These quality controls are applied at both subject samples and SNPs[101–104]. Different software can be used to conduct these QCs on GWAS data, such as PLINK[105], the R GenABEL package[106] and EIGENSTRAT[107].

4.2.1. Quality controls at the individual sample level

Five criteria are often used to identify individuals with genotypic information of insufficient quality to ensure validity of the subsequent statistical analysis.

Informative Missingness -When the percentage of SNPs with no assigned genotype for a given individual (a.k.a., missingness or genotype calling rate =1-missingness) is high, the individual is generally excluded from the GWAS analysis. This phenomenon may occur, for example, when the DNA quality is low. In general, a missingness threshold of 5% (or sometimes even 1%) is adopted, meaning that any individual with SNP call rate smaller than 95% (or 99%) of the SNPs genotyped is excluded from further analysis.

Gender Check- When the sex status reported in the clinical data file of an individual does not match the sex status inferred from the SNPs of X and Y chromosomes, the individual is usually excluded from the analysis. For example, males must generally not be heterozygous for SNPs on X chromosome (except the ones in the pseudoautosomal region). This may occur when there have been errors mislabeling the DNA samples or due to contamination (i.e., the DNA of two individuals have been mixed and genotyped together).

Heterozygosity - This test is based on the proportion of heterozygous SNPs for each individual[104] and is applied to detect those individuals with higher or smaller heterozigosity than expected according to Hardy Weinberg equilibrium (described afterwards). An excess of heterozygous SNPs could indicate a contamination of the samples whereas an excess of homozygosity (low heterozigosity) could indicate inbreeding (mating between relatives) or samples from other populations. Actually, a simple visual inspection of the distribution of heterozygosity across the whole genotyped individuals permits to detect "outlier" samples (Figure 4.4.). This heterozygosity parameter is also dependent on the missingness (discussed above), the kind of population under study and the genotyped SNPs available for that population. For that reason there is no standard threshold established to filter out heterozygosity. Every study applies the best threshold according to their data and it is indicated in the posterior report[108]. For example, The Wellcome Trust Case Control Consortium* established empirically a threshold for his reference work in GWAS[109] in which all those individuals with > 30% or < 23% of heterozygosity were excluded.

Other studies use an inbreeding coefficient instead[101], which is calculated for each individual *i*, with the proportion of 1-heterozygosity (homozygosity) as following[105]:

$$\text{Inbreeding coeff}_{\text{i}} = \frac{O_i - E_i}{T_i - E_i}$$

O is the observed homozygotes, E the expected homozygotes relaying on Hardy-Weinberg equilibrium, and T the total of genotyped autosomal polymorphisms. A positive

test suggests an excess of homozygotes and a negative test, an excess of heterozygotes. Then the test values near 0 suggest a proportion of homozygous/heterozygous as expected[101].

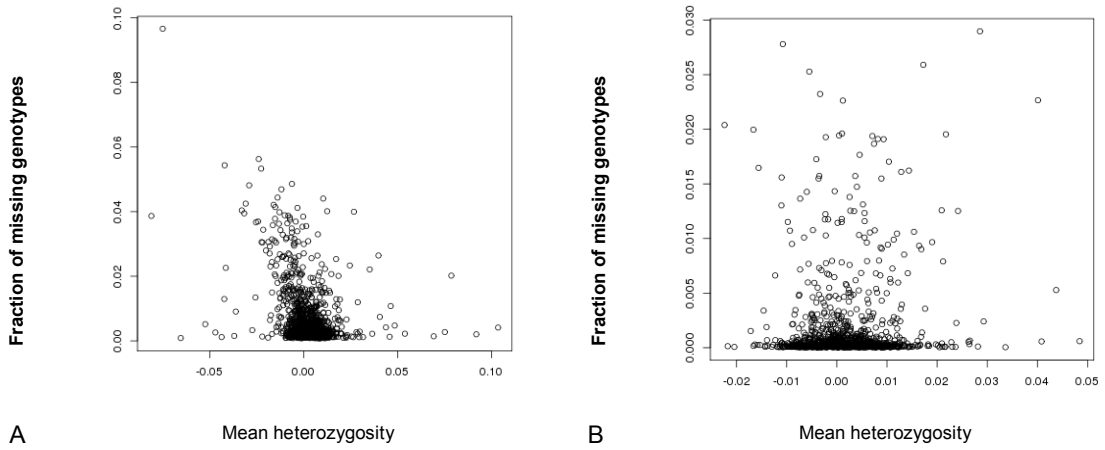


Figure 4.4. Two plots representing the relation between missingness (y-axis) and mean heterozygosity in a GWAS sample (3C study samples). The graph A individuals are displayed before any quality control. Some outliers are detected (in the up-right corner) of this graph. In B, the inbreeding coefficient based on heterozygosity of the individuals after apply all quality controls.

Relatedness - Although GWA studies have also been developed to analyze families[110], generally, they are implemented in unrelated individuals using standard generalized linear models (see Page 46). Therefore, it is important to assess that the GWAS individuals are not related beforehand. For that purpose, the genome-wide genetic information available for each subject can be used to assess the relatedness. Two individuals have a locus Identical by State (IBS) when they share, at least, half of the alleles for that locus, and that apprise about the genetic information in common. The values obtained by pair of individuals are 0,1/2 or 1 meaning 0,1 or 2 alleles in common, respectively. For the pair of individuals of a GWAS dataset, one can compute their average IBS as the addition of the IBS for each SNP divided by the total number of SNPs measured.

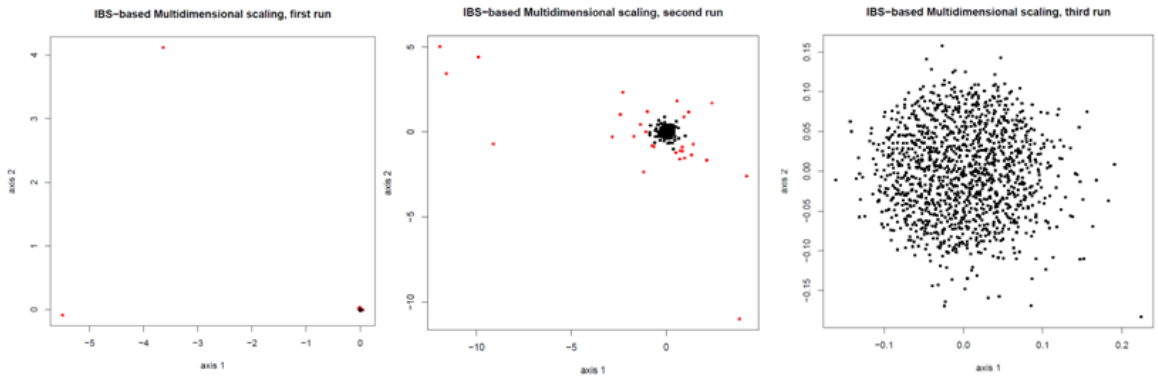
$$IBS_T = \frac{IBS2 + 0.5IBS1}{N\ pairs\ SNP}$$

Then, when a pair of individuals has an IBS> 0.95 or 0.98, the one with less call rate (less genetic information) is removed.

Outliers and population stratification- Another relevant issue in a GWAS is the possible existence of an unknown stratification of the studied, i.e., a subgroup of individuals with different genetic background (alleles and frequencies). If it is not properly detected and taken into consideration, such phenomenon can produce significant false positives. Two

strategies are commonly used to detect population stratification: the multidimensional scaling and the principal components analysis.

Multidimensional scaling (MDS) - MDS is a statistical method that represents graphically a group of samples based on a given distance matrix. In the GWAS context, the IBS value of a pair of individuals (above) is used, more exactly 1-IBS, to define the genetic distance between two individuals. The aim of this method is to observe the stratification of the samples and to detect atypical individuals that might come from a different population. The method distributes the individuals in a 2-dimension graphic, plotting together those who are homogenous or similar and drawing like a cloud of points (see Figure 4.5. C). It was considered as outliers those plotted outside the cloud with a distance of 3 times the standard deviation from the center (Figure 4.5. A and B, the red dots). Other works consider 2 or 4 times. Once such individuals are excluded from the group, the process is repeated a couple more of times until obtain a stable "cloud", i.e., homogenous group of samples (Figure 4.5. C).



A first MDS without filter **B** 2nd MDS after the first filtering **C** final MDS of homogenous samples

Figure 4.5. Multidimensional scaling graphics to detect outliers in a group of samples from Three City Study: every point is and individual. Those in black are considered homogenous within the group and the red ones the outliers, who have a distance of 3 times the standard deviation from the center-the group of samples. For instance, in A) red points appear at the left bottom and on top in the middle (MDS from 3C individuals).

Principal Components Analysis (PCA) -PCA is a method that explores and ranks the correlation structure of a sample under study. It is a dimension-reduction method widely used in biomedical research field. In the context of GWAS, the fundamental idea is to create components that can resume the general variability of the SNP data. Assuming *n* individuals genotyped for *m* SNPs, PCA distributes the SNPs in *m*-dimensions plan where each individual (with his/her m genotypes) are plotted (see an example Figure 4.6.). Individuals with similar genotypes would tend to be close to each other in this plot. Then, PCA creates "new" axis, called components, in which the genotypes will be projected characterizing a

hidden factor that explains part of the general variance of the SNP data. The first component is the one explaining the higher proportion of the variability of the data, the second component the one explaining the second higher proportion, and so on. Once the genotype data of an individual is projected into each m derived components, this individual can be characterized by his/her coordinates on each component. In an extreme case where the studied GWAS sample is composed of two populations with different genetic background, the coordinate of an individual on the first principal component can identify the subpopulation the individual belongs to.

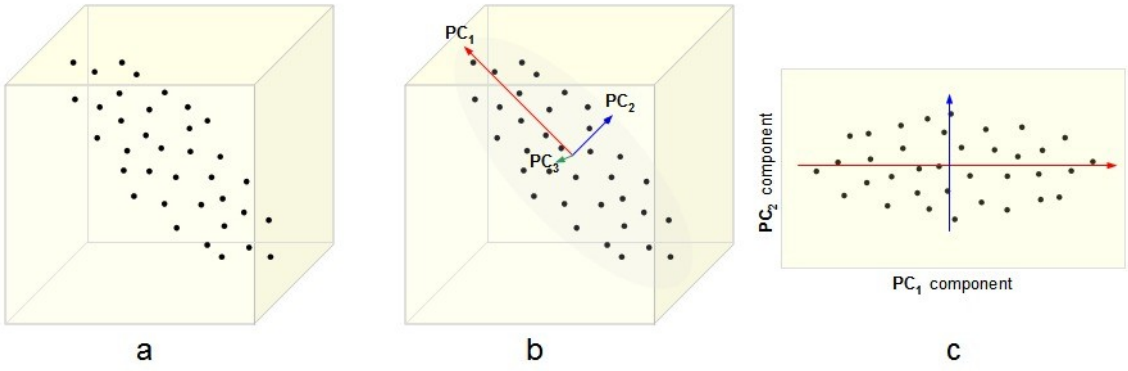


Figure 4.6. Graphic representation of a PCA[111]. In a), there is the 3-dimentional representation of a group of samples. Each axis can be a SNP that gives information about the samples. In b), 3 principal components have been created (the arrows) where the red one is the first principal component, the blue the second, and the green the 3rd principal component. In c), there is the distribution of the samples according to the 2 first principal components created that resume the information given by the 3 SNPs.

Adjusting the statistical association analysis of each SNP with the phenotype for the first *k* components (page 58) is sufficient to handle any undetected stratification of the studied populations. However, the choice for the *k* value may depend on each GWAS study. Often, it is recommended, to identify the main components (k = 1,2, ...10) that are significantly associated with the phenotype of interest and to adjust for them in the subsequent SNP association analyses.

4.2.2. Quality controls at the SNP level

Three filtering steps are also recommended for pruning the SNPs of poor quality in order to obtain reliable results:

SNP Missingness or call rate -Similar to what was described above for the informative missingness, any SNP that is correctly called (or genotyped) in less than 99% (sometimes 95%) of the samples are excluded from the analysis.

Hardy-Weinberg equilibrium (HWE) - The Hardy-Weinberg law[112] states that a population is in equilibrium when the individuals and their genes (and variants) have: a proper size to avoid genetic drift, no inbreeding but a randomly mating, no mutation events, no migration and no selection pressure. Under these ideal conditions, if the frequencies of two alleles A/a at a given locus are *p* and *q* respectively, after 1 generation of random mating the alleles shoud be in equilibrium and the genotype frequencies should be: $AA=p^2$, $Aa=2pq$ and $aa=q^2$; in the case of only two alleles per locus: $p+q = 1$ and $p^2+2qp+q^2=1$.

Deviation from HWE cannot only occur when the above conditions are not met but also when there are some genotyping errors[113]. Therefore, any SNP demonstrating strong deviation from HWE must be excluded from the subsequent GWAS analysis. HWE is often tested using a standard chi-square test with 1 degree[114] of freedom comparing the observed genotype distribution of each SNP with the expected one under equilibrium hypothesis. This test is applied to all genotyped SNPs, and for that reason the statistical threshold of 0.05 is not used for declaring statistical deviation from HWE. Rather a more stringent threshold (e.g., $p < 10^{-5}$), partially controlling for multiple testing issues, is advocated. When the GWAS sample is composed of cases and controls, HWE should be only tested in controls, as strong HWE deviation in cases samples may hide a true association with the disease which would be missed in the SNP had been excluded on the basis of HWE in cases or cases + controls).

Minor Allele Frequency (MAF)- Finally, to avoid noise (random hybridization giving signal in the genotype calling) and poor quality in the general results, it is recommended to discard from the analysis any rare SNPs, i.e., with MAF as low as 5% (or 1% according to the study sample size).

4.3. Statistical Testing for SNP - phenotype association

Once data QCs have been correctly applied, it is time to test for the association of SNPs with the phenotype of interest. There are different methods to perform an association study. The choice depends on the kind of samples (family or unrelated individuals), the background information (Bayesian or not) and the studied trait[114]. Standard methods are those based on generalized linear model (GLM) that can easily incorporate confounding factors to adjust the relation between the phenotype and genotype. For instance: age, sex, smoking and the principal components (PCs) discussed in page 43. A logistic regression model is adopted for a binary trait and a linear one for a quantitative phenotype.

By default, an additive allele effect is commonly used to test for the association of a SNP with a trait[115]. In the case of a linear model, this implies that the expected phenotypic

mean is expressed as: $E(y / G = g) = \mu + \beta g$. Where G is the variable denoting the tested genotype and takes values: 0,1 or 2 according to the number of rare or reference alleles carried by an individual, depending on the software. μ is the average phenotype mean, g represents the genotype and β is the additive effect. Other genetic models such as recessive and dominant can also be tested. In those cases, the G variable takes the values (0,0,1) and (0,1,1), respectively, instead of (0,1,2).

For each tested SNP, the β coefficient and its standard error are generally obtained by a Maximum Likelihood procedure, which also provides the corresponding p-value for association between the SNP and the phenotype.

Visualization of the GWAS results

GWAS results are normally checked for possible false positives with a Quantile-Quantile (QQ) plot[115], where the distribution of the observed p-values is compared to the expected under the null hypothesis of no SNP-phenotype association. The aim is to detect if there are more significant results than expected by chance. By definition, p-values under the null hypothesis follow a uniform distribution on the [0-1] interval. Then, the expected distribution of the *m* SNPs p-values can be obtained by computing for each *j* p-value the corresponding expected p-value E_j:

$$E_j = \frac{j}{(m+1)}$$

The observed and expected p-value distributions are then transformed to $-\log_{10}$, ordered from the lowest to the highest, and plotted one against the other into a QQ plot. Under the null hypothesis, this should lead to a diagonal line "y = x". When there is deviation from this line only at the right tail of the distribution, it indicates that a true non-null association may lie in the results (Figure 4.7.).

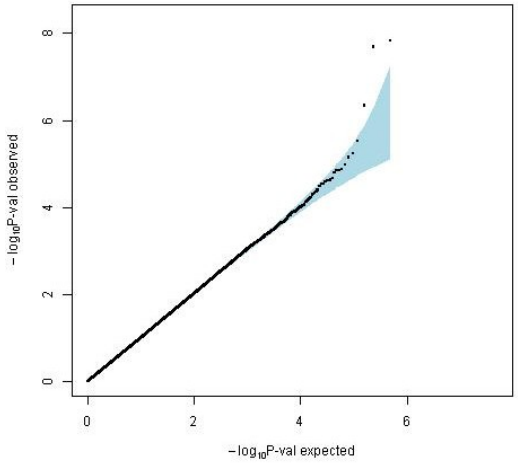


Figure 4.7. Example of QQplot of 3C study. In the Y axis the observed p-values and in the x axis the expected p-values. Both transformed: -log10 (p-values). In blue is shown the associated 95% confidence interval for the null hypothesis.

When the observed line deviates early from the straight line "y = x", there is a strong element in favor of spurious associations due to uncontrolled population structure, cryptic relatedness or other undetected phenomena, e.g., the applied test statistic is not correct[115].

This deviation can be measured by estimating the regression coefficient of the observed line "y = λ x ", also called the genomic inflation (GC) factor (Figure 4.8.). A λ value deviating from 1.00 (i.e., greater than 1.05 or lower than 0.95) indicates that something went "wrong" in the GWAS results. A Chi-Square statistic with 1 degree of freedom (df) is used to assess the SNP associations. The GC factor can also be obtained by computing the median of the observed Chi-square and dividing it by 0.455625 which corresponds to the median of a Chi-square distribution with 1 df.

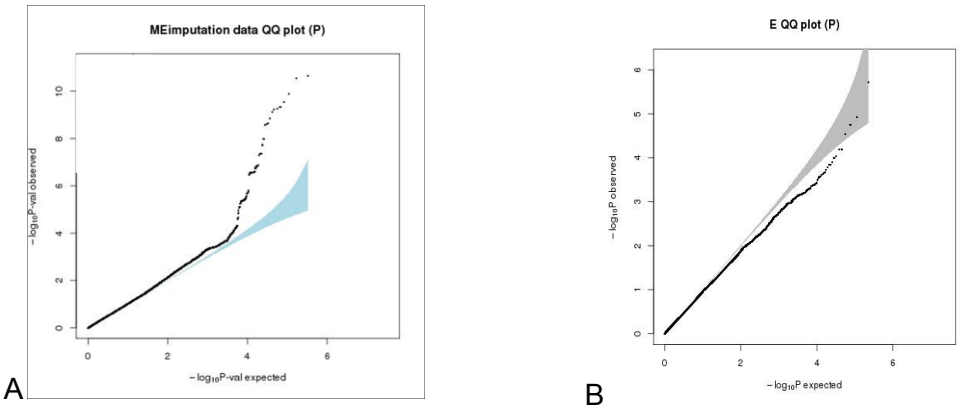


Figure 4.8. QQ plots showing deviation from what expected. In the case of A, there is a deviation showing lower p-values from what expected even though it is not too strong (λ=1.032). In the case of B, there is deviation under the diagonal, generally indicating over adjustment (λ=0.96).

Finally, once we have validated that no strong inflation lies in our GWAS results, p-values ($-\log_{10}$ (p-values), more precisely) are plotted in a graphic according to the genomic position of the associated SNP. This kind of graph is generally known as Manhattan (MH) plot (Figure 4.9.).

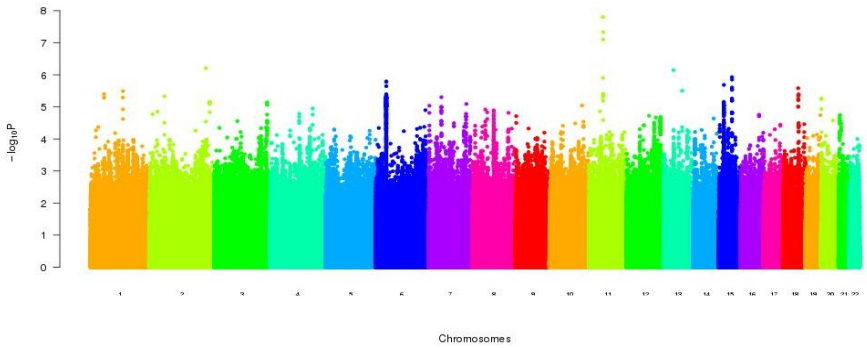


Figure 4.9. Example of Manhatan plot or MHplot of MARTHA study. The Y-axis represents the observed $-\log_{10}$ (p-values) sorted by chromosome and position(x-axis). Each dot represents a SNP p-value.

4.4. Correction for Multiple Testing

The use of a test statistic to assess a specific hypothesis can generate two different errors: The false positives (α) and false negatives (β). A common approach to control these errors is to reduce as much as possible the β, i.e., to increase the power of the test (1- β) while accepting arbitrarily a certain error α (normally α=0.05). However when the number of test increases, the probability that the test falsely declare significance increases proportionally with the number of tests (m):

$$P(\text{Type I Error})=1-(1-\alpha)^m$$

Therefore, in a GWAS where association tests are repeated over thousands of millions of SNPs, it is very important to adopt a very low statistical threshold for declaring a SNP-phenotype association significant while controlling for false positives.

Several corrections have been proposed in the context of GWAS to handle the multiple testing issues and readers interested in this topic are invited to have a look at the references[116–118] for more literature. Currently, the general recommendation is to use a genome-wide statistical threshold of $5\ 10^{-8}$ to declare genome-wide significance while controlling for a type I error of 5%. This threshold has been obtained by empirical and simulation studies estimating the number of independent common SNPs over the genome and then applying a standard Bonferroni correction on it[119–121].

The Bonferrroni correction is part of a group of test corrections called Family-wise Error Rate (FWER). It is defined as the probability to obtain at least one false positive in the case that all the tests result in the acceptation of the null hypothesis (H_0). Bonferroni[122] proposed the following correction or threshold (α') considering the independence between N tests performed:

$$\alpha' = {}^{\alpha}/_{N}$$

This correction is optimal to avoid any false positive when replicating a finite number of tests. But in some cases such as GWAS, where there are enormous numbers of tests (one for each SNP), it is not possible to assure the total independence between those tests. Then the correction becomes too stringent increasing the risk of false negatives.

4.5. Meta-Analysis of GWAS datasets

The Meta-analysis strategy is an effective approach that combines results obtained in different studies evaluating the same hypothesis when it is not possible to re-analyze all the data together at once (Mega-Analysis). This can occur when study designs are different, the confounding factors are not available in all studies, or the ethical and legal issues prevent from sharing the whole raw data. The combination of the GWAS results with meta-analysis method increases the power of the detection and, so far, this approach has provided numerous successes[115].

Generally, the combination or meta-analysis can be implemented by two different methods. One method combines the p-values obtained in each study to get an overall statistical evidence of the tested hypotheses. The other approach, widely used in the field of GWAS, provides a pooled estimate or a combined global coefficient characterizing the tested association (e.g., the regression coefficient described above in page 45) from each study-specific coefficient and its standard error. The method consists in:

Let β_i be the regression coefficient associated with a given SNP in the i^{th} study (i = 1, 2,..., N), and SE_i its standard error. The standard method consists in estimating a combined (or pooled) average for this SNP using the following formula:

$$\hat{\beta} = \frac{\sum[\beta_i/(SE_i)^2]}{\sum[1/(SE_i)^2]}$$

Then, it is possible to obtain a statistic that follows a Chi-Square distribution with 2 x N degrees of freedom.

$$\chi^2 = \sum \frac{(\beta_i - \hat{\beta})^2}{(SE_i)^2}$$

If the above statistic is significant, then a valid standard error for the resulting $\hat{\beta}$ is given by:

$$SE(\hat{\beta}) = \sqrt{\frac{1}{\sum 1/SE_i^2}}$$

The resulting coefficient is generally called the fixed-effect meta-analysis estimate. It is important to highlight that this method is valid when there is no heterogeneity between the individual regression coefficient (β_i's) across the distinct studies. The absence of heterogeneity can be tested by the Cochran-Mantel-Haenzel Q or the Thomas and Higgins I^2 test statistics. By defining

$$\omega_i = \frac{1}{(SE_i)^2},$$

then the *Q* test statistic is given by $Q = \sum \omega_i \left(\beta_i - \hat{\beta}\right)^2$. Under the null hypothesis of no heterogeneity, *Q* follows a Chi-Square distribution with *N*-1 df.

Related to the *Q* statistic is the Thomas and Higgins I^2 defined as:

$$I^2 = 100 \times \frac{(Q - df)}{Q}$$

This criterion quantifies the degree of inconsistency/heterogeneity among the studies. It represents the percentage of total variation among the studies induced by their heterogeneity and not by chance[123]. I^2 is generally positive (if not, it is fixed to 0) with evidence for heterogeneity and increasing together. When heterogeneity is suspected, the fixed-effect analysis is discarded in favor of a random-effect (RE) analysis.

The RE analysis aims at incorporating an estimation of the inter-study variability into the estimation of the pooled β estimate. This estimation, $\hat{\tau}^2$, is given by:

$$\hat{\tau}^2 = \frac{Q - (\mathrm{I} - 1)}{\sum \omega_i - \frac{\sum \omega_i^2}{\sum \omega_i}}, \text{if} Q \geq \mathrm{I} - 1$$

$$\text{and } \hat{\tau}^2 = 0, \text{if } Q < \mathrm{I} - 1.$$

Defining $\omega_i^* = \left(\omega_i^{-1} + \hat{\tau}^2\right)^{-1}$, the RE estimate of $\hat{\beta}$, the $\hat{\beta}_{RE}$, and the SE is given by

$$\beta_{SE} = \frac{\sum \omega_i^{-1} \beta_i}{\sum \omega_i^*} \quad ; \quad SE\left(\hat{\beta}_{RE}\right) = \sqrt{\frac{1}{\sum_i \omega_i^*}}$$

Once the combined $\hat{\beta}$ and its standard error are obtained, the statistical evidence for association is given by computing the p-value of the corresponding *t* (or Chi-square) test calculated as the ratio of $\hat{\beta}$ over its standard error (or the square of this ratio). When the meta-analysis p-values are obtained for all tested SNPs, the MH and QQ plots can be obtained together with the value of the genomic inflation factor (λ).

The main limitation when trying to meta-analyze the results of several GWAS analyses on the same phenotype is that the genotyped SNPs are not necessarily the same in the different GWAS studies. This can occur when different DNA microarrays with different genetic content are used across studies and/or when some SNPs did pass the QCs in some but not all studies. However, this limitation has been solved with the development of novel statistical techniques relying on "*imputation*".

4.6. Imputation Analysis

The Imputation technique is a method to infer polymorphisms, that have not been measured in the microarray. It uses the information of a reference panel of genotypes and haplotypes composed by a large number of genotyped SNPs. Sometimes it is called an *in silico* genotyping method[124,125]. The reference panel is constituted by a "model" group of individuals from whom extensive genetic sequence information is known. The most used reference panels are the HapMap[126] and the 1000 Genome Project[127]. Both tools are essential in current human genetic research by providing access to the exact genetic sequence of a large number of individuals.

The basis of the imputation method is as follows: let's assume for a moment that a known reference population has three particular SNPs (e.g., A_1/G_1, C_2/A_2 and G_3/T_3) that can only generate four haplotypes (H_1, H_2, H_3 and H_4),

H_1 ----A_1 C_2 G_3---- with frequency f_1

H_2 ----A_1 A_2 T_3----with frequency f_2

H_3 ----G_1 C_2 T_3---- with frequency f_3

H_4 ----G_1 A_2 T_3----with frequency f_4

with haplotype frequencies (f_1, f_2, f_3, f_4) whose distribution satisfies HWE. Then, let's take a group of samples under study that have been genotyped only for the SNPs A_1/G_1 and G_3/T_3.

If an individual has the following genotypes: A_1A_1 and T_3T_3, it can be deduce (or imputed) by observing the previous haplotypes that the non-genotyped C_2/A_2 SNP would be A_2A_2. Indeed, this individual carries two A_1 and T_3 alleles and this combination of alleles is only found in the H2 haplotype. Therefore, the individual is certainly homozygous for the H2 haplotype and thus it carries two A_2 alleles.

Similarly, if an individual caries the A_1A_1 and T_3G_3 genotypes, it can be estimated (or imputed) that it carries one haplotype A_1T_3 (H_2) and one A_1G_3 (H_1) and therefore this individual is C_2A_2 for the non-genotyped C_2/A_2 - SNP.

The situation becomes a bit tricky for an individual with genotypes A_1G_1 and T_3T_3. He/she can be either A_2C_2 (haplotype pair H_2H_3) or A_2A_2 (haplotype pair H_2H_4). If one has access to a precise estimate of the different haplotype frequencies in the reference population (the f_1, f_2, f_3, f_4), then it is possible to estimate that this individual has a genotype:

- A_2C_2 with a probability of $P(A_2C_2 / A_1G_1 \& T_3T_3)$
  $= P(A_1A_2T_3 \& G_1C_2T_3)/[P(A_1A_2T_3 \& G_1C_2T_3) + P(A_1A_2T_3 \& G_1A_2T_3)]$
  $= 2 f_2f_3/ (2 f_2f_3 + 2 f_2f_4)$
  $= f_3/ (f_3+f_4)$

or

- $A_2 A_2$ with a probability of $f_4/ (f_3+f_4)$.

And so on...

As a consequence, when there is a reliable estimation of the haplotype structure derived from the reference population, including haplotype frequencies estimation, it is possible to infer (impute) non-genotyped SNPs. Always assuming that the reference population has similar genetic background as the sample under study.

There are different software programs that can perform imputation analysis for unrelated samples, such as fastPHASE/BIMBAM[128], BEAGLE[129], IMPUTE[130] and MACH[125]. Generally, these software programs compute, for each non genotyped SNP, the posterior genotype probability given the measured SNP data of an individual, and then assign to this individual an expected number of imputed alleles (often called imputed dose or dosage). For example, the imputation of an individual who lacks the SNP A/C can result AA, AC or CC with a posterior probability f_{AA}, f_{AC} and f_{CC}. Then, his/her C allele imputed dose would be $2 \times f_{CC} + f_{AC}$. This dose can take values between 0 (e.g. for individuals of genotype AA with probability 1) or 2 (e.g. for individuals of genotype CC with probability 1) and are used in regression models to test for the effect of the imputed SNP on the phenotype of interest. When the allele dose is very close to 0, 1 or 2, indicates slightly uncertainty about the inferred genotype, meaning homozygous for one allele, heterozygous and homozygous for the other allele respectively. The level of certainty in the allele dosage estimation can be quantified as the squared correlation coefficient, r^2, between the expected allele count derived from the assumed known allele frequencies and the expected given the measured genotypes[131]. A r^2 value close to 1 indicates complete genotype determination of the non-measured SNPs. Conversely, a r^2 value close to 0 indicates that the measured genotypes of the samples are not enough informative to bring the genotype of the missing SNP. The imputation of a SNP is considered acceptable when $r^2>0.3$ and good when $r^2>0.8$[125,132].

Before embarking into the imputation analysis *per se*, there are some additional controls necessary to guarantee a good procedure and expect the "highest" quality possible in the results. It is important to check if the genotyped data of the samples are accordingly to the annotations used in the reference panel. Approximately every year, a new annotation of the

human genome (list of polymorphism, their names, their genomic coordinates and their alleles) is updated. Since imputation techniques require the name of the SNPs (those measured and those to be imputed), their genomic position and their condition of reference /alternative alleles, it is very important to check whether the GWAS SNP data are annotated in the same way as the reference panel in use.

4.7. Replication of the GWAS findings

Once the GWAS analysis (or the meta-analysis of GWAS) is completed and the p-values for all tested (genotyped and imputed) SNPs are available, it is time to bring the most significant associations for replication in independent samples and assess their robustness and validity to avoid claiming false-positive associations[115].

As described above, one usually assesses for replication the SNPs that have reached the pre-specified genome-wide significance threshold (~$5 \cdot 10^{-8}$). However, according to GWAS samples size and the unknown underlying genetic effects of the phenotype-associated SNPs, genuine associations may hide under the heap of genome-wide significant signals. This is related to the power of the study. This is why it is not infrequent to reduce the stringency of the statistical threshold and to adopt a less stringent threshold ($p < 10^{-7}$ or even $p < 10^{-5}$) to increase the number of selected SNPs for replication.

Then, the replication consists in investigate the association of the SNPs selected with the phenotype in another independent population. This association can be assessed by 1) wet-lab genotyping the SNP of interest, 2) wet-lab genotyping another SNP (referred to as a "proxy SNP") known to be in complete linkage disequilibrium (in particular from public reference panel database) with the candidate SNP when the latter is not technically "genotypable" using any standard protocol technique, 3) looking at the results of previously reported GWAS on the same phenotype where the SNP of interest has been measured or imputed (a.k.a *in silico* genotyping).

When several SNPs are tested for replication in independent populations, the Bonferroni correction is commonly used to address the multiple testing issues and controls for the false positive rate at the replication stage (page 49).

Chapter 5. A "GWAS Study" on Thrombin Generation Potential

In this chapter, I describe the samples I used in my firsts analysis and I also specify the details I performed to achieve my GWAS according to the steps I exposed in the previous chapter. Then I present the results I obtained followed by the discussion.

5.1. Discovery GWAS cohorts

MARseille THrombosis Association (MARTHA) project[28,133,134] coordinated by Professor Pierre-Emmanuel Morange, is composed of 1,592 unrelated Venous Thrombosis (VT) patients, mainly of French origin. Individuals were 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. 586 VT patients were typed with the Illumina Human 660W-Quad Beadchip and 1,011 patients were typed with the Illumina Human 610-Quad Beadchip. 551,141 SNPs were in common between both arrays and available in the MARTHA patients.

Thrombin generation potential was measured in a subgroup of 848 MARTHA patients using the method CAT (Chapter 2, page 22).

The Three City Consortium Study (3C Study)[135] is a population-based study carried out in 3 French cities: Bordeaux, Dijon, and Montpellier. The principal aim of the project is to find out if there are vascular factors that can account for the risk of dementia and loss of memory. It is composed of 8,707 non-institutionalized individuals aged over 65, randomly selected from the electoral rolls and free of any chronic diseases. Between March 1999 and April 2001 the 3C samples were recruited, examined and interviewed for general health information, lifestyle, cognitive tests, diagnosis of dementia and family history. For my PhD project I had access to a sample of 1,314 subjects with genome-wide genotype data typed with the Illumina Human610 DNA chips including 582,892 SNPs, and from those, 1,253 had

plasma available for TGP measurements. TGP measurements were obtained using the same CAT technique as in MARTHA.

5.1.1. Genotyping Quality Controls and Imputation

The application of genotype quality controls in MARTHA and 3C, separately (Figure 5.1.), resulted in the final selection of 491,285 and 487,154 autosomal SNP, respectively. These SNPs were then used for imputing separately 11,572,501 SNPs reported in the 1,000 Genome reference database (release 08/2010) using the MaCH program. After selecting the common SNPs with imputation quality r^2>0.3 and minor allele frequency (MAF) greater than 0.01, 6,652,054 were left for association analysis with TGP biomarkers.

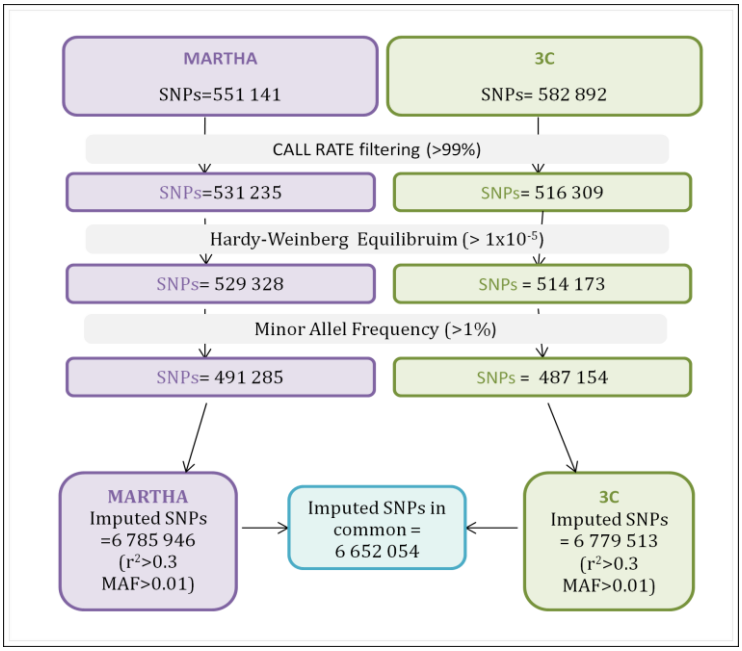


Figure 5.1. Diagram of the filters applied in MARTHA and 3C studies.

Individuals with genotyping call rate <95% were discarded from the analysis, as well as individuals demonstrating close relatedness as suspected from pairwise clustering of IBS, and genetic outliers and substructure detected by multidimensional scaling and PCA[107]. Finally, 714 and 1253 individuals from the MARTHA and 3C study were left for association analyses respectively.

5.1.2. Clinical and biological characteristics of the discovery cohorts

Detailed description of the biological and clinical characteristics observed in the MARTHA and 3C are shown below in Table 5.1. extracted from my publication **Rocanin-Arjo *et al.* 2014**[94].

In order to handle non-normal distributions, a log-transformation and a normal quantile transformation were applied to ETP and Lagtime values, respectively, separately in the 2 cohorts (Figure 5.2.).

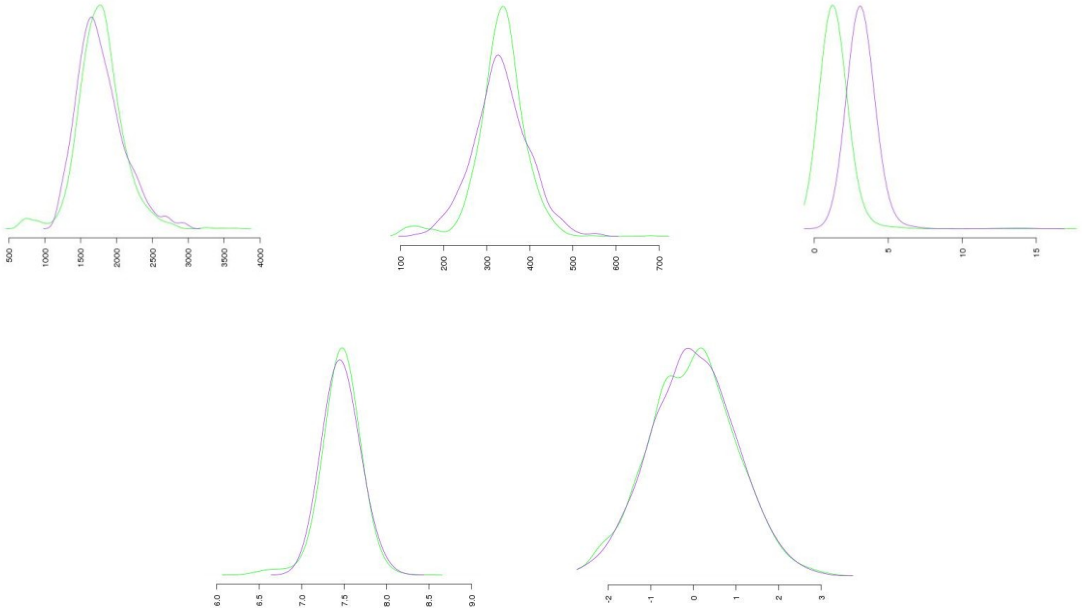


Figure 5.2. Density distributions of TGP biomarkers. On top and from left to rigth: ETP, Peak and Lagtime raw data. On the bottom, the log transformed ETP (left) and the quantile transformed Lagtime (middle) distributions. MARTHA in purle, 3C in green.

Table 5.1. Characteristics of the studied populations.

	MARTHA N = 714	3C N = 1,253
Age (SE)- yrs	46.84 (15.29)	75.05 (5.75)
Sex (% Male)	32.1%	41.6%
% VT patients	100%	0%
BMI (SE) (kg/m2)	25.1 (4.55)	25.8 (4.22)
FV Leiden (1)	153 (21.4%)	-
F2 G20210A (1)	83 (11.6%)	-
Oral coagulant	-	52 (4.2%)
ETP (nM/min) [1st - 3rd]	1780 [1554 -1958]	1775 [1586 - 1947]
Peak (nM) [1st - 3rd]	333.0 [293.5 - 372.0]	332.8 [307.0 - 364.3]
Lagtime (min) [1st - 3rd]	3.229 [2.830 - 3.500]	1.382 [1.000 - 1.670]

	Correlation between TGP markers [2]	
ETP - Peak	ρ= 0.78***	ρ= 0.77***
ETP - Lagtime	ρ = 0.14*	ρ = 0.06*
Peak -Lagtime	ρ = -0.05*	ρ = -0.10***

[1] FV Leiden and F2 G20210A mutations were genotyped in MARTHA as part of the inclusion criteria. Homozygous carriers were not included in the study.
[2] Correlations were computed on transformed values (i.e., log-transformation on ETP and quantile normalization for Lagtime) adjusted for age, sex, oral coagulant therapy, F2 G20210A (when appropriate) and BMI.

***<0.00001,**<0.0001,*<0.05

ETP and Peak biomarkers have a strong positive correlation, ρ = 0.80, both in VT patients and healthy subjects. Conversely, Lagtime exhibited very low correlation with the other two TGP biomarkers (Table 5.1.).

5.1.3. Statistical methods and adopted research strategy

The association of imputed SNPs with each TGP biomarkers was conducted separately in MARTHA and 3C. A linear regression model was applied with the imputed allele dose as covariate characterizing the imputed SNPs effects (Figure 5.3.). Analyses were adjusted for age, sex, contraception and anticoagulant therapy and the first 4th principal components derived from the GWAS QC genotypes (page 40). Association analyses were conducted by use of the mach2qtl software[125]. A fixed-effect meta-analysis was then performed on MARTHA and 3C results altogether. Homogeneity of the results across studies was assessed using the I^2 statistics[136], derived from Mantel-Haenszel method[137]. The meta-analysis was performed using the METAL software[138].

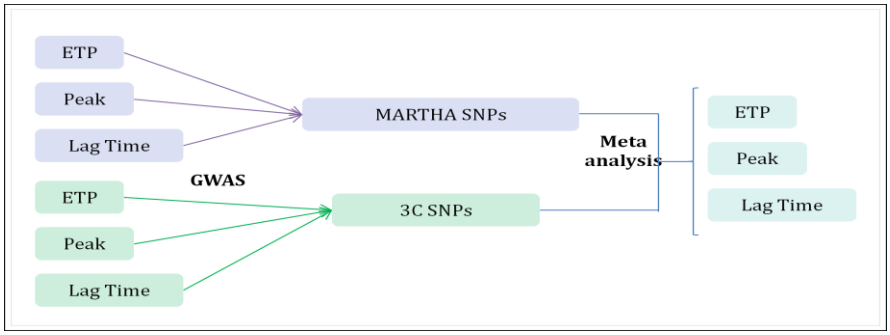


Figure 5.3. Diagram of the analysis procedure of the discovery step.

In a first step of my research strategy, I focused on SNPs that reached the genome-wide significance level of 5 10^{-8} in the meta-analysis for each TGP biomarker. In a second step, I looked for SNPs that demonstrated association p-values <10^{-5} for at least two TGP

biomarkers. Such SNPs were then tested for replication in two additional French cohorts, MARTHA12 and FITENAT, where TGP markers were measured with the same technique.

5.2. Replication cohorts

MARTHA12 The MARTHA12 study is composed of 1,245 VT patients that have been recruited at the Thrombophilia center of La Timone hospital (Marseille, France) between 2010 and 2012, following the same inclusion/exclusion criteria as those used for the MARTHA study (page 55).

FITENAT[139]. The FITENAT study was designed in Health Examination Centers (HEC) selecting healthy individuals from the French Social Security of 11 regions distributed throughout France: Bordeaux (Aquitaine), Lille (Nord-Pas de Calais), Lyon (Rhône-Alpes), Marseilles (Provence Côte d'Azur), Nancy (Lorraine), Nantes (Loire-Atlantique), Paris (Ile de France), Poitiers (Charente Poitou), Rennes (Bretagne), Strasbourg (Alsace) and Toulouse (Midi-Pyrénées). Each HEC recruited 600 individuals, 300 men and 300 women and they were grouped by age intervals of: 18–34, 35–49, 50–64, and >65 years. Other inclusion criteria were: a French origin for the individual and their parents, and European origin for grandparents. Individuals underwent a clinical examination where their personal and family history of chronic conditions was recorded, as well as details about lifestyle as: alcohol consumption, smoking habit, physical activity and medication (in case of women also contraception treatment). The FITENAT sample I used for my project was composed of 543 healthy individuals with neither history of cardiovascular disease, diabetes, hypertension, renal nor hepatic failure; and those who were not under anticoagulant therapy.

Detailed description of the biological and clinical characteristics observed in the MARTHA12 and FITENAT are shown below in Table 5.2. extracted from my publication[94].

In this case, both studies also present high correlation between ETP and Peak (ρ>0.7). However, they also present a small correlation between ETP and Lagtime, and just in the case of MARTHA12 also for Peak and Lagtime.

5.2.1. Statistical methods and adopted research strategy

Genotyping of the SNPs selected for replication was performed using allele-specific PCR in 733 MARTHA12 patients with available DNA. Association of genotyped SNPs with TGP biomarkers was assessed using a linear regression analysis under the assumption of additive allele effects. The same transformations on TGP phenotypes were applied as in the discovery GWAS cohorts to match the results. Analyses were adjusted for age, sex and oral

contraception therapy. A standard Bonferroni correction for the number of genotyped SNPs was applied to declare statistical evidence.

Table 5.2. Characteristics of the MARTHA12 and FITENAT populations.

	MARTHA12 N = 796	FITENAT N = 543
Age (SE)- yrs	49.74 (15.34)	47.81 (13.94)
Sex (% Male)	42.9%	47.9%
% VT patients	100%	0%
BMI (SE) (kg/m2)	26.0 (4.98)	24.2 (3.62)
FV Leiden [(1)]	89 (11.2%)	-
F2 G20210A [(1)]	52 (6.5%)	-
Oral coagulant	24 (3.0%)	-
ETP (nM/min) [1st - 3rd]	1892 [1654 - 2124]	1675 [1456 - 1838]
Peak (nM) [1st - 3rd]	328.0 [278.0 - 374.8]	293.5 [258.1 - 319.2]
Lagtime (min) [1st - 3rd]	3.330 [2.762 - 3.670]	2.280 [2.000 - 2.500]
	Correlation between TGP markers [(2)]	
ETP - Peak	ρ = 0.73***	ρ = 0.78***
ETP - Lagtime	ρ = 0.19***	ρ = 0.34***
Peak -Lagtime	ρ = -0.18***	ρ = 0.04

[(1)] FV Leiden and F2 G20210A mutations were genotyped in MARTHA12 as part of the inclusion criteria. Homozygous carriers were not included in the study.

[(2)] Correlations were computed on transformed values (e.g., log-transformation on ETP and quantile normalization for Lagtime) adjusted for age, sex, oral coagulant therapy, F2 G20210A (when appropriate) and BMI.

***<0.00001,**<0.0001,*<0.05

SNPs that showed statistical significant association with TGP biomarkers in MARTHA12 were further genotyped in 528 FITENAT subjects. The same statistical methods were employed in FITENAT.

Finally, results obtained in the four available cohorts were combined into a fixed-effect meta-analysis.

5.3. Main GWAS findings

A total of 6,652,054 polymorphisms were tested for association with each TGP biomarkers, ETP, Peak height and Lagtime, in a meta-analysis of 1,967 individuals from MARTHA and 3C cohorts. Corresponding QQ plots for each biomarker are shown below in Figure 5.4.

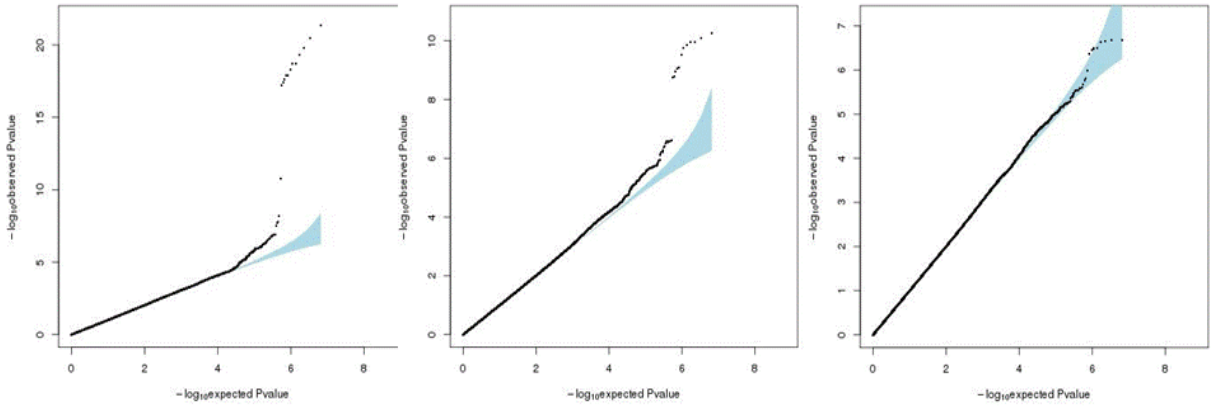


Figure 5.4. Quantile-Quantile plots of the meta-analysis p-values combining MARTHA and 3C: ETP (left), Peak height (middle) and Lag-time(right).

The associated GC inflation factors ware 0.998, 0.997 and 1.000, for ETP, Peak height and Lagtime respectively.

5.3.1. GWAS analysis on ETP

The Manhattan plot representation of the meta-analyzed GWAS findings for ETP is shown in Figure 5.5.

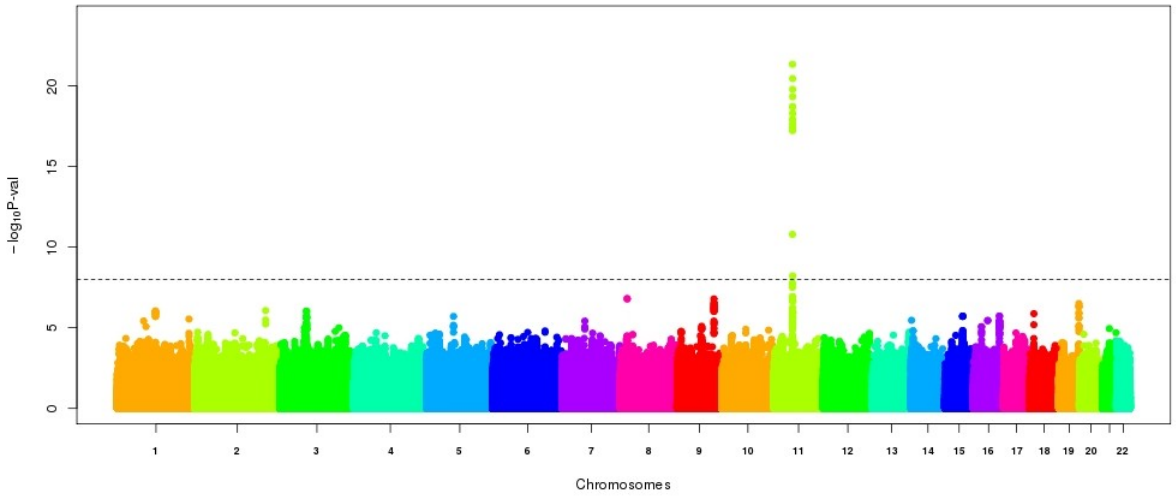


Figure 5.5: Manhattan plot representing the p-values of the GWAS meta analysis on ETP. The horizontal dashed line corresponds to the statistical threshold of $5\ 10^{-8}$.

Seventeen SNPs, all mapping to the chromosome 11p11 region, were genome-wide significantly associated with ETP. The compelling signal was observed for *MYBPC3* rs2856656 SNP ($p = 4.62x10^{-22}$). This polymorphism had already been shown to be in strong LD with the *F2* G20210A mutation (rs1799963)[140], a variant well established to influence thrombin generation. In MARTHA, the rs1799963 was genotyped as part of the inclusion criteria. In the 3C study, the quality of its imputation was of borderline quality ($r^2 = 0.274$ <0.30). As a consequence, this variant was not included in my initial GWAS meta-analysis (that was restricted to imputed SNPs with r^2> 0.3 in both cohorts). Nevertheless, I ran a second round of GWAS where I further adjusted the analysis on the rs1799963 (imputed in 3C), separately in MARTHA and 3C, and meta-analyzed the results. Resulting Manhattan plot is shown in Figure 5.6.

Fourteen SNPs were still significantly associated with ETP at $p < 5 \cdot 10^{-8}$. The strongest association was observed at the 11p11 region, but for another *F2* polymorphism, rs3136516 ($p = 5.94 \cdot 10^{-14}$). A third round of (GWAS) analysis further adjusted for rs3136516 did not reveal any other significant SNP. As a reminder, the *F2* rs3136516 is another variant known to influence thrombin generation (page 27).

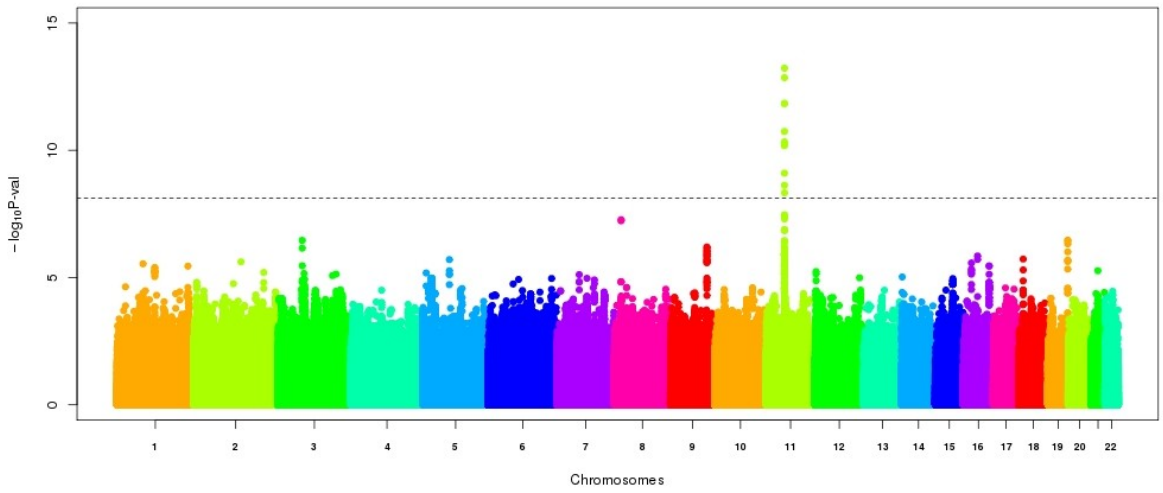


Figure 5.6. Manhattan plot representing the p-values of the meta-analyzed GWAS for ETP conditioning on the F2 rs1799963. The horizontal dashed line corresponds to the statistical threshold of $5\ 10^{-8}$.

A multivariate analysis simultaneously incorporating the *F2* rs1799963 and rs3136516 variants demonstrated that their effects on ETP were independent. In the meta-analysis of our two cohorts, the rs1799963-A ($\beta = + 0.225 \pm 0.019$, $p = 2.66 \cdot 10^{-31}$) and the rs3136516-G ($\beta = + 0.040 \pm 0.006$, $p = 5.89 \cdot 10^{-11}$) alleles were both associated with increased ETP. These effects were homogeneous in MARTHA ($\beta = + 0.224 \pm 0.019$; $\beta = + 0.042 \pm 0.009$, resp.) and in 3C ($\beta = + 0.287 \pm 0.135$, $\beta = + 0.039 \pm 0.008$, resp.), with no statistical difference ($p>0.05$) between studies. It's worth to note that, while the rs1799963-A allele was much more frequent in MARTHA patients than in 3C healthy subjects (0.058 vs 0.005), there was no difference for rs3136516-G allele frequencies (0.48 vs 0.47, resp.).

5.3.2. GWAS analysis on Peak height

The GWAS findings for Peak height (with MH plot shown in Figure 5.7.) paralleled those obtained for ETP. Twelve SNPs, all on the 11p11 region, were significantly associated with Peak height at $p < 5\ 10^{-8}$. These were as those strongly associated with ETP. In particular, the second most significant SNP was rs2856656 ($p = 8.29 \cdot 10^{-11}$). After further adjusting for rs1799963, two *F2* SNPs remained in complete association with Peak height, rs3136512 and rs3136516 (both $p = 2.91 \cdot 10^{-8}$). No other SNP was genome-wide significant after a second adjustment for rs3136516. As for the ETP biomarker, the effects of the rs1799963-A and rs3136516-G alleles were independent, $\beta = + 45.37 \pm 7.45$ ($p = 1.10 \cdot 10^{-9}$) and $\beta = + 9.79 \pm 2.03$ ($p = 1.36 \cdot 10^{-6}$), respectively. These effects were homogeneous between the MARTHA ($\beta = + 44.37 \pm 7.58$; $\beta = + 12.26 \pm 3.56$, resp.) and 3C ($\beta = + 73.22 \pm 39.97$; $\beta = + 8.62 \pm 2.46$, resp.) studies.

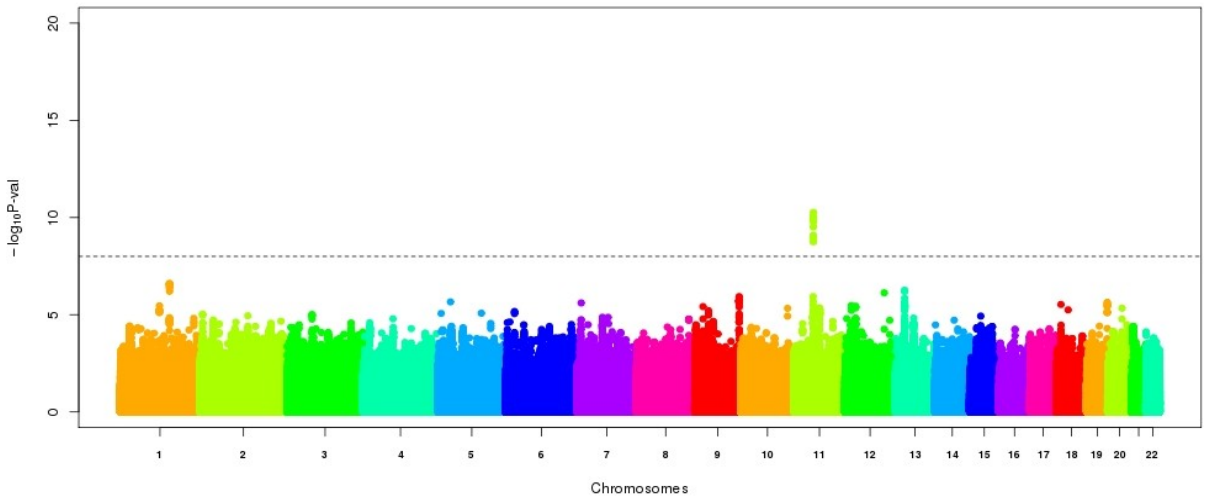


Figure 5.7. Manhattan plot representing the p-values of the GWAS meta analysis on Peak height. The horizontal dashed line corresponds to the statistical threshold of $5 \cdot 10^{-8}$.

5.3.3. GWAS analysis on Lagtime

As it can be observed on the MH plot summarizing the main GWAS findings for the Lagtime meta-analysis (Figure 5.8.), no SNP demonstrated strong evidence for association with the biomarker. Additionally, contrary to the previous results, Lagtime show no evidence for association with the two F2 SNPs, *F2* rs1799963 and rs3136516 (p = 0.096 and p= 0.451, respectively).

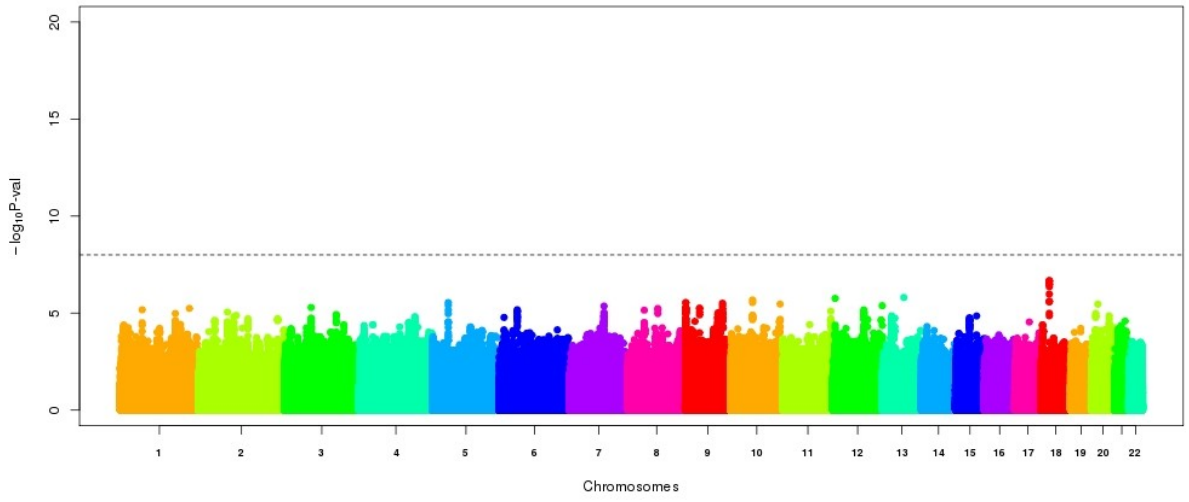


Figure 5.8. Manhattan plot representing the p-values of the GWAS meta analysis on Lagtime. The horizontal dashed line corresponds to the statistical threshold of $5 \cdot 10^{-8}$.

5.3.4. Joint GWAS analysis on two TGP biomarkers

In order to improve our detection of SNPs associated with TGP, I studied the polymorphisms with effects not strong enough to reach the pre-specified genome-wide significant threshold of $5 \cdot 10^{-8}$ but showing association at $p < 10^{-5}$ for two TGP biomarkers.

Joint analysis of ETP-Peak height - After adjusting for the *F2* rs1799963 and rs3136516, 7 SNPs mapping to chromosome 19q13, at the *RPL7AP69* locus, showed association at $p < 10^{-5}$ with both ETP and Peak height. The lowest p-value was observed for rs117368154 whose minor A allele was associated with decreased ETP ($\beta = -16.29 \pm 3.48$, $p = 2.83 \cdot 10^{-6}$) and decreased Peak height ($\beta = -0.054 \pm 0.011$, $p = 3.39 \cdot 10^{-7}$). These effects were homogeneous in the MARTHA and 3C studies, $\beta = -0.056$ $p = 6.5\ 10^{-4}$ and $\beta = -0.052$ $p = 3.4\ 10^{-4}$, respectively. After adjusting for the effect of the rs117368154, the associations at the other 6 RPL7AP69 SNPs completely vanished with p-values >0.0083 (threshold according to Bonferroni: 0.05/6).

Joint analysis of ETP - Lagtime - Four SNPs, in the promoter and coding regions of the *ORM1* gene, exhibited association at $p < 10^{-5}$ with both ETP and Lagtime. The lowest p-value was observed for rs150611042. Its minor A allele associated with lower ETP ($\beta = -0.068 \pm 0.013$, $p = 3.36 \cdot 10^{-7}$) and Lagtime ($\beta = -0.338 \pm 0.073$, $p = 4.10 \cdot 10^{-7}$). Results were again very homogeneous between MARTHA and 3C (see Table 5.3.). The rs150611042 is located in the promoter region of the *ORM1* gene and was in strong LD with other *ORM1* SNPs (minimum $r^2 > 0.86$) whose associations with Lagtime disappeared after adjusting for rs150611042 (p-values >0.013, threshold according to Bonferroni: 0.05/4)

Joint analysis of Peak height -Lagtime - No SNP was associated at $p < 10^{-5}$ with both Peak height and Lagtime.

5.4. Replication of GWAS findings

According to the results of the GWAS meta-analysis, I selected the *RPL7AP69* rs117368154 and *ORM1* rs150611042 for replication in the MARTHA12 study. No evidence for association of rs117368154 with ETP or Peak Height was observed (Table 5.4.). Conversely, I observed strong association of ORM1 rs150611042 with Lagtime ($p = 2.46\ 10^{-7}$). As in the two discovery GWAS cohorts, the rs150611042-A was associated with decreased Lagtime ($\beta = -0.439 \pm 0.084$) (Table 5.3), but not with ETP ($p = 0.147$). The rs150611042 was then further genotyped in the FITENAT study where its A allele was also associated with decreased Lagtime ($\beta = -0.280 \pm 0.099$, $p = 5.04 \times 10^{-3}$). Again, no association was found with ETP and Peak biomarkers (Table 5.3.).

Table 5.3. Association of *ORM1* rs150611042 with TGP biomarkers in four independent studies (Table extracted from Rocanin-Arjo *et al.* Blood 2014).

ORM1 rs150611042 C/A	MARTHA (N = 714)	3C (N =1,253)	MARTHA12 (N = 726)	FITENAT (N =528)	Combined [3] (N = 3,221)
Minor Allele Frequencies (A)	0.089	0.082	0.096	0.095	
Lagtime					
β[1] (SE)	-0.329 (0.086)	-0.343 (0.096)	-0.439 (0.084)	-0.280 (0.099)	-0.354 (0.045)
P[2]	p = 1.53 10^{-4}	p = 3.85 10^{-4}	p = 2.46 10^{-7}	p = 5.04 10^{-3}	p = 7.11 10^{-15}
ETP					
β (SE)	-0.041 (0.016)	-0.083 (0.018)	-0.024 (0.017)	-0.013 (0.017)	-0.038 (0.009)
P	p = 0.015	p = 9.31 10^{-6}	p = 0.147	p = 0.449	p = 8.41 10^{-6}
Peak					
β (SE)	-8.379 (6.107)	-6.557 (5.516)	8.730 (6.286)	-0.984 (4.592)	-2.013 (2.748)
P	p = 0.171	p = 0.233	p = 0.165	p = 0.830	p = 0.464

[1] Additive allele effects associated with the rs150611042-A allele.

[2] Association testing was performed by use of a linear regression model where values were adjusted for age, sex, oral contraceptive therapy (except in 3C) and on first four principal components (in MARTHA and 3C). In 3C, as well in 25 MARTHA patients, imputed allele dosage was used. Otherwise, the exact allele count derived from wet-lab genotyping was used.

[3] Combined results were derived from a meta-analysis of the four studies under the framework of a inverse-variance weighting fixed-effect model (page 49). No heterogeneity was observed across cohorts, I^2 = 1.67 (p = 0.795), I^2 = 8.59 (p = 0.072) and I^2 = 4.74 (p =0.314) for Lagtime, ETP and Peak, respectively.

Table 5.4. Association of *RPL7AP69* rs55724737 with TGP biomarkers (adapted from Rocanin-Arjo *et al*. Blood 2014).

*RPL7AP69*rs55724737	MARTHA (N = 714)	3C (N =1,253)	MARTHA12 (N = 733)	Combined [3] (N = 2,700)
Minor Allele Frequencies	0.087	0.071	0.067	
ETP β (SE) P	-0.063 (0.018) $p = 5.91\ 10^{-4}$	-0.075 (0.017) $p = 7.23\ 10^{-6}$	0.010 (0.020) p = 0.622	-0.048 (0.010)[4] $p = 5.25\ 10^{-6}$
Peak β (SE) P	-16.96 (6.675) $p = 8.95\ 10^{-3}$	-23.23 (4.841) $p = 1.79\ 10^{-6}$	-2.54 (7.494) p = 0.735	-17.23 (3.473) $p = 6.93\ 10^{-7}$
Lagtime β (SE) P	0.098 (0.096) p = 0.302	0.140 (0.085) p = 0.101	-0.003 (0.103) p = 0.974	0.087 (0.054) p = 0.107

It was technically not possible to genotype the *RPL7AP69* rs117368154 that was then substituted by its proxy rs55724737 (r^2 0.99, $p<10^{-16}$).
[1] Additive allele effects associated with the rs55724737-C allele.
[2] Association testing was performed by use of a linear regression model where values were adjusted for age, sex, oral contraceptive therapy (except in 3C) and on first four principal components (in MARTHA and 3C). Imputed allele dosage was used in MARTHA and 3C while the exact allele count derived from wet-lab genotyping was used in MARTHA12.
[3] Combined results were derived from a meta-analysis of the three studies under the framework of an inverse-variance weighting fixed-effect model.
[4] The association of rs55724737 with ETP showed significant heterogeneity (I^2 = 11.71, p = 0.003) across the three studies. No such heterogeneity was observed (p > 0.05) with Peak and Lagtime.

5.5. Further analyses at the identified ORM1 locus

When the four cohorts typed for the *ORM1* rs150611042 were combined into a fixed effect meta-analysis, the overall statistical for association with Lagtime was $p= 7.11 \cdot 10^{-15}$ with no heterogeneity across studies (p=0.795) (Table 5.4.). In the combined samples totaling 3,221 subjects, the decreasing effect on Lagtime associated with the rs150611042-A allele was β= -0.354 ± 0.045. Using a standard linear regression analysis, I further tested for any interaction between rs150611042 and F2 rs3136516 but did not observe such phenomenon (p for interaction = 0.172_MARTHA, MARTHA12 and 3C). In the two combined MARTHA and MARTHA12 VT samples enriched for *F2* rs1799963 variant, the rs150611042-A effect on Lagtime was homogeneous in patients with (β= -0.300 ± 0.255) or without (β=-0.385 ± 0.062) mutation.

In the discovery GWAS cohorts, the imputation quality for *ORM1* rs150611042 was r^2= 0.51 and r^2=0.56 in MARTHA and 3C, respectively. We also genotyped the rs150611042 SNP in 689 MARTHA with available DNA in order to get additional information about the obtained imputation quality. For this sample, I then compared the results of the association of rs150611042 with Lagtime obtained using the imputed or genotyped data. The imputed dose and the true allelic count derived from genotype data were highly correlated (Pearson

correlation ρ= 0.73). The association of rs150611042 with Lagtime was slightly stronger using true genotyped allele count (β= -0.338 ± 0.087, p = $1.11 \cdot 10^{-4}$) compared with the imputed dose (β= -0.362 ± 0.127, p = $4.71 \cdot 10^{-3}$).

5.5.1. Influence of ORM1 rs150611042 on clinical manifestations of VT

As shown in Table 5.3. the allele frequencies of the ORM1 rs150611042 were very similar in the two VT cohorts (MARTHA and MARTHA12) and the two healthy subject samples (3C and FITENAT). However, I was interested to test whether this SNP could be associated with the two different clinical manifestations of VT, deep vein thrombosis (DVT) and pulmonary embolism (PE). No difference in allele frequencies was observed between these two VT groups (Table 5.5.).

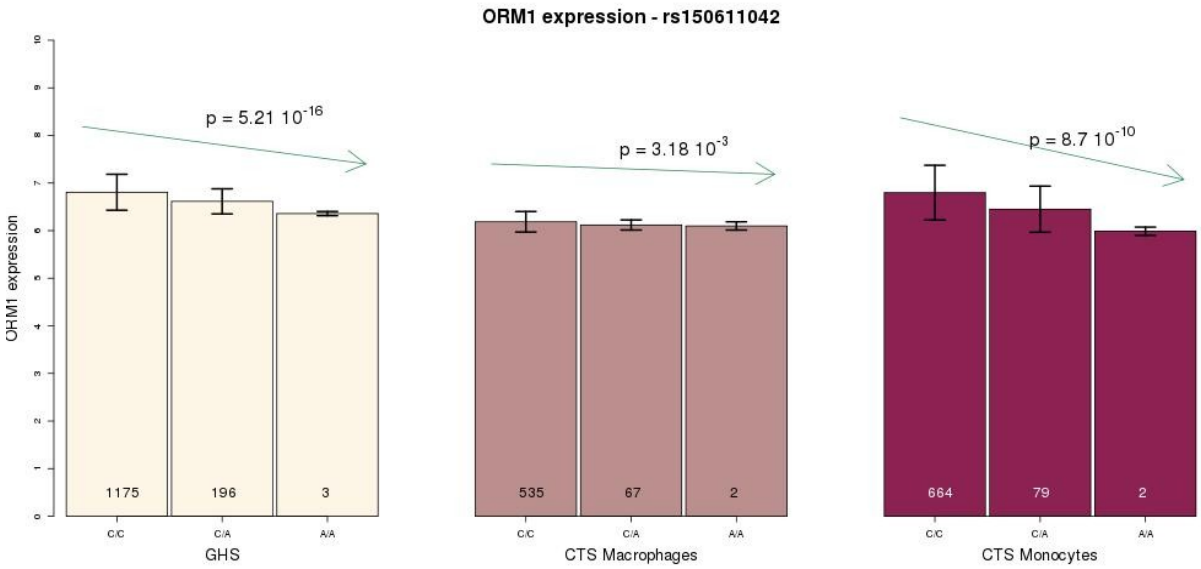
Table 5.5. Genotype frequencies of *ORM1* rs150611042 depending on the type of VT in MARTHA and MARTHA12.

ORM1 rs150611042 C/A	MARTHA $MAF_{(A)}$	MARTHA12 $MAF_{(A)}$
DVT	0.077	0.097
PE	0.056	0.079
DVT/PE	0.073	0.103
P-value[1)]	0.846	0.9585

[(1)]P-value obtained by a Fisher test to analyze the differences between the types of VT

5.5.2. *In silico* association with gene expression

To get support in favor of a potential functional effect of the *ORM1* rs150611042, I tested the association of this SNP with *ORM1* gene expression in two large epidemiological samples where the rs150611042 were imputed from GWAS data and where *ORM1* gene expression was also available. More precisely, I had access to monocyte and macrophage *ORM1* gene expression in 745 individuals of the Cardiogenics Transcriptomic Study (CTS)[141] where the rs150611042 was previously imputed (imputation r^2 = 0.64). I also had access to a second independent sample of 1,374 individuals part of the Gutenberg Health Study (GHS)[142] correctly imputed for rs150611042 (r^2 = 0.75) and measured for monocyte *ORM1* expression. In both studies, the rs150611042-A allele was significantly associated with decreased monocyte expression in a fairly additive manner (p = $8.70 \cdot 10^{-10}$ and p = $5.21 \cdot 10^{-16}$ in CTS and GHS, resp.). The ORM1 rs150611042 explained around the 5% of the variability of monocyte *ORM1* gene expression. A similar pattern of association, even though less significant (p = $3.18 \cdot 10^{-3}$), was observed in CTS on macrophage gene expression (Figure 5.9.).



(1) Mean (SD)
(2) Association test was performed between *ORM1* expression and the imputed allele dosage of rs150611042 while adjusting for age, sex and center (in CTS).

Figure 5.9. Association of ORM1 rs150611042 with *ORM1* monocyte and macrophage expression.

5.5.3. *In vitro* functional studies

To get further functional arguments for the observed *ORM1* association with Lagtime, the group of Professor Morange (INSERM UMR_S 1062) set up some in vitro experimental works. TGP biomarkers were measured in 10 poor platelet plasma samples with different concentrations of added ORM1 (see Results described in Rocanin-Arjo *et al.* Blood 2014). We showed that supplementing poor platelet plasma with orosomucoid, the protein encoded by the *ORM1* gene, was followed by a significant increase in Lagtime and significant decreases in ETP and Peak Height. These effects were dose-dependent. In these samples, the basal correlation between ORM1 levels and Lagtime was ρ = 0.646 (p = 0.049) (Figure 5.10.).

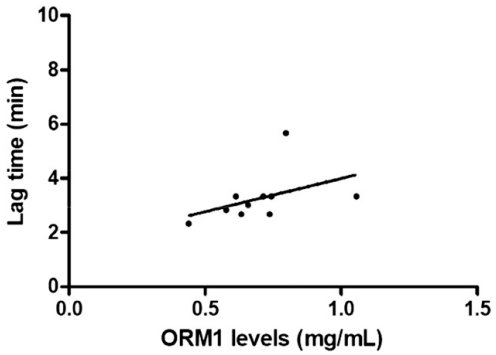


Figure 5.10. Correlation between ORM1 plasma levels and Lagtime in a sample of 10 healthy individuals.

5.6. Discussion

The (Meta) GWA study I conducted on the TGP biomarkers resulted in strong relation with two main genes: F2, affecting the variability of ETP and Peak for the same SNPs, and ORM1, a novel determinant of the variability of Lagtime.

5.6.1. The main associations with TGP

The two most significant F2 SNPs I detected in the GWAS, rs1799963 and rs3136516 were already known to influence thrombin generation by increasing the prothrombin levels. The first polymorphism, the rs1799963-A (G20210A) is located in the 3' UTR region of *F2* and so far, the studies point out that this variation at the 3'-end might be more efficient increasing mRNA stability, increasing translation efficiency, or a combination of these mechanisms[74,75,143]. The other, the rs3136516-G, also referred as A19911G, is located within the 13th intron of the gene, and affects an intronic splicing enhancer motif[69,70]. Remarkably, while the rs1799963-A allele is a rare (~2% in the general population) variant associated with a strong risk of VT, the rs3136516-G allele is common (~0.47 in all populations I have studied) and its association with VT risk is still debated[74,75,120,144]. A recent work (Segers 2010[6]) have related these two SNPs with thrombin generation, concretely with ETP measured by CAT[45] but in presence of activated protein C[20], in individuals heterozygous for FV Leiden.

The main original finding of this GWAS project is the identification of *ORM1* gene as a new locus participating somehow in the biological mechanisms of the thrombin generation pathway, mainly through an influence on the Lagtime phenotype. We detected a common *ORM1* SNP associated with both decreased Lagtime levels and monocyte ORM1 expression, and *in vitro* functional experiments confirmed the association between the encoded protein and thrombin generation. This SNP is located in the promoter region of the orosomucoid gene (Figure 5.11.).

As indicated in my first publication, ORM1, also called alpha 1 acid glycoprotein (alpha1-AGP), has been related to many processes and the most known is regulating the immune response as an acute phase protein[145]. It has also been related to dermatologic allergies such as dermatitis, psoriasis and sarcoidosis[146]. The most appealing works for this project are those who propose a possible implication of ORM1 in coagulation: as a transport protein of the blood stream, as a modulator of the synthesis of tissue factor, hence regulating the intrinsic pathway initiation. It has been associated with the normalization of platelet aggregation, coagulation factors and antithrombin activity[147]. Additionally, high doses of AGP inhibit a certain kind of stimulation of platelet aggregation (ADP and adrenaline)[148].

Apparently it interacts with plasminogen activator inhibitor 1 (PAI-1) too, stabilizing their inhibitory activity on plasminogen activators[149]. On the contrary, it reduces the activated Partial Thromboplastin Time (aPTT)[150] and stimulate the monocyte expression of tissue factor (TF)[151]. Lately, ORM1 has been proposed as a cardiometabolic biomarker related to BMI[152]. It is still unclear its exact role in all mentioned processes.

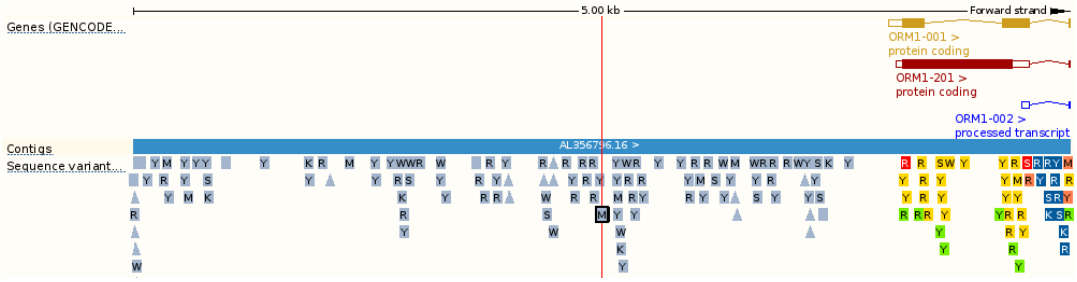


Figure 5.11. Diagram of the position of the rs150611042 SNP, the red line, marks the exact position of the SNP and in the top-left corner the ORM1 gene.

5.6.2. Limitations and comments

Despite the findings, the presented GWAS analyses suffer from some limitations that have already been acknowledged in the discussion of my first publication *A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential,* page 74)

One of the limitations is that the studies have different clinical and biological characteristics, leading to a possible lost of detection power in the statistical test. Then, the *ORM1* SNP was selected for its suggestive associations in the discovery cohorts with ETP and Lagtime. Although the correlation between Lagtime and ETP is stronger in the replication cohorts, we could not replicate the results in the case of ETP. Neither we can exclude the first association as false. Also, I analyzed the association of ORM1 expression with the rs150611042 SNP, in monocytes and macrophages, which was highly correlated. However, it would be interesting to analyze also the expression of hepatocytes, where orosomucoid is expressed. Eventually, we confirmed that ORM1 plasma levels were positively related with Lagtime in *in vitro* experiments, but also was strongly associated to ETP and Peak. That could be due to the low percentage (5% aprox, page 68) of variability that is explained by the hit SNP.

Additionally, another limitation of this study is the lack of association analysis of X chromosome genetic markers. I started the analysis just after conducting the autosomic one, but for different difficulties inherent to the properties of this chromosome, the work was paused. The chromosome X is particular compared with the other chromosomes[153], and the

most evident is that females have 2 copies, as an autosomal, and males have just one copy. This first characteristic marks already differences at the statistic analysis level making necessary to separate the samples by gender and apply differently tests or quality controls. For instance HWE is not suitable in males and imputation analysis was, at the beginnings of my PhD, recommended to be preformed separately resulting in decreasing imputation quality. It is also a chromosome with a low mutation rate, thus, it presents high linkage disequilibrium between the genes, so the tests are less independent between them. Nevertheless, it is one of the lines of TGP study that I want to continue.

Finally, aiming at complete the description of this first part studding TGP biomarkers I want to expose the following points:

- As detailed above, I observed strong evidence for association (1) between *ORM1* rs150611042 and Lagtime, (2) *ORM1* rs150611042 and *ORM1* gene expression, and (3) between orosomucoid plasma levels and Lagtime in different samples. It would be interesting to study the association of the SNP and ORM1 plasma levels. For that purpose, ORM1 plasma levels are being measured in 798 individuals of the MARTHA12 study and we should soon be able to check whether they are influenced by the rs150611042 SNP I have identified. So far, there is no functional evidence that it could be the "causal" variant. Therefore, it would be of great importance to identify the plausible causal variant(s) by deeply sequencing and characterizing the whole spectrum of the genetic variability of the locus. This may be a non-trivial task, as this locus is known to be subject of duplication[154]. Claire Perret, molecular biologist at INSERM UMR_S 1166, is currently investigating this issue and designing protocols to detect those duplications.

- The novel *ORM1* finding was identified because I did not focus only on the association signals that attained the genome-wide significant threshold of $5\ 10^{-8}$ but widened my search to SNPs with p-values as low as 10^{-5}. This emphasizes that some true associations may lie in the list of SNPs with less stringent p-values. This is why I have also undertaken a sort of candidate gene approach. I selected from all my meta-analysis GWAS results the most significant SNPs at VT candidate genes that showed association at $p < 0.05$ with any of the three TGP biomarkers (Annex, page 160). From these first lists, 11 SNPs were sent for genotyping replication in MARTHA12, which is currently under process and soon completed in Professor Pierre-Emmanuel Morange's laboratory. We should then be able to assess whether the 11 selected SNPs add to the list of SNPs that associate with thrombin generation, but with milder genetic effects than those observed at *F2* and *ORM1* loci.

blood

2014 123: 777-785
Prepublished online December 19, 2013;
doi:10.1182/blood-2013-10-529628

A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential

Ares Rocanin-Arjo, William Cohen, Laure Carcaillon, Corinne Frère, Noémie Saut, Luc Letenneur, Martine Alhenc-Gelas, Anne-Marie Dupuy, Marion Bertrand, Marie-Christine Alessi, Marine Germain, Philipp S. Wild, Tanja Zeller, Francois Cambien, Alison H. Goodall, Philippe Amouyel, Pierre-Yves Scarabin, David-Alexandre Trégouët, Pierre-Emmanuel Morange and and the CardioGenics Consortium

Blood (print ISSN 0006-4971, online ISSN 1528-0020), is published weekly by the American Society of Hematology, 2021 L St, NW, Suite 900, Washington DC 20036.

Regular Article

THROMBOSIS AND HEMOSTASIS

A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential

Ares Rocanin-Arjo,[1,2] William Cohen,[3,4] Laure Carcaillon,[5,6] Corinne Frère,[4] Noémie Saut,[3,4] Luc Letenneur,[7] Martine Alhenc-Gelas,[8] Anne-Marie Dupuy,[9] Marion Bertrand,[10] Marie-Christine Alessi,[3,4] Marine Germain,[1,2] Philipp S. Wild,[11-13] Tanja Zeller,[13,14] Francois Cambien,[1,2] Alison H. Goodall,[15] Philippe Amouyel,[16,17] Pierre-Yves Scarabin,[5,6] David-Alexandre Trégouët,[1,2] Pierre-Emmanuel Morange,[3,4] and the CardioGenics Consortium

[1]Pierre and Marie Curie University, INSERM, UMR_S 1166, Paris, France; [2]ICAN Institute for Cardiometabolism And Nutrition, Pierre and Marie Curie University, Paris, France; [3]Nutrition Obesity and Risk of Thrombosis, Aix-Marseille University, INSERM UMR_S 1062, Marseille, France; [4]Laboratory of Haematology, La Timone Hospital, Marseille, France; [5]CESP Centre for research in Epidemiology and Population Health, UMR-S1018, Hormones and Cardiovascular Disease, INSERM, Villejuif, France; [6]Université Paris Sud 11, Kremlin-Bicêtre, France; [7]INSERM U897, Bordeaux, France and University Bordeaux, ISPED, Bordeaux, France; [8]Service d'Hématologie Biologique, Hôpital Européen G Pompidou, Paris, France; [9]Hôpital La Colombière, INSERM UMR_S 1061, Montpellier, France; [10]Pierre and Marie Curie University, INSERM UMR_S 708, Paris, France; [11]Center for Thrombosis and Hemostasis, [12]Department of Medicine 2, and [13]German Center for Cardiovascular Research (DZHK), University Medical Center Mainz, Mainz, Germany; [14]Department of General and Interventional Cardiology, University Heart Center Hamburg, Hamburg, Germany; [15]Department of Cardiovascular Sciences, University of Leicester and Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, United Kingdom; [16]Institut Pasteur de Lille, Université de Lille Nord de France, INSERM UMR_S 744, Lille, France; and [17]Centre Hospitalier Régional Universitaire de Lille, Lille, France

Key Points

- Genetic variations at the *ORM1* locus and concentrations of the encoded protein associate with thrombin generation.
- These findings may guide the development of novel antithrombotic treatments.

Thrombin, the major enzyme of the hemostatic system, is involved in biological processes associated with several human diseases. The capacity of a given individual to generate thrombin, called the thrombin generation potential (TGP), can be robustly measured in plasma and was shown to associate with thrombotic disorders. To investigate the genetic architecture underlying the interindividual TGP variability, we conducted a genome-wide association study in 2 discovery samples (N = 1967) phenotyped for 3 TGP biomarkers, the endogenous thrombin potential, the peak height, and the lag time, and replicated the main findings in 2 independent studies (N = 1254). We identified the *ORM1* gene, coding for orosomucoid, as a novel locus associated with lag time variability, reflecting the initiation process of thrombin generation with a combined *P* value of $P = 7.1 \times 10^{-15}$ for the lead single nucleotide polymorphism (SNP) (rs150611042). This SNP was also observed to associate with *ORM1* expression in monocytes ($P = 8.7 \times 10^{-10}$) and macrophages ($P = 3.2 \times 10^{-3}$). In vitro functional experiments further demonstrated that supplementing normal plasma with increasing orosomucoid concentrations was associated with impaired thrombin generation. These results pave the way for novel mechanistic pathways and therapeutic perspectives in the etiology of thrombin-related disorders. (*Blood*. 2014;123(5):777-785)

Introduction

The enzyme thrombin (also called activated factor II) is a central product of the response to vascular injury, displaying procoagulant, anticoagulant, antifibrinolytic, and cellular effects; the magnitude and timing of these effects are critical to normal hemostasis.

The vast majority of thrombin is generated well after the plasma (or blood) clot formation time, which is the traditional endpoint for the activated partial thromboplastin time and prothrombin time assays.[1] These tests do not assess the whole coagulation system and also are insensitive to prothrombotic states.[2] Thrombin generation assays have recently gained in popularity and are considered useful to measure "global hemostasis," that is, capturing the complete dynamics of the coagulant response beyond initial clot formation.[3]

Patients with high levels of thrombin generation are at risk for thrombotic diseases such as acute ischemic stroke,[4] venous thromboembolism (VTE),[5-8] and myocardial infarction[9] while bleeding events are observed in presence of very low thrombin generation.[10] In addition, the role of thrombin generation extends far beyond the sole coagulation system. Several recent findings have emphasized its key impact in atherosclerosis,[11] diabetic nephropathy,[12,13] and inflammatory diseases such as sepsis,[14] Crohn disease,[15] and sickle cell disease.[16]

Altogether these observations clearly emphasize the importance of identifying factors controlling the interindividual variability of thrombin generation. Known environmental and biological determinants of thrombin generation are body mass index, estrogen-based

Submitted October 2, 2013; accepted December 7, 2013. Prepublished online as *Blood* First Edition paper, December 19, 2013; DOI 10.1182/blood-2013-10-529628.

D.-A.T. and P.-E.M. contributed equally to this work.

The online version of this article contains a data supplement.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

therapies, factor II, fibrinogen, antithrombin, tissue factor pathway inhibitor (TFPI) levels in plasma[17-19] while 2 functional genetic variants, both in the *F2* gene coding for prothrombin, rs1799963 (G20210A), and rs3136516,[20] have been robustly shown to associate with thrombin generation levels. The former is a well-established genetic risk factor for VT[21] where the rs1799963-A allele is associated with a 2.5-fold increased risk.[22] Conversely, the association of the common rs13136516 with VT risk remains questionable. It has been observed in 2 case-control studies[23,24] but was not detected in recent genome-wide association studies (GWAS).[25-29] Preliminary works also suggested that the rs3136516-G allele could associate with increased risk of systemic lupus erythematous (SLE).[30] If confirmed, this association would add support for the role of thrombin generation into inflammatory diseases.

We hypothesized that additional genetic factors, outside the *F2* gene, could modulate the potential of a given individual to generate thrombin. To test this hypothesis, we undertook a GWAS in 2 independent populations totaling 1967 subjects using 1000 Genomes based imputation techniques which allowed us to test 6 652 054 single nucleotide polymorphisms (SNPs) for association with 3 thrombin generation parameters: endogenous thrombin potential (ETP), lag time and peak thrombin generation using the calibrated automated thrombography (CAT) method.[31] The main findings were tested for replication in 2 additional independent populations gathering 1339 individuals and functional arguments derived from in silico and in vitro experiments were obtained to support the identified novel association.

Methods

Studied populations

Two independent cohorts with both GWAS data and thrombin generation measurements were used for the discovery stage: MARTHA and the Three-City (3C) Study. The main findings of the meta-analysis of these 2 GWAS datasets were tested for replication in 2 additional independent studies, MARTHA12 and FITENAT. 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 descent.

Discovery cohorts. The MARTHA study has already been extensively described.[32,33] It is composed of 1542 patients with venous thrombosis (VT) recruited from the thrombophilia center of La Timone hospital (Marseille, France). All subjects, with a documented history of VT, were free of any chronic conditions and free of any thrombophilia including anti-thrombin, protein C and protein S deficiencies, and homozygosity for factor V Leiden and factor II G20210A mutations. Patients under anticoagulant therapy were also excluded. The 3C study[34] is a population-based study carried out in 3 French cities composed of 8707 noninstitutionalized individuals aged over 65 randomly selected from the electoral rolls and free of any chronic diseases and for which biological (DNA, plasma) samples could have been obtained.

Replication cohorts. The MARTHA12 study is composed of an independent sample of 1245 VT patients that have been recruited between 2010 and 2012 according to the same criteria as the MARTHA patients. The FITENAT sample[35] consists of 543 French healthy individuals selected from health examination centers of the French Social Security. These subjects had no history of cardiovascular disease, diabetes, hypertension, renal nor hepatic failure and were not under anticoagulant therapy.

Biological measurements

In all studies, thrombin generation potential (TGP) was measured in platelet-poor plasma (PPP) using the CAT method[36] as extensively described in Lavigne-Lissalde et al.[18] Three biological TGP parameters were derived from the thrombogram analysis: the ETP (in nmol min^{-1}) which corresponds to the area under the thrombogram curve, the peak thrombin generation (peak in nmol L^{-1}) which represents the maximum amount of thrombin produced after induction by 5pM tissue factor (TF), and the lag time (LagT in minutes) which represents the time to the initial generation of thrombin after induction.

In vitro functional studies

Plasma orosomucoid levels were determined using an automated turbidimetric immunoassay based on the use of a polyclonal rabbit anti-human orosomucoid covalently attached to polystyrene microparticles resulting in a ready to use immunoparticle reagent. All reagents were from Dako A/S and all assays were performed on a Vitros 5.1 from Ortho Clinical Diagnostics.

Orosomucoid (Cell Science) was added at the concentrations of 0.2, 0.4, 0.8, 1.2 and 1.6 mg/mL of plasma to 80 μL of PPP dispensed into the wells of round-bottom 96 well-microtiter plates (Immunolon microtiter 96-well solid plates; Fischer Scientific). These concentrations were chosen for corresponding to the normal range of orosomucoid in plasma (0.6-1.6 mg/mL). Thrombin generation was then initiated by adding 20 μL of PPP reagent (Stago) containing 5 pM TF and 4 μM phospholipid mixture, and measured using the CAT method in a Fluoroskan Ascent fluorometer (Thermolab Systems OY) equipped with a dispenser. Fluorescence intensity was detected at wavelengths of 390 nm (excitation filter) and 460 nm (emission filter). The starting reagent FLUCA Kit (Stago) containing fluorogenic substrate and $CaCl_2$ was automatically dispensed by the fluorometer (20 μL per well). A dedicated software program, Thrombinoscope (Stago) was used to calculate thrombin activity against the Calibrator (Stago) and display thrombin activity vs time. ETP, peak, and lag time were calculated from the thrombogram. These experiments were conducted on normal plasma samples from healthy random individuals, Wilcoxon-paired test statistic was used to assess the association of orosomucoid with TGP biomarkers.

Genotyping and imputation

Plasma were available for TGP measurements in 848 MARTHA patients and 1314 3C subjects with genome-wide genotype data typed with Illumina Human610 and Huma660W Quad beadchips.[28] In each study, SNPs with genotyping call rate <99%, minor allele frequency (MAF) <1%, and showing significant ($P < 10^{-5}$) deviation from the Hardy-Weinberg equilibrium[37] were filtered out. This led to 491 285 and 487 154 quality-control (QC) validated autosomal SNPs in MARTHA and 3C, respectively. Individuals were excluded according to the following criteria: genotyping rate <95%, close relatedness as suspected from pairwise clustering of identity by state distances and multi-dimensional scaling implemented in PLINK,[38] and genetic outliers detected by principal components analysis as implemented in the EIGENSTRAT program.[39] Finally, 714 and 1253 individuals from the MARTHA and 3C study, respectively, were left for association analyses.

The 467 355 QC-checked SNPs common to both MARTHA and 3C were then used for imputing 11 572 501 autosomal SNPs from the 1000 Genomes 2010-08 release reference dataset. For this, the MACH (version 1.0.18.c) software was used.[40] All SNPs with acceptable imputation quality ($r^2 > 0.3$)[29,41] and MAF > 0.01 in both imputed GWAS datasets were kept for association analysis.

In the replication cohorts, genotyping was performed using allele-specific PCR in 733 MARTHA12 patients and 528 FITENAT subjects in which TGP measurements and DNA were available.

Statistical analysis

Discovery analysis. In order to handle nonnormality distributions, a log-transformation and a normal quantile transformation[42] were applied to ETP and lag time values, respectively, separately in the 2 cohorts (supplemental Figure 1, available on the *Blood* Web site). Association of imputed SNPs with TGP markers were assessed independently in each cohort by a linear regression model implemented in the MACH2QTL (version 1.1.0)[40] software. In this model, the allele dosage, a real number ranging from 0 to 2 and equal to the expected number of minor alleles computed from the posterior probabilities of possible imputed genotypes, is used for assessing the imputed SNP effect.

Table 1. Characteristics of the studied populations

	Discovery		Replication	
	MARTHA, N = 714	3C, N = 1253	MARTHA12, N = 796	FITENAT, N = 543
Age, y (SE)	46.84 (15.29)	75.05 (5.75)	49.74 (15.34)	47.81 (13.94)
Sex, % male	32.1	41.6	42.9	47.9
VT patients, %	100	0	100	0
BMI, kg/m^2 (SE)	25.1 (4.55)	25.8 (4.22)	26.0 (4.98)	24.2 (3.62)
FV Leiden (%)*	153 (21.4)	—	89 (11.2)	—
F2 G20210A (%)*	83 (11.6)	—	52 (6.5)	—
Oral anticoagulant (%)	—	52 (4.2)	24 (3.0)	—
BMI, kg/m^2	25.14 (4.55)		26.03 (4.98)	
ETP, nM/min (1st-3rd)	1780 (1554-1958)	1775 (1586-1947)	1892 (1654-2124)	1675 (1456-1838)
Peak, nM (1st-3rd)	333.0 (293.5-372.0)	332.8 (307.0-364.3)	328.0 (278.0-374.8)	293.5 (258.1-319.2)
Lag time, min (1st-3rd)	3.229 (2.830-3.500)	1.382 (1.000-1.670)	3.330 (2.762-3.670)	2.280 (2.000-2.500)

BMI, body mass index; FV, factor V.
*FV Leiden and F2 G20210A mutations were genotyped in MARTHA and MARTHA12 as part of the inclusion criteria. Homozygous carriers were not included in the study.

Analyses were adjusted for age, gender, oral contraception therapy (in MARTHA), oral coagulant therapy (3C), and the first 4 principal components. Results obtained in the 2 GWAS cohorts were entered into a fixed-effect meta-analysis relying on the inverse-variance weighted method as implemented in the METAL program.[43] Homogeneity of associations across the 2 studies was assessed using the Mantel-Haenszel method.[44] A statistical threshold of 5×10^{-8} was used to declare genome-wide significance.[41,45,46] In order to increase the sensitivity of our discovery phase, we also considered of potential interest any SNP that did not reach genome-wide significance but were nevertheless associated at $P < 10^{-5}$ with at least 2 TGP biomarkers.

Conditional analysis. A second round of GWAS analysis was performed where we further conditioned on the Prothrombin G20210A (rs1777963) mutation, a known strong genetic determinants of TGP markers. Because the rs1777963 has been genotyped in the MARTHA study as part of the inclusion criteria, the true genotypes were then used in the conditional MARTHA GWAS while, in the 3C study, the imputed allele dose were used.

Replication analysis. The same transformations as in the discovery cohorts were applied to ETP and lag time values in MARTHA12 and FITENAT. Association of tested SNPs with TGP markers was also assessed by a linear regression model under the assumption of additive allele effects, adjusted for age, sex, and oral contraception therapy.

In silico association with gene expression

The identified *ORM1* hit SNP was investigated for association with the expression of its corresponding gene in monocytes and macrophages. Two genome-wide expression studies were used, the Cardiogenics Transcriptomics Study (CTS)[47,48] and the Gutenberg Health Study.[48,49]

For this work, CTS individuals initially typed for the Illumina Sentrix Human Custom 1.2M and human 610-Quad beadchips and GHS subjects typed for Affymetrix Genome-Wide Human SNP Array 6.0 were separately imputed by the MACH (version 1.0.18.c) software according to the 1000 Genomes February 2012 reference database. RNA genome-wide expression from monocytes and macrophages was assessed in CTS using the Illumina HumanRef 8 version 3 Beadchip. In GHS, the Illumina HT-12 version 3 expression array was used to assess monocyte expression. In both datasets, *ORM1* gene expression was characterized by the ILMN_1696584 probe.

Association of the hit SNP with gene expression was tested by use of a linear regression model adjusted for age, sex, and center (in CTS).

Results

Characteristics of the 4 populations studied are given in Tables 1 and 2.

Discovery meta-analysis

A total of 6 652 054 imputed SNPs common to both GWAS cohorts satisfied pre-specified imputation quality criteria and were then tested through meta-analysis for association with the 3 TGP phenotypes, ETP, peak, and lag time. Quantile-quantile (Q-Q) plots of the association results did not reveal any inflation from what expected under the null hypothesis of no association, except for the extreme right tail distribution for ETP and peak (supplemental Figure 2) Corresponding genomic inflation coefficients were 0.998, 0.997, and 1.000 for ETP, peak, and lag time, respectively. Manhattan plot representation of the results are depicted in Figure 1.

ETP analysis. Seventeen SNPs, all mapping the chromosome 11p11 region, reached genome-wide significance for association with ETP (supplemental Table 1). The strongest signal was observed for *MYBPC3* rs2856656 ($P = 4.62 \times 10^{-22}$). As we have previously shown that this SNP tags for the *F2* G20210A (rs1799963) mutation,[29] a further GWAS meta-analysis was conducted by conditioning on the rs1799963. The imputation quality criteria of the rs1799963 was $r^2 = 1$ in MARTHA (see "Methods") and $r^2 = 0.274$ in 3C. The rare rs1799963-A allele was much more frequent in MARTHA thus allowing better imputation. As a consequence, it was not included in the initial set of imputed SNPs that entered the GWAS analysis. After adjusting for rs1799963, 14 SNPs remained significantly associated with ETP. The strongest association was observed for the *F2* rs3136516 ($P = 5.94 \times 10^{-14}$). After a further round of adjustment on the rs3136516 allele dosage, no association remained genome-wide significant. A 2-locus model incorporating the rs1799963 and rs3136516 revealed that their effects on ETP were independent and highly significant ($P = 1.02 \times 10^{-29}$). The rs1799963-A ($\beta = +0.225 \pm 0.019$, $P = 2.66 \times 10^{-31}$) and the rs3136516-G ($\beta = +0.040 \pm 0.006$, $P = 5.89 \times 10^{-11}$) alleles were both associated with increased ETP. These effects were homogeneous in MARTHA ($\beta = +0.224 \pm 0.019$; $\beta = +0.042 \pm 0.009$, respectively) and in 3C ($\beta = +0.287 \pm 0.135$, $\beta = +0.039 \pm 0.008$, respectively), with no statistical difference ($P > .05$) between studies. It is worthy of note that, while the rs1799963-A allele was much more frequent in MARTHA patients than in 3C healthy

Table 2. Correlation between TGP markers

	MARTHA	3C	MARTHA12	FITENAT
ETP – Peak, r	0.78	0.77	0.73	0.78
ETP – Lag time, r	0.14	0.06	0.19	0.34
Peak – Lag time, r	−0.05	−0.10	−0.18	0.04

Correlations were computed on transformed values (ie, log-transformation on ETP and quantile normalization for lag time) adjusted for age, sex, oral anticoagulant therapy, F2 G20210A (when appropriate), and BMI.
Abbreviations are explained in Table 1.

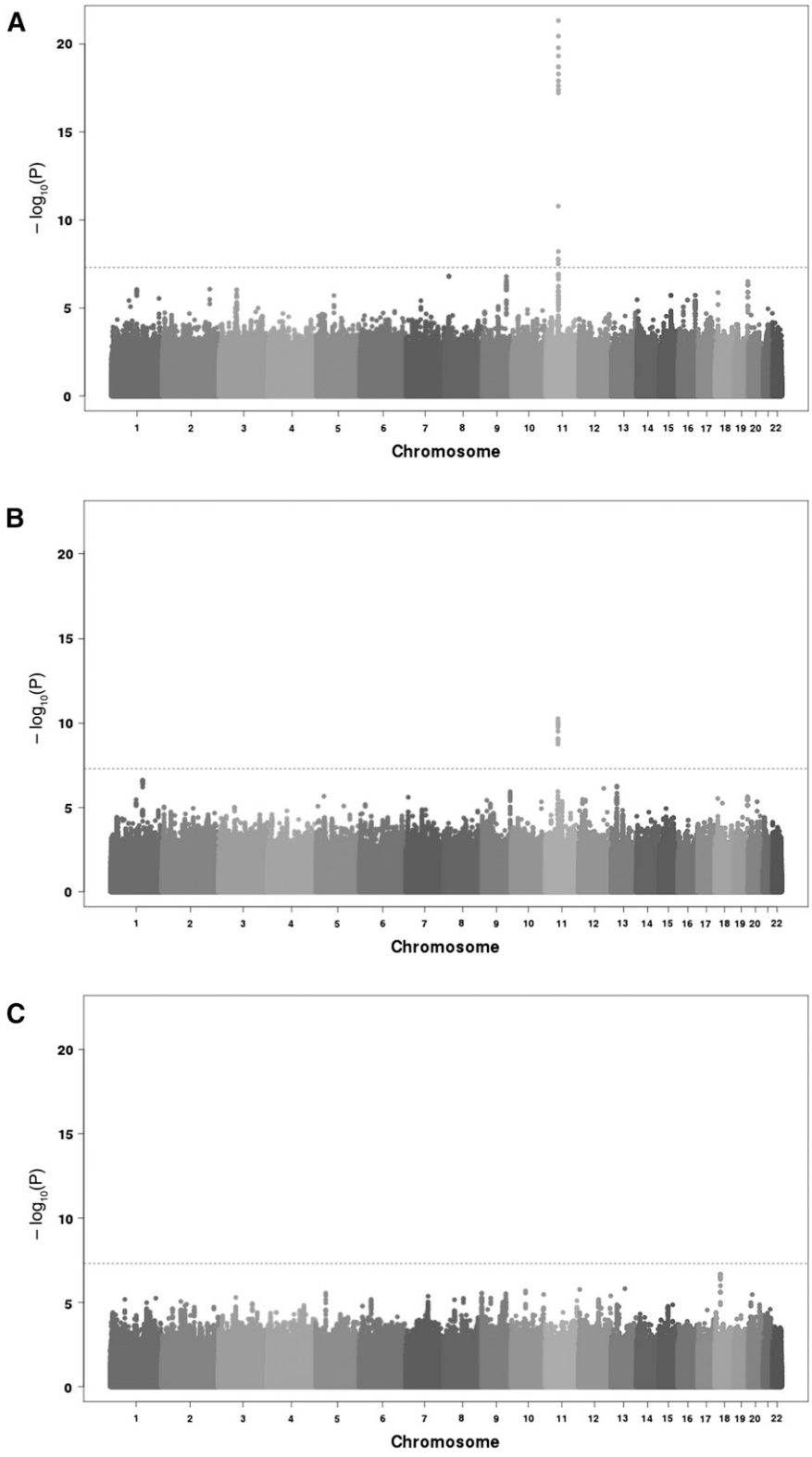


Figure 1. Manhattan plots of the association results from the meta-analysis of 2 discovery cohorts imputed for 6 652 054 SNPs on 3 TGP biomarkers. ETP (A), peak (B), and lag time (C). The horizontal line corresponds to the genome-wide significant threshold taken at 5×10^{-8}.

subjects (0.058 vs 0.005), there was no difference in the rs3136516-G allele frequencies (0.48 vs 0.47, respectively)

Peak analysis. Twelve SNPs at the 11p11.2 locus were significantly associated with peak, the strongest signal being for the rs138315285 ($P = 5.48 \times 10^{-11}$) and the second hit being the rs2856656 ($P = 8.29 \times 10^{-11}$) (supplemental Table 2). After adjusting for rs1799963, 2 associations, rs3136512 and rs3136516, remained statistically significant (both $P = 2.91 \times 10^{-8}$). These 2 SNPs were in perfect linkage disequilibrium (LD). In a joint model, the rs1799963-A and rs3136516-G alleles were independently associated with increased peak levels, $\beta = +45.37 \pm 7.45$ ($P = 1.10 \times 10^{-9}$) and $\beta = +9.79 \pm 2.03$ ($P = 1.36 \times 10^{-6}$),

Table 3. Association of ORM1 rs150611042 with biomarkers of thrombin generation in 4 independent studies

ORM1 rs150611042 C/A	MARTHA, N = 714	3C, N = 1253	MARTHA12, N = 726	FITENAT, N = 528	Combined,* N = 3221
Minor allele frequencies	0.089	0.082	0.096	0.095	
Lag time					
β† (SE)	−0.329 (0.086)	−0.343 (0.096)	−0.439 (0.084)	−0.280 (0.099)	−0.354 (0.045)
P	1.53×10^{-4}	3.85×10^{-4}	2.46×10^{-7}	5.04×10^{-3}	7.11×10^{-15}
ETP					
β (SE)	−0.041 (0.016)	−0.083 (0.018)	−0.024 (0.017)	−0.013 (0.017)	−0.038 (0.009)
P	.015	9.31×10^{-6}	.147	.449	8.41×10^{-6}
Peak					
β (SE)	−8.379 (6.107)	−6.557 (5.516)	8.730 (6.286)	−0.984 (4.592)	−2.013 (2.748)
P	.171	.233	.165	.830	.464

Abbreviations are explained in Table 1.

*Combined results were derived from a meta-analysis of the 4 studies under the framework of an inverse-variance weighting fixed-effect model. No heterogeneity was observed across cohorts, $I^2 = 1.67$ ($P = .795$), $I^2 = 8.59$ ($P = 0.072$), and $I^2 = 4.74$ ($P = .314$) for lag time, ETP, and peak, respectively.

†Additive effects associated with the rs150611042-A allele, adjusted for age, sex, oral contraceptive therapy (except in 3C), and on first 4 principal components (in MARTHA and 3C). In 3C, as well in 25 MARTHA patients, imputed allele dosage was used. Otherwise, the exact allele count derived from wet-laboratory genotyping was used. Association was tested by use of a linear regression model.

respectively. These effects were of similar amplitude in MARTHA ($\beta = +44.37 \pm 7.58$; $\beta = +12.26 \pm 3.56$, respectively) and in 3C ($\beta = +73.22 \pm 39.97$; $\beta = +8.62 \pm 2.46$, respectively).

Lag time analysis. No SNP was genome-wide significantly associated with lag time. Of note, the *F2* rs1799963 and rs3136516 did not associate with lag time ($P = .096$ and $P = .451$, respectively).

"Joint analysis" of TGP phenotypes. As the GWAS analyses of TGP phenotypes did not reveal any genome-wide significant association signal independent of the known *F2* variants (rs1799963 and rs3136516), we followed up additional SNPs that demonstrated suggestive evidence for association ($P < 10^{-5}$) with at least 2 of the 3 studied TGP biomarkers.

Four SNPs, all mapping to the chromosome 9 *ORM1* gene, demonstrated suggestive evidence for association with both ETP and lag time (supplemental Table 3). The strongest association was observed for rs150611042 whose A rare allele was associated with lower ETP ($\beta = -0.068 \pm 0.013$, $P = 3.36 \times 10^{-7}$) and lag time ($\beta = -0.338 \pm 0.073$, $P = 4.10 \times 10^{-7}$) with no evidence for heterogeneity between MARTHA and 3C ($P = .729$ and $P = .912$, respectively). After adjusting for rs150611042, the association at the other *ORM1* SNPs completely vanished confirming that these 4 SNPs were in strong LD as initially anticipated from the similarity in their allele frequencies and associated genetic effects.

Thirty-eight SNPs were suggestively associated with both ETP and peak, most of them being located in the 11p11 region discussed above. After adjusting for rs1799963 and rs3136516 *F2* variants, we observed suggestive association between a block of 7 SNPs mapping to the *RPL7AP69* locus on chromosome 19q13.43 and ETP and peak (supplemental Table 4). All these SNPs were in nearly complete association, the minor allele of the best associated SNP (rs117368154) was associated with decreased ETP ($\beta = -16.29 \pm 3.48$, $P = 2.83 \times 10^{-6}$) and peak ($\beta = -0.054 \pm 0.011$, $P = 3.39 \times 10^{-7}$).

No SNP exhibited association at $P < 10^{-5}$ with both peak and lag time.

Replication studies

Following the main findings derived from the discovery meta-analysis, *ORM1* and *RPL7AP69* hit SNPs were tested for association with TGP phenotypes in MARTHA12. No association with any TGP biomarker was observed for the *RPL7AP69* SNP (supplemental Table 5). Conversely, the *ORM1* rs150611042 demonstrated significant association ($P = 2.46 \times 10^{-7}$) with lag time but not with ETP ($P = .147$). Consistent with the discovery results, the rs150611042-A allele was associated with decreased lag time ($\beta = -0.439 \pm 0.084$). To provide additional support for this association, the rs150611042 was genotyped in the FITENAT study where its A allele was also associated with decreased lag time ($\beta = -0.280 \pm 0.099$, $P = 5.04 \times 10^{-3}$) (Table 3).

rs150611042 was imputed in the discovery GWAS. As a consequence, we de novo genotyped it in 689 patients of the MARTHA GWAS where DNA was still available. In this sample, the Pearson correlation between the imputed dose and the true genotype was $\rho = 0.73$. The association of rs150611042 with lag time was slightly stronger using true genotyped allele count ($\beta = -0.338 \pm 0.087$, $P = 1.11 \times 10^{-4}$) than that observed using imputed allele dose ($\beta = -0.362 \pm 0.127$, $P = 4.71 \times 10^{-3}$).

Finally, the meta-analysis of the 4 studies provide strong statistical evidence for the association of rs150611042 with lag time ($P = 7.11 \times 10^{-15}$) with no evidence for heterogeneity across studies ($P = .795$) (Table 3). In the combined samples totaling 3,221 subjects, the decreasing effect of the rs150611042-A allele was $\beta = -0.354 \pm 0.045$. No evidence for heterogeneity according to the presence of the rs3136516-G allele was observed ($P = .172$). Similarly, in the 2 combined VT samples enriched for *F2* G20210A mutation, the rs150611042-A effect on lag time was homogeneous (test for homogeneity $P = .746$) in patients with ($\beta = -0.300 \pm 0.255$) or without ($\beta = -0.385 \pm 0.062$) the rs1799963-A allele. Further adjustment of BMI did not alter the detected association ($\beta = -0.343 \pm 0.044$, $P = 1.08 \times 10^{-14}$)

No association of rs150611042 with peak nor ETP was observed (Table 3).

In silico association with gene expression

In the CTS and GHS studies, the rs150611042 was correctly imputed, with $r^2 = 0.64$ and $r^2 = 0.75$, respectively. In both studies, the rs150611042-A allele was significantly ($P = 8.70 \times 10^{-10}$ and $P = 5.21 \times 10^{-16}$ in CTS and GHS, respectively) associated with decreased expression in monocytes in an additive manner (Table 4) and explained ~5% of the variability of *ORM1* monocyte gene expression. A similar pattern of association, although less significant ($P = 3.18 \times 10^{-3}$), was observed for *ORM1* gene expression in macrophages from CTS (Table 4).

In vitro functional studies

As illustrated in Figure 2, orosomucoid levels were positively correlated ($r^2 = 0.646$, $P = .049$) with lag time in a plasma sample of 10

Table 4. Association of ORM1 rs150611042 with *ORM1* monocyte and macrophage expression

	CTS			GHS	
ORM1 rs150611042	N	Monocyte	Macrophage	N	Monocyte
CC*	664	6.80 (0.57)	6.19 (0.21)	1175	6.81 (0.38)
CA*	79	6.45 (0.48)	6.12 (0.11)	196	6.62 (0.26)
AA*	2	5.99 (0.09)	6.10 (0.09)	3	6.36 (0.04)
R^2, %		4.80	1.30		4.90
P†		8.70×10^{-10}	3.18×10^{-3}		5.21×10^{-16}

CTS, Cardiogenics Transcriptomics Study; GHS, Gutenberg Health Study.
*Mean (SD).
†Association testing was performed by regressing *ORM1* expression on the imputed allele dosage of rs150611042 while adjusting for age, sex, and center (in CTS).

individuals. As shown in supplemental Figure 3, there was a positive dose-dependent association between orosomucoid concentrations and lag time. For instance, at the 1.6 mg/mL concentration, the supplementation of orosomucoid to PPP was followed by a modification of the thrombin activity characterized by significant increased lag time (3.25 vs 4.50 miutesn, $P = .0057$) but also significant decreased ETP (1790 vs 1628 nmol L^{-1}, $P = .0020$) and peak (266 vs 213 nmol L^{-1} $P = .002$) (Figure 3).

Discussion

Here, we reported the results of a GWAS study aimed at identifying genetic variations associated with thrombin generation through a comprehensive analysis of 3 complementary biomarkers, ETP, peak and lag time.

For ETP and peak, associations were observed with rs1799963 and rs3136516, 2 *F2* variants already known to associate with thrombin generation and whose functionality has already been discussed.[20,50-54] Both rare alleles, rs1799963-A and rs3136516-G, were associated independently from each other with increased thrombin generation, the strongest effect being at rs1799963. These 2 mutations act on thrombin generation through increase in prothrombin levels. The mechanism by which the rs1799963-A, located in the 3′ UTR region of *F2*, influences prothrombin levels has been proposed to result from more efficient 3′-end formation, increased mRNA stability, increased translation efficiency, or a combination of these mechanisms.[50,54] The rs3136516-G, located within the 13th intron of the gene, is functional through its effect on an intronic splicing enhancer motif.[20] Of note, while the rs1799963-A allele is a rare (~2% in the general population) variant associated with a strong risk of VT, the rs3136516-G allele is common (~0.47 in all populations studied here) and its association with VT risk still warrants in-depth investigation.

The novelty of this work is the association of *ORM1* rs150611042-A allele with decreased lag time in 4 independent studies, with an overall statistical evidence of $P = 7.11 \times 10^{-15}$. Although this polymorphism explained ~2% of the lag time variability (1%, 1.7%, 3.4%, and 1.4% in MARTHA, 3C, MARTHA12 and FITENAT, respectively), it was not associated with either ETP or peak. No evidence for association with VT risk was suggested either, the frequency of the rs150611042-A allele (Table 3) being homogeneous across the 2 cohorts of VT patients (0.089 and 0.096) and the 2 cohorts of healthy individuals (0.082 and 0.095). The rs150611042 is located in the promoter region of the *ORM1* gene and was in strong LD with other *ORM1* SNPs whose association with lag time disappeared after adjusting for rs150611042. Using transcriptomic data, we further observed that the rs150611042-A allele was associated with decreased *ORM1* expression in monocytes and macrophages. Finally, in vitro functional studies revealed that plasma orosomucoid levels correlate with lag time and that supplementing the plasma of healthy individuals with orosomucoid resulted in impaired thrombin activity as characterized by increased lag time; this observation being consistent with the concomitant associations of rs150611042-A allele with both decreased lag time and decreased *ORM1* expression.

ORM1 encodes a key acute phase plasma protein, orosomucoid also called α-1-acid glycoprotein 1 (α-1-AGP)[55] whose specific function has not yet been determined; it might function as a transport protein in the blood stream, appears to modulate the immune system during the acute-phase reaction and has been shown to associate with allergic contact dermatitis, psoriasis, and sarcoidosis.[56] Several pieces of evidence support a role of α-1-AGP in coagulation. In an experimental model of peritonitis in rats, high doses of α-1-AGP normalized platelet aggregation, blood clotting parameters and antithrombin activity.[57] Increased amounts of α-1-AGP inhibit platelet aggregation induced by ADP and adrenaline.[58] α-1-AGP was also found to interact with plasminogen activator inhibitor 1, a member of the serine proteinase inhibitors (serpins) family, and to stabilize its inhibitory activity toward plasminogen activators.[59] One could speculate that α-1-AGP delays thrombin generation by playing the same role with other coagulation inhibitors belonging to the serpin family. Conversely, α-1-AGP has been observed to shorten aPTT[60] and to contribute to the cellular initiation of coagulation by inducing monocyte expression of TF.[61]

Our study provides evidence of the role of α-1-AGP in thrombin generation, in particularly on the initiation process evaluated through the lag time biomarker. Several limitations must be acknowledged. First, the 2 discovery cohorts as well as the 2 replication studies were composed of samples of VT patients and healthy individuals exhibiting different clinical and biological characteristics. This may then have introduced heterogeneity between studied individuals resulting in a decreased power for detecting genetic effects, in particular of modest effect size. Conversely, the association of *ORM1* SNPs with lag time observed in these 4 different cohorts is a strong argument in favor of a true association. The *ORM1* locus was selected for further investigation in the replication studies because of its suggestive association ($P < 10^{-5}$) with both ETP and lag time in the combined discovery cohorts. However, only the association with lag time robustly replicated. As the correlation between ETP and lag time was stronger in the replication studies than in the discovery cohorts, we could have anticipated that the association with ETP would also hold

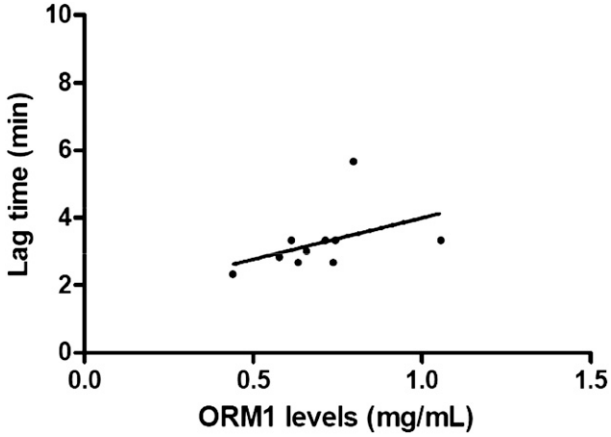


Figure 2. Association of lag time with plasma ORM1 levels. In 10 normal plasma, the correlation between ORM1 and lag time was 0.646 ($P = .049$).

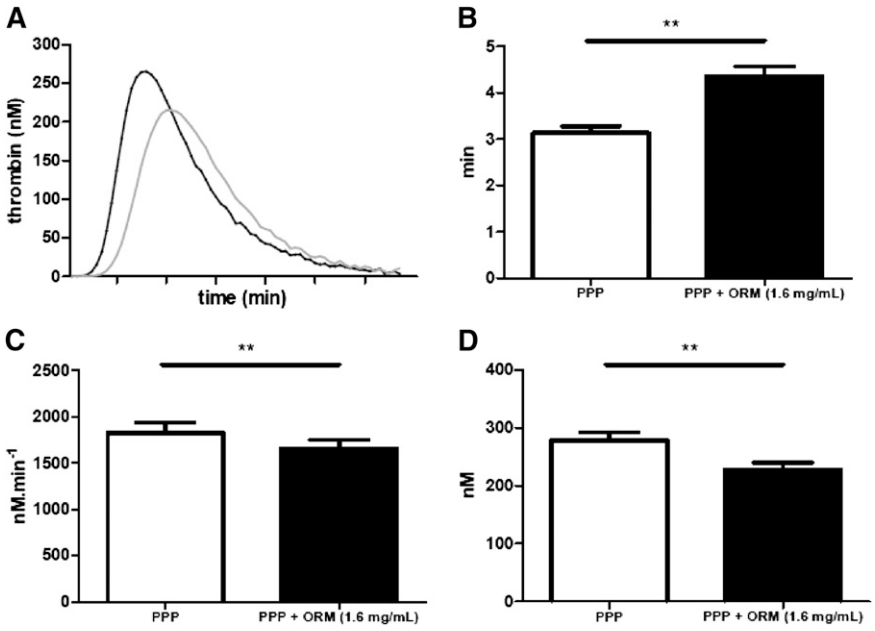


Figure 3. Influence of orosomucoid supplementation on thrombin generation. (A) Thrombin generation curves from 10 normal plasma that were (gray curves) or were not supplemented (black curves) with 1.6 mg/mL orosomucoid. Effect of supplementation of PPP with 1.6 mg/mL orosomucoid over lag time (B), ETP (C), and peak (D). Measurements were performed in duplicates and the mean values were used for each individual. $^{**}P < .01$.

in the replication. Therefore, we cannot rule out that the original association with ETP was spurious. We observed strong statistical evidence for association of our hit SNP with *ORM1* expression in monocytes and macrophages, but its influence on gene expression in other cell types would also be of great interest, in particular in hepatocytes, the main source of orosomucoid. Finally, while our in vitro experiments confirmed the association of *ORM1* with lag time, they also strongly suggested some associations with both ETP and peak biomarkers which were not robustly suspected from the GWAS investigations. This discrepancy could be explained by the fact that the identified SNP explains a modest part of the *ORM1* variability (eg, ~5% in monocytes) which could be only enough to detect an influence on lag time, whereas our in vitro experiments were able to reflect a more global and stronger effect of orosomucoid on all TGP biomarkers. This would emphasize the need for an in-depth investigation of all genetic and non-genetic (incl. epigenetic) factors influencing *ORM1* expression and their impact on thrombin generation. In particular, based on our preliminary results, it would be highly relevant to assess in larger populations the correlation between plasma ORM1 and TGP biomarkers and whether this correlation could be influenced by *ORM1* SNPs.

Despite these limitations, our results strongly support a role of *ORM1* in thrombin generation-related mechanisms. The impact of the identified SNP on thrombin generation is mild and does not seem to be sufficient to modify the risk of VT in the general population. Orosomucoid is an acute phase reaction protein which increases in concentration as much as 5-fold in acute inflammation and cancer.[59] We can speculate that its direct influence on thrombin generation might be higher during inflammation process and thus responsible for coagulation disorders observed in such circumstances. Lastly, the association of the identified *ORM1* polymorphism with thrombin-associated human diseases (eg, Crohn, SLE, diabetic nephropathy, stroke) would warrant further investigations.

Acknowledgments

A.R.-A. was supported by a grant from the Regional Ile de France (CORDDIM). The MARTHA project was supported by a grant from the Program Hospitalier de Recherche Clinique and a grant from CSL Behring. 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 Générale 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 3C Study was supported by a grant from the Agence Nationale pour la Recherche (ANR 2007-LVIE-005-01). CARDIOGENICS was funded by the European Union FP6 program (LSHM-CT-2006-037593). Collection of the Cardiogenics controls was in part supported through the Cambridge Bioresource, which is funded by the NIHR Cambridge Biomedical Research Centre. The Gutenberg Health Study is funded through the government of Rheinland-Pfalz (''Stiftung Rheinland Pfalz für Innovation,'' contract AZ 961-386261/733), the research programs ''Wissenschafft Zukunft'' and ''Schwerpunkt Vaskulare Prävention'' of the Johannes Gutenberg-University of Mainz and its contract with Boehringer Ingelheim and PHILIPS Medical Systems, including an unrestricted grant for the Gutenberg Health Study. This study was supported by the National Genome Network ''NGFNplus'' (contract A3 01GS0833 and 01GS0831) and by joint funding from the Federal Ministry of Education and Research, Germany (contract BMBF 01KU0908A), and from the Agence Nationale de la Recherche, France (contract ANR 09 GENO 106 01), for the project CARDomics. P.S.W. is funded by the Federal Ministry of Education and Research (BMBF 01EO1003).

Statistical analyses used the C2BIG Computing Centre funded by the Regional Ile de France ("CORRDIM") and the Pierre and Marie Curie University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authorship

Contribution: A.R.-A. performed all statistical analyses in the discovery and replication studies, and drafted the manuscript; W.C. contributed to patient and biological data collection; L.C., M.A.-G., M.-C.A., L.L., A.-M.D., M.B., P.A., and P.-Y.S. collected data from the discovery studies; C.F. and N.S. organized the wet-laboratory experiments, including genotyping and in vitro experiments; P.S.W., T.Z., F.C., and A.H.G. coordinated the expression analyses; M.G. and D.-A.T. supervised all statistical works; L.C., C.F., F.C., A.H.G., P.A., P.-Y.S., D.-A.T., and P.-E.M. wrote the manuscript; and D.-A.T. and P.-E.M. designed the project.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

A complete list of the members of the CardioGenics Consortium appears in "Appendix."

Correspondence: David-Alexandre Tregouet, INSERM UMR_S 1166, 91 Boulevard de l'Hopital, 75013 Paris, France; e-mail: david.tregouet@upmc.fr; and Pierre-Emmanuel Morange, Laboratory of Haematology, CHU Timone, 246, rue Saint-Pierre, 13385 Marseille Cedex 05, France; e-mail: pierre.morange@ap-hm.fr.

References

1. Hemker HC, Béguin S. Thrombin generation in plasma: its assessment via the endogenous thrombin potential. *Thromb Haemost.* 1995;74(1):134-138.
2. van Veen JJ, Gatt A, Makris M. Thrombin generation testing in routine clinical practice: are we there yet? *Br J Haematol.* 2008;142(6):889-903.
3. Brummel-Ziedins K. Models for thrombin generation and risk of disease. *J Thromb Haemost.* 2013;11(Suppl 1):212-223.
4. Carcaillon L, Alhenc-Gelas M, Bejot Y, et al. Increased thrombin generation is associated with acute ischemic stroke but not with coronary heart disease in the elderly: the Three-City cohort study. *Arterioscler Thromb Vasc Biol.* 2011;31(6):1445-1451.
5. Lutsey PL, Folsom AR, Heckbert SR, Cushman M. Peak thrombin generation and subsequent venous thromboembolism: the Longitudinal Investigation of Thromboembolism Etiology (LITE) study. *J Thromb Haemost.* 2009;7(10):1639-1648.
6. Hron G, Kollars M, Binder BR, Eichinger S, Kyrle PA. Identification of patients at low risk for recurrent venous thromboembolism by measuring thrombin generation. *JAMA.* 2006;296(4):397-402.
7. Tripodi A, Legnani C, Chantarangkul V, Cosmi B, Palareti G, Mannucci PM. High thrombin generation measured in the presence of thrombomodulin is associated with an increased risk of recurrent venous thromboembolism. *J Thromb Haemost.* 2008;6(8):1327-1333.
8. Besser M, Baglin C, Luddington R, van Hylckama Vlieg A, Baglin T. High rate of unprovoked recurrent venous thrombosis is associated with high thrombin-generating potential in a prospective cohort study. *J Thromb Haemost.* 2008;6(10):1720-1725.
9. Smid M, Dielis AW, Spronk HM, et al. Thrombin generation in the glasgow myocardial infarction study. *PLoS ONE.* 2013;8(6):e66977.
10. Rugeri L, Quélin F, Chatard B, De Mazancourt P, Negrier C, Dargaud Y. Thrombin generation in patients with factor XI deficiency and clinical bleeding risk. *Haemophilia.* 2010;16(5):771-777.
11. Borissoff JI, Spronk HM, ten Cate H. The hemostatic system as a modulator of atherosclerosis. *N Engl J Med.* 2011;364(18):1746-1760.
12. Ay L, Hoellerl F, Ay C, et al. Thrombin generation in type 2 diabetes with albuminuria and macrovascular disease. *Eur J Clin Invest.* 2012;42(5):470-477.
13. van der Poll T. Thrombin and diabetic nephropathy. *Blood.* 2011;117(19):5015-5016.
14. Petros S, Kliem P, Siegemund T, Siegemund R. Thrombin generation in severe sepsis. *Thromb Res.* 2012;129(6):797-800.
15. Bernhard H, Deutschmann A, Leschnik B, et al. Thrombin generation in pediatric patients with Crohn's disease. *Inflamm Bowel Dis.* 2011;17(11):2333-2339.
16. Noubouossie DF, Lê PQ, Corazza F, et al. Thrombin generation reveals high procoagulant potential in the plasma of sickle cell disease children. *Am J Hematol.* 2012;87(2):145-149.
17. Dielis AW, Castoldi E, Spronk HM, et al. Coagulation factors and the protein C system as determinants of thrombin generation in a normal population. *J Thromb Haemost.* 2008;6(1):125-131.
18. Lavigne-Lissalde G, Sanchez C, Castelli C, et al. Prothrombin G20210A carriers the genetic mutation and a history of venous thrombosis contributes to thrombin generation independently of factor II plasma levels. *J Thromb Haemost.* 2010;8(5):942-949.
19. Scarabin PY, Hemker HC, Clément C, Soisson V, Alhenc-Gelas M. Increased thrombin generation among postmenopausal women using hormone therapy: importance of the route of estrogen administration and progestogens. *Menopause.* 2011;18(8):873-879.
20. von Ahsen N, Oellerich M. The intronic prothrombin 19911A>G polymorphism influences splicing efficiency and modulates effects of the 20210G>A polymorphism on mRNA amount and expression in a stable reporter gene assay system. *Blood.* 2004;103(2):586-593.
21. Poort SR, Rosendaal FR, Reitsma PH, Bertina RM. 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.* 1996;88(10):3698-3703.
22. Sode BF, Allin KH, Dahl M, Gyntelberg F, Nordestgaard BG. Risk of venous thromboembolism and myocardial infarction associated with factor V Leiden and prothrombin mutations and blood type. *CMAJ.* 2013;185(5):E229-E237.
23. Martinelli I, Battaglioli T, Tosetto A, et al. Prothrombin A19911G polymorphism and the risk of venous thromboembolism. *J Thromb Haemost.* 2006;4(12):2582-2586.
24. Chinthammitr Y, Vos HL, Rosendaal FR, Doggen CJ. The association of prothrombin A19911G polymorphism with plasma prothrombin activity and venous thrombosis: results of the MEGA study, a large population-based case-control study. *J Thromb Haemost.* 2006;4(12):2587-2592.
25. Tang W, Teichert M, Chasman DI, et al. A genome-wide association study for venous thromboembolism: the extended cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium. *Genet Epidemiol.* 2013;37(5):512-521.
26. Heit JA, Armasu SM, Asmann YW, et al. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *J Thromb Haemost.* 2012;10(8):1521-1531.
27. Trégouët DA, Heath S, Saut N, et al. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood.* 2009;113(21):5298-5303.
28. Germain M, Saut N, Greliche N, et al. Genetics of venous thrombosis: insights from a new genome wide association study. *PLoS ONE.* 2011;6(9):e25581.
29. Germain M, Saut N, Oudot-Mellakh T, et al. Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. *PLoS ONE.* 2012;7(6):e38538.
30. Demirci FY, Dressen AS, Kammerer CM, et al. Functional polymorphisms of the coagulation factor II gene (F2) and susceptibility to systemic lupus erythematosus. *J Rheumatol.* 2011;38(4):652-657.
31. Hemker HC, Al Dieri R, De Smedt E, Béguin S. Thrombin generation, a function test of the haemostatic-thrombotic system. *Thromb Haemost.* 2006;96(5):553-561.
32. Antoni G, Morange PE, Luo Y, 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(12):2671-2679.
33. Oudot-Mellakh T, Cohen W, Germain M, et al. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. *Br J Haematol.* 2012;157(2):230-239.
34. 3C Study Group. Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology.* 2003;22(6):316-325.
35. Mazoyer E, Ripoll L, Gueguen R, et al; FITENAT Study Group. Prevalence of factor V Leiden and prothrombin G20210A mutation in a large French population selected for nonthrombotic history: geographical and age distribution. *Blood Coagul Fibrinolysis.* 2009;20(7):503-510.
36. Hemker HC, Kremers R. Data management in thrombin generation. *Thromb Res.* 2013;131(1):3-11.
37. Weale ME. Quality control for genome-wide association studies. *Methods Mol Biol.* 2010;628:341-372.
38. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559-575.
39. 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(8):904-909.

40. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 2010;34(8):816-834.
41. Johnson EO, Hancock DB, Levy JL, et al. Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum Genet.* 2013;132(5): 509-522.
42. Peng B, Yu RK, Dehoff KL, Amos CI. Normalizing a large number of quantitative traits using empirical normal quantile transformation. *BMC Proc.* 2007;(Suppl 1):S156.
43. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* 2010;26(17): 2190-2191.
44. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst.* 1959;22(4):719-748.
45. Panagiotou OA, Ioannidis JP; Genome-Wide Significance Project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol.* 2012;41(1):273-286.
46. Li MX, Yeung JM, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet.* 2012; 131(5):747-756.
47. Heinig M, Petretto E, Wallace C, et al; Cardiogenics Consortium. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature.* 2010;467(7314): 460-464.
48. Garnier S, Truong V, Brocheton J, et al; Cardiogenics Consortium. Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. *PLoS Genet.* 2013;9(1): e1003240.
49. Zeller T, Wild P, Szymczak S, et al. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS ONE.* 2010;5(5): e10693.
50. Gehring NH, Frede U, Neu-Yilik G, et al. Increased efficiency of mRNA 3′ end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nat Genet.* 2001;28(4): 389-392.
51. Carter AM, Sachchithananthan M, Stasinopoulos S, Maurer F, Medcalf RL. Prothrombin G20210A is a bifunctional gene polymorphism. *Thromb Haemost.* 2002;87(5):846-853.
52. Pollak ES, Lam HS, Russell JE. The G20210A mutation does not affect the stability of prothrombin mRNA in vivo. *Blood.* 2002;100(1): 359-362.
53. Ceelie H, Spaargaren-Van Riel CC, De Jong M, Bertina RM, Vos HL. Functional characterization of transcription factor binding sites for HNF1-alpha, HNF3-beta (FOXA2), HNF4-alpha, Sp1 and Sp3 in the human prothrombin gene enhancer. *J Thromb Haemost.* 2003;1(8): 1688-1698.
54. Danckwardt S, Gehring NH, Neu-Yilik G, et al. The prothrombin 3′end formation signal reveals a unique architecture that is sensitive to thrombophilic gain-of-function mutations. *Blood.* 2004;104(2):428-435.
55. Treuheit MJ, Costello CE, Halsall HB. Analysis of the five glycosylation sites of human alpha 1-acid glycoprotein. *Biochem J.* 1992;283(Pt 1):105-112.
56. Fan C. Orosomucoid types in allergic contact dermatitis. *Hum Hered.* 1995;45(2):117-120.
57. Osikov MV, Makarov EV, Krivokhizhina LV. Effects of alpha1-acid glycoprotein on hemostasis in experimental septic peritonitis. *Bull Exp Biol Med.* 2007;144(2):178-180.
58. Costello M, Fiedel BA, Gewurz H. Inhibition of platelet aggregation by native and desialised alpha-1 acid glycoprotein. *Nature.* 1979; 281(5733):677-678.
59. Boncela J, Papiewska I, Fijalkowska I, Walkowiak B, Cierniewski CS. Acute phase protein alpha 1-acid glycoprotein interacts with plasminogen activator inhibitor type 1 and stabilizes its inhibitory activity. *J Biol Chem.* 2001;276(38): 35305-35311.
60. Klatzow DJ, Vos GH. The effect of seromucoid on coagulation. *S Afr Med J.* 1981;60(11):424-427.
61. Su SJ, Yeh TM. Effects of alpha 1-acid glycoprotein on tissue factor expression and tumor necrosis factor secretion in human monocytes. *Immunopharmacology.* 1996;34(2-3): 139-145.

PARTII

INVESTIGATIONS OF DNA METHYLATION MARKS ASSOCIATED TO THE THROMBIN GENERATION POTENTIAL.

This second part describes the results related to the study of DNA methylation patterns regulating the inter-individual variability of Thrombin Generation Potential (TGP), another main objective of my PhD project.

The strategy was very similar to the one I adopted in the previous Part I (Identification of novel genetic factors influencing TGP plasma variability). Using the data from a dedicated microarray, I conducted a methylome-wide scan for DNA methylation changes associated with the three TGP biomarkers in a subsample of the MARTHA study. Afterwards, I performed a replication analysis of the most significant findings in an independent study composed of French-Canadian pedigrees. Eventually, I analyzed the complete methylome-wide scan in these pedigrees and combined with the results of MARTHA in a meta-analysis aiming to increase the statistical power to detect TGP associated methylation marks.

The structure will mimic Part I, starting by basic notions of epigenetics and the main principles of methylation wide association studies (MWAS). Then I describe the studied populations and the results I obtained. The main findings are summarized in a manuscript currently under review in Thrombosis Research (pages 120-130)

Chapter 6. DNA methylation, an epigenetic mechanism

Epigenetics is a branch of the genetics, which enclose the regulation activity not caused by genes. Arthur Riggs defined it as: "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence"[155]. As the name signs "*epi*" that in Greek means "over", they are all those marks and patterns upon or over the genes and genetic structural proteins that regulate the transcription and expression of them. These marks are stable: persist after cellular mitosis and meiosis so they can be inherited (imprinting, epigenetic memory) from one mother cell to daughter cell. However, additionally they can be modified by gene activity or environmental factors[156,157] causing changes in cells physiology or state and even cell disorders as tumour cells[158,159]. These marks could be the explanation to the question why two cells with the same genetic material will express different proteins and become different cell types. While the organism cells have the same genetic content, the epigenetic pattern is tissue specific[160]. In relation to this, this epigenetic pattern play also a role in development of the organisms, like with HOX genes, which are used several times in the development stage, in different places. Their differential expression in time and space is regulated epigenetically[156].

There are two principal types of DNA epigenetic marks (Figure 6.1.):

1. Histone code
2. DNA methylation

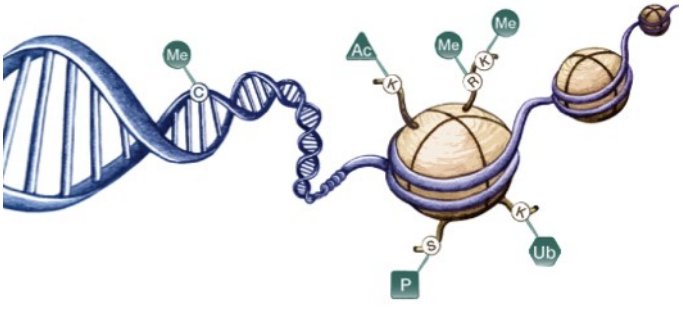


Figure 6.1. Image of the two main epigenetic types, the tails modification represented at the right of the image and the DNA methylation of C's at the left

6.1. Histone epigenetic code

Histones are nuclear proteins bound to the DNA double helix to help with its stabilization and packaging[161]. Together, they form the so-called nucleosome (Figure 6.2.) The DNA fibber makes almost 2 turns (equivalent to 2 nm) around the histone core and another histone unit (H1) will fix this complex. The histone core is formed by 8 histone units: 2xH2A, 2xH2B, 2xH3 and 2xH4.

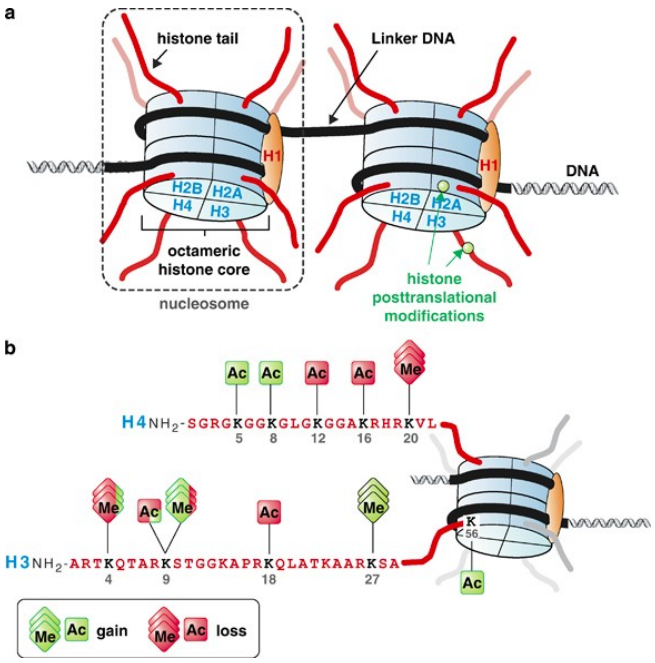


Figure 6.2. Scheme of histone cores[158]. (a) the DNA surrounds the core (formed by 8 histone units) and the histone H1 fix the proteic complex. All together form the basic unit named: Nucleosome. (b) Example of two histone tails (the sequence and line in red) modified by acetyl (Ac) and Methyl (Me) molecules. One of the most known histone modifications is the Histone 3 - Lysine 4 (H3K4), which is dimethylated and trimethylated- in this case trimethylated. It appears all along the genome in promoter regions[161].

The tails of the polypeptides (N-terminals) of these histones stand out from the nucleosome and are subject to several modifications such as acetylation, mono-, di- or/and tri-methylation of the lysine amino acids, and phosphorylation of serine amino acids (Figure 6.2.). The different combinations of these added-molecules are known as the histone code. According to this code the gene expression is enhanced (attracting i.e., transcription factors) or blocked (attracting i.e., packaging proteins). This kind of epigenetic mark is a short-term modification in the expression[162].

6.2. DNA methylation

DNA methylation consists in an enzymatic addition of a methyl group (CH_3, one carbon molecule and three of hydrogen) at the carbon 5 position of cytosine mainly in the context of the sequence 5'-cytosine-guanosine (CpG dinuclotide). This methyl group is transferred by a DNA methyltransferase from a molecule called S-adenylmethylcisteine (SAM), which

becomes a S-adenylhomocysteine (SAH) (Figure 6.3.). At the same time, the DNA methyltransferase attracts proteins to bind the DNA that will regulate the expression and the chromatin structure too[161].

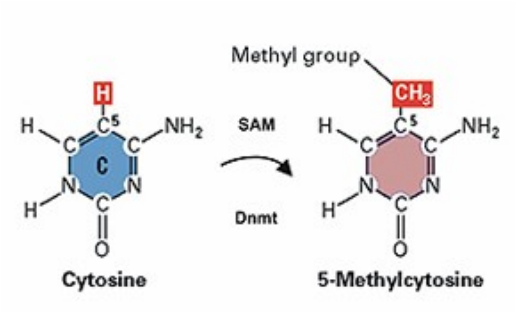


Figure 6.3. Methylation process.

Recent works have documented that DNA methylation could also occur, not only at CpG sites, but also at CpHpG sites where H can be any A, T or C nucleotide[163]. RNA methylation is also another recently observed epigenetic mechanisms[164].

The CpG dinucleotide is strongly under-represented in the human genome probably due to a mutational process in mammals that transforms the methylcytosines in thymines. Additionally, it is irregularly distributed over the genome presenting deserts interspersed by regions of high density of CpGs called CpG islands[165]. The proximal regions around those islands are called the island shores, and in turn, surrounding the shores, there are the shelves. The CpG islands contain the 7% of CpGs of the genome and they can be located either in the regulatory 5'-UTR ends or in the first exon of ~50% of human genes[162,166,167]. It has been estimated that about 70% of cytosines of CpG are subject to methylation in the human genome[168].

The unmethylated status of a cytosine is generally agreed as the standard state allowing the transcription of the downstream gene. Conversely, DNA methylation represses physically and chemically the transcription by attracting methyl-CpG binding proteins and other enzymes that modify the structure and the composition of chromatin. All these expression modifications lead to trait changes and also diseases.

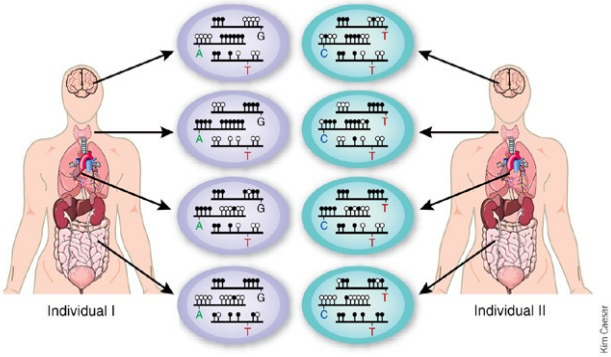


Figure 6.4. DNA methylation tissue specificity[169]**.** The black dots are for methylated sites and white for unmethylated

DNA methylation is more resistant than the histone modification, making possible its transmission from mother cells to daughter cells[170,171]. At the same time, these marks are

reversible epigenetic marks as they can be modified by gene action or by environmental influence as diet, smoking or air pollution[172]. The methylation pattern is specific of each cell type; actually it contributes to the cell differentiation making possible the development of different cell types from stem cells with the same genome[171,173] (Figure 6.4.). It can be even allele-specific, like Paternal/Maternal imprinting and X inactivation, which are two important biological mechanisms that are due to regulatory DNA methylation mechanisms[160].

6.3. DNA methylation and human diseases

In the last decade, emerging evidences have support the role of DNA methylation in human diseases. First results were observed in the field of cancer but rapidly extended to other complex human diseases such as multiple sclerosis, diabetes, inflammatory and cardiovascular diseases[174–179].

Several elements support the implication of DNA methylation in the pathophysiology of thrombotic disorders. For instance, the *F8* and *VWF* genes (two well-known quantitative biomarkers associated with VT) seem to be subject to epigenetic marks including DNA methylation[180,181]. They have also been observed regulating the expression of the methylenetetrahydrofolate reductase (*MTHFR*) gene[182,183] which influences the plasma variability of homocystein levels, another risk factor for VT[184]. Interestingly, in this last example, methylation also has been proposed as SNP-sensitive being present only when a certain polymorphism in the *MTHFR* gene.

6.4. Peripheral blood DNA methylation, a good epidemiological tool

In the study of cancer, DNA methylation patterns were identified in free floating DNA released in serum from apoptotic or necrotic tumour cells. That lead to the idea of looking for DNA methylation marks in blood serum and plasma[165]. Thereafter, this DNA blood method was followed in other successful non-cancer DNA methylation studies[185–187].

Additionally, the peripheral blood DNA is preferable to study DNA methylation changes because is easy to obtain compared to other relevant cell types that are impractical at a large epidemiological scale. Although, DNA from peripheral blood is composed of a mixture of different cell types (mainly leukocytes), the underlying hypothesis is that those DNA methylation marks are reflective of much stronger changes in specific cell-types.

A very recent but highly popular tool to measure peripheral blood DNA methylation is the Illumina HumanMethylation450K (HM450K) array, used in this project. The description of the underlying principles and of the bioinformatics and statistical analyses of the data produced are given in the following Chapter 7.

Chapter 7. How to perform a Methylation Wide Association step by step

To conduct a Methylation Wide Association (MWAS), several steps similar to those I described for a GWAS strategy must be followed. They are briefly summarized in the diagram below:

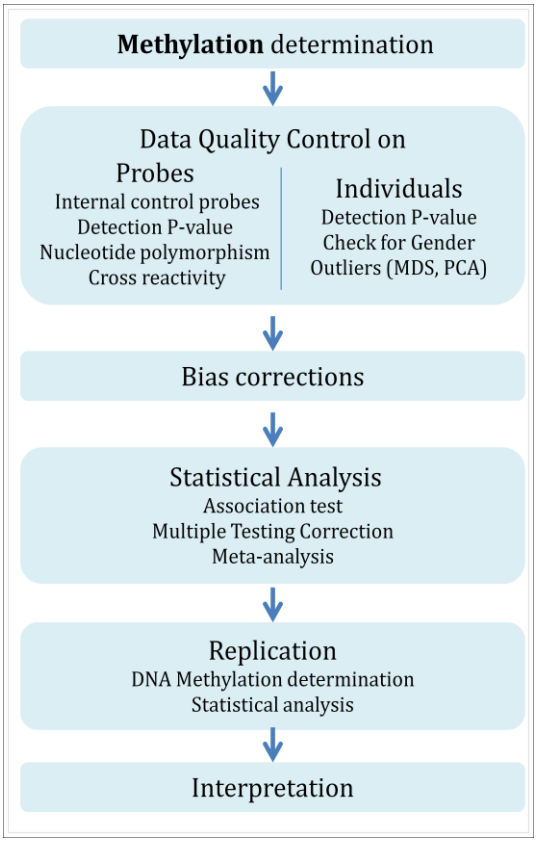


Figure 7.1. Methylation-wide association study step by step.

7.1. Methylation determination

Different methodologies are available to identify DNA methylation level: methylation-sensitive restriction enzymes, bisulfite conversion treatment and Methylated DNA ImmunoPrecipitation (MEDIP). Once detected they are analyzed by either microarray or deep-sequencing technologies, which provide very similar and reproducible results[188,189] (Figure 7.2.).

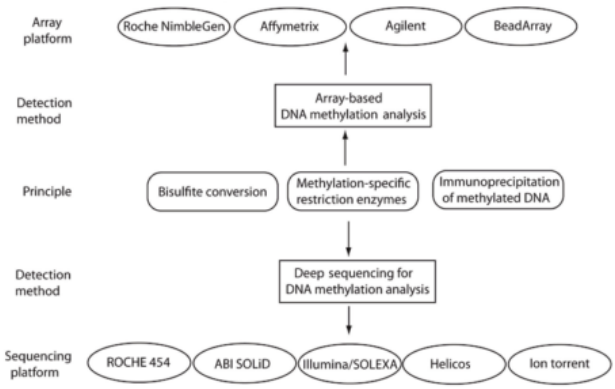


Figure 7.2. Different methods of DNA methylation measurement depending on the identification and measure approaches[190].

The Methylation sensitive restriction enzyme treatment is a very specific method that uses enzymes that break DNA at points with a specific sequence. They are also sensible to the methylation, responding differently -break or no break-depending on the methylation state. However, the detection is limited to just the sites inside the target sequences.

The immunoprecipiation method is a direct approach that doesn't need previous DNA treatment as the other two methods. That makes also skip treatment specific pruning and extra steps, making it easier to perform and analyze. It also can interrogate repetitive sequences and allow epigenetic states to be assigned to specific alleles. However it is not nucleotic specific as the bisulfite conversion.

The bisulfite conversion treatment converts the cytosines in uraciles, while the methylated-cytosines remain cytosines allowing detecting methylation at the exact nucleotide. Reason why is considered the "gold standard for detecting changes in DNA methylation"[190].

In the present work we used the bisulfite conversion and posterior analysis through the Illumina HumanMethylation450 bead array. The deep sequencing requires more complex bioinformatics tools especially for alignment mapping and is very expensive to apply at the genome-wide scale. Bioinformatics and biostatistical tools required for such DNA methylation deep-sequenced data have not yet achieved sufficient maturity for genome-wide analysis and are mainly recommended for the analysis of local regions[191].

7.1.1. The Illumina HumanMethylation 450K array

The Illumina HumanMethylation450 bead array (H450K) technique has recently gained large popularity as a robust and efficient tool for investigating, in epidemiological studies, methylation marks at CpG sites associated with environmental, genetic and biological data[185,192]. It covers 99% of RefSeq genes and all important known regions such CpG

islands, shores and shelves[193,194]. It surveys the DNA methylation levels at 482,421 CpG sites, with an average of 17 CpG sites per gene region.

The first step consists in treating DNA samples with sodium bisulfite (C-> U and Methyl-C->C) followed by an amplification step that coverts uraciles into thymines and cytosines remain as cytosines (Figure 7.3.). DNA samples are then hybridized on the H450K array using 485,577 allele specific probes, synthetic sequences of 50 base pairs that tag for the C or T of the CpG site. The hybridization needs to be perfect to permit the extension of one nucleotide that will be bound to a fluorochrome, which emits a signal. Finally a scanner measures that signal and transforms it into methylation intensity according to the used probe (see thereafter). Of note, 482,421 assessed sites correspond to true CpG sites, 3,091 correspond to CNG (where N could be any base but G) and 65 sites measured a known C/X polymorphism. Each probe is bound to a bead of 3 µm of diameter (as in the previous part I) which is in turn bound to the microarray. Each probe is repeated 15 times.

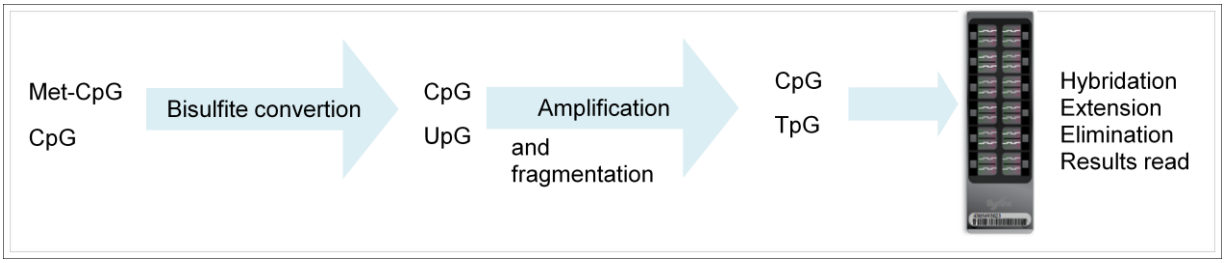


Figure 7.3. Methylation determination protocol.

The array is composed of 2 kinds of probes, Infinium I derived from the first generation of DNA methylation HumanMethylation27 array (n = 135,501, i.e., 27,9%), and Infinium II (n =350,076, i.e., 72,1%), and it additional includes 614 negative control probes that do not complement to any genomic sequence and should not hybridize. Nevertheless, these last probes produce some fluorescence signals considered as background noise, which will be used for quality control and normalization procedures (described thereafter, page 94). The H450K array uses two kinds of fluorochromes, green (Cy3) bound to adenines and thymines and red (Cy5) bound to cytosines and guanines, which in combination determine the methylation levels.

Infinium I probes. These probes use a method that consists in two different probes for the same CpG methylation site: one probe measures the methylated state (i.e., with the C nucleotide at the 50th position) and the other measures the unmethylated state (i.e., with the T nucleotide at the 50th position), and both bound to different beads. Each bead has a known pre-defined position on the array (Figure 7.4.).

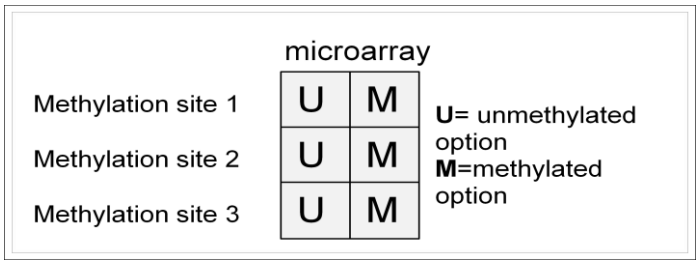


Figure 7.4. Example of the Infinium I probes display on the microarray for 3 methylation sites. For each CpG site there are two probes the unmethylated (U) and the methylated (M) placed in different points. If the samples have the site 1 methylated, the DNA will bind to the site M, and the fluorescence signal (red or green) will be in this M site but not in U site. On the contrary, if the site 1 is unmethylated, the fluorecence will appear in the U.

According to the methylation status at the target site (methylated or not), the treated DNA will bind to one of the two probes. If the hybridization is perfect, then the extension phase will be possible adding a nucleotide (i.e., the 51th nucleotide of the genomic sequence), which will produce a fluorescence signal. In this kind of probe, the Illumina strategy does not pay attention to the color of the fluorescence but to position on the array to discriminate between methylation and non-methylation (M or U, see Figure 7.4.). The signal is read by the scanner to declare whether it corresponds to a methylated or unmethylated site (Figure 7.5.). Then to quantify the intensity of the methylated or unmethlyated signal, only one of the two fluorescences (green or red) is used by the scanner according to the added nucleotide at the 51th nucleotide and some probe-specific information treated by the Illumina GenomeStudio program.

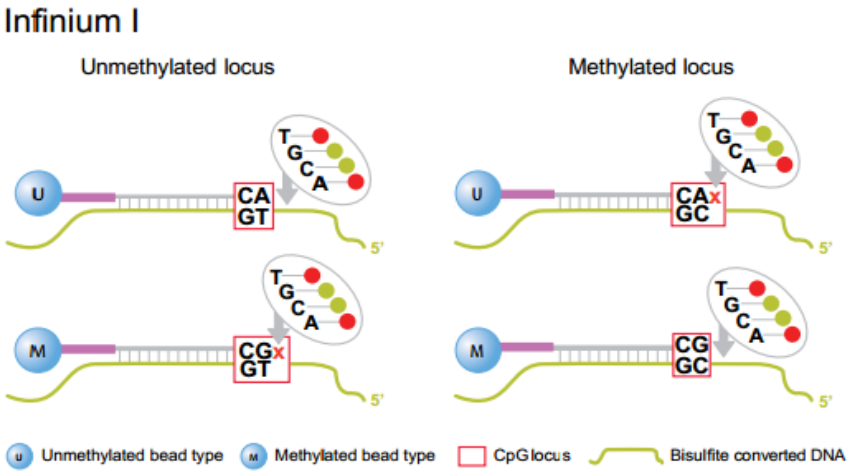


Figure 7.5. Scheme of the functioning of Infinium I (Illumina images).

Infinium II probes. For this technology, only one probe is used to assess the methylation status of a target CpG site. Each Infinium II probe is a sequence of 50 base pairs that

matches to the genomic sequence just before the position of the target CpG site. When the treated DNA perfectly hybridizes on the probe, the extension of the 51th nucleotide (that corresponds in that case) to the C or the T of the target site will produce a fluorescence signal, green (Cyanin 3) or red (Cyanin 5) according to the added nucleotide, characterizing the non-methylated and methylated states, respectively. Unlike Infinium I probe, the fluorochrome signal is the element for assessing the methylation status (Figure 7.6.).

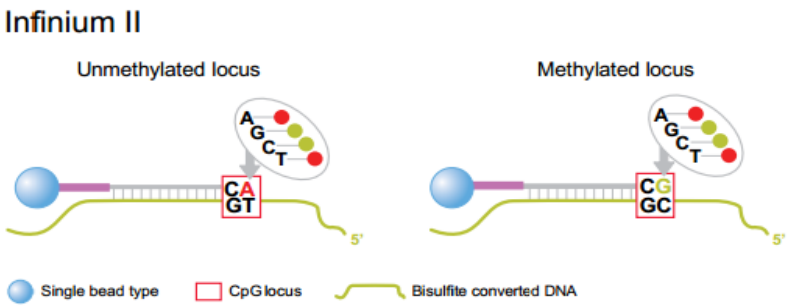


Figure 7.6. Scheme of the functioning of Infinium II (Illumina images).

Compared to the Infinium I technology, the Infinium II technology allows to increase the number of investigated CpG sites but they are less precise and less reproducible.

At the end of the H450K array process, different fluorescence signals are generated, some for methylated (Meth) and some for unmethylated (UnMeth) sites, each of them can be split according the underlying Infinium technology (I or II) (Figure 7.7.). In addition, some methylated and unmethylated signals are also produced by negative control probes.

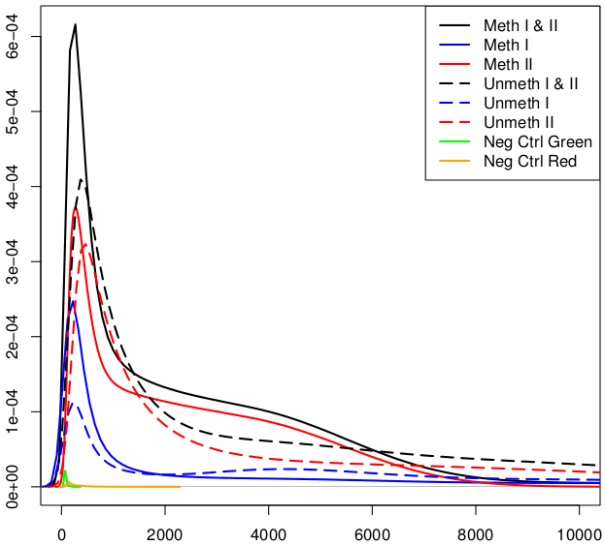


Figure 7.7. Density distribution of the intensity signals obtained from the Illumina methylation microarray (MARTHA and F5L samples).

7.1.2. Definition of the DNA methylation levels

From the intensity values, two different parameters can be defined to quantify the DNA methylation levels at a given CpG site *i*:

- **The β value** is a parameter that rangs from 0 to 1, calculated as

$$\beta_i = \frac{M_i}{(M_i + U_i + 100)}$$

where M_i is the intensity of methylated fluorescence at site *i* and U_i the intensity for the unmethylated signal. The two possible states are methylated or unmethylated for a same site, but, the DNA is extract of an individual is composed of several DNA molecules, differentially methylated for the same site. This is why the β values represent the methylated percentage of DNA molecules at a given site. The 0 value corresponds to complete unmethylated status and 1 to complete methylation. The distribution of all of the β is displayed in the Figure 7.8. where there are two main peaks, one near 0 and the other near 1.

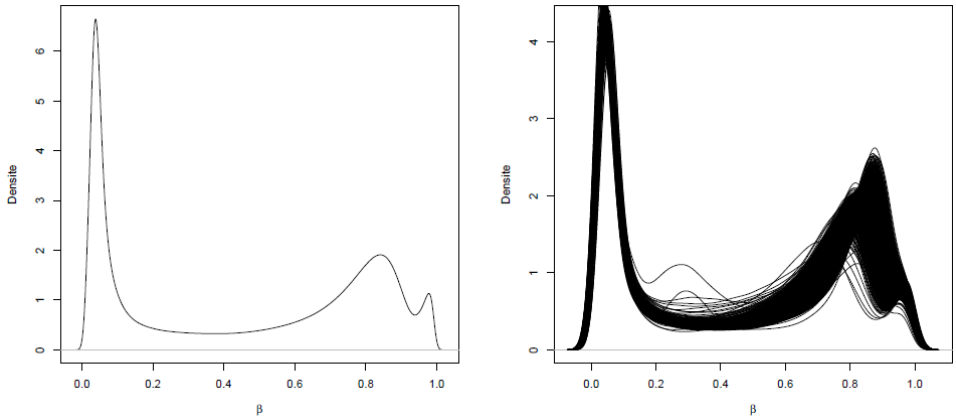


Figure 7.8. Distribution of β values. At the left, the mean β and on the right all the β from MARHTA and F5L samples.

- **The M value** is an alternative to β proposed by Du[195], that is a log transformation of β.

$$M = \log_2\left(\frac{\beta}{1-\beta}\right) = \log_2\left(\frac{M}{U+\alpha}\right)$$

Additionally M formulation has been proposed in which an *offset* α or compensation is applied in both M (methylated signal) and U (non methylated signal).

$$M = \log_2\left(\frac{M+\alpha}{U+\alpha}\right)$$

This parameter is more appreciated from a statistical point of view because the values are around 0 and the distribution is more similar to a Normal distribution when considering

methylation and non methylation separately (Figure 7.9.). It is used in some of the normalization steps. However β is more comprehensible in biological sense as a methylation percentage and a very recent work suggest that the use of β is more appropriate in the context of large epidemiological studies[196] such the MARTHA study.

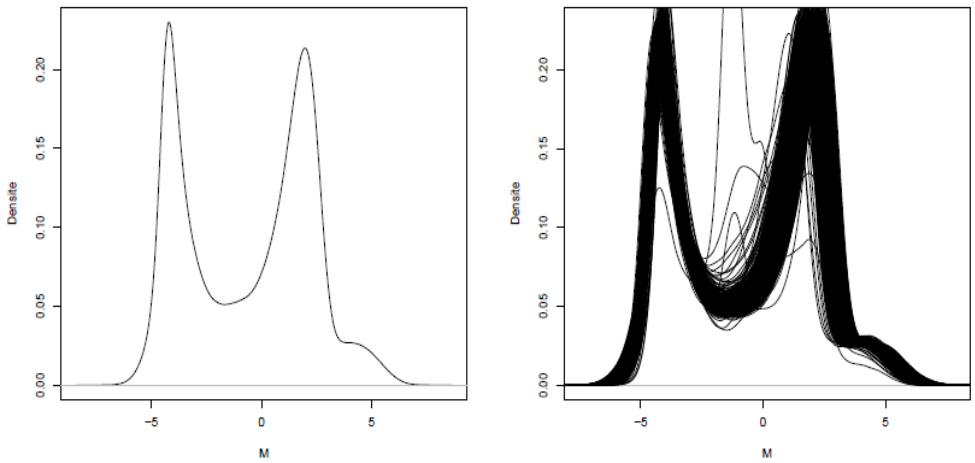


Figure 7.9. Density distribution of M value. The left distribution is the mean M value for the 577 and the right distribution is M value for each individual (MARHTA and F5L samples).

7.2 Data quality control

Any statistical and bioinformatic analysis of high-throughput data requires quality controls procedures to assess the validity of the produced data and avoid any spurious results. MWAS does not escape to the rule and many quality controls (QCs), both at the probes and the individual levels, must be applied to H450K DNA methylation data before embarking into association testing with biological/clinical data.

Most QC procedures used for H450K data analysis can be conducted using the *minfi*[197] or *methylumi* R packages[198]. All the QC steps, describes in the next paragraphs, have been performed by Dylan Aïssi (UMR_S 1166) and Jessica Dennis (Dalla Lana School of Public health, Toronto).

7.2.1. Probes filters

7.2.1.1. Control Probes

Among the 485,577 probes, 850 are designed to check for the quality of the DNA preparation and the process of the H450K array (Table 7.1.):

Table 7.1. Types and number of control probes.

Control Type	Nº Probes
Staining	6
Extension	4
Hybridization	3
Target Removal	2
Bisulfite Conversion	16 (12 +4)
Specificity	15 (12 +3)
Negatives	614
Non-polymorphic	4
Normalization	186 (32+61+32+61)
Total	850

Staining controls check the staining step from the microarray protocol, which is independent from the hybridization and extension.

Extension controls test the nucleotide extension step using a hairpin sequence that does not need the hybridization or a template for the extension.

Hybridization controls assess whether the probes hybridize correctly. For that purpose, synthetic oligonucleotides are construct that are perfectly complementary to this control probes and added instead of DNA sample. These oligonucleotides just use the green fluorochrome and they test also different concentrations of those oligonucleotides: low, medium and high. On the array sites where there are low oligonucleotide concentration there should be low intensity signal, where medium concentration-medium intensity and high concentration-high intensity (Figure 7.10.).

Target removal aims to check the correct elimination of the DNA in excess after the extension step.

Bisulfite conversion control consist in 16 probes, 12 for the Infinium I probes and 4 for the Infinium II, are used to assess whether the bisulfite treatment performed well. Infinium I probes contain a domain Hind III (5'...AAGCTT...3') for the 2 options: the complementary sequence Hind III converted with bisulfite (5'...AAG**U**TT...3') and the sequence (5'...AAG**C**TT...3'). Then synthetic oligonucleotides designed with the domain HindIII are treated with sodium bisulfite. Theoretically these oligonucleotides are not methylated, so they should be all converted to U and hybridize to the corresponding complementary probe (U/T) is emits fluorescence (Figure 7.10.). In the case of Infinium II probes, the emission of a red fluorochrome indicates that the bisulfite conversion has worked. On the contrary, the green fluorescence would indicate a problem in this step.

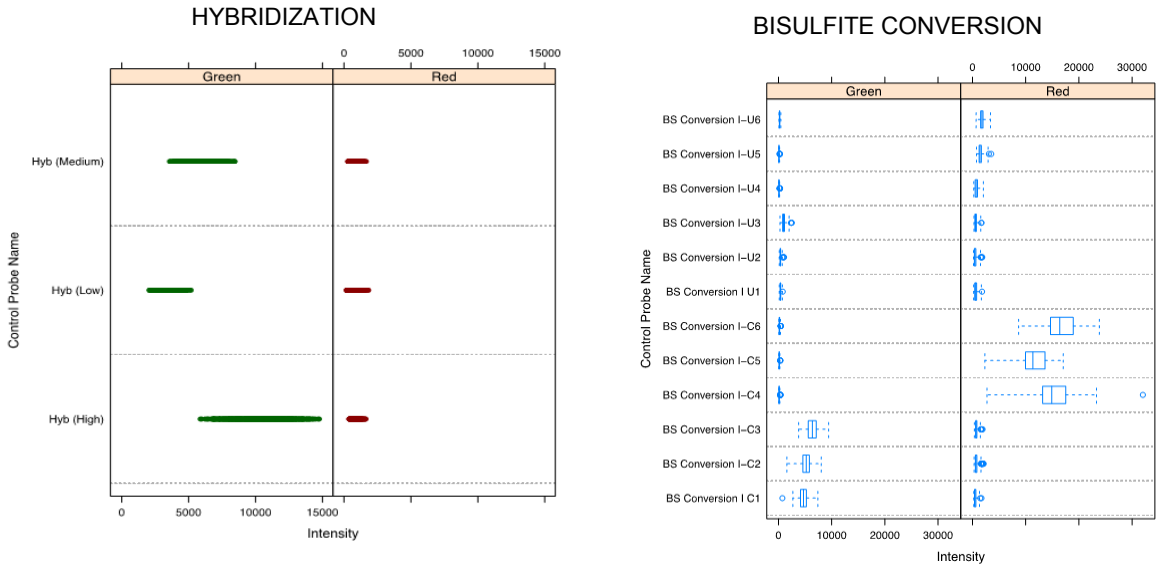


Figure 7. 10. Examples of probe controls of MARTHA and F5L samples. On the left the hybridization control probes, just tested with the green fluorochrome. It is observed how the intensities vary according to the concentrations, from the top to the bottom: medium, low, high. Also it also confirms there is no red fluorochrome signal or residual. On the right, the bisulfite conversion controls just for Infinium I. There are the six first lines, which correspond to the **U**nconverted status, where it should be no signal as showed, and the bottom lines, correspond to the **C**onverted status which do show intensity.

Specificity of the probes. There are 12 Infinium I and 3 Infinium II probes designed to control the non specific hybridization and extension. The probes correspond to sequences with non polymorphic T sites.

Negative controls probes are designed to be complementary to none DNA fragment. Therefore, there should be neither hybridization nor extension and should not produce any fluorescence signal. However, there is always signal which is used as background noise. In the previous Figure 7.7. the intensity of this background was displayed. In Figure 7.11. the negative controls are represented alone, where it is remarkable that the methylation signal is more intense than the non methylation one.

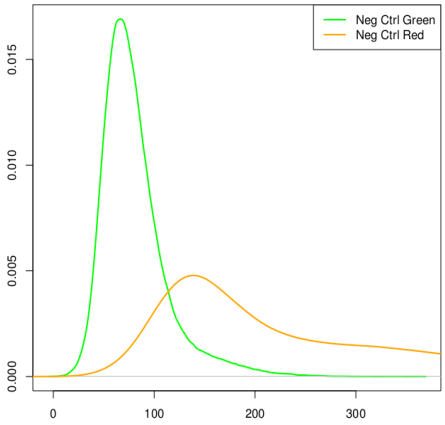


Figure 7.11. Density destribution of the Negative control probes for the 577 individuals (MARTHA and F5L samples).

Normalization probes are those complementary to housekeeping genes without CpG sites. One nucleotide is added. The extended nucleotide is an adenine, a cytosine, a guanine and a thymine for 32, 61, 32 and 62 probes, respectively.

7.2.1.2. Detection P value

Every β value has an associated p-value, which evaluates the probability that the intensity of the probe is distinguishable enough from the negative control probes intensity. A non significant pvalue (e.g., pvalue >0.05) means that the intensity of the associated probe is too low and not distinguishable from the background. This could be due to poor quality of the probe design, poor hybridization, or chromosome anomalies in the region (eg deletion/insertion). It is recommended to exclude those probes.

The standard method for calculating the detection p-value and the one used in my project is the "m+u" method implemented in the *minfi* package. For each processed individual DNA sample, it computes:

- The mean of the fluorescence of the background noise for every color red (medR) and green (medG)
- The mean absolute deviation (MAD) of the background and also by color: madR and madG
- The total intensity by probe:
  - Infinium I
    - Red: intensity of the allele 1 (methylation state) + allele 2 (unmethylation state)
    - Green: intensity of the allele 1 + allele 2
  - Infinium II: intensity red+ intensity green
- From these values, it computes the probability p that a Normal variable with mean and standard deviation given below would be greater than the observed total intensity probe.

For Infinium I: mean = medG*2 or medR*2

Standard deviation = madG*2 or madR*2

For Infinium II: mean = medG+medR

Standard deviation = madG + madR

7.2.1.3. Nucleotide Polymorphism

Probes that harbour a SNP in their 50 bp genomic sequence require great attention to avoid any artifactual results[199]. Hence, it is important to update the annotation of the probes, e.g. with the available data in 1000 genome project[127] to identify the probes with possible SNP in their sequence.

Chen *et al*. demonstrated that there are 66,877 probes on the array that contain a SNP at the methylation site of interest. A given DNA sample may then present a C or an X (another nucleotide) at the target site before the bisulfite conversion, which would turn into biased signals depending on X. These probes must be excluded from the analysis.

The same research group identified 239,238 probes with at least one SNP, not at the methylation site but in their 50 bp sequence. Such SNP may produce hybridization problems and result in biased intensity signal. These probes represent almost half of the total number of probes. For that reason, they are not filtered out from the analysis but interpreted carefully when any association is observed at such probe. When GWAS data is available, it can be easily checked whether such associations are artifacts by testing the effect of the probe SNP(s) on DNA methylation levels.

7.2.1.4. Cross-reaction

Finally, there are probes that are not specific enough and can hybridize with more than one sequence in the genome. The produced signals are then not easily interpretable. This phenomenon was also pointed out by the group of Chen *et al.*[199,200] identifying 16,532 autosomal probes that react with the X chromosome and recommending to exclude them from the analysis.

7.2.2. Individuals filters

7.2.2.1. Detection P value

Similar to the genotype missing data addressed in the GWAS context, individuals with too many probes whose intensity are not distinguishable from background noise are excluded. The strategy excludes any individual with more than 5% of their probes with a detection p-value > 0.05 (mentioned before).

7.2.2.2. Outliers

As in the GWAS QC protocol, it is important to check if there could exist some outliers in the studied samples. It is possible to apply the multidimensional scaling (MDS) method to

identify individuals that strongly differ from the others. For example, in Figure 7.12., 3 women grouped together with males and 1 man grouped together with women (see also next paragraph).

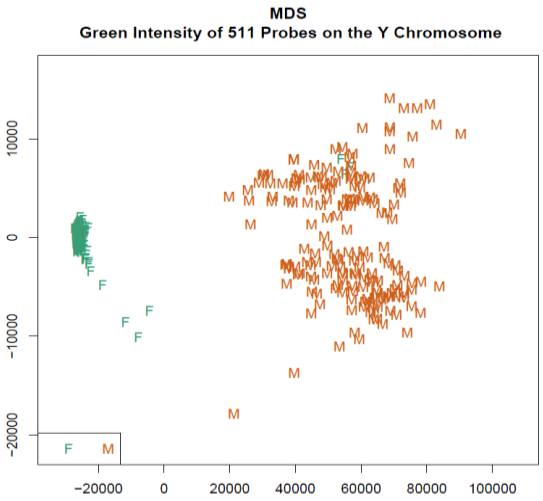


Figure 7.12. MDS of the 577 individuals (MARTHA and F5L samples). The "F" in green represent the females and the "M" in orange the males.

7.2.2.3. The X Chromosome

Similar to the "Gender Check " QC (Chapter 4, page 40), DNA methylation data can be used to check for the consistency between the reported biological sex and the methylation-derived sex. This relies on the analysis of the green fluorescence of the Y chromosome probes (Figure 7.13.). Compared to males, females should not present intensity signal above the background noise. Females presenting non negligible intensity must be excluded from the analysis.

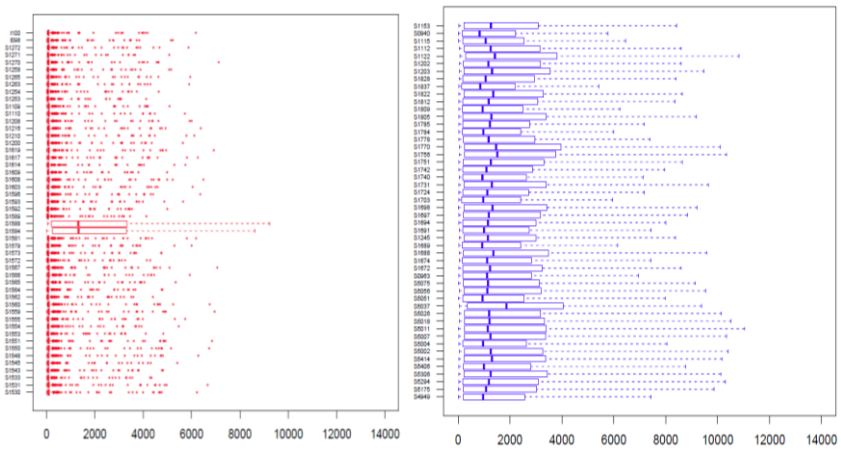


Figure 7.13. Green intensity of Y chromosome probes in females (left) and males (right). In this example, there are two females presenting intensity signal meaning a possible contamination (MARTHA and F5L samples).

7.3. Bias and corrections

The bias is a systematic error due to a poor quality sampling that can make vary the results from what it is expected. Different kind of bias can be encountered in analyzing H450K data. These are described in the next section together with the generally advocated methods to correct for them.

7.3.1. Background Bias

This kind of bias is due to the non-specific hybridization between the probes and the DNA samples. It may happen that non-specific DNA fragments bind the designed probes and are used as templates in the extension step, producing not desired fluorescence called background noise. There are two types:

- A general background that increases the general signal of all probes.
- A structured background, meaning the increase is just produced locally, in a particular place of the array (also called spatial background).

Since Illumina manufacturer have been placed the probes randomly on the array, this second kind of background is avoided. Only the general background noise must be corrected, and for that purpose, two popular methods have been proposed.

7.3.1.1. Negative control probes method

As previously mentioned, 614 negative control probes are available on the array, which should not hybridize to the treated DNA. However, there is always a degree of non specific fluorescence which is used to determine the background noise (Figure 7.11. of negative controls). The correction would consist in subtracting this background noise from the other measured signals.

In the NormExp correction model proposed by Irizarry and improved by Silver[201] the observed intensities X are modeled from a exponential S distribution with mean α. The background noise B is modeled as a normal distribution with mean μ and a standard deviation δ^2. The background corrected values of X are then taken as the expected value of S given observed x

$$\mathrm{E}(S|X=x) = \mu_{sx} + \frac{\sigma^2 \phi(0;\mu_{sx},\sigma^2)}{1-\Phi(0;\mu_{sx},\sigma^2)}$$

where ϕ and Φ are the density and the cumulative distribution function of a Normal distribution, respectively, and $\mu_{sx}=x-\mu-(\delta^2/\alpha).\mu$ and δ^2 can be directly estimated from the

negative controls and α by maximum likelihood. This correction is applied for each channel color separately (Cy3 et Cy5).

7.3.1.2. Out-of-band method

This second method takes advantage of the particularity of Infinium I that uses two probes for each tested methylation site. As mentioned above, the scanner only reads one of the two fluorochromes at each of the Infinium I, hence there are maximum 135,501 probes to quantify methylation and non methylation. These non-used intensities, for each color, serves as a surrogate for non specific nucleotide incorporation and fluorescence. Instead of using only 614 negative control probes into the NormExp model mentioned above, Triche *et al.*[202] proposed to use these 135,501 intensities in a similar NormExp model. This method is referred to as the Noob function implemented in the *methylumi* package and is the one used in this work.

In Figure 7.14. the distribution of the negative controls is represented (the 614) probes compared to the out-of-band signals (from 135,501 sites) by each fluorochrome.

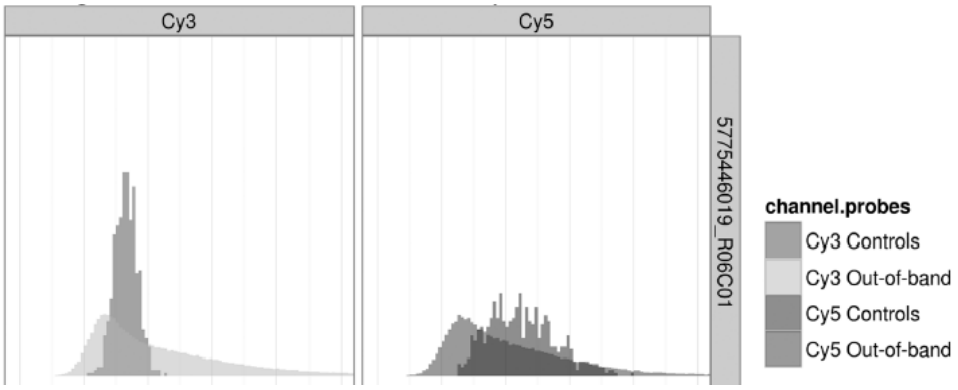


Figure 7.14. Negative controls and out-of-band intensities from MARTHA and F5L samples distributions by fluorochrome (channel).

7.3.2. Bias of two type of probes

Checking and comparing between the signals derived from the H450K array and those obtained by a bisulfite pyrosequencing method (which is also based in bisulfite technique but followed by pyro-sequencing) has demonstrated that Infinium II probes are less reproducible and less sensitive to extreme values than Infinium I[203]. Because the technology underlying the Infinium I and II probes differ, the resulting β values are not directly comparable. Figure 7.15. *A* displays the mean distribution of β overall all Infinium I and II probes (in black) which shows two peaks at the right tail of the distribution (i.e., close to 1). In blue (red) is shown the distribution for Infinium I (II) probes. While the blue and red distribution are very similar for low β values, the Infinium II distribution is shifted to the left for high β values. These

differences can be problematic when comparing probes measured by Infinium I and Infinium II. Different corrections have been proposed, the peak-based correction by Dedeurwaerder *et al.* 2011[203], the SQN method by Touleimat *et al.*[204], the Beta MIxture Quantile dilation method (BMIQ) by Teschendorff *et al.*[205], and the Subset Quantile Within-Array method (SWAN)[206]. As part of his PhD project, Dylan Aissi, supervised also by David-Alexandre Tregoüet has shown that the SWAN method was the more consistent and reproducible between replicates in the MARTHA project. His results were in complete agreement with those obtained by other works that have adopted the SWAN method for correcting this bias. Briefly, the SWAN method implemented in the *minfi* package is based on the application of quantile correction for each type of probe and separately to the methylated and non-methylated intensities. The SWAN correction makes the β distribution nearly superposable between Infinium I and II probes (Figure 7.15 b).

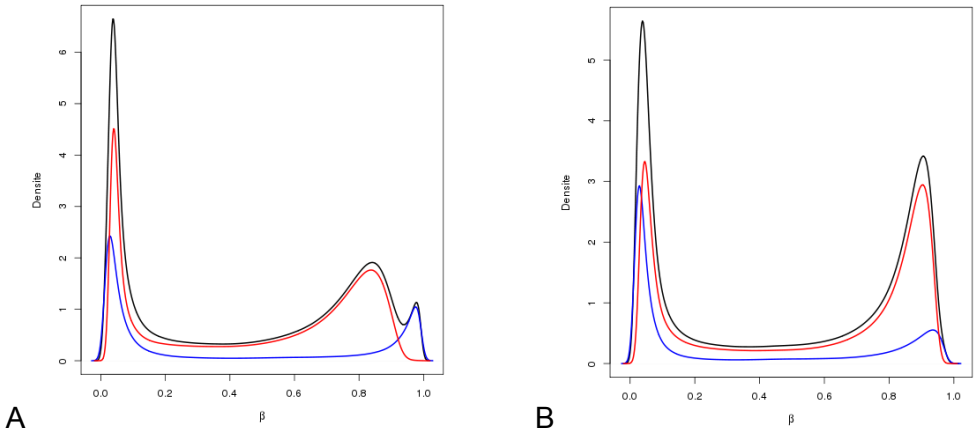


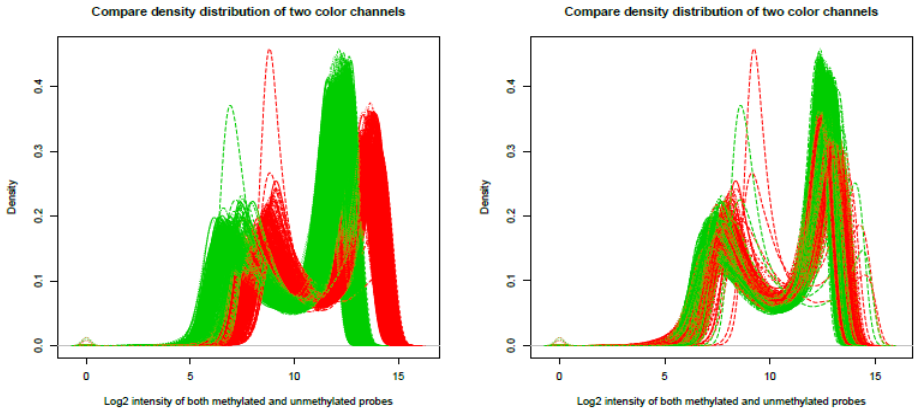
Figure 7.15. Distribution of the mean β before (A) and after (B) correction with SWAN(MARTHA and F5L samples). In black is represented the β of all the probes (I and II), in blue just the probes Infinium I and in red the probes Infinium II.

7.3.3. Bias due to the use of two fluorochromes

The use of two different fluorochromes for the Infinium II technology at the same probe can also create another bias because these fluorochromes possess different chemical properties which can produce a shift in the two color intensities, one being more "intense" than the other.

A total of 186 normalization control probes have been designed by Illumina to correct this bias. As previously described, these control probes are designed to match housekeeping genes and to be extended in the same region. These genes are present in all cells, always expressed in any individual without any kind of regulation and they do not have CpGs. Theoretically, the produced intensities should be equivalent, but they are not. The correction method is as follows:

1. For each individual, an average is calculated for each dye: green *Grn.avg* and red *Red.avg* using the normalization control probes.
2. The mean of these two average is then computed: Grn-Red.avg
3. From all the samples, one is chosen as reference. There are different methods depending on the software used. The Illumina GenomeStudio does not specify how the reference is selected. With *minfi* package, the user can choose the reference or by default will select the first individual. The *methylumi* package selects the sample with the (Gnd.avg/Red.aveg - 1) value closest to 0.
4. For each individual, the dye averages, *Grn.avg* and *Red.avg*, are then divided by the reference values resulting into a factor value, one for the red and one for the green dyes,
5. Finally the intensities are multiplied with the corresponding factor (depending on the color) in each individual for correcting the fluorochrome bias (Figure 7.16.).



A B

Figure 7.16. Distribution of M value of the Infinium II probe intensities before (A) and (B) after correction (MARTHA and F5L samples)**.** The color green Cy3 corresponds to the methylated status (nucleotides C and G) and the red or Cy5 for non methylated (T and A).

7.3.4. Plate effect Bias

The Illumina plate technology permits to determine 96 individuals simultaneously: one plate containing 8 microarrays (or chips), each of them processing 12 individuals. The individuals of MARTHA and FVL families have been processed on 7 plates. As it is impossible to reproduce the exactly same technical conditions between plates, there are always sampling differences between plates, sometimes called batch effects. Figure 7.17. displays the β distribution over all probes across the 7 plates used in the MARTHA and FVL families project. To correct for potential plate (and chip) effects, the natural way is to adjust

for such variable(s) in the regression model used when testing the association of CpG levels and a given phenotype (see below).

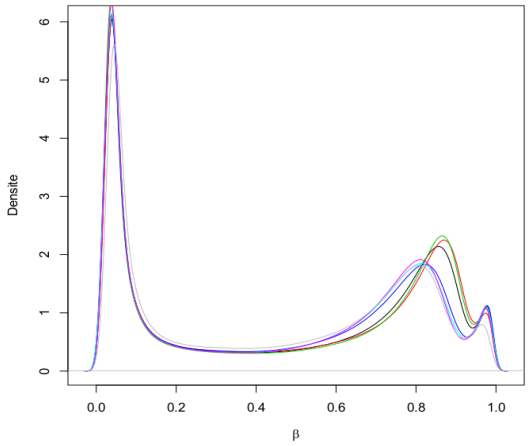


Figure 7.17. Distribution of β of each one of the 7 plates (MARTHA and F5L samples).

7.3.5. GC content bias

The GC content bias was initially reported by **Kuan *et al.*** [207] in the analysis of the first generation of Illumina DNA methylation array composed only of Infinium I probes. Kuan *et al.* observed that the intensity of the methylation signal was dependent on the probe GC content related different thermodynamic properties. The more enriched in GC the probe, the lower intensity. However, this bias has not been extensively studied in the context of the H450K array. No correction is then currently proposed.

7.4. Statistical Testing for Methylation - phenotype associations

For my MWAS project, I was interested in identifying CpG site whose levels could associate with TGP biomarkers. The standard way to address such question is to adopt a linear regression framework where the TGP biomarker is the outcome and the CpG site β value the covariate. This is similar to the GWAS strategy I used in Part I except that the genotype variable is replaced by the DNA methylation β variable (see Part II, Chapter 8, page 109).

MWAS results can be described in terms of MH and QQ plots similarly to what I have explained for the GWAS results (see Chapter 4, page 46). The standard method for correcting multiple testing errors is also the Bonferroni method as 0.05 by the number of tested CpG sites. The replication and meta-analysis strategies described for a GWAS were also applied to the present MWAS.

Chapter 8. The MWAS strategy applied to TGP biomarkers

8.1. Cohort studies

The methylation-wide association study I conducted builds upon two different datasets, a subsample of the French MARTHA study and the French-Canadian F5L pedigree study.

MARTHA

For this analysis, I had access to a subsample of 350 VT patients from the MARTHA study were epigenetically-typed with the Illumina HumanMethylation450 array. These individuals were randomly selected from the 1,542 MARTHA patients that were typed for GWAS SNP data (Part I, page 55).

French-Canadian F5L pedigrees

The F5L pedigree study is composed of five extended French-Canadian pedigrees totaling 255 relatives. These pedigrees were ascertained through an idiopathic VT proband carrying the FV Leiden mutation and seen at the Thrombosis Clinic of the Ottawa Hospital. With the exception of harboring the FV Leiden mutation, the probands were free of inherited thrombophilia (AT, PC and PS deficiencies and homozygosity of FII G20210A and FV Leiden mutation**)** and without acquired VT risk factors such as cancer, myeloproliferative disease, pregnancy, puerperium, prolonged immobilization, trauma, surgery and antiphospholipid syndrome. Once the proband was identified, the size and structure (4-5 generations) of the family were recorded, and the five largest families were enrolled in the study whose detailed description can be found in Antoni G *et al.*[28]. A total of 255 individuals were then recruited and 214 with DNA available were epigenetically -typed for the same Illumina HumanMethylation450 array.

8.1.1. Quality Controls of DNA methylation CpG probes

The Methylation patterns of both MARTHA and F5L pedigree samples were analyzed at the Toronto Centre for Applied Genomics (TCAG, http://www.tcag.ca/) using the same protocol. The same quality control and normalization procedures as in section 7.2 (page 94), were simultaneously applied to both studies. Any probe with a detection p-value (as described in the "minfi" package) greater than 0.05 in more than 5% of the total processed

samples was excluded from further analyses. The consistency assessment between the "biological" sex phenotype and the methylation patterns on sexual chromosomes (page 102) led to the exclusion of 4 individuals in the F5L pedigrees study.

The general filters and corrections applied are shown in Figure 8.1. A total of 388,120 CpG sites were finally selected for further statistical associations with TGP biomarkers.

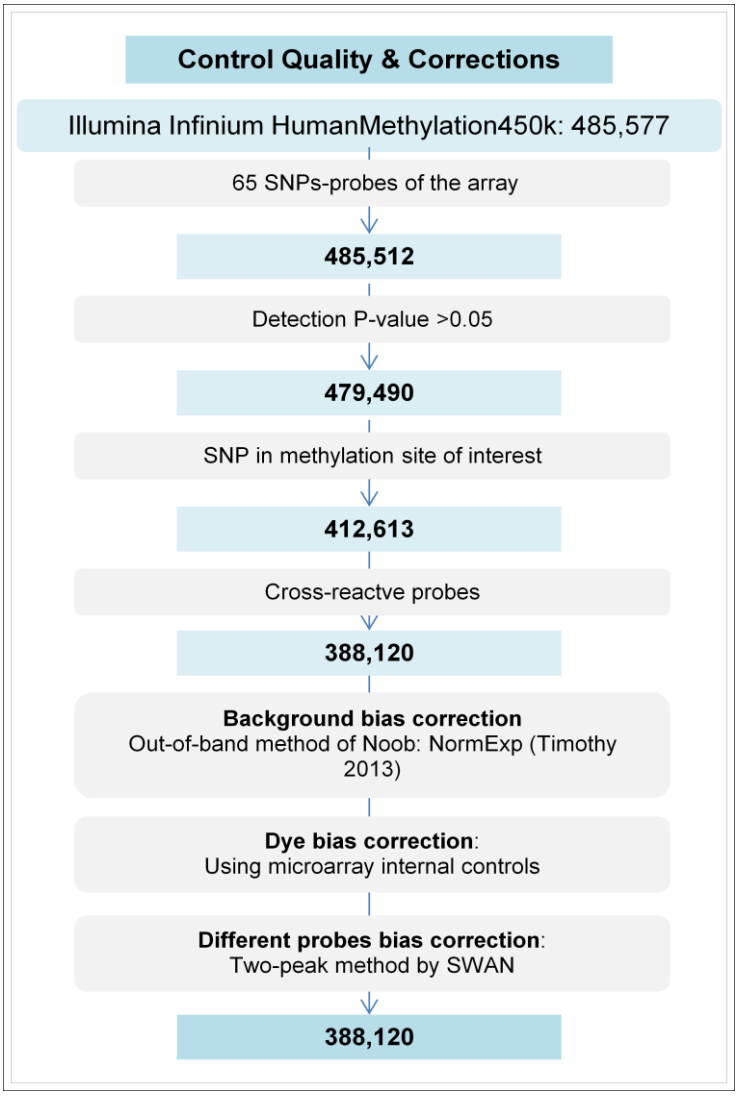


Figure 8.1. Diagram of the filters used for the quality control of the probes.

8.1.2. Clinical and biological characteristics of the discovery cohorts

TGP biomarkers were measured in 238 MARTHA patients and 187 F5L pedigree members with DNA methylation data (Chapter 2, page 22). The same methods were applied to both samples and all measurements were done at the Laboratory of Hematology of the Timone Hospital (Marseille). The biological and clinical characteristics of studied samples are

shown in the following Table 8.1. extracted from my second manuscript (Rocanin-Arjo *et al.*, page 120).

Table 8.1. Population characteristics.

	MARTHA	F5L-families
	N = 238	N = 187
Mean age in yrs ± SD	43.9 ± 14.2	39.5 ± 16.8
Males/Females	53/185	92/95
VT patients (%)	238 (100%)	4 (2.1%)
Heterozygote+*F5* rs6025	68 (28.6%)	41 (21.9%)
Heterozygote+*F2 G20210A*	35 (14.7%)	-
Contraception treatment	21/185	15/95
log-vWF[1]	4.9 (0.37)	1.29 (0.48)
BMI kg/m2	24.1 (5.4)	26.8 (5.9)
ETP[2] nM/min (1st-3rd)	1751.8 (1502- 1934)	1881.7 (1649-2068)
Peak (nM) (1st-3rd)	320.5 (279.8-359.8)	339.3 (307.1-377.9)
Lagtime (min) (1st-3rd)	3.09 (2.67-3.33)	3.16 (2.83-3.48)
Correlations		
ETP*Peak	0.758*	0.809*
ETP*Lagtime	0.036	-0.112
Peak*Lagtime	-0.107	-0.400*

[1] 220 ind in MARTHA and 183 ind in FVL
[2] 185 ind in FVL
*p>0.05, VT, venous thrombosis; vWF, von Willebrand factor; BMI, body mass index.
+ FV Leiden and F2 G20210A mutations were genotyped in MARTHA as part of the inclusion criteria. Homozygous carriers were not included in the study.

8.1.3. Statistical methods and adopted research strategy

For this MWAS project I adopted two research strategies:

First, I performed a MWAS on each TGP biomarker in the MARTHA samples and selected for further replication in the F5L pedigree study any CpG sites that reached the genome-wide significant threshold of $1.29\ 10^{-7}$, which corresponds to the standard Bonferroni threshold for the number of tested CpGs (= 0.05 / 388,120).

In a second step, I conducted a similar MWAS investigation in the F5L pedigrees and combined the results with those obtained in MARTHA through a random-effect meta-analysis. The GWAMA program[208] was used for the meta-analysis.

Similar to the GWAS project (Chapter 5, page 55), ETP values were log-transformed before association analyses while a quantile normal transformation was applied to Lagtime, separately in each cohort (see Figure 8.2.).

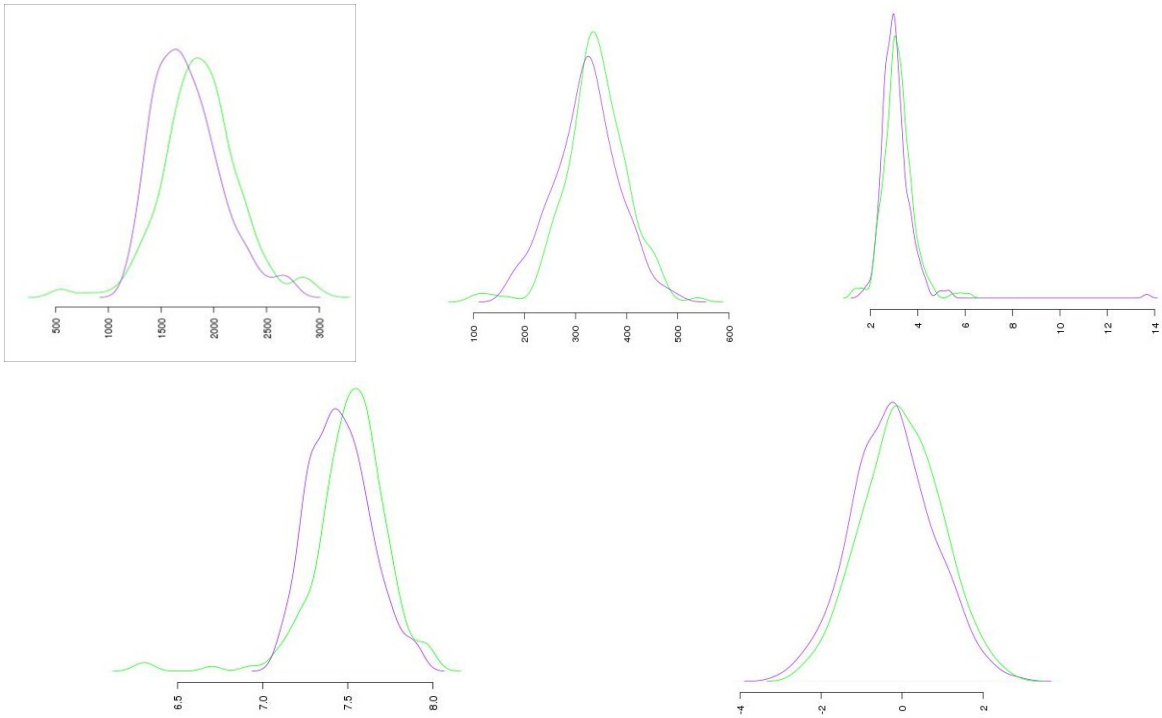


Figure 8.2. Distribution of TGP biomarkers in MARTHA (purple) and F5L pedigree study (green). On top and from left to rigth: ETP, Peak and Lagtime raw data. On the bottom, the log transformed ETP (left) and the quantile transformed Lagtime (right) distributions. MARTHA in purle, 3C in green.

In the MWA study, linear regression analyses were used to assess the influence of CpG site variability, as covariates, on each TGP biomarker, as the outcome. In the F5L pedigree study, a linear mixed-model accounting for the non-independence between family members was used. All analyses were adjusted for age, sex, contraception therapy, body mass index, and batch and chip effects. The two last variables correspond to microarray plate on which samples were processed and their position on the plate, respectively to account for experimental variability between position and plates[209]. Because DNA methylation levels measured in peripheral blood DNA reflect the average level of DNA methylation in different cell types including lymphocytes, monocytes, neutrophils, basophils and eosinophils, all analyses were also adjusted for cell type composition to avoid any contamination bias[197,210,211]. For MARTHA samples we used available specific biological counts of lymphocytes, monocytes, neutrophils, eosinophils and basophils to characterize their cell type composition. In F5L pedigrees the cell type counts were not available, however we handled adjustment for cell type composition using the method described in Koestler *et al.*[212].

8.2. Strategy 1 MWAS findings

Quantile-quantile (Q-Q) plots representing the results of the MARTHA MWAS performed on the three TGP biomarkers are shown in Figure 8.3. There was no observable deviation from what expected under the assumption of no association between CpG sites and TGP biomarkers. The corresponding genomic inflation coefficients were 1.004, 1.012, and 0.957 for ETP, Peak, and Lagtime, respectively. Corresponding Manhattan plot representations of the 3 MWAS scan in MARTHA are shown in Figure 8.4.

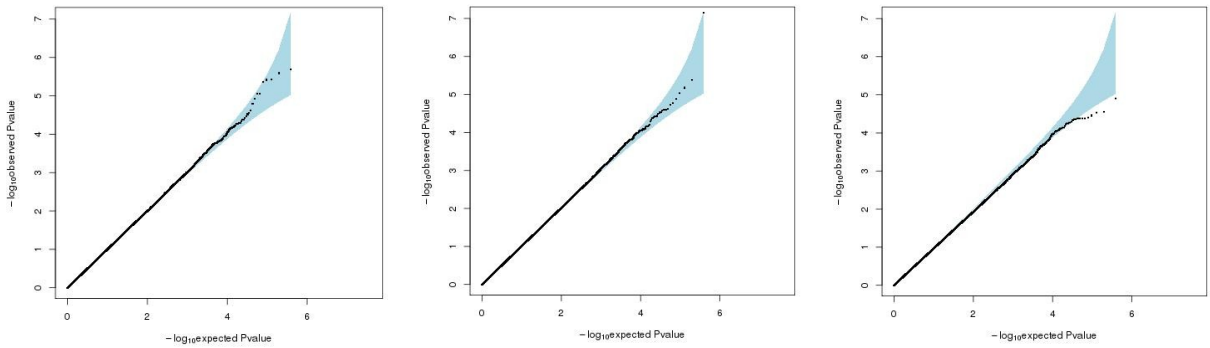


Figure 8.3. Quantile-Quantile graphics comparing the p-values resulted from the MWAS in MARTHA with the expected p-values. From left to right ETP. the second is Peak and the last Lagtime.

None of the tested 388,120 CpG sites passed the Bonferroni correction threshold for declaring statistical significant association with ETP or Lagtime. The lowest p-values observed for these two phenotypes were $p = 2.08 \cdot 10^{-6}$ for CpG cg06613480 (SOX2OT locus) and $p = 1.23 \cdot 10^{-5}$ for CpG cg20240757 (PTH locus), respectively. On the other hand, genome-wide significance was reached for the association of CpG cg26285502 mapping to C15orf41 on chromosome 15 with Peak. A 1% increase in cg26285502 DNA methylation levels was associated with an 11.96 ± 2.12 nmol/L increase ($p = 7.15\ 10^{-8}$) in Peak values.

I then looked for replication of the cg26285502 association signal in the F5L pedigree study. Unfortunately, in this sample, a 1% increase in cg268285502 DNA methylation was associated with a non significant (p = 0.40) increase of 2.16 ± 2.57 nmol/Lin Peak values.

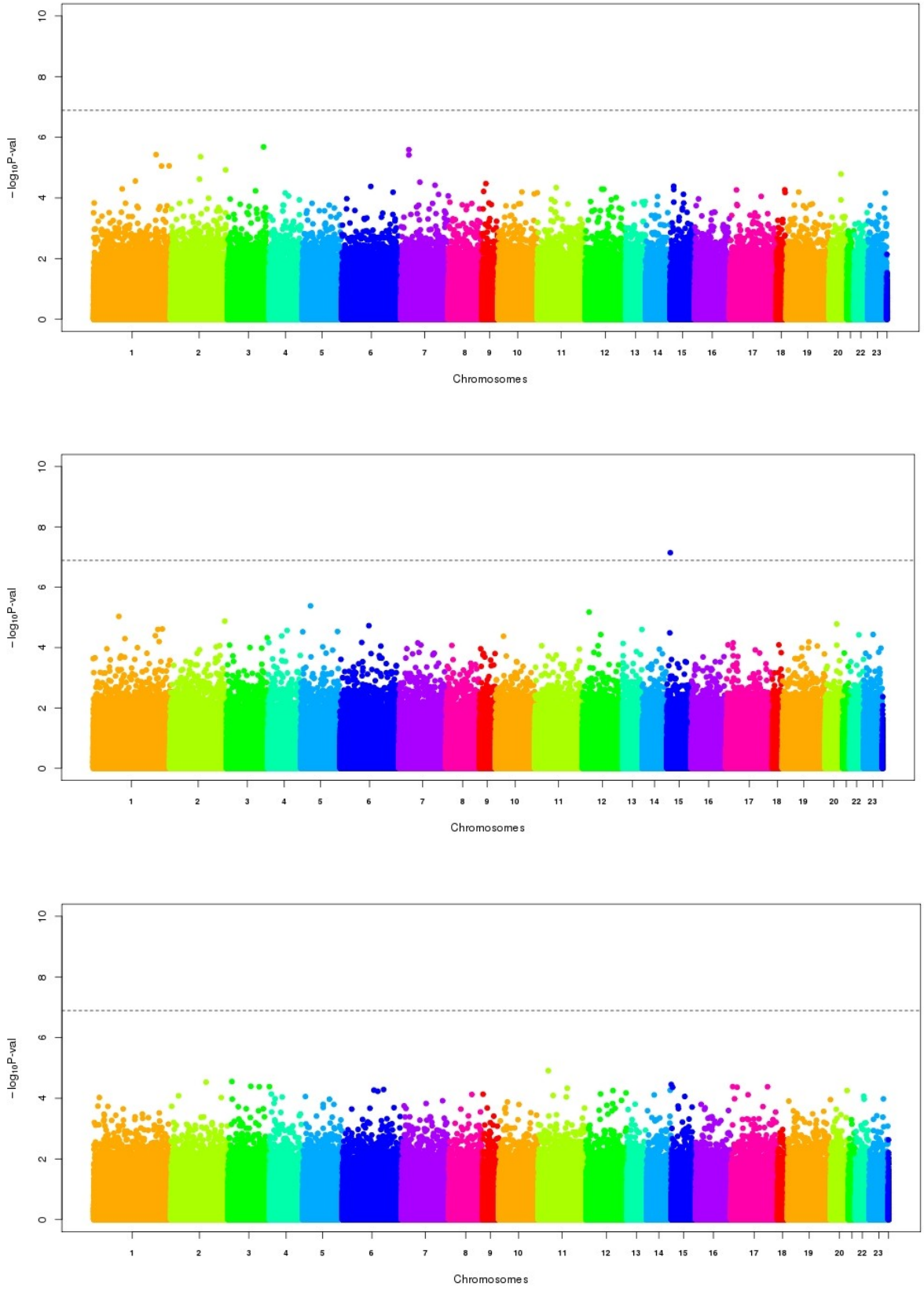


Figure 8.4. Manhattan plots from the MWAS results observed in MARTHA where 388,120 CpG sites were tested for association with ETP (top), Peak (middle), and Lagtime (bottom). The horizontal line corresponds to the Bonferroni significant threshold taken at $0.05/388120= 1.29\ 10^{-7}$.

Afterwards, following the "joint analysis" strategy I have applied for my GWAS project, I further focused on the CpG sites that demonstrated significant associations at $p < 10^{-4}$ with at least two TGP biomarkers. There was no such CpG site sharing these features between Peak and Lagtime or between ETP and Lagtime. However, 7 CpG sites were associated at $p < 10^{-4}$ with both ETP and Peak (Table 8.2.), and one was the *C15orf41* cg26285502 probe discussed above.

Table 8.2. CpG sites showing evidence for association at $p < 10^{-4}$ with both ETP and Peak in the MARTHA MWAS.

					ETP		Peak	
CpG	Chrom	Position	Strand	Gene	BETA (%)	P	BETA (%)	P
cg05970398	4	95129288	F	SMARCAD1	0.054 (0.013)	$8.62\ 10^{-5}$	19.53 (4.53)	$2.71\ 10^{-5}$
cg21583690	6	33385779	F	CUTA	0.075 (0.018)	$4.18\ 10^{-5}$	26.34 (5.99)	$1.88\ 10^{-5}$
cg24750156	12	58025315	F	B4GALNT1	0.014 (0.003)	$5.16\ 10^{-5}$	4.66 (1.10)	$3.70\ 10^{-5}$
cg24849736	1	201084428	F		0.011 (0.002)	$3.73\ 10^{-6}$	3.21 (0.76)	$4.02\ 10^{-5}$
cg25181043	20	47013841	R		-0.007 (0.002)	$1.63\ 10^{-5}$	-2.33 (0.53)	$1.66\ 10^{-5}$
cg26285502	15	36871420	R	C15orf41	0.027 (0.007)	$4.10\ 10^{-5}$	11.96(2.13)	$7.15\ 10^{-8}$
cg26957150	7	50861739	R	GRB10	0.015 (0.004)	$3.01\ 10^{-5}$	4.85 (1.19)	$6.99\ 10^{-5}$

I sought for replication of these additional promising signals in the F5L pedigree study. As illustrated in Table 8.3. below, but no signal was replicated.

Table 8.3. Replication in the F5L pedigree study of the Table II.XX results observed in MARTHA.

					ETP		Peak	
Sondes	CHR	Position	Strand	Gène	BETA E(%)	P	BETA P(%)	P
cg05970398	4	95129288	F	SMARCAD1	0.0054 (0.012)	0.6623	1.84 (4.27)	0.6672
cg21583690	6	33385779	F	CUTA	-0.0120 (0.025)	0.6375	1.67 (8.87)	0.8510
cg24750156	12	58025315	F	B4GALNT1	-0.0037 (0.004)	0.2946	-1.15 (1.22)	0.3499
cg24849736	1	201084428	F		-0.0025 (0.003)	0.3300	-0.38 (0.89)	0.6684
cg25181043	20	47013841	R		0.0005 (0.002)	0.7964	1.01 (0.68)	0.1397
cg26285502	15	36871420	R	C15orf41	-0.0014 (0.007)	0.8530	2.16 (2.57)	0.4013
cg26957150	7	50861739	R	GRB10	-0.0080 (0.004)	0.0494	-3.04 (1.41)	0.0324

8.3. Strategy 2 MWAS findings

In order to increase statistical power to detect CpG sites associated with any of the three studied TGP biomarkers, I further conducted a MWAS on each TGP phenotype in the F5L pedigree study and combined the results with those obtained in MARTHA in the framework of a random-effect meta-analysis. The resulting QQ plots and Manhattan plots are shown in the next two figures (Figure 8.5. and 8.6.).

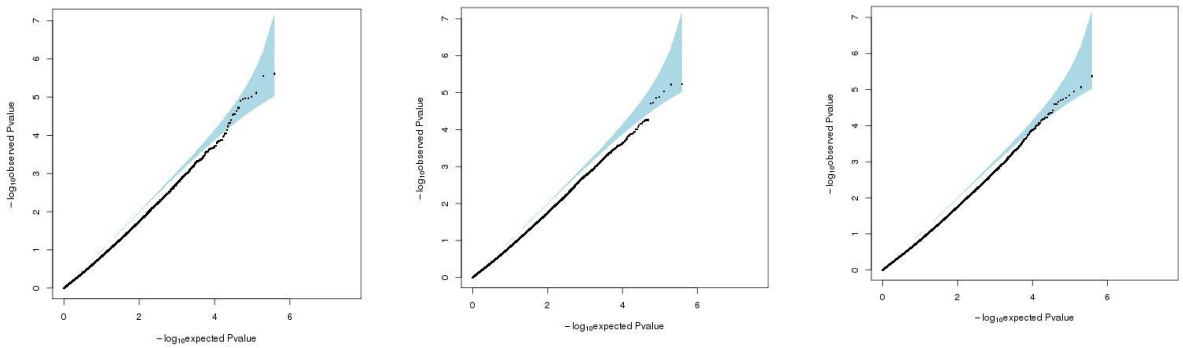


Figure 8.5. Quantile-Quantile graphics from the Meta analysis comparing the p-values resulted with the expected p-values. From left to right: ETP, Peak, and Lagtime.

No novel CpG site associations genome-wide significants were revealed from the meta-analysis of the two MWAS studies. The signal *C15orf41* cg26285502 CpG discussed above ($p = 2.51\ 10^{-8}$) was not observed. The best association with ETP was detected in the *SDAD1* cg27589582 probe with a $p = 2.44\ 10^{-6}$ where a 1% increase in DNA methylation corresponded to a 0.087 ± 0.027 nmol min/L and a 0.086 ± 0.025 nmol min/L increase of ETP, in MARTHA and F5L pedigrees, respectively. For Peak, the probe with the lowest p-value ($p = 5.87\ 10^{-6}$) was the cg18197092 probe located in the chromosome 4, at the position 117,626,456 (a 1% increase in DNA methylation corresponded to 9.564 ± 3.3 nmol/L and 12.147 ± 3.4 nmol/L decrease of Peak, in MARTHA and F5L, respectively). Finally, for Lagtime, the most significant result ($p = 4.23\ 10^{-6}$) was observed with *MITF* cg06070625 CpG (a 1% increase in DNA methylation corresponded to 0.114 ± 0.049 min and a 0.127 ± 0.031 min increase of Lagtime, in MARTHA and F5L, respectively).

8.4. Further analyses

Given the known interplay between DNA methylation and sequence variants, I interrogated the results of my meta-GWAS project on TGP biomarkers (part I) to investigate whether there could exist some SNP at the *C15orf41*, *SDAD1* and *PTH* loci with statistical

evidence for association with Peak, ETP and Lagtime, respectively. If such associations exist, this could provide additional support to my preliminary MWAS findings and guide us towards novel hypotheses of novel genetic factors.

However, in my meta-analysis GWAS study for TGP biomarker performed in 1,967 individuals, the strongest association of SNP at the *C15orf41* locus (i.e., ±100kb from the CpG cg26285502) with Peak was rather modest, $p = 1.63\ 10^{-3}$ at the rs34280623. The lowest p-value observed at the *SDAD1* locus with respect to ETP was p = 0.026 for the rs112868025. Lastly, the rs147594032 was the SNP showing the strongest association with Lagtime at the *MITF*locus (p = 0.036).

To complete my MWAS investigations, I further looked into the results of the MWAS meta-analysis for CpG sites at the *ORM1* and *F2* loci, the main two loci identified in my GWAS project, with suggestive statistical evidence for association with TGP biomarkers. No probe mapping to ORM1 was available in the Illumina HumanMethylation450K array. Conversely, ten probes typed by this array mapped to the *F2* locus. Unfortunately, none of them showed suggestive evidence for association with TGP biomarkers. The lowest p-values observed in the meta-MWAS analysis were p = 0.075, p = 0.118 and p = 0.061 for ETP, Peak and Lagtime, respectively.

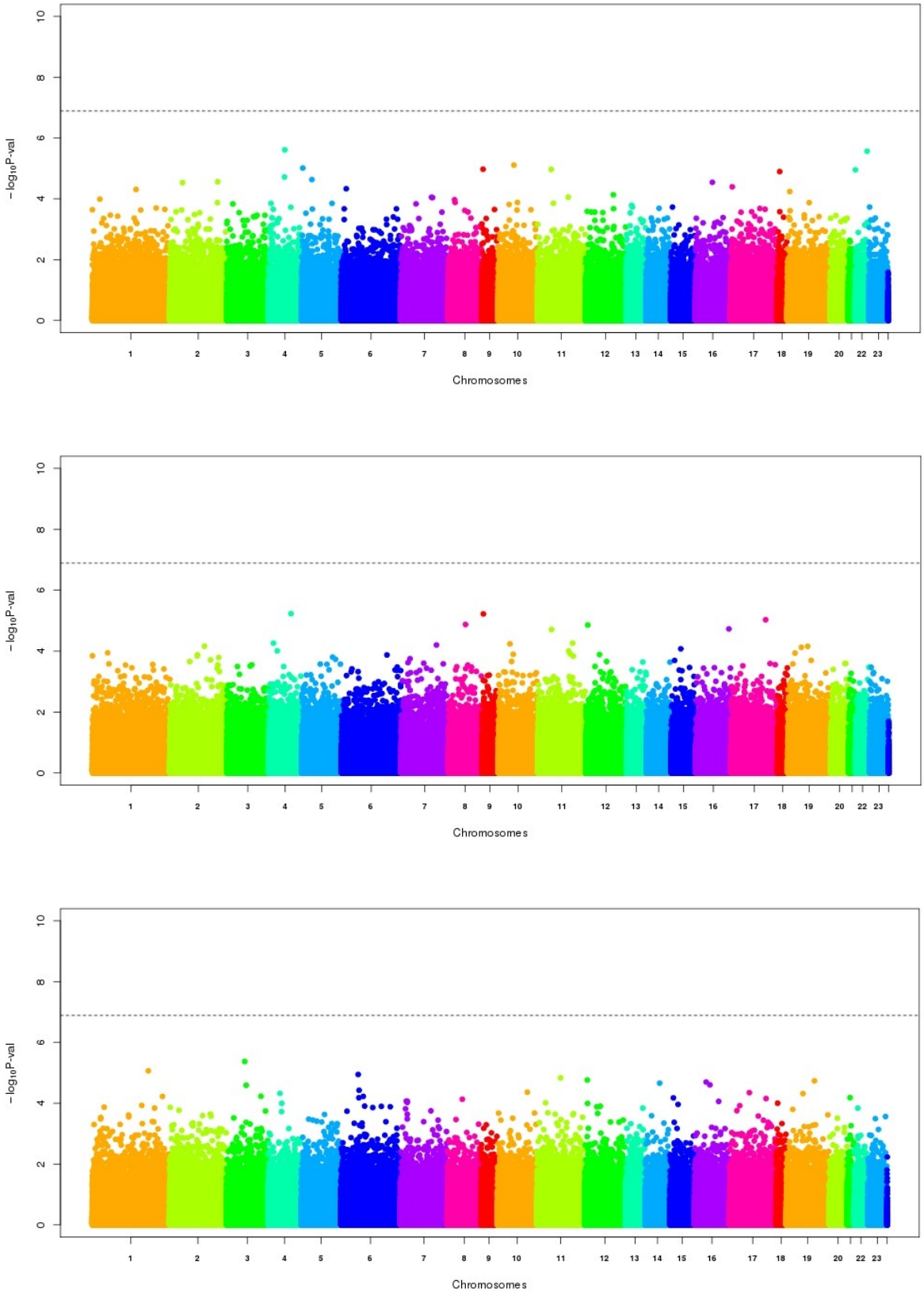


Figure 8.6. Manhattan plots of the results from the meta-analysis of the MARTHA and F5L pedigree MWAS on ETP (top), Peak (middle), and Lagtime (bottom). The horizontal dashed line corresponds to the Bonferroni significant threshold taken at 0.05/388120= 1.29 10^{-7}.

8.5. Discussion

The second part of my PhD project was to conduct a DNA methylome-wide scan to identify DNA methylation changes in relation to the plasma variability of three TGP biomarkers. Unlike my GWAS project, my MWAS strategy was less successful as I did not identify robust methylation signals associated with the studied phenotypes. Even after selecting probes by joint -analysis, the results were not significant. Several limitations can be addressed.

- While extremely dense, the used Illumina array does not cover all sites of the genome that could be subject to DNA methylation. Therefore, we cannot exclude that some relevant methylation association were missed.

- I restricted my MWAS investigations to single CpG site analysis. However, a multi site analysis of CpG located in the same region could be a more powerful approach. This could be particularly useful in presence of several correlated CpG sites, for example in promoter regions with high density of CpGs, similar to what is sometimes advocated in the context of association analyses of SNPs in LD. There are statistical methods for analyzing simultaneously several CpG sites in the context of MWAS:

- **Bump hunting**[213].
- **GAMP** which a package designed to analyze of DNA methylation data including the combination of correlated p-values[214].
- Methods **Comb-**p combining p-values based on the Stouffer, S.A. approach [215]
- **WGCNA** to create networks based on the CpG sites obtaining functional related genes but far located from each other[216]

However, they are not yet robust enough and have not standard methods to handle correctly this issue as it is done with haplotype analysis with SNP data.

- According to an independent work performed by Dylan Aïssi[217], another PhD student working under David-Alexandre Trégouët 's supervision, the sample size of the MARTHA MWAS is large enough to detect a 0.05 increase in DNA methylation β-value at the Bonferroni corrected threshold of $1.29\ 10^{-7}$. Whether an increase of smaller magnitude could be biologically relevant remains an open question, especially given that methylation was measured in peripheral blood DNA, which reflects the average DNA methylation level of different cell types. Therefore, the cell subtype and tissue specific methylation marks would show a weaker effect in whole blood compared to levels that could be observed in more effector cells relevant to thrombin generation.

- TGP technique was applied at the same laboratory for both MARTHA and F5L pedigree studies, and both were also typed for epigenetic patterns in another laboratory. Nevertheless, one cannot exclude heterogeneity caused by design differences between the two studies, decreasing our power to identify robust signals.

In conclusion, as highlighted in my manuscript (page 120), *it may be speculated that 1 - whole blood DNA may not be a good model for identifying differentially methylated regions associated with thrombin generation or 2- variability of DNA methylation levels measured in peripheral blood cells has weak effects on plasma levels of thrombin generation, effects that would require much larger sample size in order to be detected.*

Thrombin Generation Potential and Whole-Blood DNA methylation

Ares Rocañín-Arjó, Jessica Dennis, Pierre Suchon, Dylan Aïssi, Vinh Truong, David-Alexandre Trégouët, France Gagnon, Pierre-Emmanuel Morange

PII: S0049-3848(14)00650-1
DOI: doi: 10.1016/j.thromres.2014.12.010
Reference: TR 5776

To appear in: *Thrombosis Research*

Received date: 13 October 2014
Revised date: 2 December 2014
Accepted date: 6 December 2014

Please cite this article as: Rocañín-Arjó Ares, Dennis Jessica, Suchon Pierre, Aïssi Dylan, Truong Vinh, Trégouët David-Alexandre, Gagnon France, Morange Pierre-Emmanuel, Thrombin Generation Potential and Whole-Blood DNA methylation, *Thrombosis Research* (2014), doi: 10.1016/j.thromres.2014.12.010

Thrombin Generation Potential and Whole-Blood DNA methylation.

Ares Rocañín-Arjó, [1,2,3] Jessica Dennis,[4] Pierre Suchon,[5,6,7] Dylan Aïssi, [1,2,3] Vinh Truong, [4] David-Alexandre Trégouët, [1,2,3], France Gagnon, [4] Pierre-Emmanuel Morange[5,6,7]

[1] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, Team *Genomics & Pathophysiology of Cardiovascular Diseases*, F-75013, Paris,[2] INSERM, UMR_S 1166, Team *Genomics & Pathophysiology of Cardiovascular Diseases*, F-75013, Paris,[3] ICAN Institute for Cardiometabolism and Nutrition, F-75013, Paris, France, [4] Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada, [5] Aix-Marseille University, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, F-13385, Marseille, [6] INSERM, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis,F-13385, Marseille, [7]Laboratory of Haematology, La Timone Hospital, F-13385, Marseille, France.

Correspondence to: Pr Pierre-Emmanuel Morange, Laboratory of Haematology, CHU Timone, 246, rue Saint-Pierre, 13385 Marseille, Cedex 05, France; E-mail: pierre.morange@ap-hm.fr. Tél: +33 4 91 49 24 49

Short running title: DNA methylation and thrombin generation

Financial support:ARA was supported by a PhD grant from the Région Ile de France (CORDDIM).The MARTHA project was supported by a grant from the Program Hospitalier de Recherche Clinique. The F5L Thrombophilia French-Canadian 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 Human450Meethylationepityping was partially funded by the Canadian Institutes of Health Research (grant MOP 86466) and by the Heart and Stroke Foundation of Canada (grant T6484).FG holds a Canada Research Chair. Statistical analyses were performed using the C2BIG computing cluster, funded by the Région Ile de France, Pierre and Marie Curie University, and the ICAN Institute for Cardiometabolism and Nutrition (ANR-10-IAHU-05).

Introduction

A complex cascade of coagulation proteins underlies hemostasis and prevents life-threatening blood loss from damaged blood vessels. Determining the inter-individual variability in plasma levels of the collective effect of these proteins is of great importance. In the last decade, agnostic genome-wide association studies have contributed to discovering novel genes and pathways associated with biomarkers of the coagulation cascade. However, the identified genetic factors account for a minor proportion of the heritability of these biomarkers, suggesting that other contributing factors remain to be identified. DNA methylation is one of the compelling sources that could contribute to the so-called missing heritability of quantitative biological traits. DNA methylation is an epigenetic mechanism that participates in the regulation of gene expression, generally through gene silencing. Several key molecules participating in the coagulation pathway, such as von Willebrand Factor (1), factor VII (2) and factor VIII (3), have been shown to be under the influence of DNA methylation marks. While the former finding was observed in endothelial and kidney cells, the latter two were derived from the analysis of DNA methylation measured in blood.

Intense discussions are emerging about the use of whole blood as a model tissue to discover methylation marks associated with biological phenotypes through methylation-wide association studies (MWAS)(4). In this context, we explored whether the whole blood DNA based MWAS strategy may help discover novel mechanisms associated with hemostasis regulation. We applied this strategy to quantitative biomarkers for thrombotic disorders characterizing the overall hemostatic status of individuals as evaluated by a global thrombin generation assay (5).

Materials and Methods

MARTHA study - We measured genome-wide DNA methylation in whole blood samples of 238 individuals of the MARTHA study (6) using the dedicated Illumina HumanMethylation450 array. A detailed description of the quality controls and the normalization procedures applied to the methylation array data has previously been published (7).

MARTHA subjects were assessed for a thrombin generation assay using the calibrated automated thrombography method (5) as extensively described in (8). Three biological biomarkers were derived from the thrombin generation potential (TGP) measurements produced by the assay: the endogenous thrombin potential (ETP),the lag time and peak height (Supplementary Table 1).

A total of 388,120 CpG sites were then tested for association with three TGP biomarkers. For each TGP biomarker, association analyses were performed using a linear regression approach and adjusted for age, sex, oral contraception therapy, body-mass index, batch and chip effects (7) and cell type composition determined by specific biological counts of lymphocytes, monocytes, neutrophils, eosinophils and basophils.

Any association that was genome-wide significant at the Bonferroni threshold of $1.29\ 10^{-7}$ (=0.05/388,120) was sought for independent replication in the F5L pedigree study.

F5L pedigree study - A total of 187 related individuals of the F5L pedigree study described in (9) were phenotyped for the three TGP biomarkers (Supplementary Table 1) with exactly the same protocol

and by the same laboratory; and simultaneously processed for the DNA methylation array with the MARTHA samples. Association analysis of CpG sites with TGP phenotypes was performed using a linear mixed effect model accounting for the familial relatedness and adjusted for the same factors as in the MARTHA study. As cell counts were not available in the F5L pedigree study, we used the method described in (10) to adjust for cell type composition.

Finally, a MWAS investigation was also carried out in the F5L pedigree study and the results were combined to those obtained in MARTHA through a random-effect meta-analysis using the GWAMA program (11).

Results and Conclusions

None of the tested 388,120 CpG sites was significantly associated with ETP or Lag Time at the Bonferroni threshold of $1.29\ 10^{-7}$. Conversely, one CpG, cg26285502, which mapped to *C15orf41* on chromosome 15q14, reached genome-wide significance for association with Peak height ($P=7.15\ 10^{-8}$). A 1% increase in cg26285502 DNA methylation was associated with an 11.96±2.13nmol L^{-1} increase (increase±SE) in adjusted Peak values. We sought to replicate this finding in an independent sample of 187 individuals of theF5L pedigree study. In F5L, a 1% increase in cg26285502 DNA methylation was associated with anon significant (P=0.40) increase of 2.16±2.57 nmol L^{-1}increase in adjusted Peak values.

In order to increase statistical power to detect variability in CpG sites associated with TGP biomarkers, we conducted a MWAS on each TGP biomarker in the F5L-pedigree study and combined the resultswith those obtained in MARTHA. A Manhattan plot representation of these results is shown in Figure 1. No novel CpG association signal emerged at the pre-specified genome-wide threshold. The smallest observed p-values were $P=2.44\ 10^{-6}$ (*SDAD1* cg27589582), $P=5.87\ 10^{-6}$ (cg18197092 mapping to a region with no gene on chromosome 4q26) and $P=4.23\ 10^{-6}$ (*MITF* cg06070625) for ETP, Peak and Lag Time, respectively.

Recently, it has been suggested that decreased levels of leukocyte DNA methylation at *NOS3* and *EDN1* genes could associate with higher ETP (12). We did not observe strong statistical support of these findings in our combined datasets. Indeed, the *NOS3* CpG site showing the strongest association with ETP was the cg03469471 with combined p-value=0.165. For the *EDN1* locus, the strongest association was observed with cg07714708 (p=0.031). However, this association was not consistent with the results observed in (12). In our samples, a 1% increase in cg03469471 DNA methylation was associated with a 5.42±2.50 increase in ETP (7.84±3.77 and 3.51±3.35 in MARTHA and F5L studies, respectively).

To our knowledge, this work is the first large-scale epidemiological investigation of DNA methylation marks measured in whole blood in relation to quantitative traits of the coagulation cascade. Even though DNA methylation measured in blood reflects average DNA levels from different cell types, the application of the MWAS strategy in whole blood cells has been proposed to detect differentially methylated regions (DMRs). Coagulation-relevant DNA methylation marks detected in blood could reflect stronger effects from effector cells(7), such as hepatocytes and endothelial cells. Such an

approach may be particularly suited when access to relevant cell types/tissues is still not feasible at a large epidemiological scale.

We observed a genome-wide significant association of DNA methylation levels at *C15orf41* with Peak height variability in the MARTHA study but did not validate it in our replication study. *C15orf41* was an interesting candidate as it has very recently been found mutated in patients with congenital dyserythropoietic anaemia (13). We investigated whether *C15orf41* methylation levels could associate with biological criteria of anemia (red blood cells counts and hemoglobin levels)in MARTHA but did not observe any evidence for such associations (data not shown).Nevertheless, further investigations would be warranted to assess whether this original methylation signal was due to random fluctuations or could suggest unsuspected biological links between thrombin generation, *C15orf41* methylation and anaemia.

According to our previous work(14), our study was well powered to detect, at the statistical threshold of $1.29\ 10^{-7}$,any DNA methylation change explaining at least 6% of the variability of a quantitative trait. As a consequence, it may then be speculated that: 1- whole blood DNA may not be a good model for identifying DMRs associated with thrombin generation or 2- variability of DNA methylation levels measured in whole blood cells has weaker effects, if any, on plasma levels of thrombin generation, effects that would require a much larger sample size in order to be detected. In that perspective, the summary statistics derived from the meta-analysis of our two MWAS cohorts on TGP are available upon request for those who would like to combine their own results with ours. In addition, Illumina HumanMethylation450 array and TGP phenotype data from MARTHA participants used in this work are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-3127.

Authorship contributions

A. Rocañín-Arjó conducted the statistical analysis of genome-wide methylation data and drafted the short report. J. Dennis, P. Suchon, D Aïssi and V. Truong collected, pre processed and analyzed data. PE Morange, F. Gagnon and DA. Trégouët designed the research studies, coordinated the analyses and wrote the manuscript.

Conflict of interest disclosure

No conflict of interest to disclose for the study.

References

1. Peng Y, Jahroudi N. The NFY transcription factor inhibits von Willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases. J Biol Chem 2003; 278: 8385–94.

2. Friso S, Lotto V, Choi S-W, et al. Promoter methylation in coagulation F7 gene influences plasma FVII concentrations and relates to coronary artery disease. J Med Genet 2012; 49: 192–9.

3. El-Maarri O, Becker T, Junen J, et al. Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. Hum Genet 2007; 122: 505–14.

4. Murphy TM, Mill J. Epigenetics in health and disease: heralding the EWAS era. Lancet 2014 383: 1952–4.

5. Hemker HC, Giesen P, Al Dieri R, et al. Calibrated Automated Thrombin Generation Measurement in Clotting Plasma. Pathophysiol Haemost Thromb 2003; 33: 4–15.

6. Oudot-Mellakh T, Cohen W, Germain M, et al. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. Br J Haematol 2012; 157: 230–9.

7. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: a genome-wide analysis. Lancet 2014; 6736: 1–9.

8. Rocanin-Arjo A, Cohen W, Carcaillon L, et al. A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential. Blood 2014; 123: 777–85.

9. Antoni G, Morange P-E, Luo Y, 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: 2671–9.

10. Koestler DC, Christensen B, Karagas MR, et al. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. Epigenetics 2013; 8: 816–26.

11. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 2010; 11: 288.

12. Tarantini L, Bonzini M, Tripodi A, et al.Blood hypomethylation of inflammatory genes mediates the effects of metal-rich airborne pollutants on blood coagulation. Occup Environ Med. 2013;70:418-25.

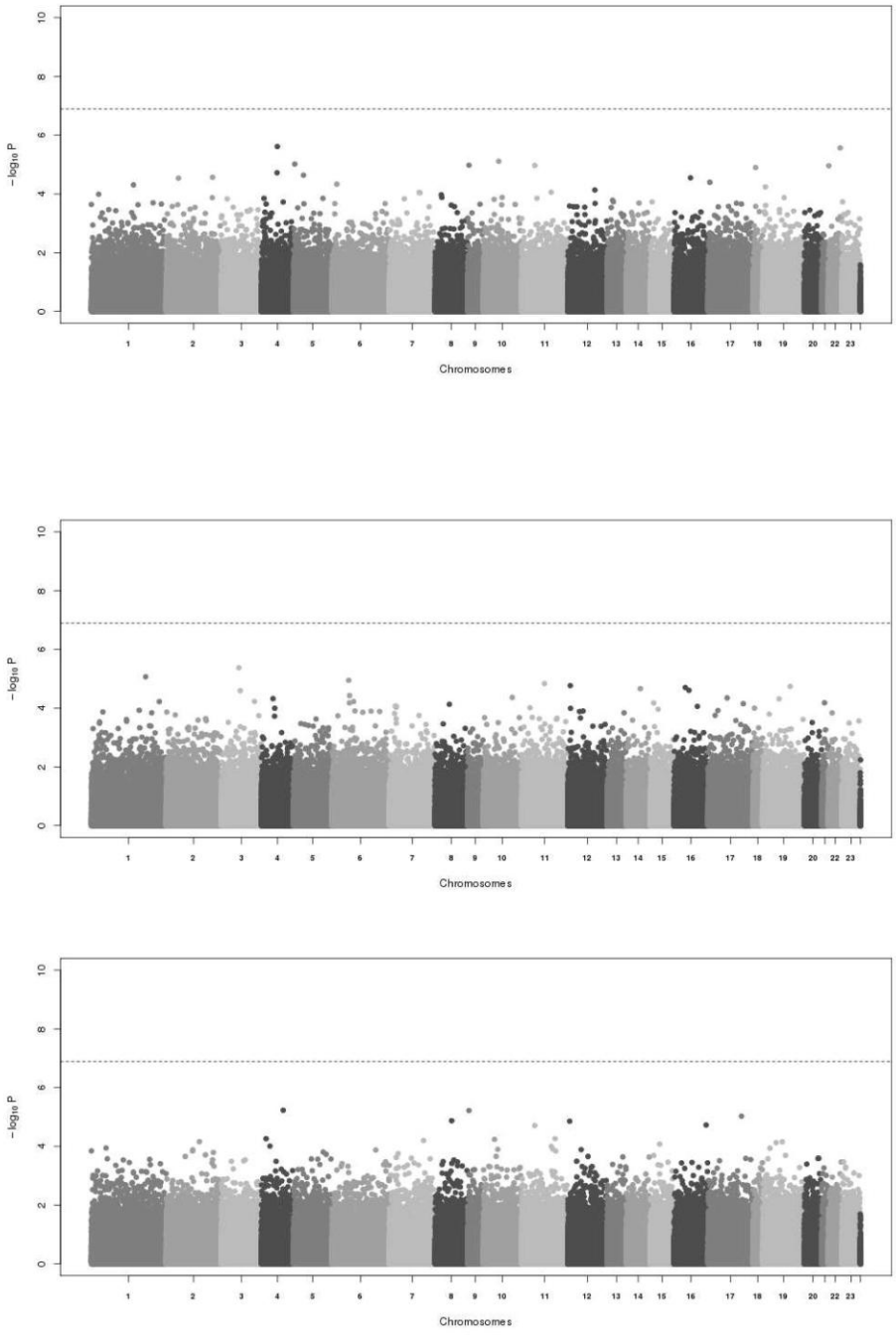
13. Babbs C, Roberts NA, Sanchez-Pulido L, et al. Homozygous mutations in a predicted endonuclease are a novel cause of congenital dyserythropoietic anemia type I.Haematologica. 2013;98:1383-7.

14. Aïssi D, Dennis J, Ladouceur M, et al. Genome-wide investigation of DNA methylation marks associated with FV Leiden. PLoS One 2014; e108087.

Figure Legend

Meta MWAS results of MARTHA and F5L pedigrees.
Manhattan plot representation of the results derived from the meta-analysis of 2 methylome-wide association studies assessing the influence of DNA methylation levels at 388,120 CpG sites on TGP biomarkersin 425 individuals. ETP (Top), peak height (Middle), and lag time (Bottom). The horizontal linecorresponds to the genome-wide significant threshold fixed at $1.29\ 10^{-7}$.

Figure 1



GENERAL CONCLUSIONS

Chapter 9. General discussion, balance and perspectives

My thesis project focused on the identification of new genetic factors and DNA methylation marks that could influence in Thrombin Generation Potential (TGP). TGP is a new promising test measuring the activation of thrombin during the coagulation process. This measurement is an appealing tool for investigating novel determinants of the thrombin activation cascade and of the physiopathological anomalies resulting from abnormal thrombin plasma levels. From a TGP measurement, a curve that displays the activation and production of thrombin throughout the time, it is possible to extract different biomarkers, the main ones being the endogenous thrombin potential (ETP), the total amount of thrombin produced after activation, the maximum amount of thrombin that can be produced at a given time (Peak); and the time that takes to start the thrombin production after activation (Lagtime). The objectives of my project were to identify new genetic and epigenetic factors contributing to modulate the inter-individual variability of these 3 biomarkers with the ultimate goals of better understanding the physiology of the thrombin activation cascade and, if possible, to bring new targets for diagnosis of coagulation/thrombotic disorders.

My work was divided in two main parts according to my objectives. In the first part, I focused my research on screening the whole genome for finding novel common genetic polymorphisms associated with TGP biomarkers using a genome-wide association based strategy. Second, I sought for DNA methylation marks measured from peripheral blood DNA that could also influence TGP measurements using a methylome-wide association based strategy.

My GWAS findings first confirmed that the main genetic factor controlling TGP measurements is the F2 gene coding for prothrombin with two SNPs, rs1799963 (G20210A) and rs3136516 (A19911G) independently influencing ETP and Peak plasma levels. The association of these two SNPs with TGP was already known before the start of my project. While the rs1799963 is also a well-established risk factor for venous thrombosis, the impact of rs3136516 on venous thrombosis is still debated even if some associations have been reported[74,75]. The most important finding of my GWAS project is the identification of the *ORM1* gene as a new susceptibility locus contributing to Lagtime variability. This locus emerged from my GWAS not because it reached genome-wide significance but because it

exhibited strong association signal at $p < 10^{-5}$ with two TGP biomarkers, LagTime and ETP. But only the association with Lagtime robustly replicated in two additional independent studies. *ORM1* codes for orosomucoid (also known as α-1-acid glycoprotein 1 or α-1-AGP) that is protein linked to inflammatory responses such as acute phase response or dermatitis[145,146]. However, it has also been studied in relation to the blood stream transporting proteins and may have a role in the initiation phase of coagulation as a transport protein[147]. Even though we demonstrated *in vitro* that increasing ORM1 levels in plasma is associated with increased Lagtime, there is a lot to be done to understand how *ORM1* acts to influence Lagtime and which genetic variant(s) are the functional one(s) modulating Lagtime variability. Plasma orosomucoid levels are currently under measurement in the MARTHA12 study and, once completed, we will be able to assess whether the identified Lagtime associated *ORM1* SNP is associated with plasma orosomucoid levels. Of note, the allele frequency of the identified SNP was very homogeneous across the venous thrombosis patients and healthy individuals sample I studied suggesting that, even if influencing TGP measurement, the *ORM1* SNP is not a candidate for venous thrombosis. The strategy I used and which led to the *ORM1* findings clear illustrates the interest of widening the search for suggestive association signals among the SNPs that do not pass the genome wide significance threshold, pending the adoption of a clear and rational research strategy based on biological and/or statistical criteria. See for instance the project I initiated and that relies on a candidate genes strategy (Annex I, page 153).

I have not fully exploited all the genetic information I had access to in my GWAS project and there are still some additional research paths to follow in order to detect novel genetic factors contributing to TGP variability. First, X chromosome SNPs were excluded from my GWAS project, as their analysis requires special care and methodologies[153]. This analysis could add novel insights since some key genes participating to the coagulation cascade (F8, F9) are located on the X chromosome. Second, it could be interesting to assess the existence of SNP x SNP interactions modulating TGP measurements. A large number of methodologies have been proposed to search for such interaction in the context of GWAS and none of them is considered as the panacea. Nevertheless, the strategy proposed by Deng[218] to select for entering into pairwise interactions the SNPs exhibiting the strongest inter-genotype variability for a studied phenotype seems appealing. This strategy was the one that was successfully adopted by Nicolas Greliche, a former PhD student at INSERM UMR_S 1166, in the context of SNP x SNP interaction influencing monocyte gene expression[219].

The second component of my PhD project was to conduct a methylome wide association study (MWAS) for TGP biomarkers on DNA methylation measured from peripheral blood DNA. This work was very original as such MWAS approach is just starting to emerge as a promising tool to identify epigenetic marks associated with complex traits. Unfortunately, this part of my work was not as successful as my GWAS project. I looked at both "methylome-wide" significant results and suggestive associations with at least two TGP biomarkers, but was not able to detect any robust methylation signals in the two datasets I have access to. As already discussed in Part II, two main issues could explain why no such signal was found. This could be due to power considerations and additional samples with both TGP and methylome wide data similar to those I used in my project would be required to increase the size/power of the study. Second, while peripheral blood DNA has been shown to be a good model for studying human traits such as lipids and BMI, it may not be a good model for investigating DNA methylation marks associated with thrombin-related phenotypes. Studying DNA methylation in specific tissues or cell types with key roles in coagulation, such as hepatocytes or epithelial cells, would be likely more relevant. Conversely, it would much more difficult to have access to such data at a large-scale epidemiological level.

The MWAS part of my work was partially independent of the GWAS component. Another strategy that deserves to be adopted would be to integrate the MWAS results to dig into the GWAS findings (or vice-versa) in order to improve the power to detect SNP associated with both methylation levels and TGP phenotypes. For example, the CpG - TGP association p-values could be used as weight to prioritize the p-values of the SNP - TGP associations through stratified or weighted- false discovery rate approaches[220,221].

Finally, my GWAS and MWAS researches were conducted on three TGP biomarkers, ETP, Peak and Lagtime. From the global TGP measurement, it is also possible to extract other parameters such as velocity and time to peak (Chapter 2, page 26) that may bring additional information about the haemostasis dynamics of the coagulation response. It would then be interesting to apply the same GWAS and MWAS strategies to these two phenotypes or to multivariate phenotypes combining all 5 TGP biomarkers that could be derived for example through principal component analysis[222].

In conclusion, the study of thrombin physiology is not yet finished; new insights can be found thanks to thrombin generation potential that will help to understand its mechanisms, and hopefully, to better detect anomalies.

BIBLIOGRAPHY

Bibliography

1. Mann, K. G. Thrombin formation. *Chest* **124,** 4S–10S (2003).

2. Mann, K. G. The Dynamics of Thrombin Formation. *Arterioscler. Thromb. Vasc. Biol.* **23,** 17–25 (2002).

3. Licari, L. G. & Kovacic, J. P. Thrombin physiology and pathophysiology. *J. Vet. Emerg. Crit. care* **19,** 11–22 (2009).

4. Crawley, J. T. B., Zanardelli, S., Chion, C. K. N. K. & Lane, D. a. The central role of thrombin in hemostasis. *J. Thromb. Haemost.* **5,** 95–101 (2007).

5. Maragoudakis, M. E. & Tsopanoglou, N. E. *Thrombin: Physiology and Disease*. 266 pages (Springer, 2009).

6. Segers, O., van Oerle, R. Van, ten Cate, H. Ten, Rosing, J. & Castoldi, E. Thrombin generation as an intermediate phenotype for venous thrombosis. *Thromb. Haemost.* **103,** 114–122 (2010).

7. Huntington, J. A. Natural inhibitors of thrombin. *Thromb. Haemost.* **111,** 583–589 (2014).

8. Borissoff, J., Spronk, H. M. H. & Ten Cate, H. The hemostatic system as a modulator of atherosclerosis. *N. Engl. J. Med.* **364,** 1746–1760 (2011).

9. Furie, B. & Furie, B. C. Mechanisms of thrombus formation. *N. Engl. J. Med.* **359,** 938–949 (2008).

10. Mann, K., Orfeo, T. & Butenas, S. Blood Coagulation Dynamics in Hemostasis. *Hamostaseologie* **29,** 7–16 (2009).

11. Esmon, C. T. Inflammation and thrombosis. *J. Thromb. Haemost.* **1,** 1343–1348 (2003).

12. Gruber, A. The role of the contact pathway in thrombus propagation. *Thromb. Res.* **133,** S45–S47 (2014).

13. Jiang, P. *et al.* The extrinsic coagulation cascade and tissue factor pathway inhibitor in macrophages: a potential therapeutic opportunity for atherosclerotic thrombosis. *Thromb. Res.* **133,** 657–66 (2014).

14. Giangrande, P. Six characters in search of an author: the history of the nomenclature of coagulation factors. *Br. J. Haematol.* ***121,*** 703–712 (2003).

15. Esmon, C. Targeting factor Xa and thrombin: impact on coagulation and beyond. *Thromb. Haemost.* **111,** 1–9 (2014).

16. Mann, K. G., Brummel, K. & Butenas, S. What is all that thrombin for? *J. Thromb. Haemost.* **1,** 1504–1514 (2003).

17. Sabater-Lleal, M. *et al.* Combined cis-regulator elements as important mechanism affecting FXII plasma levels. *Thromb. Res.* **125,** e55–60 (2009).

18. Mackman, N. Triggers, targets and treatments for thrombosis. *Nature* **451,** 914–918 (2008).

19. Wood, J. & Bunce, M. Tissue factor pathway inhibitor-alpha inhibits prothrombinase during the initiation of blood coagulation. *PNAS* **110,** 17838–17843 (2013).

20. Castoldi, E. *et al.* Similar hypercoagulable state and thrombosis risk in type I and type III protein S-deficient individuals from families with mixed type I/III protein S deficiency. *Haematologica* **95,** 1563–1571 (2010).

21. Eichinger, S., Hron, G., Kollars, M. & Kyrle, P. A. Prediction of recurrent venous thromboembolism by endogenous thrombin potential and D-dimer. *Clin. Chem.* **54,** 2042–2048 (2008).

22. Wang, W., Boffa, M. B., Bajzar, L., Walker, J. B. & Nesheim, M. E. A Study of the Mechanism of Inhibition of Fibrinolysis by Activated Thrombin-activable Fibrinolysis Inhibitor. *J. Biol. Chem.* **273,** 27176–27181 (1998).

23. Popović, M. *et al.* Thrombin and vascular inflammation. *Mol. Cell. Biochem.* **359,** 301–313 (2012).

24. Kahn, M. L., Nakanishi-Matsui, M., Shapiro, M. J., Ishihara, H. & Coughlin, S. R. Protease-activated receptors 1 and 4 mediate activation of human platelets by thrombin. *J. Clin. Invest.* **103,** 879–887 (1999).

25. Danckwardt, S., Hentze, M. W. & Kulozik, A. E. Pathologies at the nexus of blood coagulation and inflammation: thrombin in hemostasis, cancer, and beyond. *J. Mol. Med. (Berl).* **91,** 1257–1271 (2013).

26. Chen, B. *et al.* Thrombin activity associated with neuronal damage during acute focal ischemia. *J. Neurosci.* **32,** 7622–7631 (2012).

27. Prandoni, P. Venous and arterial thrombosis: two aspects of the same disease? *Eur. J. Intern. Med.* **20,** 660–661 (2009).

28. Antoni, G. *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.* **8,** 2671–2679 (2010).

29. Morange, P.-E. *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.* **86,** 592–595 (2010).

30. Brummel-Ziedins, K. Models for thrombin generation and risk of disease. *J. Thromb. Haemost.* **11 Suppl 1,** 212–223 (2013).

31. Goodeve, a C., Rosén, S. & Verbruggen, B. Haemophilia A and von Willebrand's disease. *Haemophilia* **16 Suppl 5,** 79–84 (2010).

32. Buil, A. *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* **115,** 4644–4650 (2010).

33. Levi, M. & Schultz, M. Coagulopathy and platelet disorders in critically ill patients. *Minerva Anestesiol.* **76,** 851–859 (2010).

34. Daan de Boer, J., Majoor, C. J., van 't Veer, C., Bel, E. H. D. & van der Poll, T. Asthma and coagulation. *Blood* **119,** 3236–3244 (2012).

35. Smid, M. *et al.* Thrombin generation in the Glasgow Myocardial Infarction Study. *PLoS One* **8,** e66977 (2013).

36. Kalz, J., ten Cate, H. & Spronk, H. M. H. Thrombin generation and atherosclerosis. *J. Thromb. Thrombolysis* **37,** 45–55 (2014).

37. Hori, Y., Gabazza, E. & Yano, Y. Insulin resistance is associated with increased circulating level of thrombin-activatable fibrinolysis inhibitor in type 2 diabetic patients. *J. Clin. Endocrinol. Metab.* **87,** 660–665 (2002).

38. Jeong, J. C., Kim, J.-E., Ryu, J. W., Joo, K. W. & Kim, H. K. Plasma haemostatic potential of haemodialysis patients assessed by thrombin generation assay: hypercoagulability in patients with vascular access thrombosis. *Thromb. Res.* **132,** 604–609 (2013).

39. Monagle, P. *Haemostasis: Methods and Protocols*. 427 pages (Humana Press. Springer, 2013).

40. Brodin, E. *et al.* Endogenous thrombin potential (ETP) in plasma from patients with AMI during antithrombotic treatment. *Thromb. Res.* **123,** 573–579 (2009).

41. Hemker, H., Dieri, R. Al, Smedt, E. De & Béguin, S. Thrombin generation , a function test of the haemostatic- thrombotic system. *Thromb Haemost* **96,** 553–561 (2006).

42. Aoki, I. *et al.* Platelet-dependent thrombin generation in patients with hyperlipidemia. *J. Am. Coll. Cardiol.* **30,** 91–96 (1997).

43. Whelihan, M. F., Kiankhooy, A. & Brummel-Ziedins, K. E. Thrombin generation and fibrin clot formation under hypothermic conditions: an in vitro evaluation of tissue factor initiated whole blood coagulation. *J. Crit. Care* **29,** 24–30 (2014).

44. Hemker, H. *Thrombin*. 36 pages (Synapse b.v., 2013).

45. Hemker, H. C. *et al.* Calibrated Automated Thrombin Generation Measurement in Clotting Plasma. *Pathophysiol. Haemost. Thromb.* **33,** 4–15 (2003).

46. Rugeri, L. *et al.* Thrombin generation in patients with factor XI deficiency and clinical bleeding risk. *Haemophilia* **16,** 771–777 (2010).

47. Cimenti, C., Schlagenhauf, A. & Leschnik, B. Low endogenous thrombin potential in trained subjects. *Thromb. Res.* **131,** 281–285 (2013).

48. Hemker, H. C. *et al.* The calibrated automated thrombogram (CAT): a universal routine test for hyper- and hypocoagulability. *Pathophysiol. Haemost. Thromb.* **32,** 249–253 (2002).

49. Macfarlane, R. & Biggs, R. A Thrombin Generation Test. The application in haemophilia and Thrombocytopenia. *J. Clin. Pathol.* **6,** 3–9 (1953).

50. Herbert, F. K. The estimation of prothrombin in human plasma. *Biochem. J.* **34,** 1554–1568 (1940).

51. Pitney, W. R. & Dacie, J. V. A simple method of studying the generation of thrombin in recalcified plasma; application in the investigation of haemophilia. *J. Clin. Pathol.* **6,** 9–14 (1953).

52. Beguin, S. & Hemker, H. Method for determining the endogenous thrombin potential of plasma and blood. *US Patent. 5,192,689* (1993).

53. Ramjee, M. K. The use of fluorogenic substrates to monitor thrombin generation for the analysis of plasma and whole blood coagulation. *Anal. Biochem.* **277,** 11–18 (2000).

54. Lavigne-Lissalde, G. *et al.* Prothrombin G20210A carriers the genetic mutation and a history of venous thrombosis contributes to thrombin generation independently of factor II plasma levels. *Haemostasis* **8,** 942–949 (2010).

55. Castoldi, E. & Rosing, J. Thrombin generation tests. *Thromb. Res.* **127 Suppl ,** S21–S25 (2011).

56. Chantarangkul, V., Clerici, M. & Bressi, C. Thrombin generation assessed as endogenous thrombin potential in patients with hyper-or hypo-coagulability. *Haematologica* **88,** 547–554 (2003).

57. Brummel-Ziedins, K. E., Everse, S. J., Mann, K. G. & Orfeo, T. Modeling thrombin generation: plasma composition based approach. *J. Thromb. Thrombolysis* **37,** 32–44 (2014).

58. Dielis, A. W. J. H. *et al.* Coagulation factors and the protein C system as determinants of thrombin generation in a normal population. *J. Thromb. Haemost.* **6,** 125–131 (2008).

59. Hemker, H. C., Al Dieri, R. & Béguin, S. Thrombin generation assays: accruing clinical relevance. *Curr. Opin. Hematol.* **11,** 170–175 (2004).

60. Sonnevi, K. *et al.* Obesity and thrombin-generation profiles in women with venous thromboembolism. *Blood Coagul. fibrinolysis* **24,** 1–7 (2013).

61. Brummel-Ziedins, K. The plasma hemostatic proteome: thrombin generation in healthy individuals. *J. Thromb. Haemost.* **3,** 1472–1481 (2005).

62. Haidl, H., Cimenti, C. & Leschnik, B. Age-dependency of thrombin generation measured by means of calibrated automated thrombography (CAT). *Thromb. Haemost.* **95,** 772–775 (2006).

63. Brummel-Ziedins, K. E. *et al.* The prothrombotic phenotypes in familial protein C deficiency are differentiated by computational modeling of thrombin generation. *PLoS One* **7,** e44378 (2012).

64. Chaireti, R., Gustafsson, K. M., Byström, B., Bremme, K. & Lindahl, T. L. Endogenous thrombin potential is higher during the luteal phase than during the follicular phase of a normal menstrual cycle. *Hum. Reprod.* **28,** 1846–1852 (2013).

65. Scarabin, P.-Y., Hemker, H. C., Clément, C., Soisson, V. & Alhenc-Gelas, M. Increased thrombin generation among postmenopausal women using hormone therapy: importance of the route of estrogen administration and progestogens. *Menopause* **18,** 873–879 (2011).

66. Bloemen, S., Pieters, M., Hemker, H. & Dieri, R. Al. No Effect of Ethanol Intake on Thrombin Generation Parameters. *Thromb. Res.* **129,** 10–11 (2012).

67. Sanchez, C. *et al.* Diet modulates endogenous thrombin generation, a biological estimate of thrombosis risk, independently of the metabolic status. *Arterioscler. Thromb. Vasc. Biol.* **32,** 2394–2404 (2012).

68. Ceelie, H., Bertina, R. M., van Hylckama Vlieg, a, Rosendaal, F. R. & Vos, H. L. Polymorphisms in the prothrombin gene and their association with plasma prothrombin levels. *Thromb. Haemost.* **85,** 1066–1070 (2001).

69. Von Ahsen, N. & Oellerich, M. The intronic prothrombin 19911A>G polymorphism influences splicing efficiency and modulates effects of the 20210G>A polymorphism on mRNA amount and expression in a stable reporter gene assay system. *Blood* **103,** 586–593 (2004).

70. Demirci, F. Y. K. *et al.* Functional polymorphisms of the coagulation factor II gene (F2) and susceptibility to systemic lupus erythematosus. *J. Rheumatol.* **38,** 652–657 (2011).

71. Poort, S. R., Rosendaal, F. R., Reitsma, P. H. & Bertina, R. M. 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* **88,** 3698–3703 (1996).

72. Kyrle, P. a. *et al.* Clinical Studies and Thrombin Generation in Patients Homozygous or Heterozygous for the G20210A Mutation in the Prothrombin Gene. *Arterioscler. Thromb. Vasc. Biol.* **18,** 1287–1291 (1998).

73. Sode, B., Allin, K. & Dahl, M. Risk of venous thromboembolism and myocardial infarction associated with factor V Leiden and prothrombin mutations and blood type. *Can. Med. Assoc. J.* **185,** 229–237 (2013).

74. Martinelli, I. *et al.* Prothrombin A19911G polymorphism and the risk of venous thromboembolism. *J. Thromb. Haemost.* **4,** 2582–2586 (2006).

75. Chinthammitr, Y., Vos, H. L., Rosendaal, F. R. & Doggen, C. J. M. The association of prothrombin A19911G polymorphism with plasma prothrombin activity and venous thrombosis: results of the MEGA study, a large population-based case-control study. *J. Thromb. Haemost.* **4,** 2587–2592 (2006).

76. Tang, W. *et al.* A genome-wide association study for venous thromboembolism: the extended cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium. *Genet. Epidemiol.* **37,** 512–521 (2013).

77. Germain, M. *et al.* Genetics of venous thrombosis: insights from a new genome wide association study. *PLoS One* **6,** e25581 (2011).

78. Germain, M. *et al.* Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. *PLoS One* **7,** e38538 (2012).

79. Trégouët, D., König, I. & Erdmann, J. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.* **41,** 2008–2010 (2009).

80. Lutsey, P. L., Folsom, A. R., Heckbert, S. R. & Cushman, M. Peak Thrombin Generation and Subsequent Venous Thromboembolism: The Longitudinal Investigation of Thromboembolism Etiology (LITE). *October* **7,** 1639–1648 (2010).

81. Hron, G., Kollars, M., Binder, B. R., Eichinger, S. & Kyrle, P. a. Identification of patients at low risk for recurrent venous thromboembolism by measuring thrombin generation. *JAMA* **296,** 397–402 (2006).

82. Tripodi, A. *et al.* High thrombin generation measured in the presence of thrombomodulin is associated with an increased risk of recurrent venous thromboembolism. *J. Thromb. Haemost.* **6,** 1327–33 (2008).

83. Carcaillon, L. *et al.* Elevated plasma fibrin D-dimer as a risk factor for vascular dementia: the Three-City cohort study. *J. Thromb. Haemost.* **7,** 1972–1978 (2009).

84. Orbe, J. *et al.* Increased thrombin generation after acute versus chronic coronary disease as assessed by the thrombin generation test. *Thromb. Haemost.* **99,** 382–387 (2008).

85. Ten Cate, H. Thrombin generation in clinical conditions. *Thromb. Res.* **129,** 367–370 (2012).

86. Al Dieri, R. *et al.* The thrombogram in rare inherited coagulation disorders: its relation to clinical bleeding. *Thromb. Haemost.* **88,** 576–582 (2002).

87. Vijil, C., Hermansson, C., Jeppsson, A., Bergström, G. & Hultén, L. M. Arachidonate 15-lipoxygenase enzyme products increase platelet aggregation and thrombin generation. *PLoS One* **9,** e88546 (2014).

88. Ay, L., Hoellerl, F., Ay, C. & Brix, J. Thrombin generation in type 2 diabetes with albuminuria and macrovascular disease. *Eur. J. Clin. Invest.* **42,** 470–477 (2012).

89. Van der Poll, T. Thrombin and diabetic nephropathy. *Blood* **117,** 5015–5016 (2011).

90. Potze, W. *et al.* Differential in vitro inhibition of thrombin generation by anticoagulant drugs in plasma from patients with cirrhosis. *PLoS One* **9,** e88390 (2014).

91. Bernhard, H. *et al.* Thrombin generation in pediatric patients with Crohn's disease. *Inflamm. Bowel Dis.* **17,** 2333–2339 (2011).

92. Petros, S., Kliem, P., Siegemund, T. & Siegemund, R. Thrombin generation in severe sepsis. *Thromb. Res.* **129,** 797–800 (2012).

93. Noubouossie, D. F. *et al.* Thrombin generation reveals high procoagulant potential in the plasma of sickle cell disease children. *Am. J. Hematol.* **87,** 145–149 (2012).

94. Rocanin-Arjo, A. *et al.* A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential. *Blood* **123,** 777–785 (2014).

95. Potter, J. D. At the interfaces of epidemiology, genetics and genomics. *Nat. Rev. Genet.* **2,** 142–147 (2001).

96. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

97. Zemunik, T. & Boraska, V. Chapter 23 Genetics of Type 1 Diabetes in *Type 1 Diabetes. Pathogenesis, Genetics and Immunotherapy*. pages 529–548 (InTech, 2011).

98. Manolio, T. a. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* **14,** 549–558 (2013).

99. National Human Genome Research Institute- NHGRI. at <http://www.genome.gov/>

100. Illumina. at <http://supportres.illumina.com/documents/>

101. Weale, M. E. Quality control for genome-wide association studies. *Methods Mol. Biol.* **628,** 341–372 (2010).

102. Pluzhnikov, A. *et al.* Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *Am. J. Hum. Genet.* **87,** 123–128 (2010).

103. Ioannidis, J. P. a, Thomas, G. & Daly, M. J. Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.* **10,** 318–329 (2009).

104. Anderson, C. a *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5,** 1564–1573 (2010).

105. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

106. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).

107. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

108. Anderson, C. Chapter7 Data Quality Control in *Analysis of Complex Disease Association Studies.* pages 95–108 (Elsevier Inc., 2010).

109. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).

110. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* **12,** 465–474 (2011).

111. Kavraki, L. E. Dimensionality Reduction Methods for Molecular Motion. at <http://cnx.org/content/m11461/latest/>

112. Hardy, G. Mendelian Proportions in a Mixed Population. *Science (80-.).* **XXVIII,** 49 (1908).

113. Attia, J. *et al.* How to use an article about genetic association: A: Background concepts. *JAMA* **301,** 74–81 (2009).

114. Cordell, H. & Clayton, D. Genetic association studies. *Lancet* **366,** 1121–1131 (2005).

115. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9,** 356–369 (2008).

116. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practival and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).

117. Gordon, A., Glazko, G., Qiu, X. & Yakovlev, A. Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Ann. Appl. Stat.* **1,** 179–190 (2007).

118. Carvajal-Rodríguez, A., de Uña-Alvarez, J. & Rolán-Alvarez, E. A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. *BMC Bioinformatics* **10,** 209 (2009).

119. Amos, C. I. Successful design and conduct of genome-wide association studies. *Hum. Mol. Genet.* **16 Spec No,** R220– R225 (2007).

120. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32,** 381–385 (2008).

121. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32,** 227–234 (2008).

122. Abdi, H. *The Bonferonni and Šidák corrections for multiple comparisons* in *Encycl. Meas. Stat.* (Salkind, N.) 1–9 (Thousand Oaks (CA): Sage, 2007).

123. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21,** 1539–1558 (2002).

124. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).

125. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).

126. Consortium, T. I. H. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

127. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

128. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78,** 629–644 (2006).

129. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).

130. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).

131. Stram, D. O. *et al.* Modeling and E-M Estimation of Haplotype-Specific Relative Risks from Genotype Data for a Case-Control Study of Unrelated Individuals. *Hum. Hered.* **55,** 179–190 (2003).

132. De Bakker, P. I. W. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17,** R122– R128 (2008).

133. Morange, P.-E. *et al.* KNG1 Ile581Thr and susceptibility to venous thrombosis. *Blood* **117,** 3692–3694 (2011).

134. Oudot-Mellakh, T. *et al.* Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. *Br. J. Haematol.* **157,** 230–239 (2012).

135. Group, T. 3C S. Vascular Factors and Risk of Dementia : Design of the Three-City Study and Baseline Characteristic of the Study Population. *Neuroepidemiology* **22,** 316–325 (2003).

136. Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327,** 557–560 (2003).

137. Mantel, N. & Haenszel, W. illia. Statical Aspects of the Analysis of Data From Retrospective Studies of Disease. *J Natl Cancer Inst* **22,** 719– 748 (1959).

138. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190– 2191 (2010).

139. Mazoyer, E. *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 Coagul. Fibrinolysis* **20,** 503–510 (2009).

140. Germain, M. *et al.* Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. *PLoS One* **7,** e38538 (2012).

141. Shah, S. *et al.* Four genetic loci influencing electrocardiographic indices of left ventricular hypertrophy. *Circ. Cardiovasc. Genet.* **4,** 626–635 (2011).

142. Zeller, T. *et al.* Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One* **5,** e10693 (2010).

143. Poort, S., Rosendaal, F., Reitsma, P. & Bertina, R. 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* **88,** 3698–3703 (1996).

144. Demirci, F. Y. K. *et al.* Functional polymorphisms of the coagulation factor II gene (F2) and susceptibility to systemic lupus erythematosus. *J. Rheumatol.* **38,** 652–657 (2011).

145. Treuheit, M. J., Costellot, C. E. & Halsall, H. B. Analysis of the five glycosylation sites of human alpha1-acid glycoprotein. *Biochem. J.* **283,** 105–112 (1992).

146. Fan, C. No Orosomucoid types in allergic contact dermatitis. *Hum. Hered.* **45,** 117–120 (1995).

147. Osikov, M., Makarov, E. & Krivokhizhina, L. Effects of alpha1-acid glycoprotein on hemostasis in experimental septic peritonitis. *Butlletin Exp. Biol. Med.* **144,** 178–180 (2007).

148. Costello, M., Fiedel, B. & Gewurz, H. Inhibition of platelet aggregation by native and desialised alpha-1 acid glycoprotein. *Nature* **281,** 677–678 (1979).

149. Boncela, J., Papiewska, I., Fijalkowska, I., Walkowiak, B. & Cierniewski, C. S. Acute phase protein alpha 1-acid glycoprotein interacts with plasminogen activator inhibitor type 1 and stabilizes its inhibitory activity. *J. Biol. Chem.* **276,** 35305–35311 (2001).

150. Klatzow, D. & Vos, G. The effect of seromucoid on coagulation. *South African Med. J.* **60,** 424–427 (1981).

151. Su, S. & Yeh, T. Effects of alpha 1-acid glycoprotein on tissue factor expression and tumor necrosis factor secretion in human monocytes. *Immunopharmacology* **34,** 139–145 (1996).

152. Wiklund, P. K. *et al.* Serum metabolic profiles in overweight and obese women with and without metabolic syndrome. *Diabetol. Metab. Syndr.* **6,** 40 (2014).

153. Castagné, R. *et al.* Influence of sex and genetic variability on expression of X-linked genes in human monocytes. *Genomics* **98,** 320–326 (2011).

154. Katori, N. & Sai, K. Genetic variations of orosomucoid genes associated with serum alpha-1-acid glycoprotein level and the pharmacokinetics of paclitaxel in Japanese cancer patients. *J. Pharm. Sci.* **100,** 4546–4559 (2011).

155. Bird, A. Perceptions of epigenetics. *Nature* **447,** 396–398 (2007).

156. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33,** 245–254 (2003).

157. Duncan, E. J., Gluckman, P. D. & Dearden, P. K. Epigenetics, plasticity, and evolution: How do we link epigenetic change to phenotype? *J. Exp. Zool. B. Mol. Dev. Evol.* **322,** 208–220 (2014).

158. Füllgrabe, J., Kavanagh, E. & Joseph, B. Histone onco-modifications. *Oncogene* **30,** 3391–403 (2011).

159. Jones, P. a & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3,** 415–428 (2002).

160. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* **20,** 883–889 (2010).

161. Strachan, T. & Read, A. *Human Molecular Genetics*. 781 (Garland Science/Taylor & Francis Group, 2010).

162. Miranda, T. & Jones, P. DNA methylation: the nuts and bolts of repression. *J. Cell. Physiol.* **213,** 384–390 (2007).

163. Laird, P. W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11,** 191–203 (2010).

164. Fu, Y., Dominissini, D., Rechavi, G. & He, C. Gene expression regulation mediated through reversible m(6)A RNA methylation. *Nat. Rev. Genet.* **15,** 293–306 (2014).

165. Laird, P. W. The power and the promise of DNA methylation markers. *Nat. Rev. Cancer* **3,** 253–266 (2003).

166. Bell, C. G. *et al.* Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS One* **5,** e14040 (2010).

167. Antequera, F. & Bird, a. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 11995–11999 (1993).

168. Dai, W. *et al.* Methylation Linear Discriminant Analysis (MLDA) for identifying differentially methylated CpG islands. *BMC Bioinformatics* **9,** 337 (2008).

169. Brena, R. M., Huang, T. H.-M. & Plass, C. Toward a human epigenome. *Nat. Genet.* **38,** 1359–1360 (2006).

170. Rakyan, V. K., Down, T. a, Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12,** 529–541 (2011).

171. Bagci, H. & Fisher, A. G. DNA demethylation in pluripotency and reprogramming: the role of tet proteins and cell division. *Cell Stem Cell* **13,** 265–269 (2013).

172. Zeilinger, S. *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One* **8,** e63812 (2013).

173. Irizarry, R. a *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41,** 178–186 (2009).

174. Bell, C. G. *et al.* Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC Med. Genomics* **3,** 33 (2010).

175. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366,** 362–365 (1993).

176. Matouk, C. C. & Marsden, P. a. Epigenetic regulation of vascular endothelial gene expression. *Circ. Res.* **102,** 873–887 (2008).

177. Hodge, D., Peng, B., Cherry, J. & Hurt, E. Interleukin 6 Supports the Maintenance of p53 Tumor Suppressor Gene Promoter Methylation. *Cancer Res.* **65,** 4673–4682 (2005).

178. Movassagh, M. *et al.* Differential DNA methylation correlates with differential expression of angiogenic factors in human heart failure. *PLoS One* **5,** e8564 (2010).

179. Stenvinkel, P. *et al.* Impact of inflammation on epigenetic DNA methylation - a novel risk factor for cardiovascular disease? *J. Intern. Med.* **261,** 488–499 (2007).

180. Peng, Y. & Jahroudi, N. The NFY transcription factor inhibits von Willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases. *J. Biol. Chem.* **278,** 8385–8394 (2003).

181. El-Maarri, O. *et al.* Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. *Hum. Genet.* **122,** 505–514 (2007).

182. Stern, L., Mason, J., Selhub, J. & Choi, S. Genomic DNA Hypomethylation , a Characteristic of Most Cancers , Is Present in Peripheral Leukocytes of Individuals Who Are Homozygous for the C677T Polymorphism in the Methylenetetrahydrofolate Reductase Gene. *Cancer Epidemiol. Biomarkers Prev.* **9,** 849–853 (2000).

183. Friso, S. *et al.* A common mutation in the 5,10-methylenetetrahydrofolate reductase gene affects genomic DNA methylation through an interaction with folate status. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 5606–5611 (2002).

184. Den Heijer, M., Lewington, S. & Clarke, R. Homocysteine, MTHFR and risk of venous thrombosis: a meta-analysis of published epidemiological studies. *J. Thromb. Haemost.* **3,** 292–299 (2005).

185. Dick, K. J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383,** 1–9 (2014).

186. Gagnon, F., Aïssi, D., Carrié, A., Morange, P.-E. & Trégouët, D.-A. Robust validation of methylation levels association at CPT1A locus with lipid plasma levels. *J. Lipid Res.* **55,** 1189–1191 (2014).

187. Terry, M. B., Delgado-Cruzata, L., Vin-Raviv, N., Wu, H. C. & Santella, R. M. DNA methylation in white blood cells: Association with risk factors in epidemiologic studies. *Epigenetics* **6,** 828–837 (2011).

188. Irizarry, R., Ladd-Acosta, C. & Carvalho, B. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* **18,** 780–790 (2008).

189. Harris, R., Wang, T. & Coarfa, C. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **28,** 1097–1105 (2010).

190. Gupta, R., Nagarajan, A. & Wajapeyee, N. Advances in genome-wide DNA methylation analysis. *Biotechniques* **49,** iii–xi (2010).

191. Korshunova, Y. & Maloney, R. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Res.* **18,** 19–29 (2008).

192. Manolio, T. a *et al.* New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.* **39,** 1045–1051 (2007).

193. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6,** 692–702 (2011).

194. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98,** 288–295 (2011).

195. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587 (2010).

196. Wahl, S. *et al.* On the potential of models for location and scale for genome-wide DNA methylation data. *BMC Bioinformatics* **15,** 232 (2014).

197. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31,** 142–147 (2013).

198. Davis, S., Du, S., Bilke, S., Triche, T. J. & Bootwalla, M. Methylumi: Handle Illumina methylation data. R package version 2.10.0. (2014). at <http://www.bioconductor.org/packages/release/bioc/html/methylumi.html>

199. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8,** 203–209 (2013).

200. Chen, Y. *et al.* Cross-reactive DNA microarray probes lead to false discovery of autosomal sex-associated DNA methylation. *Am. J. Hum. Genet.* **91,** 762–764 (2012).

201. Silver, J. D., Ritchie, M. E. & Smyth, G. K. Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. *Biostatistics* **10,** 352–363 (2009).

202. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41,** e90 (2013).

203. Dedeurwaerder, S., Defrance, M. & Calonne, E. Evaluation of the Infinium Methylation 450K technology. *Futur. Med.* **3,** 771–784 (2011).

204. Touleimat, N. & Tost, J. Complete pipeline for Infinium ® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation T echnology R eport. *Epigenomics* **4,** 325–341 (2012).

205. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29,** 189–196 (2013).

206. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol.* **13,** R44 (2012).

207. Kuan, P. F., Wang, S., Zhou, X. & Chu, H. A statistical framework for Illumina DNA methylation arrays. *Bioinformatics* **26,** 2849–2855 (2010).

208. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11,** 288 (2010).

209. Wilhelm-Benartzi, C. S. *et al.* Review of processing and analysis methods for DNA methylation array data. *Br. J. Cancer* **109,** 1394–1402 (2013).

210. Houseman, E. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86–102 (2012).

211. Jaffe, A. E. & Irizarry, R. a. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15,** R31 (2014).

212. Koestler, D. C. *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* **8,** 816–826 (2013).

213. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41,** 200–209 (2012).

214. Zhao, S.-X. *et al.* A Refined Study of FCRL Genes from a Genome-Wide Association Study for Graves' Disease. *PLoS One* **8,** e57758 (2013).

215. Pedersen, B. S., Schwartz, D. a, Yang, I. V & Kechris, K. J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* **28,** 2986–2988 (2012).

216. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

217. Aïssi, D. *et al.* Genome-Wide Investigation of DNA Methylation Marks Associated with FV Leiden Mutation. *PLoS One* **9,** e108087 (2014).

218. Deng, W. Q. & Paré, G. A fast algorithm to optimize SNP prioritization for gene-gene and gene-environment interactions. *Genet. Epidemiol.* **35,** 729–738 (2011).

219. Greliche, N. *et al.* A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis. *BMC Med. Genet.* **14,** 36 (2013).

220. Dalmasso, C., Génin, E. & Trégouet, D.-A. A weighted-Holm procedure accounting for allele frequencies in genomewide association studies. *Genetics* **180,** 697–702 (2008).

221. Yoo, Y. J., Bull, S. B., Paterson, A. D., Waggott, D. & Sun, L. Were genome-wide linkage studies a waste of time? Exploiting candidate regions within genome-wide association studies. *Genet. Epidemiol.* **34,** 107–118 (2010).

222. Aschard, H. *et al.* Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am. J. Hum. Genet.* **94,** 662–676 (2014).

223. Carcaillon, L. *et al.* Increased thrombin generation is associated with acute ischemic stroke but not with coronary heart disease in the elderly: the Three-City cohort study. *Arterioscler. Thromb. Vasc. Biol.* **31,** 1445–1451 (2011).

224. Borissoff, J. I. *et al.* Early atherosclerosis exhibits an enhanced procoagulant state. *Circulation* **122,** 821–830 (2010).

225. Potze, W. *et al.* Decreased tissue factor pathway inhibitor (TFPI)-dependent anticoagulant capacity in patients with cirrhosis who have decreased protein S but normal TFPI plasma levels. *Br. J. Haematol.* **162,** 819–826 (2013).

226. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13,** R97 (2012).

227. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6,** 597–610 (2005).

228. Frazier-Wood, A. C. *et al.* Methylation at CPT1A locus is associated with lipoprotein subfraction profiles. *J. Lipid Res.* **55,** 1324–1330 (2014).

229. Buil, A. *et al.* Protein C levels are regulated by a quantitative trait locus on chromosome 16: results from the Genetic Analysis of Idiopathic Thrombophilia (GAIT) Project. *Arterioscler. Thromb. Vasc. Biol.* **24,** 1321–1325 (2004).

ANNEXES

ANNEX 1

Following the results of my metaGWAS analysis on TGP biomarkers, I further investigated whether SNPs located within some candidate genes showed suggestive evidence for statistical association with TGP biomarkers even if they have not reached genome-wide significance. Selected candidates were genes with strong support for association with venous thrombosis for which TGP phenotypes are risk factor. Nineteen genes derived from^{28,29,77,229} were considered (Table A1.1).

GENE	CHR	Start*	End*
FV	1	169431192	169605769
AT_SERPINC1	1	173822938	173936516
C4BPA	1	207227607	207368317
PC	2	128125996	128236822
PS	3	93541881	93742934
KNG1	3	186385098	186512199
FGG_FGA	4	155475286	155551897
FXI	4	187137118	187250835
HIVEP1	6	11962724	12215232
BAI3	6	69295632	70149403
STXBP5	6	147475494	147761612
ABO	9	136080563	136200630
FII	11	46690743	46811056
vWF	12	6008040	6283836
STAB2	12	103931069	104210502
TC2N	14	92196266	92383880
SERPINF2	17	1596130	1708559
GP6	19	55475073	55599632
PROCR	20	33709774	33815165

Table A1.1 Candidate Genes list

*Annotation corresponding to GRCh37/hg19

At each candidate locus (± 50kb), I focused on SNPs with marginal association p-values ($p < 0.05$) with at least two TGP biomarkers and with consistent association between MARTHA and 3C. I further selected for replication in MARTHA12 the one with the lowest p-values (Table A1.2).

Gene	SNP	ETP	Peak	LagTime
BAI3	rs17691415	$P = 1.59\ 10^{-4}$	P =0.0465	
C4BPA	rs2012296	P = 0.014	$P = 7.19\ 10^{-4}$	
F2	rs1799963	$P < 10^{-16}$	$P = 1.60\ 10^{-8}$	P = 0.034
F2	rs3136516	$P = 5.94\ 10^{-14}$	$P = 2.92\ 10^{-8}$	P > 0.05
F5	rs6015	$P = 2.94\ 10^{-4}$	$P = 4.05\ 10^{-7}$	P > 0.05
F5	rs6025	P = 0.026	$P = 3.46\ 10^{-4}$	P > 0.05
FGG	rs2066865	$P = 2.46\ 10^{-4}$	P = 0.0212	P > 0.05
PROS	rs12629037	P > 0.05	P = 0.050	$P = 1.52\ 10^{-3}$
TC2N	rs10138058	P = 0.040	$P = 9.59\ 10^{-4}$	P = 0.012
STAB2	rs2292687	P >0.05	P = 0.050	$P = 2.04\ 10^{-3}$

Table A1.2 Selected SNPs for replication

Abstract

Thrombin Generation Potential (TGP) is a promising *in vitro* measurement that allows quantifying thrombin activity, in a close way to what happens *in vivo*. It is sensitive to coagulation factors deficiencies, anticoagulant proteins and is associated to thrombotic disorders. There exists two polymorphisms located in the F2 (prothrombin) gene known to influence TGP levels, and altogether they explain 11.3% of the TGP inter-individual variability. With the aims of identifying novel genetic and epigenetic factors that influence TGP variability, I have performed two different studies in the present work. First, I conducted the first genome-wide association study for the three TGP biomarkers (ETP, Peak and Lagtime) using imputation data from two French studies. The most significant single nucleotide polymorphisms (SNPs) were then replicated in two independent French studies. This analysis lead to the discovery of ORM1 as a new gene participating to the control of TGP. Second, I followed a similar strategy using this time whole blood DNA methylation levels at CpG sites to identify DNA methylation marks involved in TGP variability. I analyzed the association between methylation-wide patterns from a French study and a French-Canadian families measured for TGP. Unfortunately, I did not identify robust associations between whole DNA methylation levels and thrombin generation.

Keywords: Thrombin Generation Potential, TGP, Meta-GWAS, ORM1, DNA Methylation, epidemiology.

Résumé

Le potentiel de génération thrombine (TGP en anglais) est une nouvelle mesure qui permet de quantifier *in vitro* l'activité globale de la thrombine reflétant bien les mécanismes *in vivo* de la coagulation. Ce méthode de dosage est sensible aux déficits de facteurs de coagulation, à la prise d'anti-coagulants et à de nombreux troubles de la coagulation. Au moment où j'ai débuté ma thèse, seuls deux polymorphismes génétiques, tous les deux situés dans le gène F2 codant pour la prothrombine, étaient connus pour influencer la variabilité plasmatique du TGP. Mon projet de thèse avait pour objectifs d'identifier de nouveaux facteurs génétiques, mais également épigénétiques, pouvant influencer les taux plasmatiques de TGP. Dans une première partie, j'ai mené la toute première étude d'association génome-entier (*GWAS* pour Genome Wide Association Study en anglais) sur 3 biomarqueurs (temps de latence, quantité totale de thrombine produite et niveau maximal de thrombine produite) du TGP dans deux études françaises rassemblant 1267 sujets et j'ai répliqué les résultats les plus significatifs dans deux autres études françaises indépendantes de 1344 sujets. Cette stratégie a permis de mettre en évidence qu'un polymorphisme génétique du gène ORM1 était associé de manière robuste au temps de latence, biomarqueur caractérisant le temps nécessaire pour initier la coagulation après induction. Dans la seconde partie de ma thèse, en suivant une stratégie similaire mais cette fois-ci en étudiant non plus des polymorphismes génétiques mais des marques de méthylation d'ADN, j'ai recherché si des niveaux de méthylation de site CpG, mesurés à partir d'ADN sanguin et couvrant l'ensemble du génome, pouvaient être associés à la variabilité des 3 mêmes biomarqueurs de TGP. Malheureusement, à partir de deux échantillons mis à ma disposition et rassemblant 425 sujets, je n'ai pas pu mettre en évidence d'association robuste entre des marques de méthylation sanguine et la génération trombine.

Mots-clés: Potentiel Génération Thrombine, TGP, Meta-GWAS, ORM1, Methylation, épidémiologie.
