Méthodologies statistiques pour l'analyse des déterminants génétiques de la maladie thromboembolique veineuse et de sa récidive

Gaëlle Munsch

université de BORDEAUX

THÈSE PRÉSENTÉE POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SOCIÉTÉS, POLITIQUE, SANTÉ PUBLIQUE

SPÉCIALITÉ Santé Publique – OPTION Biostatistiques

Méthodologies statistiques pour l'analyse des déterminants génétiques de la maladie thromboembolique veineuse et de sa récidive

Par Gaëlle MUNSCH

Sous la co-direction de David-Alexandre TRÉGOUËT et Hélène JACQMIN-GADDA

Soutenue publiquement le 2 mars 2023

Membres du jury :

M. BERTOLETTI Laurent	PU-PH, Université Jean Monnet Saint-Etienne	Président, Rapporteur
Mme. MAUCORT-BOULCH Delphine	PU-PH, Université Claude Bernard Lyon 1	Rapportrice
Mme. GÉNIN Emmanuelle	DR Inserm, UMR1078, Université de Bretagne Occidentale	Examinatrice
M. GUEDJ Mickaël	Directeur Biométrie & Science des données, Nanobiotix SA	Examinateur
Mme. JACQMIN-GADDA Hélène	DR Inserm, U1219, Université de Bordeaux	Co-directrice
M. TRÉGOUËT David-Alexandre	DR Inserm, U1219, Université de Bordeaux	Co-directeur

Résumé

Titre : Méthodologies statistiques pour l'analyse des déterminants génétiques de la maladie thromboembolique veineuse et de ses complications

Résumé : Dans le cadre des analyses pangénomiques, plusieurs millions de tests statistiques sont réalisés pour identifier des polymorphismes génétiques associés à un phénotype d'intérêt. Des logiciels bio-informatiques dédiés à l'analyse de ces données ont été développés permettant d'optimiser l'implémentation de modèles de régression linéaires ou logistiques. Cependant, ces logiciels ne permettent pas d'utiliser d'autres types de modélisation qui peuvent être nécessaires, notamment lorsque le phénotype d'intérêt ne présente pas une distribution gaussienne ou lorsque des évènements récurrents doivent être étudiés. Dans la première partie de cette thèse, j'ai proposé une stratégie de modélisation reposant sur un modèle de Cox pondéré permettant d'étudier des facteurs génétiques de récidive de la maladie thromboembolique veineuse (MTEV) dans l'étude MARTHA, en intégrant à la fois des récidives pré- et post-inclusion. La MTEV est une pathologie fréquente dont la prévalence dans les pays développés augmente avec le vieillissement de la population et qui représente la troisième cause de décès parmi les pathologies cardiovasculaires, après les accidents vasculaires cérébraux et l'infarctus du myocarde. La récidive de MTEV est elle aussi fréquente puisque près de 30 % des patients présentent une récidive dans les 10 ans suivant leur premier évènement. L'identification des facteurs prédictifs de la récidive de MTEV pourrait permettre, à terme, d'établir un score afin de pouvoir discriminer les patients selon leur risque de récidive et ainsi adapter la durée de leur traitement anticoagulant. Ce travail a permis de clarifier les associations entre certaines variables cliniques et génétiques avec le risque de récidive de MTEV et de souligner des différences existant entre la MTEV et sa récidive. De plus, la modélisation proposée pourra être utilisée pour réaliser l'analyse pangénomique des déterminants génétiques de la récidive de MTEV dans l'étude MARTHA. Dans la seconde partie de cette thèse, j'ai comparé les propriétés de différents modèles adaptés aux distributions semicontinues (caractérisées par un excès de valeurs en zéro suivi d'une distribution continue avec asymétrie à droite) et je les ai utilisés pour conduire des analyses pangénomiques des taux plasmatiques de *Neutrophil Extracellular Traps* (NETs), un biomarqueur émergeant dans les pathologies cardiovasculaires et impliqué dans l'immunothrombose. Cette thèse souligne l'importance de choisir une modélisation adaptée à la variable d'intérêt permettant d'étudier toute l'information disponible et ce malgré la complexité de mise en œuvre et les importants temps de calcul dus au manque de logiciels d'analyses bio-informatiques.

Mots clés : Analyse de données génétiques ; modélisation statistique ; évènements récurrents ; distribution semicontinue ; maladie thromboembolique veineuse.

Unité de recherche : Bordeaux Population Health Research Center, Inserm U1219. Equipes Biostatistiques et Eleanor (*Molecular epidemiology of vascular and brain disorders*).
146 rue Léo Saignat, 33076 Bordeaux, France.

Abstract

Title: Statistical methodologies for the analysis of genetic determinants of venous thromboembolic disease and its complications

Abstract: In the context of genome-wide association studies (GWAS), several millions of statistical tests are performed to identify genetic polymorphisms associated with a phenotype of interest. Bioinformatics softwares dedicated to the analysis of these data have been developed to optimize the implementation of linear or logistic regression models. However, these softwares do not allow the use of other types of modeling that may be necessary, especially when the phenotype of interest does not have a Gaussian distribution or when recurrent events are studied. In the first part of this thesis, I proposed a modeling strategy based on a weighted Cox model to study the genetic risk factors of venous thromboembolic disease (VTE) recurrence in the MARTHA study, incorporating both pre- and post-inclusion recurrences. VTE is a frequent pathology whose prevalence in developed countries is increasing with population aging and which represents the third cause of death among cardiovascular pathologies, after stroke and myocardial infarction. Recurrence of VTE is also frequent, with nearly 30% of patients having a recurrence within 10 years of their first event. The identification of predictive factors of VTE recurrence could eventually allow the establishment of a score to discriminate patients according to their risk of recurrence and thus adapt the duration of their anticoagulant treatment. This work allowed to clarify the associations between some clinical and genetic variables with the risk of VTE recurrence and highlighted the differences between VTE and its recurrence. In addition, the proposed modeling can be used to perform genome-wide analysis of genetic determinants of VTE recurrence in the MARTHA study. In the second part of this thesis, I compared models adapted to semicontinuous distributions (characterized by an excess of values in zero followed by a continuous distribution with right skewness) and I used them to perform GWAS on the plasma levels of Neutrophil Extracellular Traps (NETs), an emerging biomarker in cardiovascular pathologies and involved in immunothrombosis. This thesis underlines the importance of selecting a model adapted to the variable of interest that allows maximizing the use of the available information despite the complexity of implementation and the important computation time due to the lack of bioinformatics analysis software.

Keywords: Genetic data analysis; statistical modeling; recurrent events; semicontinuous distribution; venous thromboembolic disease.

Remerciements

A David-Alexandre Trégouët, mon directeur de thèse,

Merci David pour ton soutien et pour la confiance que tu m'as accordée depuis mon stage de master. Bientôt quatre ans que nous travaillons ensemble et j'ai appris énormément de toi, au niveau scientifique mais pas seulement. Désormais je me tire une bûche lorsque je viens dans ton bureau, j'aime lorsqu'il y a de la sérendipité et je m'émoustille devant mon ordinateur. Je te remercie également pour toutes les opportunités que tu m'as offertes et de m'avoir emmenée avec toi « en touriste » aux congrès. Tu n'es peut-être pas un grand fan de l'enseignement mais je trouve que tu as été très pédagogue avec moi (même si ce n'était pas toujours facile en zoom – merci covid) et que tu as su stimuler ma curiosité scientifique. Je t'en suis très reconnaissante. Merci+++++

A Hélène Jacqmin-Gadda, ma co-directrice de thèse,

Merci Hélène pour votre bienveillance tout au long de ces trois années de thèse. Vous m'avez énormément appris notamment en termes de rigueur statistique, et je suis très fière d'avoir réalisé ma thèse à vos côtés. J'espère que notre collaboration continuera car il reste énormément de choses à faire dans la continuité de cette thèse et j'ai encore beaucoup à apprendre.

Aux membres de mon jury de thèse,

Je remercie tout d'abord les professeurs **Delphine Maucort-Boulch** et **Laurent Bertoletti** pour avoir accepté de relire ce travail. Merci également à **Emmanuelle Génin** et **Mickaël Guedj** qui ont accepté de participer à ce jury en tant qu'examinateurs.

Aux membres de mon comité de suivi de thèse,

Je tiens à remercier **Cécile Proust-Lima** et **Thierry Couffinhal** qui ont participé à mon comité de suivi de thèse. Je vous remercie pour votre bienveillance ainsi que tous les conseils et avis que vous avez pu me donner au cours de nos réunions annuelles.

Aux membres de l'équipe ELEANOR,

Je remercie tous les membres de l'équipe ELEANOR (anciennement VINTAGE) avec qui j'ai pu passer de bons moments durant ces trois ans. Je remercie **Stéphanie Debette** et **Cécilia Samieri** pour leurs précieuses remarques et conseils lors de nos réunions d'équipe. Une pensée particulière à ceux qui ont quitté l'équipe : **Dylan** alias « chouchou » qui m'avait dit que la thèse donnait des cheveux blancs (je confirme !), **Misbah** mon ancienne collègue de bureau qui a appris qu'en France il fallait râler (merci Dylan…), et **Florian** qui m'a précédée en tant que thésard de David et que j'ai eu l'occasion de croiser à quelques reprises.

Merci aux membres du « groupe David », notamment **Carole**, **Omar**, **Charlène**, **Surya**, **Blandine**, ainsi que **Marine** et **Caroline** avec qui j'ai le plaisir de partager mon bureau actuel. Enfin, je remercie **Nathalie** qui est toujours disponible pour répondre aux questions administratives.

Aux collaborateurs marseillais et brestois,

Je tenais à remercier également les collaborateurs de Marseille (**Pierre-Emmanuel Morange**, **Manal Ibrahim-Kosta** et **Louisa Goumidi**) et de Brest (**Francis Couturaud** et **Lénaïck Gourhant**) avec qui j'ai eu de nombreux échanges tout au long de ces trois années.

A l'ensemble des doctorants et post-doctorants du BPH,

Je remercie tous les doctorants et post-doctorants du centre avec qui j'ai pu échanger durant ces trois ans. Je remercie en particulier **Anthony**, mon binôme de projets durant le master 2, avec qui j'ai partagé de nombreuses pauses café. Je te souhaite tout le meilleur pour ton aventure en Australie. Je remercie également **Kateline** et **Tiphaine**, avec qui j'ai partagé quelques jours à Marseille. Merci aussi à **Bénédicte**, pour ta gentillesse et tes conseils. J'espère qu'on pourra se revoir bientôt à Montréal.

A toutes les personnes qui ont contribué au bon déroulement de cette thèse, je remercie en particulier **Emmanuelle Génin** et **Gaëlle Marenne** qui m'ont fait découvrir la statistique génétique à Brest lors de mon stage de master.

A mes amis, et tout particulièrement à **Maïwenn**. Merci pour ton amitié de longue date et le plaisir de se retrouver à chacun de mes retours en Bretagne.

A ma famille,

Pour finir, je tenais à remercier **mes parents** pour m'avoir soutenue depuis toutes ces années pour arriver à la fin de cette thèse. A mes petits frères (**Guillaume** et **Baptiste**), qui désormais sont grands et ont quitté le cocon familial. Je remercie également mes **grands-parents**, **oncles**, **tantes**, **cousins**, **cousines**, que j'apprécie toujours autant de retrouver l'été au terrain de Kerlouan, pour les fêtes de fin d'année ou du côté de Landrezac.

Enfin, je remercie ma **belle-famille** pour m'avoir fait découvrir votre petit coin de la Pointe de Trévignon et surtout merci à **Quentin** avec qui je partage mon quotidien depuis quelques années maintenant. Partager cette aventure qu'est la thèse à deux l'a rendue plus facile à vivre et elle n'aurait pas eu la même saveur si tu n'avais pas été là pour me remonter le moral dans les moments difficiles. Un grand merci.

Table des matières

Valorisation scientifique

Articles de la thèse

Munsch G, Proust C, Labrouche-Colomer S, […], Smadja M D, Jacqmin-Gadda H*, Trégouët D-A*. *Genome-wide association study of a semicontinuous trait: Illustration of the impact of the modeling strategy through the study of neutrophil extracellular traps levels.* **medRxiv**. 2022 Jan 1;2022.09.19.22279929

Cet article est présenté en *section **6.1*** et s'accompagne de sept figures, neuf tableaux et un texte supplémentaire.

Munsch G, Goumidi L, van Hylckama Vlieg A, […], Jacqmin-Gadda H, Morange P-E*, Trégouët D-A*. *Modelling of time-to-events in an ambispective study: illustration with the analysis of ABO blood groups on venous thrombosis recurrence*. **medRxiv**. 2021 Jan 1;2021.11.20.21266583

Cet article est présenté en *section **5.1*** et s'accompagne de sept figures et de cinq tableaux.

Abstracts publiés

Munsch G, Ibrahim-Kosta M, Gouimidi L, Morange P, Trégouët D, Jacqmin-Gadda H. *Modélisation rétro-prospective d'évènements récurrents*. **Rev DÉpidémiologie Santé Publique**. 2021 Jun 1;69:S18

Munsch G, Aïssi D, James C, Jacqmin-Gadda H, Morange P-E, Deleuze J-F, Trégouët D-A, Smadja M D. *Genome wide association analysis of neutrophils extracellular traps*. **Research and Practice in Thrombosis and Haemostasis** ; 5(SUPPL 2), 2021.

Autres articles sur la thématique

Sanchez-Rivera L†, Iglesias MJ†, Ibrahim-Kosta M, […], **Munsch G**, […], Butler LM*, Trégouët D-A*, Odeberg J*. *Elevated plasma Complement Factor H Regulating Protein 5 is associated with venous thromboembolism and COVID-19 severity*. **medRxiv**. 2022 Jan 1;2022.04.20.22274046

Razzaq M, Goumidi L, Iglesias MJ, **Munsch G**, […], Odeberg J, Morange P-E, Trégouët D-A. *Explainable Artificial Neural Network for Recurrent Venous Thromboembolism Based on Plasma Proteomics*. In: Cinquemani E, Paulevé L, editors. **Computational Methods in Systems Biology**. Cham: Springer International Publishing; 2021. p. 108–21. (Lecture Notes in Computer Science).

Thibord F, **Munsch G**, Perret C, […], Deleuze J-F, Morange P-E*, Trégouët D-A*. *Bayesian network analysis of plasma microRNA sequencing data in patients with venous thrombosis.* **Eur Heart J Suppl**. 2020 Apr 1;22(Supplement_C):C34–45.

Communications orales

European Mathematical Genetics Meeting (EMGM) – **avril 2021** – en ligne

Retro-prospective modelling of recurrent events

Conférence francophone d'EPIdémiologie CLInique (EPICLIN) – **juin 2021** – Marseille

Modélisation Rétrospective d'évènements récurrents

International Society on Thrombosis and Haemostasis conference (ISTH) – **juillet 2021** – en ligne

Genome Wide Association Analysis of Neutrophil Extracellular Traps (NETs)

International Society For Clinical Biostatistics conference (ISCB) – juillet 2021 – en ligne

Ambispective modelling of recurrent events

Programme transversal de l'Inserm "Genomics variability in health & Disease" **(GOLD) – octobre 2022 –** en ligne

Genome-wide association study of a semicontinuous trait: Illustration of the impact of the modeling strategy through the study of Neutrophil Extracellular Traps levels

Activités annexes

Co-encadrement de stagiaires avec David-Alexandre Trégouët

Avril – Juin 2021 :

Jade Dupin (Master 1 Santé Publique – ISPED Bordeaux) :
Impact du syndrome métabolique sur le risque de récidive de la maladie thromboembolique veineuse : Approche par score de risque génétique.

Perrine Lafaye (Ecole d'Ingénieurs Informatique et Systèmes d'Information pour la Santé – ISIS Castres)
Identification de polymorphismes génétiques associés à des taux plasmatiques par une approche de régression pénalisée.

Bertille Ségier (Master 1 Santé Publique – ISPED Bordeaux) :
Recherche de facteurs de risque génétiques de la maladie thromboembolique veineuse dans une population de femmes sous contraception orale.

Avril – Juin 2022 :

Hugo Motyl (Master 1 Santé Publique – ISPED Bordeaux) :
Identification de polymorphismes génétiques associés au risque d'embolie pulmonaire chez des patients atteints d'une maladie thromboembolique veineuse : Apport des forêts aléatoires

Floriane Samaria (Master 1 Santé Publique – ISPED Bordeaux) :
Identification de facteurs de risque génétiques associés aux séquelles perfusionnelles à la suite d'une embolie pulmonaire.

Activités d'enseignement

Participation à la mise en place d'une Unité d'Enseignement dans les Masters 2 Santé Publique parcours Biostatistiques et Epidémiologie à l'ISPED (Bordeaux) : *Analyse de données génomiques*

Participation à la création d'un module d'enseignement dans l'Ecole d'Eté de l'ISPED : *Basics in Genetics Epidemiology (BIGGY)*

2020-2021 : 40h d'enseignement en statistiques (Travaux pratiques-dirigés, cours magistraux) aux étudiants en Licence de Psychologie et aux étudiants de l'Ecole de Sage-Femmes.

2021-2022 : 32h d'enseignement en statistiques (Travaux pratiques-dirigés, cours magistraux) aux étudiants du Master Santé Publique de l'Isped et aux étudiants de l'Ecole de Sage-Femmes.

2022-2023 : 14h d'enseignement en statistique génétique (Travaux Pratiques et Cours Magistraux) dans l'unité d'enseignement et le module mis en place.

Liste des abréviations

ADN Acide DésoxyriboNucléique

ARN Acide RiboNucléique

EP Embolie Pulmonaire

GRS *Genetic Risk Score*

GWAS *Genome Wide Association Study*

HR *Hazard Ratio*

IMC Indice de Masse Corporelle

MTEV Maladie ThromboEmbolique Veineuse

NETs *Neutrophil Extracellular Traps*

RC *Rapport de Cotes*

RR *Risque Relatif*

SNP *Single Nucleotide Polymorphim*

TVP Thrombose Veineuse Profonde

Liste des tableaux de la thèse (hors articles)

Liste des figures de la thèse (hors articles)

1 La maladie thromboembolique veineuse

La maladie thromboembolique veineuse (MTEV) regroupe à la fois la thrombose veineuse profonde (TVP), appelée plus communément phlébite, et sa complication immédiate l'embolie pulmonaire (EP). La MTEV résulte de la formation d'un caillot de sang (ou thrombus) dans une veine entraînant un ralentissement ou une stagnation du flux sanguin (*voir section **1.4***). Dans le cadre de la TVP, le caillot se forme généralement dans une veine des membres inférieurs. Lorsque ce caillot se détache, partiellement ou dans sa totalité, de la paroi du vaisseau, il peut alors migrer via la circulation sanguine dans une artère pulmonaire pour l'obstruer, et être responsable d'une EP (**Figure 1**).

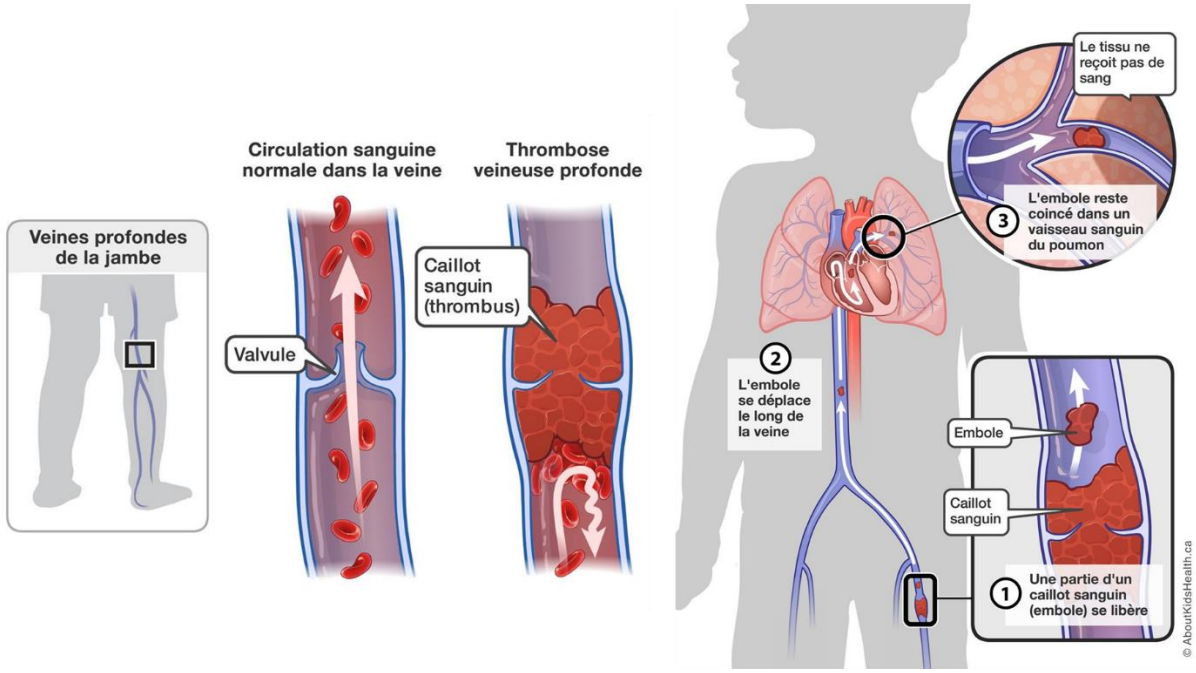


Figure 1 : Illustration de la formation d'une thrombose veineuse profonde pouvant évoluer en embolie pulmonaire *(www.aboutkidshealth.ca)*

1.1 Epidémiologie

La MTEV est une pathologie fréquente dans les pays développés puisque son incidence est estimée à 1-2 cas pour 1 000 personnes-années (Delluc et al., 2016). L'incidence de la MTEV est plus élevée en Europe, aux Etats-Unis ainsi que dans les populations d'origine Africaine, alors qu'elle reste relativement faible dans les pays d'Asie où l'incidence est estimée à 0,8 cas pour 1 000 personnes-années (Lutsey & Zakai, 2022; Zakai & McCLURE, 2011). En France, cette incidence correspond à 40 000 EP et 50-100 000 TVP qui surviennent chaque

année (Inserm, La science pour la santé, 2017). Chez les personnes de plus de 70 ans la MTEV est plus fréquente, comme le montre la **Figure 2**, et touche près de 5 cas pour 1 000 personnes-années, ce qui en fait un réel enjeu de santé publique avec le vieillissement de la population (Duthé et al., 2021). Par ailleurs avec le vieillissement de la population, le nombre de cas de MTEV aux Etats-Unis devrait doubler d'ici 2050 et concerner ainsi plus de 1,8 millions de personnes (Deitelzweig et al., 2011). En France en 2010, plus de 50 000 personnes ont été hospitalisées pour une MTEV. Ce chiffre passe à 120 000 lorsque les hospitalisations pour lesquelles la MTEV n'était pas la cause principale d'hospitalisation sont comptabilisées. Près de la moitié de ces évènements de MTEV sont des EP et cette complication majeure de la thrombose veineuse est responsable de plus de 12 000 décès chaque année en France. Le taux de mortalité à 6 mois de l'EP est estimé à près de 20 % (Lutsey & Zakai, 2022). Ces éléments font de la MTEV (principalement sous la forme EP) la troisième cause de mortalité cardiovasculaire après les accidents vasculaires cérébraux et l'infarctus du myocarde (Waheed et al., 2022).

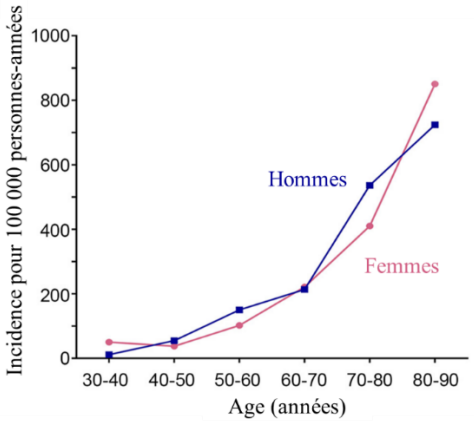


Figure 2 : Evolution de l'incidence de la MTEV au cours de la vie adulte séparément chez les hommes et les femmes (adapté de (Arshad et al., 2017))

La MTEV est une maladie multifactorielle résultant de l'effet de facteurs environnementaux et/ou génétiques. Environ 50 % des cas de MTEV surviennent en présence de facteurs de risque exogènes ou environnementaux, comme les immobilisations prolongées, les interventions chirurgicales ou les cancers. On parle alors de MTEV provoquée. Les 50 % restants sont appelés MTEV non provoquées parmi lesquelles on retrouve les thrombophilies familiales qui se caractérisent par une tendance génétiquement déterminée à développer une MTEV. Ces thrombophilies peuvent être considérées comme des anomalies congénitales de l'hémostase incluant des anomalies héréditaires comme la mutation du Facteur V Leiden, la

mutation du Facteur II G20210A, et les déficits en antithrombine, protéine C et protéine S (Bauer, 2003; Roldan et al., 2009). En population générale, les fréquences des mutations du Facteur V Leiden et du Facteur II G20210A sont de l'ordre de 4 % et 2 %, respectivement. En ce qui concerne les déficits en inhibiteurs naturels de la coagulation, leur fréquence est de 0,2 % pour les protéines S et C, et de l'ordre de 0,02 % pour l'antithrombine (Jadaon, 2011; Obaid et al., 2020). Les thrombophilies représentent moins de 50 % des cas de MTEV non provoquées, ce qui laisse à penser que d'autres facteurs encore non identifiés contribueraient à expliquer plus de la moitié des cas de MTEV non provoquées (Kreidy, 2015; Obaid et al., 2020). Certaines études utilisent ce terme (thrombophilie) pour caractériser uniquement les MTEV survenant chez un individu présentant une histoire familiale de MTEV, et donc compatibles avec l'hypothèse d'un (ou de quelques) variations génétiques avec des effets assez forts (comme l'un des facteurs majeurs cités ci-dessus). D'autres l'utilisent pour faire référence à toute forme de MTEV non provoquée résultant de l'effet combiné de plusieurs polymorphismes associés à des augmentations de risque relativement modestes. Cette nuance peut donc être à l'origine d'une surestimation de la proportion de thrombophilies parmi les cas de MTEV non provoquées.

1.2 Prise en charge et complications

Après un évènement de MTEV, un traitement anticoagulant est administré au patient afin de fluidifier sa circulation sanguine et de réduire son risque immédiat de récidive et de décès. Néanmoins, ce traitement expose le patient à un risque accru d'hémorragie avec une incidence estimée à 7 cas pour 100 personnes-années et qui serait fatale dans environ 9 % des cas (Klok & Huisman, 2020). A ce jour, il n'existe pas d'outil validé permettant de déterminer les profils de patients les plus à risque d'hémorragie.

La durée du traitement anticoagulant administré est déterminée en tenant compte du caractère provoqué ou non de la MTEV. Ainsi, lorsque la MTEV est provoquée par un facteur de risque transitoire (comme une grossesse ou une immobilisation temporaire, *voir section **1.3***), le traitement est administré pour 3 mois alors qu'en présence d'un facteur de risque persistant (comme le cancer) la durée est étendue à 6 mois voire à vie (Kearon et al., 2016). Pour ce qui est des MTEV non provoquées, les recommandations internationales préconisent de traiter à vie ces patients (Authors/Task Force Members et al., 2014; Delluc et al., 2016). Le rôle du traitement est de faciliter la dissolution du caillot sanguin (*voir section **1.4.1***) mais également de prévenir la récidive de MTEV.

En effet, après un premier épisode thromboembolique veineux, la récidive est fréquente et est responsable d'une morbidité et d'une mortalité importantes. Il est estimé que dans les 10 années suivant l'arrêt du traitement pour une première MTEV, 30 % des patients présentent une récidive (Lutsey & Zakai, 2022). Il est admis que certains facteurs de risque associés au premier épisode de MTEV, comme le sexe masculin ou le cancer (*voir section **1.3***), sont également associés au risque de récidive (Authors/Task Force Members et al., 2014). En revanche, aucun consensus n'est établi en ce qui concerne le lien entre risque de récidive et divers facteurs tels que l'âge ou le type de MTEV (à savoir TVP ou EP). Certaines études ont tout de même montré que les patients ayant eu un premier événement de MTEV non provoqué sont plus à risque de récidiver que les autres (Prandoni et al., 2007). Même si quelques biomarqueurs, comme le dosage des D-dimères (molécules issues de la dégradation du caillot de fibrine), ont été proposés comme facteurs prédictifs du risque de récidive de MTEV, à l'heure actuelle le profil des sujets les plus à risque de récidive n'a pas été clairement établi (Avnery et al., 2020). Certains scores clinico-biologiques (HERDOO, DASH, VIENNA) ont été développés afin de prédire le risque de récidive mais ils sont généralement spécifiques à certains sous-groupes d'individus (ex : HERDOO concerne uniquement les femmes de plus de 18 ans avec MTEV non provoquée) et leurs résultats manquent de validation dans des échantillons indépendants ce qui explique pourquoi leurs utilisations en pratique clinique restent débattues (Rodger et al., 2008; Tosetto et al., 2012; Eichinger et al., 2010; Houghton & Moll, 2017). Les éléments communs dans ces scores qui augmentent le risque de récidive de MTEV sont le sexe masculin et des taux élevés de D-dimères suite à l'arrêt du traitement anticoagulant. Les D-dimères représentent le produit de la dégradation de la fibrine mais sont peu spécifiques à la MTEV puisque des valeurs élevées peuvent également être observées en cas d'infection ou de fibrillation auriculaire (Siegbahn et al., 2016).

Il existe également d'autres complications survenant à la suite d'une MTEV. Le syndrome post-thrombotique apparaît chez 20-50 % des cas de TVP, et il est sévère dans 5 à 10 % des cas. Il se caractérise par une insuffisance veineuse chronique qui peut entraîner des ulcères veineux chez environ 4 % des patients (Goldhaber, 2012; Lutsey & Zakai, 2022; Winter et al., 2017). Après une EP, près de 4 % des patients développent une hypertension pulmonaire thromboembolique chronique qui est causée par la dissolution incomplète du caillot qui continue à obstruer partiellement les artères pulmonaires. On parle alors d'obstruction vasculaire pulmonaire résiduelle (Picart et al., 2020; Tromeur et al., 2018).

1.3 Facteurs de risque

Comme mentionné précédemment, la MTEV est une maladie multifactorielle résultant de l'effet combiné de facteurs environnementaux et génétiques. Les principaux facteurs de risque environnementaux sont présentés dans le **Tableau 1** ci-dessous.

Tableau 1 : Facteurs de risque environnementaux de la MTEV

Démographiques	Age avancé
	Sexe masculin*
Anthropométriques	Taille élevée
	IMC (>30kg/m^2)
Comportementaux	Tabagisme
	Manque d'activité physique
	Alimentation non équilibrée
Transitoires	Immobilisation (chirurgie, voyage, ...)
	Etat hormonal (contraceptif oral, traitement hormonal substitutif, grossesse, post-partum)
	Infection aiguë (hépatite)
Cardiovasculaires	Hypertension
	Diabète
	Dyslipidémie
	Comorbidités (infarctus du myocarde, fibrillation auriculaire, ...)
Acquis	Cancer
	Maladies auto-immunes (syndrome des anti-phospholipides)
	Maladies inflammatoires chroniques
	Syndrome de May-Thurner
	Antécédents personnels de MTEV

**varie selon l'âge*

Comme évoqué précédemment, l'âge avancé est un facteur bien établi du risque de MTEV. Une étude récente, conduite dans plusieurs cohortes de population générale comptabilisant plus d'un million de participants et plus de 3 000 cas de MTEV, a estimé que le risque de MTEV serait doublé tous les 10 ans (Gregson, Kaptoge, Bolton, Pennells, Emerging Risk Factors Collaboration, et al., 2019). Les explications les plus souvent avancées à ces observations sont que les taux des facteurs de la coagulation, principalement produits par le foie, augmentent avec l'âge et qu'un âge avancé s'accompagne généralement d'une santé plus fragile et d'une vie plus sédentaire contribuant au surpoids et à une augmentation de la stase veineuse (*voir section **1.4.1***).

La taille élevée d'un individu, et en particulier la longueur de ses jambes, augmenterait le risque MTEV. Les hypothèses biologiques sous-jacentes sont que la plus grande surface veineuse, la présence de plus de valvules et une plus forte pression hydrostatique pourraient augmenter le risque de MTEV (Lutsey et al., 2011).

Le sexe masculin est également un facteur de risque connu de MTEV, néanmoins cette association n'est pas homogène tout au long de la vie puisqu'elle s'inverse dans la tranche d'âge 20-45 ans (**Figure 2**), période durant laquelle les femmes sont en âge de procréer. Durant cette période, ces femmes peuvent prendre des contraceptifs à base d'œstrogènes voire des traitements hormonaux substitutifs, des traitements contre l'infertilité ou être enceintes. Tous ces facteurs sont responsables d'une augmentation du risque de MTEV. Par exemple, la grossesse est un état pro-coagulant qui est associé à un taux d'hormones élevé et une augmentation de la stase veineuse ce qui augmente le risque de MTEV avant l'accouchement mais aussi durant la période de post-partum. En France, la MTEV complique plus d'une grossesse sur 1 000 et représente la deuxième cause directe de mortalité maternelle après les hémorragies (Olié et al., 2015).

Les évènements conduisant à l'immobilisation sont un autre facteur de risque important de MTEV. En effet, les opérations, fractures osseuses qui immobilisent les membres, et les voyages de longue durée favorisent la stase veineuse (Chalal & Demmouche, 2013; Domeij-Arverud et al., 2015). Les chirurgies nécessitant la pose d'un cathéter veineux central permettant notamment de surveiller la pression veineuse centrale augmenteraient le risque de MTEV et particulièrement d'EP (Wang et al., 2020). De la même façon, le manque d'activité physique diminuerait le débit de circulation sanguine et augmenterait les taux des facteurs de la coagulation. Le manque d'activité physique influence les marqueurs d'adiposité comme le tour de taille ainsi que l'indice de masse corporelle (IMC), pour lesquels des valeurs élevées sont également associées à l'augmentation de la stase veineuse et des concentrations des biomarqueurs hémostatiques et inflammatoires, ainsi qu'à d'autres pathologies qui augmentent elles-mêmes le risque de MTEV (Rahmani et al., 2020). Parmi elles, on retrouve notamment les maladies rénales, les infarctus du myocarde et les fibrillations auriculaires (Christiansen et al., 2014; Lutsey et al., 2018).

Les associations entre certains facteurs de risque de maladies cardiovasculaires tels que l'hypertension, le diabète ou encore l'hyperlipidémie, et la MTEV ont été débattues pendant longtemps. Ils sembleraient tout de même augmenter le risque de MTEV indépendamment du sexe et de l'âge (Ayed Alanazi et al., 2017; García Raso et al., 2014; Jiao et al., 2022). De la même façon, pour l'effet du tabagisme actif sur le risque de MTEV, les résultats ont longtemps

été inconsistants entre les études. Néanmoins, des analyses prospectives menées en population générale ont permis de montrer que le tabagisme actif est associé à une augmentation d'environ 50 % du risque de MTEV (Enga et al., 2012). Concernant la consommation d'alcool, il semblerait qu'une consommation régulière soit protectrice du risque de MTEV (Gregson, Kaptoge, Bolton, Pennells, Emerging Risk Factors Collaboration, et al., 2019). Cependant, la consommation d'alcool est souvent évaluée à l'aide d'auto-questionnaires et l'exactitude des renseignements peut être remise en cause. Pour autant, au niveau physiopathologique l'alcool altère le foie ce qui l'empêche de synthétiser les facteurs de la coagulation et diminue donc le risque de MTEV mais expose l'individu à un risque plus élevé d'hémorragie. Quant aux profils nutritionnels qui sont parfois difficiles à évaluer, peu d'éléments ont été mis en évidence à l'exception de quelques études montrant que la supplémentation en vitamine E serait protectrice de MTEV et qu'un déficit en vitamine B serait un facteur de risque de MTEV (Glynn et al., 2007; Oger et al., 2006).

Les maladies inflammatoires chroniques et les infections aiguës sont également des facteurs de risque de MTEV. En effet, plusieurs études ont mis en évidence une association entre des taux élevés de protéine C-réactive, qui est une protéine sécrétée par le foie en réponse à une infection ou inflammation de l'organisme, et le risque de MTEV (Gregson, Kaptoge, Bolton, Pennells, for the Emerging Risk Factors Collaboration, et al., 2019; Grimnes et al., 2018). Les infections virales telles que les grippes (en particulier le virus H1N1) ou plus récemment le covid-19, sont également des facteurs de risque de MTEV et surtout d'EP, notamment lorsque les symptômes sont sévères et nécessitent un alitement du patient, favorisant ainsi la stase veineuse (Bunce et al., 2011). Durant la pandémie du covid-19, près d'un tiers des patients hospitalisés pour ce virus ont développé une MTEV. Les Rapports de Cotes (RC) de MTEV estimés étaient de l'ordre de 3 et 6 pour les patients hospitalisés avec des symptômes non sévères et sévères, respectivement (Angelini et al., 2022).

Les maladies auto-immunes, comme le syndrome des anti-phospholipides qui engendre une altération des phospholipides qui protègent la paroi des vaisseaux sanguins, sont également considérées comme des facteurs de risque majeurs de MTEV (Farmer-Boatwright & Roubey, 2009). D'autres pathologies vasculaires plus rares sont aussi connues comme étant des facteurs de risque importants de MTEV, comme par exemple le syndrome de May-Thurner qui provoque une compression par l'artère iliaque droite de la veine iliaque gauche contre la colonne lombaire ce qui entraîne une insuffisance veineuse (Harbin & Lutsey, 2020).

Le cancer est un facteur de risque majeur de MTEV et, après le cancer lui-même, l'EP est la deuxième cause de mortalité chez les patients souffrant d'un cancer. L'incidence annuelle

de la MTEV chez ces patients varie de 0,5 à 20 % selon le type de cancer et les traitements associés (Lutsey & Zakai, 2022). Dans certains cas, c'est la survenue de la MTEV qui permet de détecter le cancer sous-jacent.

Enfin, les antécédents familiaux de MTEV sont également un facteur de risque de MTEV ce qui peut refléter la présence de facteurs génétiques dans une famille (Bezemer et al., 2009). En effet, le caractère multifactoriel de la MTEV se traduit par son héritabilité qui est estimée entre 40 et 60 % dans les études familiales (Baylis et al., 2021). Par ailleurs, le risque de MTEV d'un individu est doublé lorsque l'un de ses frères ou sœurs a déjà présenté une MTEV, ce qui montre l'importance des facteurs génétiques dans la survenue de cette pathologie (Zöller et al., 2011).

Les principaux facteurs de risque ou de susceptibilité génétique de la MTEV sont résumés dans la **Figure 3** ci-dessous. Cette figure représente les découvertes des gènes impliqués dans la MTEV depuis 1965, selon le RC de MTEV associé aux allèles à risque identifiés au sein de ces gènes. Les premières études comptabilisaient généralement peu de sujets, mais suffisamment pour détecter des effets très forts (RC > 5). L'augmentation du nombre de sujets étudiés dans les études d'associations pangénomiques (*sur lesquelles je reviendrai en section **3.4***) et leurs méta-analyses a permis d'identifier des facteurs de susceptibilité génétique assez fréquents ayant des effets plus faibles sur la MTEV (RC = 1,05). Ces observations sont valables pour d'autres pathologies complexes comme la maladie d'Alzheimer ou les accidents vasculaires cérébraux (Bellenguez et al., 2022; Mishra et al., 2022).

Les premières études visant à identifier des facteurs de risque génétiques de la MTEV remontent aux années 1960 et reposaient sur des analyses de liaison génétique dans des familles présentant plusieurs apparentés atteints. Ces études ont permis l'identification de variations génétiques extrêmement rares avec des effets très forts dans les gènes *SERPINC1*, *PROC* et *PROS1*, qui sont responsables de formes familiales de MTEV via des déficits des principaux inhibiteurs naturels de la coagulation que sont respectivement l'antithrombine, la protéine C et la protéine S (Bucciarelli et al., 2012; Comp & Esmon, 1984; Egeberg, 1965; Griffin et al., 1981). En 1969, le gène *ABO* du groupe sanguin a pour la première fois été proposé comme impliqué dans le risque de MTEV, les individus de groupe sanguin non O (A, AB, B) étaient plus à risque de MTEV que les autres (Jick et al., 1969). Par ailleurs, de récents travaux ont permis d'affiner ces observations en précisant les risques associés aux différents haplotypes A1, A2 et B, et en montrant qu'au sein d'un même groupe sanguin de type A, B et AB, il existait également une grande variabilité individuelle du risque de MTEV (Goumidi et al., 2020).

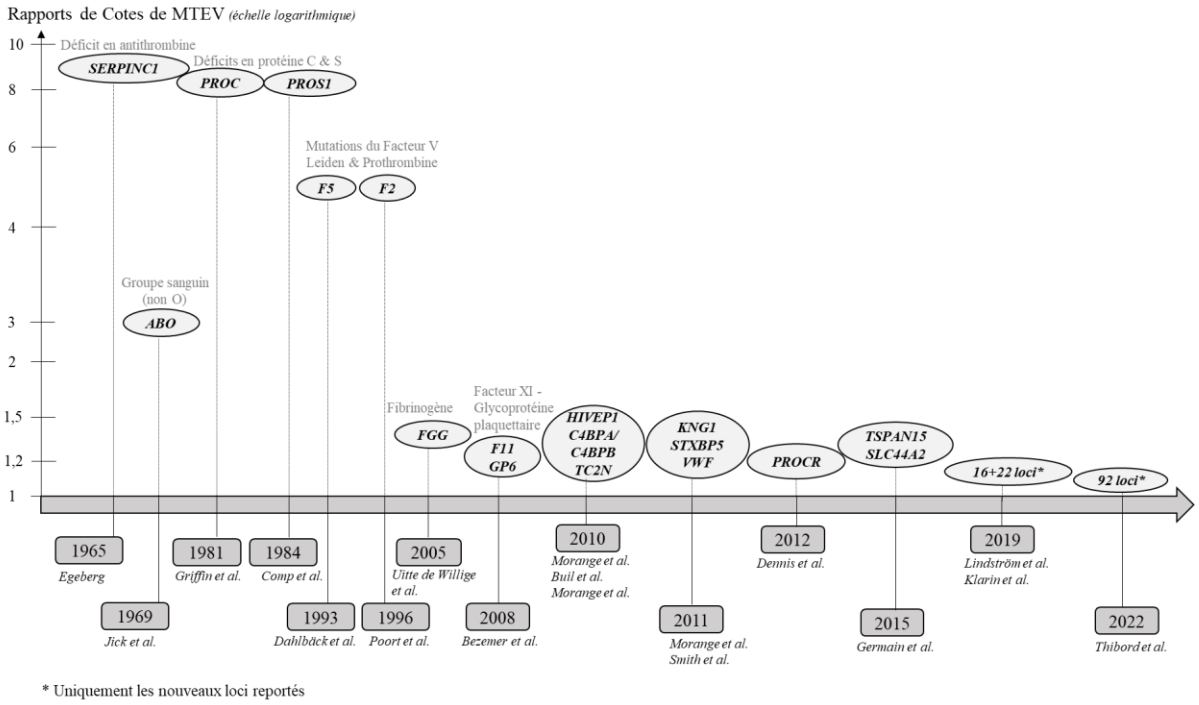


Figure 3 : Représentation historique de la découverte des facteurs de risque génétiques de la MTEV

Dans les années 1990, les stratégies de recherche des facteurs de risque génétiques se sont orientées vers l'approche dite « gènes-candidats » dont je reparlerai en *section **3.3***. Elles ont permis d'identifier des variants génétiques au sein du gène du *F5* (FV R506Q, communément appelé la mutation du FV Leiden), du gène *F2* (FII G20201A, mutation du gène de la prothrombine) et du gène *FGG* (codant pour une composante du fibrinogène) (Dahlbäck et al., 1993; Poort et al., 1996; Uitte de Willige et al., 2005). A noter que les individus homozygotes pour la mutation du FV Leiden ou du FII G20201A sont particulièrement à risque de MTEV avec des RC de MTEV passant de 5 à 10. A ce jour, les recherches des déficits en antithrombine, protéines C, S et l'homozygotie pour le FV Leiden ou le FII G20201A, constituent le bilan de thrombophilie couramment réalisé en clinique lors du diagnostic de MTEV.

A partir de 2008, la recherche de nouveaux variants génétiques associés au risque de MTEV s'est réalisée à partir d'approches d'associations pangénomiques (*Genome Wide Association Study - GWAS*) que je décrirai en *section **3.4*** (Bezemer et al., 2008; Buil et al., 2010; Dennis et al., 2012; Germain et al., 2015; Morange et al., 2010; Morange, Oudot-Mellakh, et al., 2011; Morange, Saut, et al., 2011; Smith et al., 2011). La première étude de grande ampleur reposant sur ce type de stratégie a été réalisée par *Germain et al.* et analysait plus de

7 500 cas de MTEV et 52 600 témoins (Germain et al., 2015). D'autres méta-analyses sur la MTEV ont ensuite été conduites par *Klarin et al.*, *Lindström et al.* et en 2022, la plus grande méta-analyse menée à ce jour sur la MTEV et réalisée par *Thibord et al.* recensait plus de 80 000 cas de MTEV (Klarin et al., 2019; Lindström et al., 2019; Thibord et al., 2022). Cette méta-analyse co-coordonnée par l'équipe ELEANOR du centre Inserm U1219 de Bordeaux et soutenue par le consortium *International Network against VENous Thrombosis (INVENT),* a permis d'identifier près d'une centaine de nouveaux loci génétiques (régions génomiques) associés à la MTEV. Il existe désormais plus de 150 loci présentant des polymorphismes génétiques associés au risque de MTEV dont la plupart sont impliqués dans les mécanismes ou se situent dans des gènes de la coagulation, processus intervenant dans la formation du caillot sanguin. Par ailleurs, certains loci génétiques ont également été retrouvés associés à des facteurs environnementaux de MTEV, comme par exemple l'IMC, illustrant la complexité et la fine barrière entre facteurs environnementaux et facteurs génétiques.

1.4 Physiopathologie

Au cours de ces 10 dernières années, les études pangénomiques ont permis d'identifier de nombreux nouveaux déterminants moléculaires associés au risque de MTEV. Certains travaux reposent sur des hypothèses considérant l'implication d'altérations biologiques telles que l'inflammation, l'immunité innée, les dysfonctionnements rénaux et plaquettaires. Néanmoins, les mécanismes physiopathologiques exacts mis en jeu dans le développement de la MTEV sont encore mal caractérisés à ce jour.

Un mécanisme de MTEV bien établi depuis longtemps est la cascade de la coagulation, représentée en **Figure 4**, qui fait intervenir de nombreux facteurs comme les protéines de l'antithrombine, C et S ainsi que le TFPI (pour inhibiteur de la voie du facteur tissulaire) qui sont des inhibiteurs naturels de la coagulation (Sallah, 1997; Sandset, 1996). En situation non pathologique les composants de la cascade de la coagulation s'équilibrent avec les mécanismes de la circulation sanguine. La coagulation peut s'initier par la voie intrinsèque lorsqu'il y a une lésion interne ou la voie extrinsèque lorsque la lésion est externe.

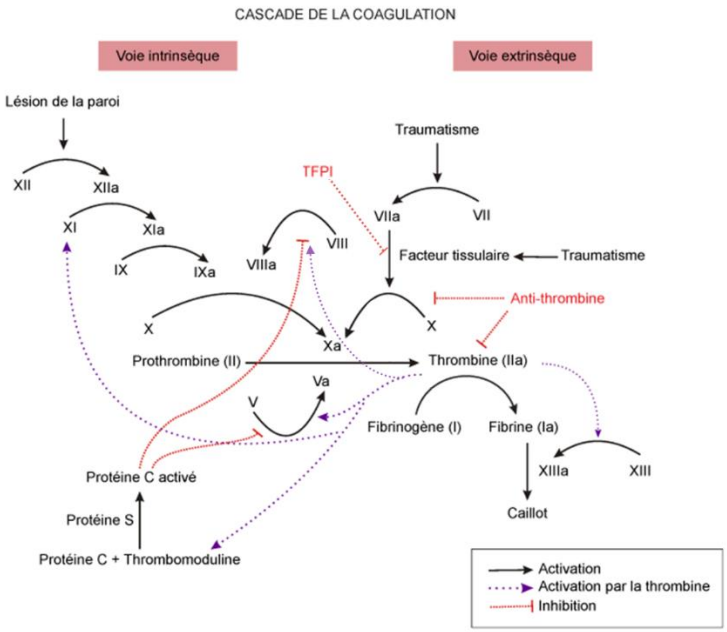


Figure 4 : Schéma de la cascade de la coagulation (adapté de (Pitte, 2019))

1.4.1 Hémostase et coagulation

En cas de brèche vasculaire plusieurs mécanismes cellulaires et moléculaires vont s'activer pour arrêter le saignement grâce à la formation d'un thrombus, c'est le principe de l'hémostase. Tout d'abord, le flux sanguin va ralentir afin de limiter la perte de sang (vasoconstriction) et favoriser le rassemblement de plaquettes sanguines et de facteurs de la coagulation au niveau de la brèche. Les plaquettes vont alors s'agréger sur la paroi du vaisseau sanguin, appelée également endothélium, afin de former une fine membrane (ou clou plaquettaire) pour obstruer la brèche. La membrane est par la suite renforcée par des filaments de fibrine qui emprisonnent des globules rouges (érythrocytes) et blancs (leucocytes, monocytes, neutrophiles) ainsi que des plaquettes, pour former le caillot sanguin (**Figure 5**). Une enzyme, la thrombine préalablement activée par différents facteurs de la coagulation (Xa : Facteur X activé ; Va : Facteur V activé), va alors permettre de transformer le fibrinogène soluble en fibrine insoluble. Lorsque la cicatrisation du vaisseau sanguin est terminée, le processus de fibrinolyse engendre la dissolution du caillot pour permettre à la circulation sanguine de reprendre son cours normal. L'acteur principal de la fibrinolyse est le plasminogène qui va se transformer en plasmine sous l'action de l'activateur du plasminogène (ou tPA, qui est synthétisé par les cellules endothéliales au niveau de la lésion) et de l'urokinase (produite au niveau du caillot sanguin). Une autre enzyme, la plasmine, aura elle le rôle de dégrader le réseau de fibrine ainsi que plusieurs facteurs de la coagulation et le fibrinogène. Les D-dimères

sont des fragments de fibrine qui sont libérés lors du processus de fibrinolyse et utilisés uniquement pour écarter un diagnostic de MTEV à cause de leur manque de spécificité (Kelly et al., 2002).

Dans l'hémostase, il y a un équilibre permanent entre les processus de coagulation et la fibrinolyse. Néanmoins, un caillot peut se former malgré l'absence de saignement. En 1856, Rudolph Virchow décrivait pour la première fois trois mécanismes physiopathologiques favorisant la survenue de la MTEV : (i) déséquilibre du système de la coagulation, (ii) ralentissement de la circulation sanguine, appelé également stase veineuse, (iii) présence de lésions vasculaires. Ces trois éléments sont désormais connus sous le nom de « triade de Virchow ». Toutefois, d'autres facteurs plus complexes n'appartenant pas à la triade de Virchow et non décrits dans la cascade de la coagulation influencent la survenue de la MTEV, dont ceux impliqués dans l'immunothrombose.

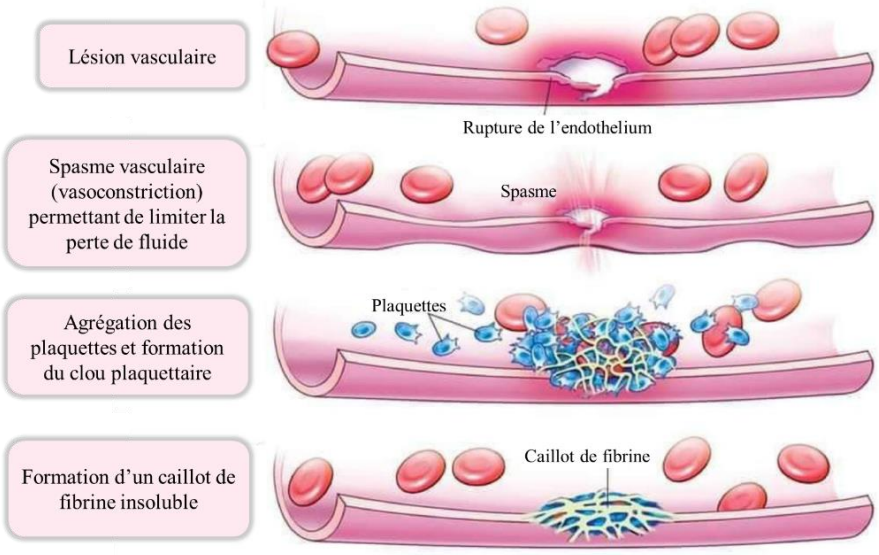


Figure 5 : Illustration de la formation du caillot de fibrine en cas de brèche vasculaire (adaptée de (*Basic Steps in Hemostasis - Nurseinfo*, 2020))

1.4.2 Le concept d'immunothrombose

Le concept d'immunothrombose a été introduit pour la première fois en 2013, suggérant que des interactions entre les mécanismes de défense immunitaire et de formation de MTEV pourraient exister dans certaines circonstances (Engelmann & Massberg, 2013).

En effet, durant l'inflammation, les cellules endothéliales sécrètent une protéine de type *P-sélectine* sur laquelle des monocytes et des neutrophiles viennent se lier via la glycoprotéine PSGL-1 (**Figure 6**). Le facteur tissulaire présent à la surface des monocytes favorise la

formation du thrombus en initiant la formation de la fibrine et le recrutement des globules rouges. Par ailleurs, l'activation des neutrophiles leur permet de libérer des pièges extracellulaires de neutrophiles (NETs pour *Neutrophil Extracellular Traps*), constitués de fragments d'ADN recouverts de plusieurs protéines. Le rôle des NETs est de piéger et détruire les agents pathogènes extracellulaires mais ils sont également impliqués dans l'activation et l'agrégation des plaquettes ainsi que dans la formation d'un thrombus (Bonaventura et al., 2021). Ils contribuent au concept d'immunothrombose grâce à leurs propriétés biologiques qui inactivent les anticoagulants naturels et favorisent la coagulation par le biais des facteurs de la voie intrinsèque. Par ailleurs, la fibrine produite par les NETs permet au caillot d'être plus résistant à la fibrinolyse, ce qui peut donc retarder voire empêcher sa dissolution (Vayne et al., 2017).

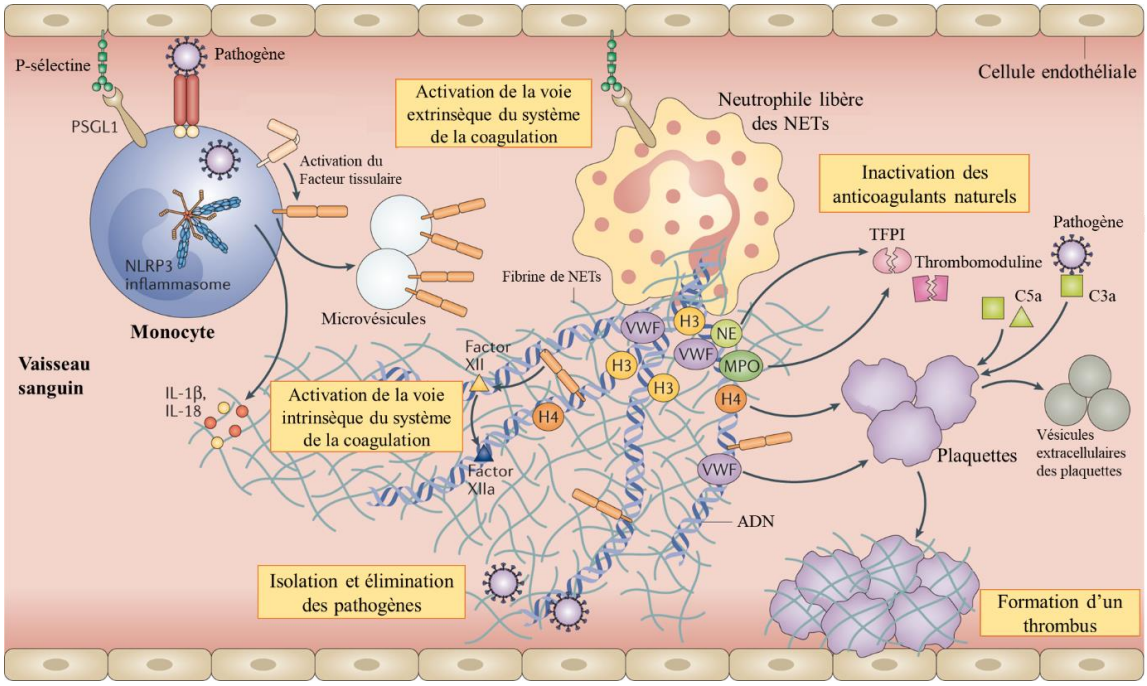


Figure 6 : Schéma des mécanismes physiopathologiques de l'immunothrombose (adaptée de (Bonaventura et al., 2021))

1.4.3 Le système du complément

Autour de ce concept d'immunothrombose, un autre système biologique interagit avec les facteurs de la coagulation : le système du complément. Le système du complément est constitué d'une trentaine de protéines plasmatiques produites par le foie et appartenant au système d'immunité innée. Elles ont notamment pour fonction de faciliter la phagocytose des agents pathogènes, de promouvoir l'inflammation et de créer un complexe d'attaque membranaire au niveau de la membrane de l'agent pathogène pour permettre sa lyse (Warwick

et al., 2021). Il existe trois voies d'activation de la cascade du complément : la voie classique, la voie des lectines et la voie alterne présentées en **Figure 7**. Les facteurs de la coagulation se trouvent à gauche de la figure et les protéines du complément (Complément C1 à C5) à droite.

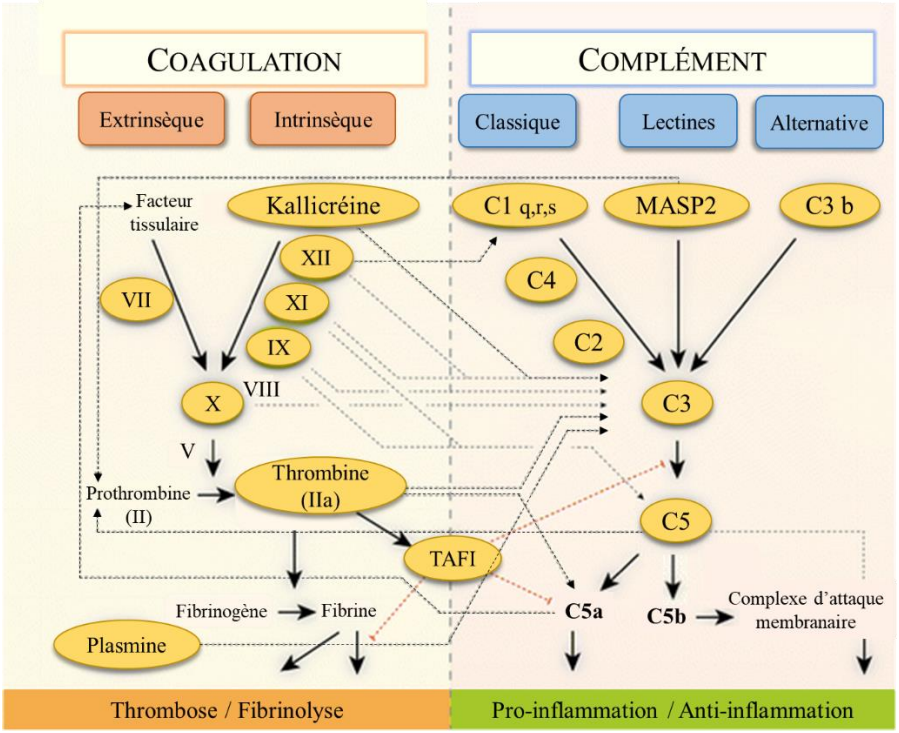


Figure 7 : Schéma de l'interaction entre les cascades de la coagulation et du complément (adaptée de (Danckwardt et al., 2013))

Les cascades du complément et de la coagulation interagissent à plusieurs niveaux. L'activation du Facteur XII permet d'activer la voie classique du complément en initiant le clivage de la protéine du complément C1 en C1q, C1r et C1s. De façon similaire, les enzymes et protéines de la coagulation (thrombine, kallicréine, plasmine) permettent l'activation du clivage de la protéine C3 en C5. La thrombine peut également cliver la protéine C5 en C5a, indépendamment du C3, c'est-à-dire en contournant l'activation d'une des trois voies du complément. Le système du complément permet également d'amplifier la coagulation en activant le facteur tissulaire et la prothrombine via C5 et C5a. De plus, l'enzyme de la lectine sérine protéase 2 (MASP2) déclenche la coagulation en transformant la prothrombine en thrombine (Danckwardt et al., 2013). Le système du complément, comprenant notamment C5, serait associé à l'augmentation de l'agrégation plaquettaire ainsi qu'à l'activation du système de coagulation et soutiendrait ainsi le concept d'immunothrombose (Mizuno et al., 2017).

2 Difficultés méthodologiques et objectif

2.1 Difficultés méthodologiques

Les GWAS sont une stratégie d'analyse pour identifier simultanément l'effet individuel de plusieurs millions de variables génétiques sur un trait d'intérêt. Il s'agit de la stratégie la plus communément adoptée pour identifier de nouveaux loci de susceptibilité génétique aux maladies complexes. Le développement d'outils bio-informatiques permettant de réaliser ces analyses dans des délais raisonnables a conduit à un engouement pour l'étude des facteurs de risque génétiques de maladies multifactorielles, telles que la MTEV ou les accidents vasculaires cérébraux, et de leurs biomarqueurs (Mishra et al., 2022; Thibord et al., 2022). Certaines de ces pathologies peuvent se manifester par des évènements aigus qui peuvent parfois se répéter dans le temps. L'analyse des facteurs de risque génétiques de ces récurrences constitue un champ d'intérêt émergeant pour comprendre les voies biologiques sous-tendant ces maladies (de Haan et al., 2018; Williams et al., 2016).

Actuellement, la plupart des logiciels disponibles, comme Plink (Purcell et al., 2007), proposent de réaliser les GWAS en utilisant uniquement des modèles de régression linéaire ou logistique. Des extensions et de nouveaux logiciels ont été proposés pour ajouter des effets aléatoires dans les modèles permettant, dans le cadre de l'analyse de données génétiques, de tenir compte des liens familiaux entre des individus apparentés et des différences d'origines ethniques (Mbatchou et al., 2021; Yang et al., 2011, 2014). D'autres logiciels comme ProbABEL ont été développés pour permettre l'implémentation du modèle de Cox pour l'analyse de données de survie et peuvent s'appliquer pour l'étude des facteurs génétiques des évènements récurrents (Aulchenko et al., 2010; Syed et al., 2017).

Cependant, toutes les spécificités de certains modèles (notamment pour le modèle de Cox) ne sont pas implémentées dans les logiciels ce qui les rend inutilisables lorsque la modélisation nécessaire ne rentre pas dans le cadre standard, notamment en présence d'un schéma d'étude de type ambispectif sur lequel je reviendrai en *section **4.1***. De plus, lors de l'analyse des facteurs génétiques d'une variable continue, il faut impérativement que cette variable soit gaussienne puisque seul le modèle linéaire est implémenté. Pour autant, de nombreuses distributions de biomarqueurs dévient d'une loi gaussienne et il n'existe à ce jour aucun logiciel permettant d'étudier leurs facteurs génétiques au travers d'une étude GWAS de manière efficiente. La complexité de mise en œuvre ainsi que l'important temps de calcul qui sont nécessaires pour la réalisation des GWAS en dehors des outils standards, incitent à la

transformation ou la simplification des variables. Par exemple, même si le modèle de Cox a une meilleure efficacité pour analyser un trait binaire dans une cohorte prospective qu'un simple modèle logistique qui ne prend pas en compte l'information sur le délai d'apparition de l'évènement d'intérêt, c'est malgré tout ce dernier qui est le plus souvent utilisé (en comparant les individus à la fin du suivi), car les logiciels d'analyse sont mieux implémentés (de Haan et al., 2018; van der Net et al., 2008). Dans d'autres situations où la variable d'intérêt est continue mais ne présente pas de distribution gaussienne, des transformations (comme la transformation inverse-normale qui se base sur les quantiles de la distribution) ou des dichotomisations à des seuils d'intérêt sont généralement appliquées pour faciliter l'implémentation (Kals et al., 2022; McCaw et al., 2020). Néanmoins, ces stratégies d'analyses alternatives ne sont pas toujours pertinentes selon les données étudiées.

2.2 Objectif

Ce projet de thèse s'inscrit dans une initiative nationale portée par le réseau FCRIN-INNOVTE (*Investigation Network On Venous Thrombo-Embolism*) qui soutient la recherche clinique et translationnelle sur la MTEV. Au cours de ces dernières années, les études GWAS ont permis de mettre en évidence de nouveaux facteurs impliqués dans les mécanismes thrombotiques. Néanmoins, ces facteurs ne permettent pas d'expliquer la totalité des cas de MTEV et il reste encore beaucoup d'éléments à découvrir afin d'améliorer la compréhension, la prévention et le traitement de la MTEV.

L'objectif de ce projet de thèse était de développer des méthodologies statistiques permettant de répondre aux deux problématiques particulières en lien avec la MTEV soulevées par le réseau INNOVTE, à savoir le développement de modélisations statistiques permettant d'étudier l'association entre des polymorphismes génétiques et 1) le risque de récidive d'un évènement dans le cadre d'une étude basée sur un schéma d'étude de type ambispectif et 2) un biomarqueur présentant une distribution semicontinue.

Après de succincts rappels de quelques notions d'épidémiologie génétique (*Chapitre **3***) et une description des cohortes sur lesquelles mon projet s'est basé (*Chapitre **4***), je décrirai dans les chapitres ***5*** et ***6*** les deux développements que j'ai menés et les applications qui les ont motivés. Le dernier chapitre (***7***) sera consacré à un travail annexe auquel j'ai fortement contribué, une méta-analyse d'études GWAS sur les taux plasmatiques d'une protéine de la cascade du complément et l'étude de l'association entre ces mêmes taux et le risque de récidive de MTEV.

3 Notions d'épidémiologie génétique

3.1 De l'ADN à la protéine

Le génome humain est constitué de 22 paires de chromosomes autosomiques et une paire de chromosomes sexuels qui sont chacun composés d'une macromolécule en double hélice d'Acide Désoxyribonucléique (ADN). L'ADN est le support de l'information génétique contenant des milliers de gènes et de régions régulatrices sous la forme de successions de nucléotides, également appelées bases. Il existe quatre différents nucléotides : l'Adénine (A), la Cytosine (C), la Guanine (G) et la Thymine (T). Au sein des gènes, les séquences nucléotidiques forment des exons et introns qui sont généralement présents en alternance, avec en amont la région 5' non traduite (5'-UTR) et en aval la région du 3' non traduite (3'-UTR) (**Figure 8**). Selon le projet *« The Human Genome »* il existerait près de 25 000 gènes codant pour la production de protéines qui représenteraient seulement 1 à 2 % des trois milliards de bases du génome humain (Salzberg, 2018). Les autres 98 % des bases du génome humain sont situées dans des régions inter-géniques qui peuvent tout de même influencer l'expression des gènes.

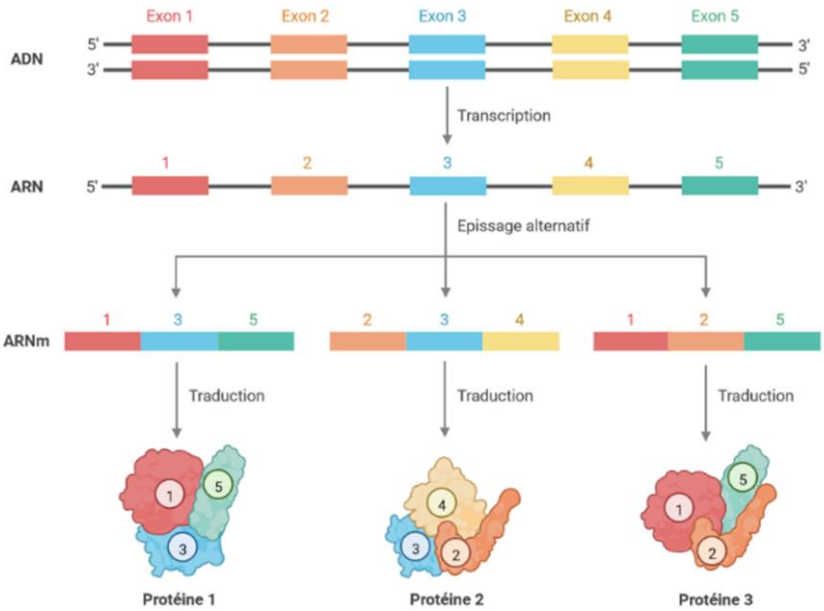


Figure 8 : De l'ADN à la protéine (créée avec BioRender.com)

Le processus de transcription est la première étape de l'expression génique qui permet de transformer un segment de la double hélice d'ADN en un simple brin d'Acide Ribonucléique (ARN) (Inserm, La science pour la santé, 2022). L'épissage permet de garder uniquement les

exons qui composent l'ARN messager (ARNm) et sont utilisés pour la synthèse des protéines (traduction).

3.2 Les variations génétiques

La quasi-totalité du génome de deux individus est identique. C'est donc la petite part de variations génétiques entre les génomes des individus qui est à l'origine de la diversité des phénotypes. On appelle polymorphisme toute variation dans la séquence d'ADN présente dans au moins 1 % de la population générale. Un polymorphisme est dit bi-allélique lorsqu'il n'est retrouvé que sous deux formes (ou allèles) dans la population. Lorsque ces deux formes ne différent que par la substitution d'une base nucléotidique par une autre, un polymorphisme est communément appelé un SNP pour *Single Nucleotide Polymorphism*. Les SNPs sont le type de variation génétique le plus fréquent. L'allèle le plus fréquemment rencontré en population générale est qualifié d'allèle majeur et le moins fréquent d'allèle mineur. Etant donné qu'un individu possède deux copies de chaque chromosome (une copie maternelle et une copie paternelle), il présente donc deux allèles pour chaque SNP, ces deux allèles formant ce que l'on nomme le génotype. Les individus qui portent deux copies du même allèle d'un SNP sont dits homozygotes et sont à l'inverse appelés hétérozygotes lorsque les deux allèles sont différents. Il existe bien d'autres types de variations génétiques (insertions, délétions, réarrangements, etc.) dont il sera peu discuté dans ce travail. Les combinaisons d'allèles de plusieurs SNPs situés sur un même chromosome s'appellent des haplotypes (**Figure 9**).

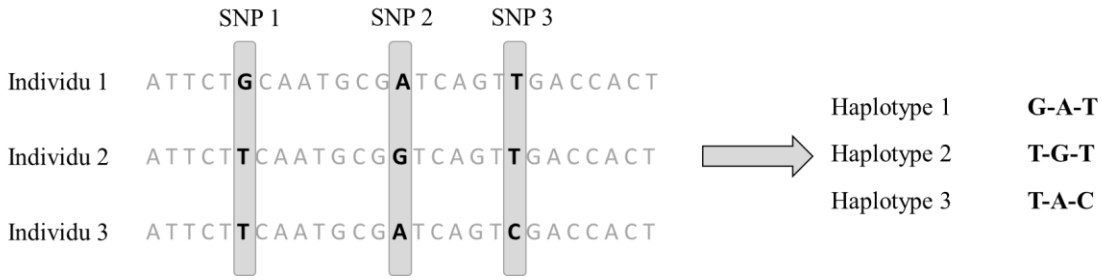


Figure 9 : Illustration de la notion d'haplotypes

La plupart du temps, la présence d'un SNP n'a pas de conséquence fonctionnelle mais dans certains cas elle peut influencer certains traits humains ou le développement de certaines maladies. Le génome humain contient environ 10 millions de SNP et, en moyenne, 3 millions de SNP sont retrouvés lorsque les génomes de deux individus sont comparés (Børsting & Morling, 2013; National Human Genome Research Institute, 2018).

Le génome humain est composé de plus de 3 milliards de bases et les avancées technologiques de ces dernières années ont permis de développer des technologies de génotypage et séquençage de l'ADN permettant de l'étudier de manière très approfondie. Le séquençage de l'ADN de plusieurs milliers de personnes dans le monde entier a notamment permis de construire des catalogues, appelés panels de référence, recensant les variations génétiques observées ainsi que leur fréquence dans les différents groupes d'origine ethnique. Les panels les plus couramment utilisés sont *HapMap*, *1000 Genomes*, *Haplotype Reference Consortium (HRC)* et *Trans-Omics for Precision Medicine (TOPMed)* (1000 Genomes Project Consortium et al., 2010; Gibbs et al., 2003; McCarthy et al., 2016; Taliun et al., 2021). La création de ces panels a également permis d'établir des cartes de régions où les variants génétiques sont en déséquilibre de liaison et la structure des haplotypes fréquents au sein de ces régions. Le déséquilibre de liaison entre deux polymorphismes est représenté par l'association préférentielle de leurs allèles, c'est-à-dire qu'ils ne s'associent pas de façon aléatoire dans chaque gamète des individus d'une population. De façon simplifiée, le déséquilibre de liaison peut être considéré comme la corrélation entre les allèles observés pour deux polymorphismes.

3.3 Analyse des variations génétiques

Jusque dans les années 1980, l'étude de la génétique humaine s'est principalement focalisée sur l'analyse de données familiales en utilisant des analyses de liaison permettant d'observer des régions chromosomiques qui ségrégaient de façon non aléatoire au sein de familles présentant une agrégation familiale (ou corrélation) forte pour le phénotype étudié. L'hypothèse sous-jacente était que cette agrégation pouvait être la résultante de la présence d'un variant génétique avec un effet fort sur le phénotype d'intérêt. Par la suite, une étude plus approfondie de ces régions pouvait permettre d'affiner la région chromosomique d'intérêt.

Cependant, de nombreuses maladies fréquentes et ayant un impact important en santé publique, telles que les pathologies cardiovasculaires (par exemple l'infarctus du myocarde ou la maladie thromboembolique veineuse) ou encore la démence, ne sont généralement pas la conséquence d'une seule variation génétique et les analyses de liaison n'étaient pas assez puissantes pour identifier des effets de taille modeste. Ces pathologies sont qualifiées de « maladies complexes » car elles ne présentent pas de forte agrégation familiale et sont influencées par un ensemble de facteurs à la fois génétiques et environnementaux. La notion d'héritabilité d'un trait permet de quantifier la proportion de sa variance phénotypique qui serait attribuable à des facteurs génétiques (Robette et al., 2022).

Afin d'améliorer la compréhension de ces traits complexes, les premières études génétiques procédaient par une approche dite « gène candidat » où les hypothèses *a priori* permettaient d'identifier un ou plusieurs gènes qui seraient le plus probablement impliqués. Cette approche a notamment permis d'identifier dès les années 1990 l'implication du gène *APOE4* dans les pathologies du vieillissement cognitif (Corder et al., 1993). Néanmoins, cette approche nécessitait que les hypothèses *a priori* soient suffisamment solides et pertinentes pour obtenir des résultats qui soient répliqués dans des échantillons indépendants (Duncan et al., 2019).

3.4 Les études d'associations pangénomiques

Le développement des plateformes de génotypage à haut débit et des puces à ADN permettant de génotyper des ensembles prédéfinis de SNPs, ont permis l'apparition des analyses d'associations pangénomiques pour la première fois en 2002. Ces méthodes consistent à comparer les fréquences des variants selon un phénotype d'intérêt binaire (ou en utilisant un modèle linéaire si le phénotype est continu) en utilisant une approche agnostique puisque les SNPs étudiés sont répartis dans le génome entier, sans sélection préalable (Ozaki et al., 2002). Le nombre important de tests statistiques réalisés (un test par SNP) engendre une inflation de l'erreur de type I qui doit être corrigée et dans le cadre des analyses GWAS, le seuil de significativité pangénomique est fixé à 5×10^{-8}. Ce seuil a été établi en estimant qu'il y aurait un million de SNPs indépendants répartis tout au long du génome, auquel la correction de Bonferonni (qui consiste à diviser le seuil de significativité nominal ($\alpha = 0{,}05$) par le nombre de tests indépendants) a été appliquée (Bland & Altman, 1995). Les premières puces à ADN permettaient de mesurer environ 100 000 SNPs alors qu'à l'heure actuelle les plus grandes puces à ADN permettent de mesurer précisément plus d'un million de polymorphismes. A partir des données génotypées, il est possible d'inférer environ 9 millions de polymorphismes avec une fréquence supérieure à 1 % à partir des méthodes dites d'imputation. Comme illustré sur la **Figure 10**, à partir des panels de référence et en se basant sur les haplotypes et le déséquilibre de liaison avec les SNPs génotypés, il est possible d'imputer les SNPs manquants (Marchini & Howie, 2010). Dans cet exemple, trois SNPs ont été génotypés et en regardant les haplotypes correspondants dans le panel de référence, il est possible d'inférer les SNPs non génotypés avec une certitude plus ou moins élevée. En effet, dans les deux premières séquences du panel de référence, on peut voir que lorsqu'il y a un nucléotide A en quatrième position, il est toujours précédé de la séquence de nucléotides TCT. Il est donc possible d'inférer avec une bonne

certitudes ces trois nucléotides lorsqu'il y a un nucléotide A en quatrième position. A l'inverse, lorsqu'il y a un nucléotide T, il peut être précédé des séquences CCC ou TCC, il y aura donc plus d'incertitude lors de l'inférence du premier nucléotide qui sera soit un C soit un T. Lors de l'imputation, un score compris entre 0 et 1 est généré pour chacun des polymorphismes, permettant ainsi de quantifier le degré de confiance en ces valeurs imputées, 1 représentant une qualité d'imputation parfaite.

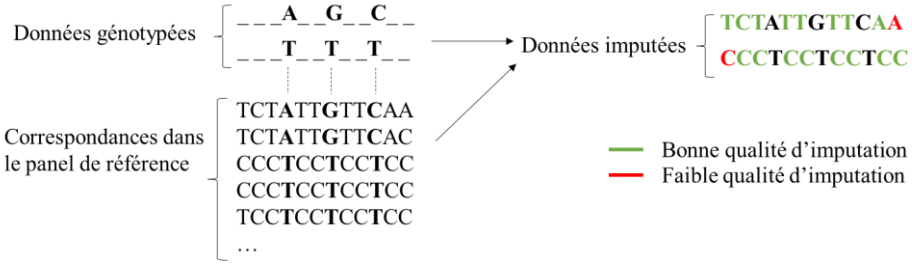


Figure 10 : Illustration du principe de l'imputation

Grâce au développement des outils bio-informatiques et la réduction du coût des puces de génotypage, les GWAS sont désormais une stratégie de recherche très commune pour identifier des variations génétiques fréquentes associées à un trait complexe. Cependant, leurs résultats nécessitent systématiquement une réplication dans un échantillon indépendant afin d'éviter que les signaux significatifs obtenus soient dus au hasard ou à une spécificité non-identifiée de l'échantillon étudié.

Les GWAS réalisées sur plusieurs milliers d'individus permettent d'avoir une puissance statistique suffisante pour identifier des effets génétiques de plus en plus modeste. La première étude GWAS sur la MTEV a été réalisée en 2009 sur un échantillon d'environ 450 sujets de l'étude française EOVT (Trégouët et al., 2009). Depuis, le nombre d'études a considérablement évolué, témoignant ainsi de la démocratisation de cette stratégie, pour arriver désormais à une méta-analyse intégrant près de 80 000 cas (Thibord et al., 2022). A l'échelon national, les études MARTHA, FARIVE et EDITH, que je détaillerai dans le chapitre suivant, ont eu et jouent encore un rôle majeur dans ce domaine de la recherche.

4 Populations d'étude

4.1 MARTHA

L'étude MARTHA (MARseille THrombosis Association) a été mise en place en 1994 par le Professeur Pierre-Emmanuel Morange à Marseille. Cette étude avait pour objectif principal d'améliorer les connaissances des mécanismes physiopathologiques de la MTEV en étudiant ses facteurs génétiques et biologiques. Les patients ont été recrutés à l'occasion d'une visite médicale au Centre d'Exploration des pathologies Hémorragiques et Thrombotiques de l'Hôpital de La Timone à Marseille à partir de janvier 1994 (Antoni et al., 2011; Oudot-Mellakh et al., 2012). Les critères d'inclusion étaient : origine caucasienne, au moins un antécédent personnel de MTEV documenté par veinographie, échographie Doppler, angiographie par tomodensitométrie spirale, et/ou scintigraphie pulmonaire de ventilation/perfusion, pas d'antécédent personnel de cancer ni de syndrome des anti-phospholipides, sans déficience en antithrombine, protéine C ou protéine S et non homozygote mineur pour les mutations FV R506Q et du FII G20201A.

L'étude MARTHA est composée de deux échantillons indépendants de cas de MTEV : MARTHA08 (N = 1 006 cas recrutés entre 1994 et 2005) et MARTHA10 (N = 586 cas recrutés entre 2005 et 2008), totalisant ainsi 1 592 cas de MTEV. Après vérification des critères d'éligibilité, 17 individus avec un cancer ont été exclus. Une prise de sang a été effectuée pour chacun des sujets permettant notamment de génotyper les 1 575 sujets de MARTHA (à l'aide d'une puce à ADN *Illumina Human610-Quad Beadchip* mesurant 567 589 SNPs pour les individus de MARTHA08, et d'une puce à ADN *Illumina 660W-Quad Beadchip* mesurant 547 886 SNPs pour les individus de MARTHA10).

Des contrôles de qualité standards ont été appliqués sur les données génétiques et sont décrits dans *Antoni et al.* (Antoni et al., 2011). Au final, 472 123 SNPs (communs entre MARTHA08 et MARTHA10) et 1 525 individus ont passé les contrôles de qualité. Par la suite, ces données ont été imputées avec le logiciel minimac sur le panel 1 version 2 de 1000 Genomes et plus récemment sur le panel 3 version 5 de 1 000 Genomes, permettant ainsi d'obtenir plus de 9,5 millions de SNPs avec une fréquence allélique supérieure à 1 % et une qualité d'imputation satisfaisante (supérieure à 0,3) (Howie et al., 2012).

A l'inclusion des patients dans l'étude, les évènements antérieurs de MTEV ont été collectés rétrospectivement puis, entre 2013 et 2018, les individus de MARTHA ont été recontactés afin de collecter des informations prospectives sur une éventuelle nouvelle MTEV

depuis leur inclusion dans l'étude. Ce schéma d'étude comprenant à la fois des informations rétrospectives et prospectives est appelé ambispectif et est illustré dans la **Figure 11**. Pour chaque évènement de MTEV, les dates, types (TVP, EP) et caractères provoqués ont été collectés.

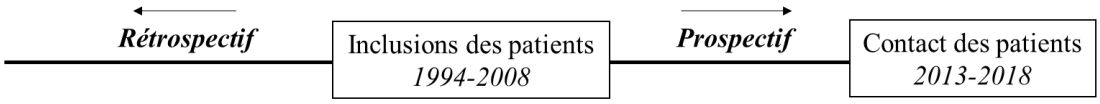


Figure 11 : Illustration du schéma d'étude ambispectif de l'étude MARTHA

A la fin de la phase de contact des patients, les informations sur le statut vital des participants ont été obtenues via le Répertoire National d'Identification des Personnes Physiques ou les bases de données hospitalières. Les profils de suivi de patients et leurs évènements de MTEV sont présentés dans la **Figure 12**. Les trois premiers profils représentent les patients qui avaient eu une seule MTEV avant l'inclusion. Les patients de la situation n°1 ont pu être recontactés et n'ont pas eu de récidive, ceux de la n°2 ont indiqué une récidive de MTEV post inclusion et les patients de la situation n°3 n'ont pas pu être recontactés (pas de réponse ou décès). Les trois autres profils représentent les patients qui avaient eu au moins une récidive avant l'inclusion, seules les informations concernant la 1ère MTEV et la 1ère récidive ont été collectées. Le profil n°4 représente les patients qui ont été contactés et qui n'ont pas eu une autre récidive post-inclusion, ceux du profil n°5 ont eu une récidive post-inclusion et ceux du profil n°6 n'ont pas pu être contactés.

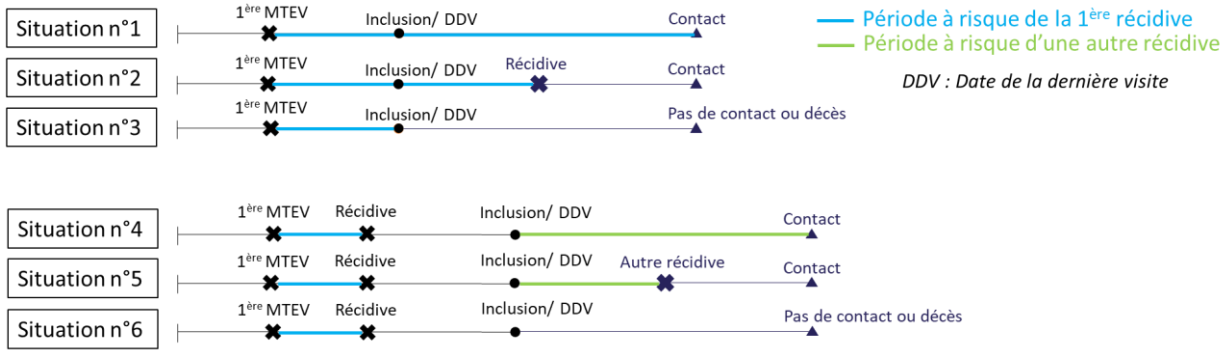


Figure 12 : Schéma des profils de suivi et des évènements de MTEV dans l'étude MARTHA

Les données de l'étude MARTHA ont été analysées dans le cadre du premier objectif de ma thèse concernant les déterminants génétiques de la récidive de MTEV présenté en *section **5***.

4.1 MEGA

L'étude MEGA (*Multiple Environmental and Genetic Assessment of risk factors for venous thrombosis*) a été mise en place en 1999 par le Professeur Frits R. Rosendaal (Blom et al., 2005). Entre 1999 et 2004, près de 4 900 cas incidents de 1ère MTEV, âgés de 18 à 70 ans, ont été recrutés dans six cliniques régionales aux Pays-Bas. Dans une étude ancillaire, les ADN de 1 289 patients ont été génotypés avec une puce à ADN *Illumina Human660-Quad v.1 Beadchip*. Des contrôles de qualité standards ont été appliqués et 497 563 SNPs ont été utilisés pour l'imputation réalisée avec le logiciel minimac sur le panel 1 version 2 de 1000 Genomes et plus récemment sur le panel 3 version 5 de 1 000 Genomes, permettant ainsi d'obtenir près de 9,5 de millions SNPs avec une fréquence allélique supérieure à 1 % et une qualité d'imputation satisfaisante (supérieure à 0,3) (de Haan et al., 2018). Entre 2008 et 2009, des questionnaires ont été envoyés aux participants afin de collecter des informations sur une possible récidive de MTEV post-inclusion.

L'étude MEGA a été utilisée comme échantillon de réplication dans le cadre du premier objectif de thèse concernant les déterminants génétiques de la récidive de MTEV présenté en *section **5***.

4.2 EDITH

L'étude EDITH (Etude des Déterminants et Interactions de la THrombose veineuse) est une cohorte mise en place en 2000 par le Professeur Francis Couturaud (Lacut et al., 2004). Entre 2000 et 2019, 3 169 cas incidents de MTEV de la région de Brest ont été inclus dans EDITH. Les participants d'EDITH devaient être âgés de plus de 18 ans et présenter une MTEV documentée. Les informations concernant leurs antécédents personnels de MTEV ont également été collectées. Les patients sont suivis depuis leur inclusion et ils sont régulièrement contactés par mail ou courrier pour collecter des informations sur leurs éventuelles récidives de MTEV post-inclusion ou la survenue d'évènements cardiovasculaires ou du décès. Cette étude a un design similaire à celui de MARTHA puisque les individus n'étaient pas toujours inclus pour leur premier évènement de MTEV. Néanmoins, toutes les informations sur les évènements pré-inclusion ont été recueillies et le suivi était plus régulier.

Les individus d'EDITH ont été génotypés avec une puce à ADN *Illumina Infinium Human Global Screening Array GSAMD-24v3,* permettant de mesurer 730 059 SNPs. Suite aux contrôles de qualité standards, 180 individus ont été exclus et 475 305 SNPs ont été gardés pour

l'imputation réalisée sur le panel de référence 1000 Genomes phase 3 version 5. Au final, plus de 9,5 millions de génotypes ont été imputés pour 2 989 individus.

Les données de l'étude EDITH ont été utilisées suite à la soumission du premier article de cette thèse pour affiner et apporter plus de poids aux résultats obtenus sur les déterminants génétiques de la récidive de MTEV présentés en *section **5.2.1***.

4.3 FARIVE

L'étude FARIVE (FActeurs de RIsque et de récidives de la maladie thromboembolique Veineuse) est une étude cas-témoins multicentrique française mise en place en 2003 par le Professeur Joseph Emmerich (Trégouët et al., 2009). Entre 2003 et 2007, 607 cas incidents de MTEV (documentée) et 607 témoins hospitaliers appariés sur l'âge et le sexe, d'origine caucasienne ont été recrutés. Tous les individus de l'étude devaient être âgés de plus de 18 ans et ne pas avoir d'antécédents personnels de MTEV ou de cancer. Les participants de FARIVE ont été génotypés avec une puce à ADN *Illumina Infinium Global Screening Array v3.0 (GSAv3.0) microarray* permettant de mesurer 730 059 SNPs et des contrôles de qualité standards décrits dans *Thibord et al.* ont été appliqués (Thibord et al., 2022). L'échantillon final était composé de 1 077 individus et 535 105 variants ont été utilisés pour l'imputation sur le panel de référence 1000 Genomes phase 3 version 5, permettant ainsi d'obtenir plus de 9,6 millions de SNPs avec une fréquence allélique supérieure à 1 % et une qualité d'imputation satisfaisante (supérieure à 0,3).

Les niveaux plasmatiques de plusieurs protéines ont été mesurés dans l'étude FARIVE dans la semaine suivant l'inclusion des patients (Bruzelius et al., 2016). Pour effectuer ces mesures (réalisées par l'équipe du Professeur Odeberg en Suède), des anticorps ciblant des protéines candidates ont été ajoutés aux échantillons sanguins afin d'incuber pour pouvoir se coupler aux protéines en question. Après un jour d'incubation, une solution a été ajoutée pour mesurer les niveaux des protéines grâce à une détection de l'intensité de la fluorescence par un laser. Les protéines mesurées ont été sélectionnées pour être soit (i) impliquées dans les cascades de la coagulation et de la fibrinolyse, qui représentent les processus de formation et de dissolution des caillots sanguins ; (ii) fortement exprimées dans les cellules endothéliales qui peuvent être associées à des anomalies de la coagulation ou (iii) synthétisées par des gènes récemment identifiés comme étant associés à des pathologies cardiovasculaires ou à leurs facteurs de risque. Des variables biologiques pertinentes dans le cadre de la MTEV, comme le

dosage des D-dimères ou les taux des facteurs de la coagulation, ont également été mesurées dans cet échantillon.

Pour les cas, un second prélèvement sanguin a également été réalisé à la suite de l'arrêt du traitement anticoagulant (environ 7 mois après l'inclusion) permettant ainsi de mesurer d'autres variables biologiques d'intérêt. C'est notamment à partir de ces échantillons plasmatiques que les taux de NETs (voir *section **1.4.2***) ont été mesurés et feront l'objet de mon deuxième article sur l'analyse GWAS d'un biomarqueur présentant une distribution semicontinue (*section **6.1***).

5 Analyse des facteurs génétiques de la récidive de la maladie thromboembolique veineuse

Le premier développement méthodologique auquel je me suis intéressée au cours de ma thèse s'inscrivait dans le contexte de l'étude des déterminants génétiques de la première récidive de la MTEV. Comme présenté dans l'introduction de cette thèse, la récidive de MTEV est fréquente et l'identification de ses facteurs de risque pourrait permettre d'établir le profil des patients les plus à risque de récidiver afin d'adapter leur prise en charge, notamment en prolongeant leur traitement anticoagulant. A l'inverse, la durée de ce traitement pourrait être diminuée pour les patients les moins susceptibles de récidiver afin de leur éviter de s'exposer à un risque d'hémorragie.

Au moment où je débutais cette thèse, deux études européennes composées de patients MTEV avec et sans récidive avaient été génotypés par puce à ADN, les études MARTHA et MEGA. L'étude MEGA reposait sur un plan d'expérience classique de type cohorte prospective, puisque les individus ont été inclus pour leur première MTEV et ils ont par la suite été recontactés pour étudier la survenue de leur première récidive (*voir section **4.1***). L'étude MARTHA présentait quant à elle un schéma ambispectif puisqu'à l'inclusion dans l'étude certains individus avaient déjà présenté une récidive de MTEV (*voir section **4.1***). L'objectif de mon travail était de proposer une méthodologie statistique pour l'étude des facteurs de risque de récidive de MTEV dans l'étude MARTHA qui permette d'inclure dans l'analyse à la fois les récidives collectées rétrospectivement (survenues avant l'inclusion) et les récidives collectées prospectivement, afin de maximiser la puissance statistique tout en évitant le biais de sélection lié à la survie des individus. L'objectif final était de tester l'association entre des SNPs et le risque de récidive en méta-analysant les résultats obtenus dans MARTHA et MEGA.

Cette méthode devait permettre de réaliser une GWAS dans l'étude MARTHA afin d'identifier de nouveaux déterminants génétiques du risque de récidive de MTEV. Dans un premier temps, ce modèle a été évalué et utilisé pour estimer l'effet génétiquement déterminé du groupe sanguin *ABO*, l'un des facteurs les plus importants de la MTEV en raison de l'ampleur des effets génétiques associés aux groupes sanguins à risque A1 et B et de leur prévalence dans la population générale, sur le risque de récidive de MTEV (Goumidi et al., 2020; Trégouët et al., 2009).

Pour analyser les facteurs de risque associés à la survenue d'un évènement, le modèle semi-paramétrique à risques proportionnels de Cox est le plus utilisé (Abd ElHafeez et al.,

2021; Cox, 1972). Dans le cadre classique de ces analyses, les variables d'exposition sont mesurées à l'inclusion, c'est-à-dire avant la survenue de l'évènement d'intérêt afin d'éviter un biais de causalité inverse. Néanmoins, lorsque les variables d'exposition ne sont pas dépendantes du temps, comme c'est le cas pour les variables génétiques qui sont fixes dès la naissance, ce biais est maîtrisé même si l'évènement a eu lieu avant l'inclusion et la mesure de l'exposition. Cependant, l'analyse d'évènements survenant avant l'inclusion des sujets induit un biais de sélection puisque l'étude de ces individus est conditionnelle à leur survie jusqu'à l'inclusion. Afin de tenir compte de ce biais, des poids se basant sur la méthode de probabilité inverse (appelée également IPW - *Inverse Probability Weighting*) ont été attribués aux individus permettant ainsi de donner plus de poids dans l'analyse aux individus qui étaient les moins susceptibles d'avoir été inclus dans l'étude, comme décrit dans l'article en *section **5.1*** (Alonso et al., 2006).

Pour appliquer cette stratégie aux patients de MARTHA, j'ai tout d'abord estimé le risque de décès dans cette population de cas ayant tous survécu à au moins un évènement de MTEV. Pour ce faire, un modèle de Cox du risque de décès avec troncature à gauche au moment de l'inclusion dans l'étude avec l'âge des participants comme temps base a été utilisé. Grâce aux estimations des paramètres de ce modèle, il est possible d'estimer les probabilités de survie des individus jusqu'à la date à laquelle leurs informations sur la présence d'une récidive de MTEV ont été collectées. Ces probabilités ont ensuite été inversées puis standardisées, définissant ainsi le poids statistique de chacun des individus.

Pour l'analyse ambispective de MARTHA, un modèle de Cox du risque de première récidive de MTEV intégrant les poids estimés et considérant comme temps de base le délai depuis le premier évènement de MTEV a été utilisé. Les résultats de ce modèle ont été comparés à l'approche classique prospective qui aurait pu être considérée en restreignant l'analyse aux patients inclus pour une première MTEV puis recontactés. Le diagramme de flux des différents échantillons de MARTHA considérés dans ce projet sont présentés en **Figure 13**.

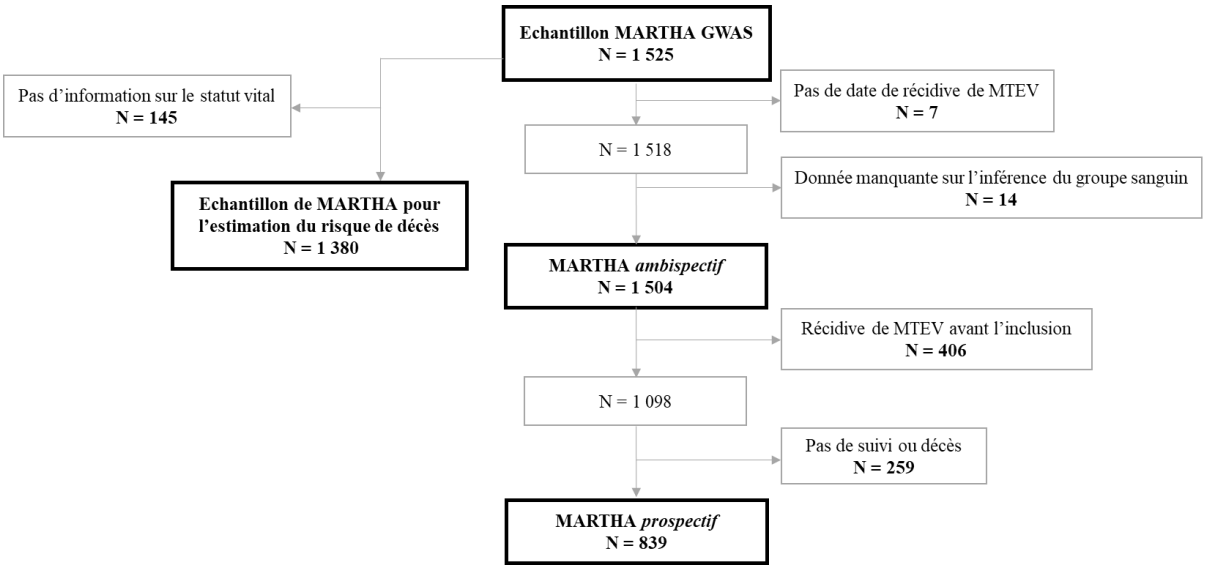


Figure 13 : Diagramme de flux pour l'étude du groupe sanguin sur le risque de récidive dans l'étude MARTHA

Cette stratégie d'analyse a permis d'étudier les facteurs cliniques non dépendants du temps, comme le sexe ou les caractéristiques de la première MTEV, ainsi que les haplotypes du groupe sanguin (A1, A2, O1, O2, B) avec le risque de récidive de MTEV. Les analyses ont été réalisées dans les échantillons de MARTHA (en prospectif et ambispectif) et MEGA, puis les résultats de MARTHA ambispectif et MEGA ont été méta-analysés combinant ainsi 2 752 cas de MTEV dont 993 récidives. Cette étude a permis d'affiner des observations précédentes indiquant que les individus non O seraient plus à risque de récidiver en identifiant les haplotypes A1 et A2 comme des facteurs de risque de récidive de MTEV en comparaison à l'haplotype O1.

5.1 Article 1 : Association of *ABO* blood groups with venous thrombosis recurrence in middle-aged patients: insights from a weighted Cox analysis dedicated to ambispective design

Article 1 soumis (1er auteur) :

Munsch G, Goumidi L, van Hylckama Vlieg A, […], Jacqmin-Gadda H, Morange P-E*, Trégouët D-A*.

Contribution :

- Analyses statistiques
- Ecriture du manuscrit

Cet article a été déposé en 2021 sur le serveur d'archives medRxiv (https://doi.org/10.1101/2021.11.20.21266583) sous le nom *"Modelling of time-to-events in an ambispective study: illustration with the analysis of ABO blood groups on venous thrombosis recurrence"*. Par la suite, des modifications ont été apportées au titre ainsi que dans le contenu de l'article dont la version actuellement soumise est présentée dans la section suivante.

Association of ABO blood groups with venous thrombosis recurrence in middle-aged patients: insights from a weighted Cox analysis dedicated to ambispective design

Gaëlle Munsch[1], Louisa Goumidi[2], Astrid van Hylckama Vlieg[3], Manal Ibrahim-Kosta[2,4], Maria Bruzelius[5,6], Jean-François Deleuze[7,8], Frits R. Rosendaal[3], Hélène Jacqmin-Gadda[1+], Pierre-Emmanuel Morange[2,4*], David-Alexandre Trégouët[1*]

[1] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France

[2] INSERM UMR_S 1263, Nutrition Obesity and Risk of Thrombosis, Center for CardioVascular and Nutrition research (C2VN), Aix-Marseille University, Marseille 13385, France

[3] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, Netherlands.

[4] Laboratory of Haematology, La Timone Hospital, Marseille 13385, France

[5] Department of Medicine Solna, Karolinska Institute, Stockholm, Sweden

[6] Department of Hematology, Karolinska University Hospital, Stockholm, Sweden

[7] Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, 91057, Evry, France

[8] Centre d'Etude du Polymorphisme Humain, Fondation Jean Dausset, Paris, France

* These authors equally contributed to the work.

+ Corresponding author:

Dr Hélène Jacqmin-Gadda,

Université de Bordeaux, Inserm U1219, 146 rue Léo Saignat, 33076 Bordeaux, France

helene.jacqmin-gadda@u-bordeaux.fr

ABSTRACT

Background: In studies of time-to-events, it is common to collect information about events that occurred before the inclusion in a prospective cohort. When the studied risk factors are independent of time, including both pre- and post-inclusion events in the analyses, generally referred to as relying on an ambispective design, increases the statistical power but may lead to a selection bias. Motivated by the study of the association of genetically determined ABO blood groups with venous thromboembolism (VT) recurrence in the MARTHA study built on an ambispective design, a dedicated methodology is here proposed to optimise the statistical power while taking into account the selection bias due to mortality.

Methods: This work relies on two independent cohorts of VT patients, the French MARTHA study built on an ambispective design and the Dutch MEGA study built on a standard prospective design. While the impact of ABO blood groups on VT recurrence was assessed using a standard Cox model in MEGA, a dedicated weighted Cox model was developed in MARTHA where weights were defined by the inverse of the survival probability at the time of data collection about the events. Thanks to the collection of information on the vital status of patients at the end of the follow-up, we could estimate the survival probabilities using a left-truncated Cox model on the death risk. Finally, results obtained in both studies were then meta-analysed.

Results: In the combined sample totalling 2,752 patients including 993 recurrences, the A1 blood group has an increased risk (Hazard Ratio (HR) of 1.18, $p=4.2x10^{-3}$) compared with the O1 group, homogeneously in MARTHA and in MEGA. The same trend (HR=1.19, p=0.06) was observed for the less frequent A2 group.

Conclusion: In conclusion, the methodology proposed for studies built on ambispective design allows to clarify the association of ABO blood groups with the risk of VT recurrence. Besides, this methodology has an immediate field of application in the context of genome wide association studies.

Key-words: Ambispective design; Survival analysis; Venous thrombosis; Recurrence; ABO blood groups; Genetic association studies

INTRODUCTION

Venous Thrombosis (VT) is a common cardiovascular disease with an annual incidence of ~1 to 3 per 1,000 in the general population which increases with age (1). This pathology can manifest as either deep vein thrombosis (DVT) or pulmonary embolism (PE) with a mortality rate within a month of diagnosis at 6% and 12%, respectively (2).

After a first VT, the recurrence rate is approximately 30% within 10 years (3). VT recurrence could be prevented by a continued anticoagulant treatment but this therapy leads to a substantial risk of bleeding and a significant cost to society (4). Understanding the pathophysiological mechanisms of VT recurrence may facilitate the identification of groups of patients at lower risk of recurrence who do not require these treatments. While about 30 loci are now well established to be associated with the genetic susceptibility to VT (5,6), less is known about the genetic susceptibility to VT recurrence which likely differs from that of first VT (7). Among VT disease loci, the *ABO* locus, coding for the *ABO* blood groups, is one of the most important due to the magnitude of the genetic effects associated with the A1 and B at-risk blood groups and their prevalence in the general population (8–10). To date, few studies have explored the effect of *ABO* blood types on VT recurrence risk (11–14). These analyses have generally been conducted in studies of moderate size with few recurrent events and have often relied on serological measurement of blood groups. Recently, our group showed that molecularly defined blood groups are more reliable than serological measurements (15). In this work, we wish to investigate the effect of molecularly defined *ABO* blood groups on the risk of recurrence in two large VT cohorts, the MARTHA and MEGA studies (16,17).

In MEGA, participants were included at the time of their first VT which represents the beginning of the at risk period for the recurrence. To study the risk of first recurrence a standard time-to-event analysis among which the Cox model is the most popular one (18) can be used. The MARTHA study has a different design since it included all subjects who visited a Thrombophilia centre in Marseille (France) between 1994 and 2012 and had a history of VT (possibly many years before inclusion). Information on recurrence post-inclusion was collected at a follow-up visit several years later but many participants had already experienced a VT recurrence at the time of inclusion. The standard practice is to consider only the information about events that occurred post-inclusion, while excluding patients that have experienced the event of interest (VT recurrence) before their inclusion (19,20). In the framework of recurrent events, it has also been proposed to analyse all recurrences that fall within the observation window after the inclusion and to stratify the analyses according to the number of events before inclusion (21). Others have proposed to study only the occurrence of a recurrence during the observation time without excluding patients with a recurrence prior to inclusion and without differentiating between them (22). In that case, the analysis focused on the association between risk factors measured at inclusion and the risk of recurrence (not necessarily the first one) during the observation time window.

When the risk factors are time-dependent variables such as biological measurements, the analysis including only the events occurring after the collection of the studied variables is needed as the exposure must be measured before the event occurrence in order to avoid bias due to reverse causality. However, when the explanatory variables do not change over time, as genetic factors, this bias is avoided. When the dates of the events that occurred before the inclusion in the study are known, considering this information in the analysis could greatly increase the statistical power. However, specific data analysis procedures should be considered to avoid selection bias by death.

In order to be able to efficiently analyse the impact of *ABO* blood groups on first VT recurrence in MARTHA, a weighted survival analysis is here proposed to enable the joint analysis of patients with or without recurrence prior inclusion in the study.

MATERIALS AND METHODS

MARTHA study

This work was motivated by the identification of genetic risk factors for VT recurrence in the MARseille THrombosis Association (MARTHA) study (23,24). MARTHA includes 2,837 unrelated VT patients who had a consultation visit at the Thrombophilia centre of La Timone Hospital in Marseille (France) between 1994 and 2012. The inclusion date of patients refers to this visit. All patients with at least one documented VT and free of any chronic conditions and of any well characterized genetic risk factors including homozygosity for Factor V Leiden or Factor II 20210A, protein C, protein S and antithrombin deficiencies, and lupus anticoagulant, were eligible. As an ancillary genetic study, a subsample of 1,592 MARTHA patients have been typed by a high density genotyping arrays, referred thereafter as the MARTHA GWAS subsample (where GWAS stands for Genome Wide Association Study) (25). The MARTHA GWAS sub-study was further extended over the 2013-2018 period and patients were re-contacted to gather information on post-inclusion VT events.

MARTHA GWAS sub-study and VT recurrence

The previous application of standard quality control procedures on the genome wide genotype data of MARTHA participants has led to the selection of 1,542 VT patients for genetic analyses (25). From these remaining individuals, we further excluded patients with autoimmune disease or cancer at inclusion, or with missing information on time to VT recurrence for concerned patients. Finally, 1,518 VT patients were left for the VT recurrence analysis. Among these patients, 411 already had at least one VT recurrence before inclusion. The dates, types (DVT or PE) and provoked characters of the first VT and first recurrence were collected. During the 2013-2018 period, patients were re-contacted via phone call, mail questionnaire or medical visit, to gather information on post-inclusion VT. Among the 1,107 patients with a unique VT at inclusion, 846 (76%) could be re-contacted which led to the identification of 160 additional first recurrences. At the end of the second phase, information on the vital status of

non-responders was obtained either through the Répertoire National d'Identification des Personnes Physiques (RNIPP) or medical data. Vital status was finally available for 1,380 individuals including 73 deaths.

Ambispective design

As the current project aims to assess the effect of *ABO* polymorphisms on the risk of recurrence after a first VT, 4 different types of MARTHA participants can be distinguished (**Fig. 1**). Case 1 corresponds to patients who had a single VT before inclusion in the study and who were followed up during the recontact phase within which no recurrent event was observed. Case 2 represents patients who had a single VT before inclusion and experienced a recurrent event which was collected during the recontact phase. Case 3 corresponds to patients with a single VT at inclusion for whom no follow-up information was collected during the recontact phase (i.e lost to follow-up). Finally, Case 4 represents patients who had both a first VT and a recurrence before the inclusion in the study. For each of these 4 situations, the at-risk period, a key element in the analysis of recurrent data that represents the period of time that contributes to the estimation of recurrence risk, is shown in grey in **Fig. 1**.

In a standard cohort analysis, only the post-inclusion period of patients from Cases 1&2 (represented with dotted lines), thereafter referred to as the "*prospective sample*", would be used to investigate risk factors for recurrence. However, since genetic polymorphisms are fixed at birth, all cases of patients can contribute to the analysis of the genetic susceptibility of VT recurrence, considering the first VT as the starting point of the analysis (non-solid lines). This last comment also holds for non-genetic variables available at the time of first VT that are fixed over time such as sex and age at first VT. In the following, we will refer to the "*ambispective sample*" (26) when the four situations are simultaneously considered as it includes both pre- and post-inclusion VT recurrences, that is recurrences which occurred before or within the observation window.

Finally, the MARTHA *prospective sample* was composed of 846 patients including 160 VT recurrences and the extended *ambispective sample* involved 1,518 patients including 571 recurrent VT.

Statistical modelling of recurrent events using weighted Cox model

The Cox proportional-hazards model is a popular semi-parametric model proposed by Cox in 1972 (27). The relationship between the instantaneous risk function (or hazard function) associated with the occurrence of an event and the vector Z of explanatory variables can be written as follows: $\lambda(t, Z, \beta) = \lambda_0(t)\exp(\beta^\top Z)$ where β is the vector of regression coefficients and $\lambda_0(t)$ represents the baseline hazard function.

In order to account for a possible selection bias due to mortality induced by the selection of MARTHA participants, we are proposing a weighted Cox analysis with weights defined by the inverse of the survival probability of individuals up to the time when the information on their possible recurrent event was collected. These weights are used to assign a higher weight to individuals who were less likely to

be observed, e.g. individuals at high risk of death before collection of information on VT recurrence (28). To estimate these weights, we had to model the risk of death in the MARTHA population of VT survivors, using the information on the vital status available for 1,380 patients. As age is the main risk factor for death, we estimated a delayed-entry Cox model with age as time-scale allowing non-parametric modelling of the age effect. For this analysis, subjects contributed from their age of inclusion in the study to their death or last information on the vital status. Estimated parameters from this model were used to compute for all MARTHA patients their individual survival probabilities up to the appropriate time point according to their own clinical and covariates information. While for Cases 1&2 patients, the collection of information on VT recurrence is conditional to the survival of patients up to the recontact date, the collection of this information for Case 3&4 patients is conditional to their survival up to their inclusion in the study.

Once these weights are computed, they can be used in a weighted Cox model for the risk of first recurrence, with delay since the first VT as time scale, to analyse both *prospective* and *ambispective* samples. As the prospective analysis considers only post-inclusion events, a model with delayed-entry at the time of inclusion was estimated. This is not necessary for the *ambispective* analysis as all available information since the first VT is then considered. Once Hazard Ratio (HR) association parameters are obtained, their variance can be estimated using the robust method accounting for the within-subject correlation induced by the weights (29). The weights were computed and standardized using the survival probabilities so that their sum corresponds to the studied sample size with the following formula:

$$w_i = \frac{1/s_i}{\sum_{i=1}^{N} \frac{1}{s_i}}$$

With w_i the weight of i^{th} individual, s_i the survival probability of the i^{th} individual at its own data collection time and N the studied sample size. This approach is implemented with the *survival* package of the R version 3.6.1 environment (30,31).

The MEGA study

Briefly, MEGA is a case-control study for VT that includes almost 4,900 patients who were included for their first VT between 1999 and 2004 (32). Among them, 1,289 VT cases had available genetic data. Between 2008 and 2009, questionnaires were sent to patients to gather information on a possible VT recurrence. From the 1,289 VT patients, we excluded 9 individuals who died before the re-contact phase, 17 individuals with missing information on the provoked character of the first VT event and 9 individuals who were homozygous for the factor V Leiden in order to match to MARTHA exclusion criteria. Six patients from the MEGA study (0.5%) were further excluded as it was not possible to unambiguously determine their *ABO* blood group (see next paragraph). Finally, 1,248 MEGA patients including 428 recurrences were included in the analysis. As these patients were included for their first event, a Cox model in which the delay since the first event was employed as time scale to investigate risk factors of first VT recurrence.

ABO blood groups genetic determination

Five *ABO* polymorphisms were investigated in order to infer *ABO* blood groups. Following recent recommendations (15), the rs8176719-delG was used to tag for O1, the rs41302905-T allele for O2, the rs2519093-T for A1, the rs1053878-A for A2 and the rs8176743-T allele for B.

As MARTHA and MEGA participants have been typed by high-throughput genotyping arrays and imputed on the 1000G Phase I Integrated Release Version 2 Haplotypes, we used best-guessed genotypes from imputed data to infer *ABO* blood groups (25,33). Note that all 5 polymorphims have imputation quality greater than 0.9 in MARTHA and in MEGA. It was possible to infer *ABO* haplotypes and pair of haplotypes without ambiguity for 1,504 (99.1%) and 1,248 (99.5%) MARTHA and MEGA participants, respectively. Finally, the MARTHA *prospective sample* was composed of 839 individuals including 159 recurrences, the extended MARTHA *ambispective sample* included 1,504 among which 565 recurrences were observed and the MEGA sample was composed of 1,248 individuals including 428 recurrences. A detailed flow chart of the MARTHA sub- samples is presented in **Supplementary Figure S1**.

Modelling strategy

Association of *ABO* blood groups with first recurrence was tested assuming additive effects of *ABO* tagging polymorphims, using the O1 group as a reference. Analyses were adjusted for sex, provoked status of the first VT (corresponding to the presence of a risk factor which temporarily promotes VT such as pregnancy or surgery), age at first VT, type of the first VT (DVT or PE) and the first 4 principal components derived from the GWAS genotypic data, in accordance with the literature (7).

Finally, *ABO* association parameters from the MARTHA *ambispective* and MEGA analyses were meta-analysed using a fixed-effects model (Mantel-Haenszel methodology) to highlight the observed trends (34).

RESULTS

Population characteristics

The main characteristics of the MARTHA *prospective* and extended *ambispective* samples are shown in **Table 1**. MARTHA patients were included in the study at approximately 47 years old with a mean age at the first VT around 41. Patients were included in average 6 years after their first VT. The distribution of the age at enrolment and the delay between enrolment and the first VT are provided in **Supplementary Figures S2 & S3**. For approximately 80% of patients, the first VT was DVT and in two-thirds of individuals the first VT was provoked. One-third of patients were male, with a higher proportion of men in those with recurrences. On average, patients were followed for 9 years and this time was longer for patients without recurrence in both samples, since follow-up ends at the occurrence

of a recurrence. Regarding *ABO* blood groups, O1 was the most frequent (~50%) followed by A1 (~33%), B (~9%), A2 (~6%) and finally O2 (~2%).

A description of the principal characteristics of the MEGA participants is provided in **Table 2**. The main differences with MARTHA sample are a higher proportion of men (49%), a higher age at first VT (~48yrs), a lower rate of DVT (61%) and a shorter (~5yrs) follow-up. The delay between inclusion and the first VT is not presented as only incident VT cases were recruited.

The sample used to estimate the risk of death in MARTHA was composed of 1,380 patients among whom 73 deaths were observed (**Supplementary Table S1**). The mean time of follow-up according to the last known vital status was around 12 years and other characteristics were similar to the MARTHA *ambispective* sample.

Risk of death estimation

We estimated the risk of death in MARTHA with a delayed-entry Cox model (**Supplementary Figure S4**). The explanatory variables of this model were sex, provoked character of the first VT, age at the first VT and the first four principal components of the population stratification. Men had an HR (95% Confidence Interval) for death of 1.44 (0.87-2.40) whereas the provoked character of the first VT (HR=0.42 (0.24-0.74)) and a higher age of first VT (HR=0.98 (0.96-1.00) per year) were associated with reduced risk of death.

Using this model, we estimated the survival probability of patients up to the time at which information on their possible VT recurrence was collected. As described in *Methods* section, weights were based on survival probabilities and their range varied between 0.9 and 2.3 (**Supplementary Figure S5**).

Clinical variables and VT recurrence risk

As a first step, we assessed the association of non-time dependent clinical variables on the risk of first VT recurrence in the MARTHA *prospective* and *ambispective* samples (**Table 3**). In the *prospective* analysis of 839 subjects including 159 recurrences, male sex was associated with an increased risk of VT recurrence (HR=1.47 (1.03-2.09)). Other variables were not significantly associated with recurrence, but a trend was observed for the provoked character of the first VT that tends to be protective (HR=0.69 (0.48-1.00)).

The analysis of the same variables performed in the extended *ambispective* sample allows to refine some of these observations with a higher power. Male sex was still associated with a higher risk of recurrence HR=1.65 (1.36-2.01); and older age at first VT appeared as deleterious (HR=1.08 (1.02-1.15) for a 10 years increase). Conversely, we did not find any trend for the provoked status of the first VT (HR=0.99 (0.80-1.23)).

In MEGA, male sex was also associated with an increased risk of recurrence (HR=1.81) but older age was not. Besides, the provoked status of the first VT was significantly associated with a decreased risk

of recurrence (HR=0.61), as initially observed in the MARTHA *prospective* analysis but not confirmed in the *ambispective* analysis.

Lastly, even if the type of first VT (DVT vs PE) was not significantly associated with VT recurrence in neither of the two studies, the same trend for a higher risk of recurrence associated with DVT was observed in the MARTHA *ambispective* (HR=1.17) and the MEGA (HR=1.15) samples. The meta-analysis of these two HRs yielded a combined HR of 1.16.

ABO blood groups

In the MARTHA *prospective* sample, we observed a significant association of A1 blood group compared with O1 on the risk of first VT recurrence (HR=1.32 (1.02-1.70); p=0.035) which was confirmed in the analysis of the *ambispective* sample (HR=1.15 (1.00-1.32); p=0.045) (**Table 3**). The same trend was observed for the A2 group but did not reach statistical significance (HR=1.27 (0.98-1.64); p=0.061 in the *ambispective* analysis). In MEGA, only A1 was significantly associated with a higher risk of VT recurrence (HR=1.21; p=0.018). No evidence for association with VT recurrence was observed for B and O2 groups in either MARTHA or MEGA.

Finally, based on the meta-analysis of the results observed in the MARTHA *ambispective* and MEGA samples, the risk of VT recurrence associated with *ABO* blood groups compared with O1 were HR=1.18 ($p=4.2x10^{-3}$), HR=1.19 (p=0.06), HR=1.01 (p=0.90) and HR=1.03 (p=0.88) for A1, A2, B and O2, respectively.

Since MARTHA *ambispective* and MEGA samples slightly differed with respect to the proportion of DVT events at first VT, the age at first VT, and the delay of follow-up (**Tables 2 & 3**), we further assessed whether the observed association of ABO blood groups was consistent according to these variables. No evidence for heterogeneity was observed whether for type of first VT (**Supplementary Table S2**), for age at first VT (**Supplementary Table S3**) or delay of follow-up (**Supplementary Table S4**).

DISCUSSION

The objective of this work was to investigate the risk of VT recurrence associated with *ABO* blood groups in two large cohorts of middle-aged VT patients, MARTHA and MEGA, the former being built upon an ambispective design.

To achieve this goal, we had first to propose a weighted approach to analyse non-time dependent risk factors (such as genetic polymorphisms) of an event which could have occurred in patients before their inclusion in the study. This approach was mandatory to maximize the power of the MARTHA study as about 70% of first VT recurrences in MARTHA occurred in patients before their inclusion in the study. The proposed modelling relies on a weighted Cox model where the use of weights allows to limit the selection bias associated with the use of pre-inclusion events and thus to gain statistical power by jointly analysing pre- and post-inclusion recurrent events. This method differs from the weighting approach for

repeated events proposed to deal with event-dependent sampling (35). Indeed, the inclusion in MARTHA depends on an event, the first VT, which is not the outcome of interest; the first VT defines the beginning of the period at risk for the recurrence. Our weighting approach handles potential bias due to mortality until the time of data collection for the recurrence.

Our proposed weighted estimation approach is unbiased if the weights are well-specified which means that the Cox model for death is correct. In this work, the death model has two main limitations. First, as the information on VT recurrence was often missing for subjects who died, it was not possible to include VT recurrence (and other possible unknown variables) as a risk factor in our model for death risk. Second, we assumed the proportionality of the risk of death and did not account for the calendar time that could modify either the baseline risk of death or the association with risk factors for death. Moreover, as the number of death during the follow-up in MARTHA is quite small, a Monte Carlo analysis was performed to evaluate the sensitivity of the results to the uncertainty on the weights (description is available in the Supplement). Despite some slight variability in HRs' estimates (especially for O2 group), the overall results remain unchanged.

In this work, we were interested in the association of ABO blood groups with the risk of first VT recurrence and not with the risk of multiple VT recurrences. Indeed, at inclusion in MARTHA, only detailed information on first VT and possibly first recurrence was collected, preventing us from investigating the association with multiple recurrences. Besides, the analysis of such multiple events would require more complex modeling that would take into account the correlation of repeated events (36).

The analysis of these two studies, totaling 2,752 VT patients including 993 recurrences, revealed that the A1 and A2 blood groups were both associated with a moderate increased risk of VT recurrence, HR ~1.20 for both, compared with O1. Note that, likely because of the modest frequency of the A2 blood group (~5%), the association was only marginally significant (p=0.06). Some studies have already investigated the association of *ABO* blood groups with the risk of VT recurrence (11–14), but often with a moderate sample size or using serological *ABO* phenotypes whereas we here used genetically defined *ABO* blood groups which has been shown to be more efficient to capture the effect of *ABO* on VT risk (15). Our results are consistent with those showing a higher risk of recurrence in non-OO patients (12–14). However, they are discordant with the study of *Baudouy et al.* who found a higher risk of VT recurrence in B blood group patients (11), while no association (HR=1.01, p=0.90) was observed in our work. This lack of association in our work is unlikely due to a power issue as the B blood group was more frequent than A2 which was significantly associated here with VT recurrence. Beyond its rather modest size (N=100) and the analysis of serological *ABO* phenotypes, the work of *Baudouy et al.* focused on patients with PE as first VT. A stratified analysis of *ABO* blood groups with recurrence according to the type of the first event (DVT or PE) did not reveal in our work any evidence for specific sub-group *ABO* effects (**Supplementary Table S2**).

MARTHA and MEGA are composed of middle-aged VT patients, with average age at first VT event ~45 yrs. While the association of ABO blood groups with VT recurrence was consistent between patients with age at first event lower and higher than 45yrs (**Supplementary Table S3**), our study is not well-suited to assess whether the observed ABO association also holds in older ages. Our results cannot then be generalizable to older populations and further studies are mandatory to investigate this issue.

As the proposed *ambispective* modelling is only valid for analysing non-time dependent variables, we could not adjust the *ABO* blood group's effect on biological variables that have only been measured at the time of inclusion, such as von Willebrand Factor (vWF). Adjusting for vWF plasma levels would have allowed us to determine whether the effect of *ABO* on VT recurrence is mediated by vWF. This is however unlikely as the observed pattern of association of *ABO* blood groups with recurrence does not match the known associations between *ABO* blood groups and vWF plasma levels (15). Conversely, the observed pattern matches with the one observed between ABO blood groups and plasma levels of Intercellular Adhesion Molecule 1 (ICAM1) where both A1 and A2 groups associate with ICAM1 levels, but not B (15). These observations suggest that the biological factors involved in the association of ABO blood groups with VT recurrence differ from those involved in their relation with incident VT. More than 50 plasma proteins have been shown to be under the genetic influence of the *ABO* locus (37). Determining which of them are associated with the risk of incident and/or recurrent VT merit further deep investigations. Finally, we could not adjust our analysis for the familial history of VT as the available information in MARTHA refers to the presence of a history at the time of inclusion and many recurrences arose earlier.

Nevertheless, our modelling enabled us to assess the impact on the risk of VT recurrence of several clinical variables that are fixed after the first VT event such as age at first VT and the type of first VT (DVT vs PE; provoked vs unprovoked). Consistent in MARTHA and in MEGA were the associations of male sex and DVT as first VT with an increased risk of first VT recurrence, confirming previous observations (7,38). However, we did not observe consistent results with respect to age at first VT nor with the provoked status of the first VT. For the effect of the provoked character on VT recurrence, the different trends observed can be due to the different design and sample selection between MARTHA and MEGA. Indeed, participants in MARTHA were included for at least one previous VT which may have occurred more than fifty years before their inclusion (Mean=6years; Standard Error=10years) whereas in MEGA, patients were recruited at the time of their first VT. Besides, the definition of the provoked character slightly differs between MARTHA and MEGA (**Supplementary Table S5**). We also observed some differences between MARTHA *prospective* and *ambispective* that might be due to the calendar time which has not been taken into account in our work. We are aware that the differences in the management, prevention and identification of VT events may have masked the association between the provoked character of the first event and VT recurrence in the *ambispective* analysis. Indeed, among the 25% of MARTHA patients that had their first VT before the start of the study (1994), for 80% of them the VT was provoked. Whereas in the remaining sample of 75% of MARTHA patients,

the first VT was provoked in only 62% of cases. Of note, when we restrict the analysis to recurrent events that occur within less than 2 years after the first event, the provoked status of first VT was protective against recurrence, consistently in MARTHA and MEGA (**Supplementary Table S4**). These results are in line with previous findings from a 2-years follow-up study (39). Furthermore, the association between A1 and A2 blood groups with VT recurrence remain unchanged when focusing on patients whose first VT occurred after the start of the MARTHA study. Altogether, we feel that such differences may have modest impact when one is interested in genetic factors as illustrated here with the consistent patterns observed for *ABO* blood groups in both MARTHA *prospective* and *ambispective* samples as well as in MEGA (**Supplementary Table S6**).

CONCLUSION

This study demonstrated that both A1 and A2 blood groups are associated with increased risk of VT recurrence in middle-aged patients. This finding was made possible thanks to a weighting approach to study non-time dependent risk factors while integrating not only post-inclusion events but also those that occurred before inclusion. This modelling finds an immediate field of application to genetic association studies for time-to-events in cohorts where follow-up information for deaths is available and events before inclusion are collected.

LIST OF ABBREVIATIONS

DVT: Deep Vein Thrombosis
GWAS: Genome Wide Association Study
HR: Hazard Ratio
PE : Pulmonary Embolism
RNIPP: Répertoire National d'Identification des Personnes Physiques
VT: Venous Thrombosis
vWF: von Willebrand Factor

STATEMENTS AND DECLARATIONS

Ethics approval

Research have been performed in accordance with the Declaration of Helsinki. The MARTHA study was initially approved by the local ethic committee "Mediterranean I Committee for the Protection of Individuals" (reference: 12 61). The MEGA study was approved by the local ethic committee "Medical Ethics Committee of the Leiden University Medical Center".

All experimental protocols to study the genetics of VT recurrence were approved by the local ethic committee "Mediterranean I Committee for the Protection of Individuals" (reference: 12 61) for

MARTHA and by the local ethic committee "Medical Ethics Committee of the Leiden University Medical Center" for MEGA.

Consent to participate

Written informed consent to participate was obtained from all MARTHA and MEGA participants.

Availability of data and materials

Summary statistic of the data analyzed in this work are all provided in the main manuscript document and its supplements.

Competing interests

The authors declare that they have no competing interests.

Fundings

GM and D-AT are supported by the EPIDEMIOM-VT Senior Chair from the University of Bordeaux initiative of excellence IdEX.
The MARTHA project was supported by a grant from the Program Hospitalier de la Recherche Clinique and the GENMED Laboratory of Excellence on Medical Genomics [ANR-10-LABX-0013], a research program managed by the National Research Agency (ANR) as part of the French Investment for the Future.
The MEGA (Multiple Environmental and Genetic Assessment of risk factors for venous thrombosis) study was supported by the Netherlands Heart Foundation (NHS98.113 and NHS208B086), the Dutch Cancer Foundation (RUL 99/1992), and the Netherlands Organization for Scientific Research (912-03-033|2003).

Authors' contributions

GM performed statistical analyses and wrote the first draft of the paper. HJ-G and D-AT supervised the statistical analyses and revised the paper. LG, AvHV, MI-K, MB, J-FD, FR and P-EM participated to data collection. FR, P-EM and D-AT designed the study. All authors read and approved the final manuscript.

Acknowledgments

GM benefited from the EUR DPH, a PhD program supported within the framework of the PIA3 (Investment for the future). Project reference 17-EURE-0019.
This project was carried out in the framework of the INSERM GOLD Cross-Cutting program (P-EM, D-AT).
Statistical analyses benefited from the CBiB computing centre of the University of Bordeaux.

REFERENCES

1. Oger E. Incidence of venous thromboembolism: a community-based study in Western France. EPI-GETBP Study Group. Groupe d'Etude de la Thrombose de Bretagne Occidentale. Thromb Haemost. 2000 May;83(5):657–60.

2. White RH. The epidemiology of venous thromboembolism. Circulation. 2003 Jun 17;107(23 Suppl 1):I4-8.

3. Heit JA. Epidemiology of venous thromboembolism. Nat Rev Cardiol. 2015 Aug;12(8):464–74.

4. Ruppert A, Steinle T, Lees M. Economic burden of venous thromboembolism: a systematic review. J Med Econ. 2011 Jan;14(1):65–74.

5. Lindström S, Wang L, Smith EN, Gordon W, van Hylckama Vlieg A, de Andrade M, et al. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. Blood. 2019 Nov 7;134(19):1645–57.

6. Klarin D, Busenkell E, Judy R, Lynch J, Levin M, Haessler J, et al. Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. Nat Genet. 2019 Nov;51(11):1574–9.

7. Authors/Task Force Members, Konstantinides SV, Torbicki A, Agnelli G, Danchin N, Fitzmaurice D, et al. 2014 ESC Guidelines on the diagnosis and management of acute pulmonary embolism. Eur Heart J. 2014 Nov 14;35(43):3033–80.

8. 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 Can Med Assoc J J Assoc Medicale Can. 2013 Mar 19;185(5):E229-237.

9. Trégouët D-A, Heath S, Saut N, Biron-Andreani C, Schved J-F, Pernod G, et al. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. Blood. 2009 May 21;113(21):5298–303.

10. Franchini M, Mannucci PM. ABO blood group and thrombotic vascular disease. Thromb Haemost. 2014;112(12):1103–9.

11. Baudouy D, Moceri P, Chiche O, Bouvier P, Schouver E-D, Cerboni P, et al. B blood group: A strong risk factor for venous thromboembolism recurrence. Thromb Res. 2015 Jul;136(1):107–11.

12. Dentali F, Franchini M. Recurrent venous thromboembolism: a role for ABO blood group? Thromb Haemost. 2013;110(12):1110–1.

13. Limperger V, Kenet G, Kiesau B, Köther M, Schmeiser M, Langer F, et al. Role of prothrombin 19911 A>G polymorphism, blood group and male gender in patients with venous thromboembolism: Results of a German cohort study. J Thromb Thrombolysis. 2021 Feb 1;51(2):494–501.

14. Gándara E, Kovacs MJ, Kahn SR, Wells PS, Anderson DA, Chagnon I, et al. Non-OO blood type influences the risk of recurrent venous thromboembolism: A cohort study. Thromb Haemost. 2013;110(12):1172–9.

15. Goumidi L, Thibord F, Wiggins KL, Li-Gao R, Brown MR, van Hylckama Vlieg A, et al. Association of ABO haplotypes with the risk of venous thrombosis: impact on disease risks estimation. Blood. 2020 Dec 22;(blood.2020008997).

16. Morange P-E, Oudot-Mellakh T, Cohen W, Germain M, Saut N, Antoni G, et al. KNG1 Ile581Thr and susceptibility to venous thrombosis. Blood. 2011 Mar 31;117(13):3692–4.

17. Rosendaal F. Air travel and thrombosis. Pathophysiol Haemost Thromb. 2002 Sep 1;32:341–2.

18. de Haan HG, van Hylckama Vlieg A, Germain M, Baglin TP, Deleuze J-F, Trégouët D-A, et al. Genome-Wide Association Study Identifies a Novel Genetic Risk Factor for Recurrent Venous Thrombosis. Circ Genomic Precis Med. 2018 Feb;11(2).

19. Ahmad A, Sundquist K, Palmér K, Svensson PJ, Sundquist J, Memon AA. Risk prediction of recurrent venous thromboembolism: a multiple genetic risk model. J Thromb Thrombolysis. 2019 Feb;47(2):216–26.

20. Hara M, Sakata Y, Nakatani D, Suna S, Usami M, Matsumoto S, et al. Reduced risk of recurrent myocardial infarction in homozygous carriers of the chromosome 9p21 rs1333049 C risk allele in the contemporary percutaneous coronary intervention era: a prospective observational study. BMJ Open. 2014 Aug 13;4(8):e005438–e005438.

21. Giorda CB, Avogaro A, Maggini M, Lombardo F, Mannucci E, Turco S, et al. Recurrence of cardiovascular events in patients with type 2 diabetes: epidemiology and risk factors. Diabetes Care. 2008 Nov;31(11):2154–9.

22. Ianotto J-C, Chauveau A, Mottier D, Ugo V, Berthou C, Lippert E, et al. JAK2V617F and calreticulin mutations in recurrent venous thromboembolism: results from the EDITH prospective cohort. Ann Hematol. 2017 Mar;96(3):383–6.

23. Oudot-Mellakh T, Cohen W, Germain M, Saut N, Kallel C, Zelenika D, 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 Apr;157(2):230–9.

24. Trégouët DA, Delluc A, Roche A, Derbois C, Olaso R, Germain M, et al. Is there still room for additional common susceptibility alleles for venous thromboembolism? J Thromb Haemost JTH. 2016 Sep;14(9):1798–802.

25. Germain M, Saut N, Greliche N, Dina C, Lambert J-C, Perret C, et al. Genetics of Venous Thrombosis: Insights from a New Genome Wide Association Study. PLOS ONE. 2011 Sep 27;6(9):e25581.

26. Modak A, Suthar R, Sharawat IK, Sankhyan N, Sahu JK, Malhi P, et al. An Ambispective Cohort Study to Assess Seizure Recurrences in Children with Calcified Parenchymal Neurocysticercosis. Am J Trop Med Hyg. 2019 Oct 2;101(4):812–20.

27. Cox DR. Regression Models and Life-Tables. J R Stat Soc Ser B Methodol. 1972;34(2):187–220.

28. Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias: Epidemiology. 2004 Sep;15(5):615–25.

29. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiol Camb Mass. 2000 Sep;11(5):561–70.

30. Therneau TM, Elizabeth A, Cynthia C. survival: Survival Analysis. 2020.

31. R Development Core Team. a language and environment for statistical computing: reference index. Vienna: R Foundation for Statistical Computing; 2010.

32. Chinthammitr Y, Vos HL, Rosendaal FR, Doggen CJM. 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):2587–92.

33. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng L-C, et al. Meta-analysis of 65,734 Individuals Identifies TSPAN15 and SLC44A2 as Two Susceptibility Loci for Venous Thromboembolism. Am J Hum Genet. 2015 Apr 2;96(4):532–42.

34. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959 Apr;22(4):719–48.

35. Kvist K, Andersen PK, Angst J, Kessing LV. Event dependent sampling of recurrent events. Lifetime Data Anal. 2010 Oct;16(4):580–98.

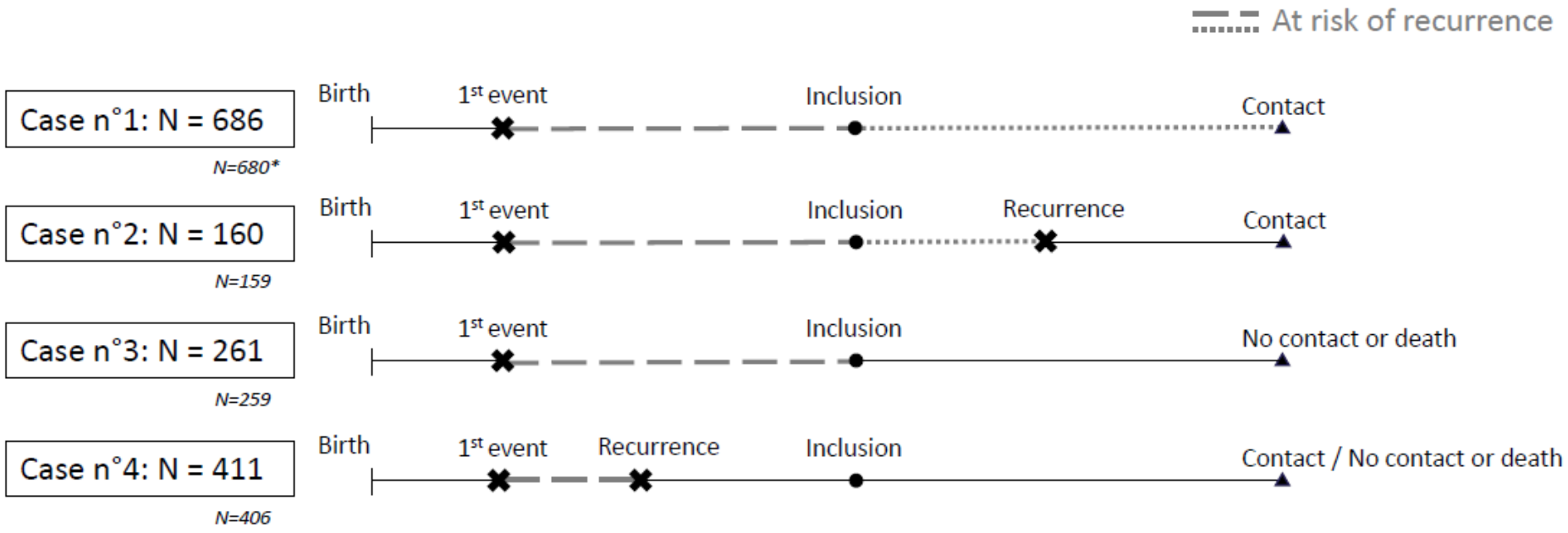
36. Amorim LD, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015 Feb;44(1):324–33.

37. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, et al. Co-regulatory networks of human serum proteins link genetics to disease. Science. 2018 Aug 24;361(6404):769–73.

38. Zhu T, Martinez I, Emmerich J. Venous Thromboembolism: Risk Factors for Recurrence. Arterioscler Thromb Vasc Biol. 2009 Mar;29(3):298–310.

39. Baglin T, Luddington R, Brown K, Baglin C. Incidence of recurrent venous thromboembolism in relation to clinical and thrombophilic risk factors: prospective cohort study. The Lancet. 2003 Aug 16;362(9383):523–6.

Fig1: Illustration of the 4 scenarii of patients included in the MARTHA study



**Corresponding numbers used in the association analysis of ABO blood groups*

Case 1: patients who had a single VT before inclusion in the study and who were followed up during the recontact phase within which no recurrent event was observed. Case 2: patients who had a single VT before inclusion and experienced a recurrent event which was collected during the recontact phase. Case 3: patients with a single VT at inclusion for whom no follow-up information was collected during the recontact phase (i.e lost to follow-up). Case 4: patients who had both a first VT and a recurrence before the inclusion in the study. For all 4 situations, the at-risk period is shown in grey with dotted lines for the post-inclusion period and with dashed lines for the retrospective period.

Table 1: Description of the main characteristics in the prospective and ambispective MARTHA samples

Variables	MARTHA *Prospective sample*			MARTHA *Ambispective sample*		
	Total N=839	**Recurrences N=159**	**Non-recurrences N=680**	**Total N=1,504**	**Recurrences N=565**	**Non-recurrences N=939**
	N (%)	**N (%)**	**N (%)**	**N (%)**	**N (%)**	**N (%)**
Gender						
Men	275 (32.8%)	63 (39.6%)	212 (31.2%)	509 (33.8%)	232 (41.1%)	277 (29.5%)
Age at inclusion (mean ± SD)	45.2 ± 14.9	44.0 ± 13.9	45.5 ± 15.4	47.1 ± 15.4	50.0 ± 14.7	45.3 ± 15.5
Age at the first VT (mean ± SD)	42.3 ± 15.5	41.5 ± 14.4	42.5 ± 15.8	41.0 ± 15.7	40.5 ± 15.1	41.3 ± 16.0
Delay between inclusion and first VT (In years, mean ± SD)	2.9 ± 6.2	2.5 ± 5.3	3.0 ± 6.4	6.1 ± 9.85	9.5 ± 11.3	4.0 ± 8.2
Type of the first VT						
DVT	653 (77.8%)	122 (76.7%)	531 (78.1%)	1,189 (79.1%)	454 (80.4%)	735 (78.3%)
Characteristic of the first VT						
Provoked	544 (64.8%)	93 (58.5%)	451 (66.3%)	993 (66.0%)	368 (65.1%)	625 (66.6%)
Age at the collection of information on recurrence (mean ± SD) †	54.9 ± 15.2	54.9 ± 13.9	54.9 ± 15.5	52.5 ± 15.6	43.0 ± 14.2	52.1 ± 16.4
Delay of follow-up in years ‡ (In years, mean ± SD)	8.8 ± 5.5	6.4 ± 5.3	9.4 ± 5.4	9.7 ± 9.6	7.9 ± 8.9	10.8 ± 9.8
ABO haplotypes						
A1	32.8%	37.4%	31.7%	33.5%	35.8%	32.2%
A2	5.5%	5.3%	5.5%	5.9%	6.9%	5.3%
O1	51.0%	47.5%	51.8%	49.8%	46.9%	51.5%
O2	1.7%	1.6%	1.7%	1.5%	1.5%	1.5%
B	9.1%	8.2%	9.3%	9.3%	8.9%	9.5%

† Since inclusion in the prospective sample, since the first VT in the ambispective sample ;
‡ Refers to the recontact for Cases 1 & 2 and inclusion for Cases 3 & 4 (see Fig1)

Table 2: Description of the main characteristics in the MEGA sample

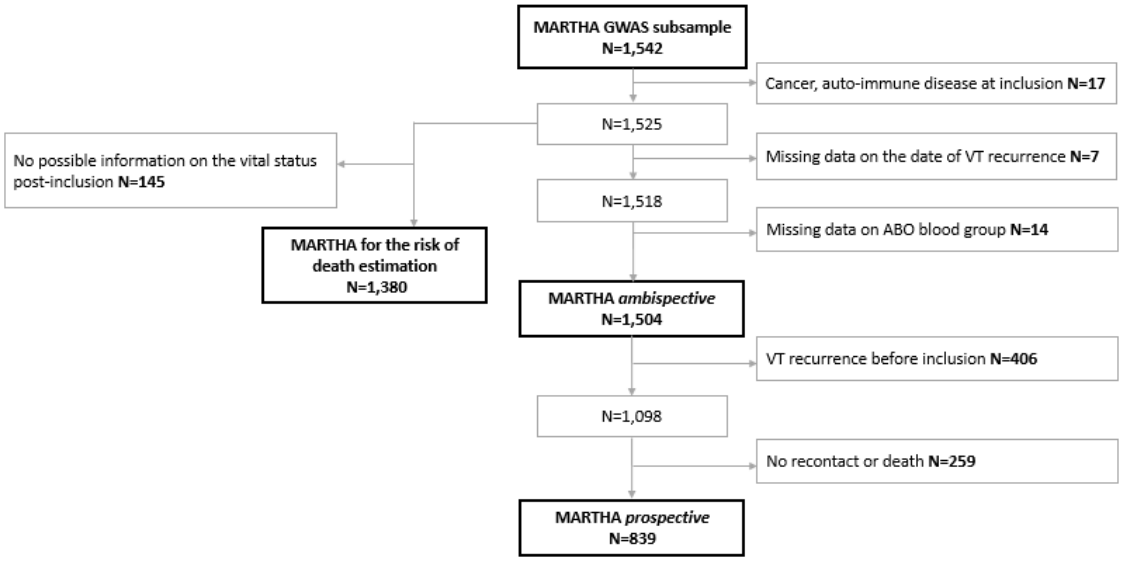
Variables	MEGA *sample*		
	Total N=1,248	**Recurrences N=428**	**Non-recurrences N=820**
	N (%)	**N (%)**	**N (%)**
Gender			
Men	661 (49.0%)	272 (63.6%)	389 (47.4%)
Age at the first VT (mean ± SD)	48.0 ± 12.8	49.8 ± 12.8	47.1 ± 12.8
Type of the first VT			
DVT	763 (61.1%)	270 (63.1%)	493 (60.1%)
Characteristic of the first VT			
Provoked	847 (67.9%)	237 (55.4%)	610 (74.4%)
Delay of follow-up since inclusion (In years, mean ± SD)	5.2 ± 2.9	3.1 ± 2.2	6.3 ± 2.7
ABO haplotypes			
A1	28.9%	31.8%	27.4%
A2	7.0%	7.7%	6.6%
O1	53.4%	50.9%	54.7%
O2	1.8%	1.5%	2.0%
B	8.9%	8.1%	9.3%

Table 3: Association of clinical variables and ABO haplotypes with VT recurrence in MARTHA (prospective and ambispective) and MEGA

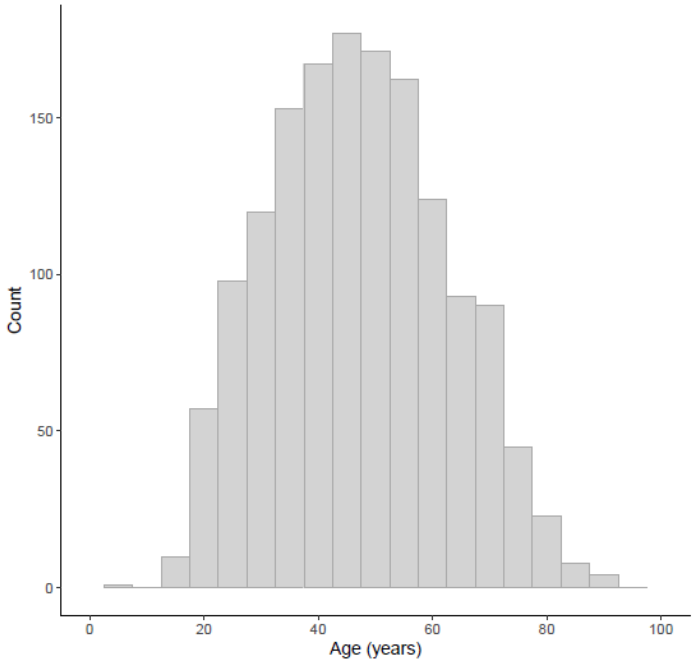
Variables	MARTHA *Prospective* N=839 Nb recurrences=159		MARTHA *Ambispective* N=1,504 Nb recurrences=565		MEGA N=1,248 Nb recurrences=428		Meta-analysis MARTHA Ambispective & MEGA	
	HR (95% CI)	**P**	**HR (95% CI)**	**P**	**HR (95% CI)**	**P**	**HR (95% CI)**	**P**
Gender								
Men	1.47 (1.03-2.09)	0.034	1.65 (1.36-2.01)	4.0×10^{-7}	1.81 (1.46-2.25)	5.9×10^{-8}	1.72 (1.47-2.01)	3.0×10^{-12}
Age at the first VT (10 years increase)	0.91 (0.81-1.02)	0.105	1.08 (1.02-1.15)	0.020	0.99 (0.92-1.07)	0.810	1.05 (0.99-1.11)	0.107
Type of the first VT								
DVT	0.85 (0.60-1.21)	0.368	1.17 (0.96-1.42)	0.140	1.15 (0.95-1.40)	0.160	1.16 (1.01-1.33)	0.036
Characteristic of the first VT								
Provoked	0.69 (0.48-1.00)	0.059	0.99 (0.80-1.23)	0.920	0.61 (0.49-0.76)	6.7×10^{-6}	0.78 (0.67-0.91)	1.2×10^{-3}
ABO haplotypes								
A1	1.32 (1.02-1.70)	0.035	1.15 (1.00-1.32)	0.045	1.21 (1.03-1.42)	0.018	1.18 (1.05-1.33)	4.2×10^{-3}
A2	1.13 (0.68-1.88)	0.644	1.27 (0.98-1.64)	0.061	1.11 (0.86-1.43)	0.409	1.19 (1.00-1.42)	0.062
O1	Reference		Reference		Reference		Reference	
O2	1.05 (0.47-2.35)	0.896	1.19 (0.73-1.94)	0.476	0.86 (0.50-1.49)	0.584	1.03 (0.70-1.52)	0.880
B	1.00 (0.64-1.57)	0.998	1.02 (0.82-1.27)	0.874	1.00 (0.78-1.29)	0.987	1.01 (0.85-1.21)	0.900

HR: Hazard Ratio
CI: Confidence Interval

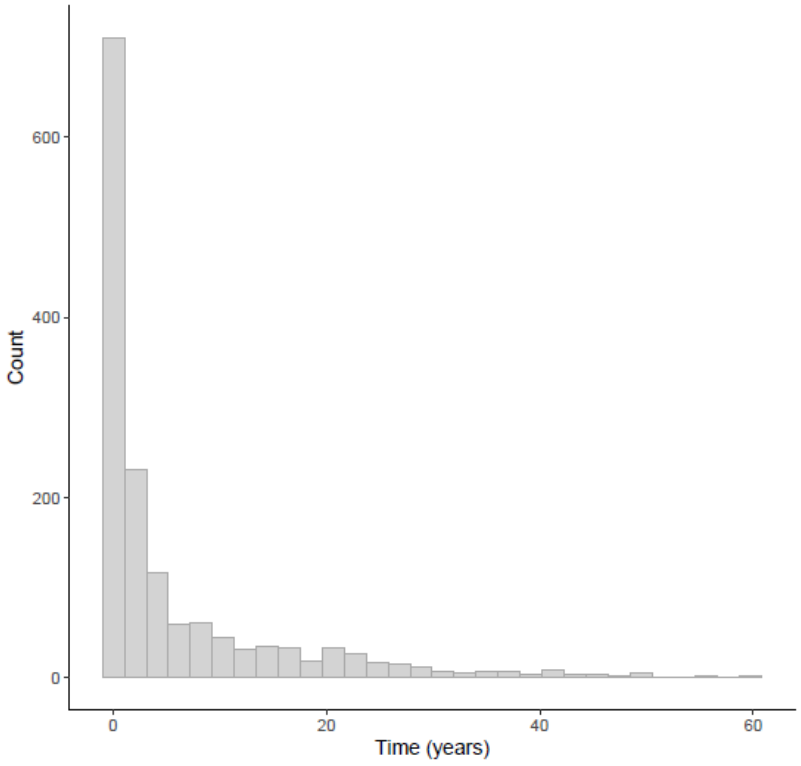
Supplementary Figure S1. Flow chart of the MARTHA sub-samples



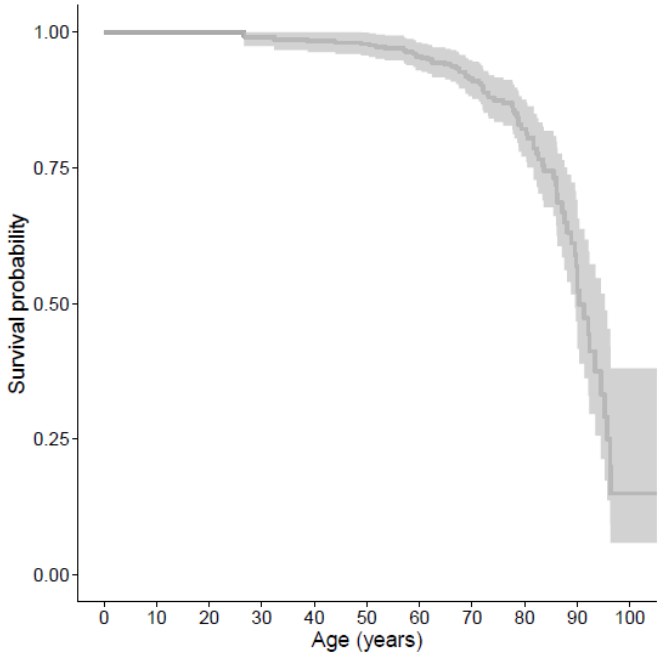
Supplementary Figure S2. Distribution of the age at enrolment in MARTHA participants (N=1,504)



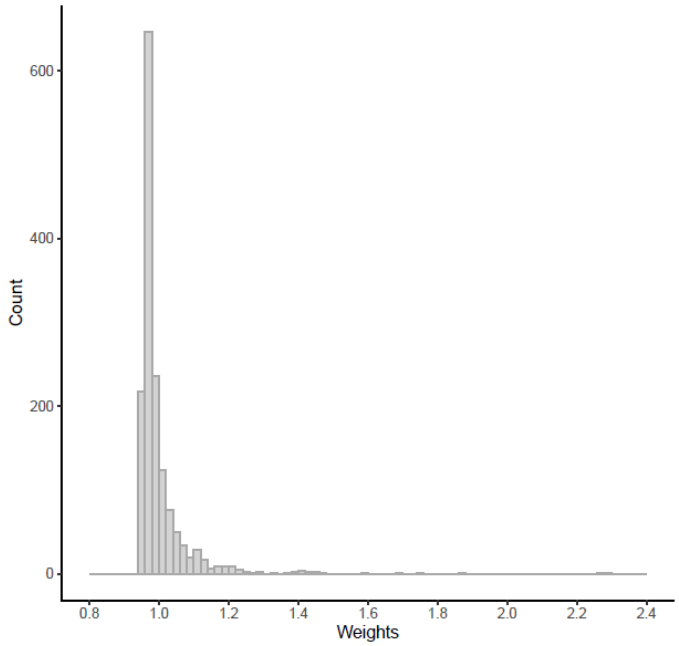
Supplementary Figure S3. Distribution of the delay between enrolment and the first VT in MARTHA participants (N=1,504)



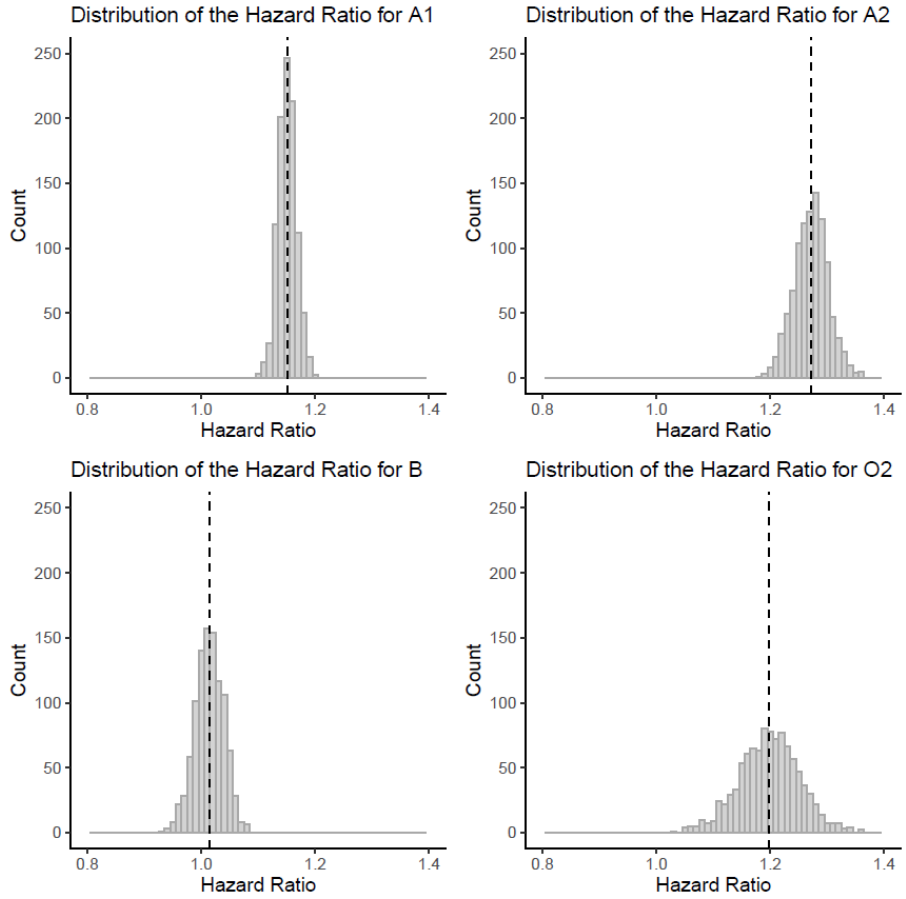
Supplementary Figure S4. Kaplan Meier plot of the survival probability in MARTHA participants with an available follow up (N=1,380 including 73 deaths)



Supplementary Figure S5. Distribution of the estimated weights for the MARTHA participants (N=1,504)



Supplementary Figure S6. Sensitivity of the association of *ABO* blood groups with recurrence according to the weights estimation in MARTHA (N=1,504)



Note: The 4 panels show the distribution of the Hazard Ratio in the Monte Carlo resampling analysis (See Supplementary Text). The dashed line corresponds to the estimated value in the initial model

Supplementary Table S1. Description of the MARTHA sample for the death risk estimation

Variables	Total N=1,380
	N (%)
Gender	
Men	468 (33.9%)
Age at inclusion (mean ± Standard Deviation (SD))	47.1 ± 15.3
Age at the first VT (mean ± SD)	41.3 ± 15.7
Delay between inclusion and first VT (In years, mean ± SD)	5.8 ± 9.6
Type of the first VT	
DVT only	1,087 (78.8%)
Characteristic of the first VT	
Provoked	911 (66.0%)
Delay of follow-up in years* (In years, mean ± SD)	11.8 ± 5.3

*According to the death event

Supplementary Table S2. Association of *ABO* haplotypes with first VT recurrence in MARTHA *ambispective* and MEGA stratified on the type of the first VT

Variables	MARTHA *Ambispective* N=1,504 Nb recurrences=565		MEGA N=1,248 Nb recurrences=428		Meta-Analysis Fixed-effects	
	HR (95% CI)	**P**	**HR (95% CI)**	**P**	**HR (95% CI)**	**P**
ABO haplotypes – PE as first VT	**N=315 ; 111 recurrences**		**N=485 ; 158 recurrences**			
A1	1.10 (0.82-1.48)	0.536	1.38 (1.05-1.82)	0.020	1.24 (1.02-1.51)	0.029
A2	1.87 (1.04-3.37)	0.039	0.79 (0.49-1.27)	0.329	1.11 (0.80-1.55)	0.554
O1	Reference		Reference		Reference	
O2	0.65 (0.13-3.24)	0.600	0.70 (0.28-1.72)	0.434	0.69 (0.36-1.32)	0.250
B	0.85 (0.47-1.53)	0.574	0.83 (0.51-1.36)	0.447	0.84 (0.59-1.20)	0.318
ABO haplotypes – DVT as first VT	**N=1,189 ; 454 recurrences**		**N=763 ; 270 recurrences**			
A1	1.16 (0.99-1.36)	0.063	1.14 (0.94-1.39)	0.211	1.15 (1.00-1.32)	0.045
A2	1.20 (0.91-1.58)	0.180	1.29 (0.94-1.77)	0.104	1.24 (1.00-1.54)	0.059
O1	Reference		Reference		Reference	
O2	1.41 (0.85-2.35)	0.180	0.94 (0.46-1.90)	0.872	1.23 (0.75-2.01)	0.422
B	1.06 (0.84-1.34)	0.612	1.08 (0.79-1.48)	0.637	1.07 (0.86-1.33)	0.566

HR: Hazard Ratio
CI: Confidence Interval

Supplementary Table S3. Association of *ABO* haplotypes with first VT recurrence in MARTHA *ambispective* and MEGA stratified on age at the first VT

Variables	MARTHA *Ambispective* N=1,504 Nb recurrences=565		MEGA N=1,248 Nb recurrences=428		Meta-Analysis Fixed-effects	
	HR (95% CI)	**P**	**HR (95% CI)**	**P**	**HR (95% CI)**	**P**
ABO haplotypes – First VT before 45 years	**N=932 ; 349 recurrences**		**N=487 ; 144 recurrences**			
A1	1.12 (0.94-1.33)	0.201	1.31 (0.99-1.72)	0.055	1.17 (0.97-1.42)	0.109
A2	1.33 (0.95-1.87)	0.098	1.44 (0.94-2.22)	0.098	1.37 (1.01-1.86)	0.043
O1	Reference		Reference		Reference	
O2	1.58 (0.85-2.93)	0.148	0.80 (0.29-2.20)	0.673	1.31 (0.64-2.77)	0.454
B	1.13 (0.88-1.44)	0.344	0.81 (0.50-1.33)	0.411	1.06 (0.75-1.49)	0.762
ABO haplotypes – First VT after 45 years	**N=572 ; 216 recurrences**		**N=761 ; 284 recurrences**			
A1	1.21 (0.97-1.50)	0.091	1.17 (0.96-1.42)	0.129	1.19 (1.03-1.36)	0.018
A2	1.21 (0.84-1.75)	0.310	1.00 (0.72-1.38)	0.999	1.09 (0.87-1.36)	0.476
O1	Reference		Reference		Reference	
O2	0.78 (0.36-1.71)	0.537	0.87 (0.44-1.71)	0.689	0.83 (0.52-1.34)	0.448
B	0.84 (0.55-1.29)	0.430	1.12 (0.82-1.53)	0.485	1.01 (0.81-1.26)	0.921

HR: Hazard Ratio
CI: Confidence Interval

Supplementary Table S4: Associations with VT recurrence according to a censored time of follow-up, in MARTHA *Ambispective* and MEGA

	Censoring after 1 year of follow-up					
	MARTHA **N=1,504** **109 recurrences**		MEGA **N=1,248** **92 recurrences**		META-ANALYSIS **N=2,752** **201 recurrences**	
Variables	HR (95%CI)	P	HR (95%CI)	P	HR (95%CI)	P
Gender (Men)	0.94 (0.60-1.45)	0.766	2.33 (1.42-3.84)	$8.8x10^{-4}$	1.39 (0.98-1.98)	0.066
Age at the first VT (10 years increase)	1.14 (0.99-1.31)	0.074	0.89 (0.75-1.06)	0.202	1.03 (0.91-1.17)	0.600
Type of the first VT (DVT)	1.25 (0.76-2.06)	0.373	1.41 (0.91-2.20)	0.127	1.34 (0.98-1.83)	0.068
Characteristic of the first VT (Provoked)	0.59 (0.38-0.94)	0.026	0.59 (0.37-0.93)	0.024	0.59 (0.43-0.82)	0.001
ABO haplotypes						
A1	1.06 (0.78-1.44)	0.714	1.33 (0.94-1.87)	0.109	1.17 (0.92-1.50)	0.203
A2	1.19 (0.70-2.00)	0.524	1.05 (0.59-1.86)	0.868	1.12 (0.75-1.68)	0.578
O1	Reference		Reference		Reference	
O2	1.30 (0.52-3.26)	0.578	1.05 (0.33-3.37)	0.933	1.20 (0.53-2.73)	0.669
B	1.05 (0.67-1.65)	0.815	0.94 (0.52-1.67)	0.825	1.01 (0.67-1.52)	0.965
	Censoring after 2 years of follow-up					
	MARTHA **N=1,504** **178 recurrences**		MEGA **N=1,248** **169 recurrences**		META-ANALYSIS **N=2,752** **347 recurrences**	
Variables	HR (95%CI)	P	HR (95%CI)	P	HR (95%CI)	P
Gender (Men)	1.23 (0.87-1.73)	0.240	1.95 (1.37-2.77)	$2.2x10^{-4}$	1.54 (1.20-1.97)	$7.4x10^{-4}$
Age at the first VT (10 years increase)	1.11 (0.90-1.38)	0.063	0.96 (0.84-1.09)	0.534	1.00 (0.91-1.10)	0.997
Type of the first VT (DVT)	1.41 (0.93-2.14)	0.110	1.39 (1.01-1.93)	0.046	1.40 (1.11-1.76)	0.004
Characteristic of the first VT (Provoked)	0.66 (0.46-0.95)	0.027	0.68 (0.49-0.96)	0.028	0.67 (0.53-0.85)	0.001
ABO haplotypes						
A1	1.07 (0.84-1.37)	0.591	1.28 (0.99-1.65)	0.062	1.17 (0.97-1.40)	0.096
A2	1.27 (0.85-1.91)	0.247	1.26 (0.85-1.87)	0.257	1.27 (0.96-1.67)	0.099
O1	Reference		Reference		Reference	
O2	1.58 (0.82-3.04)	0.171	0.97 (0.40-2.38)	0.947	1.33 (0.71-2.52)	0.375
B	0.85 (0.57-1.27)	0.427	1.03 (0.68-1.56)	0.886	0.93 (0.70-1.25)	0.641

	Censoring after 3 years of follow-up					
	MARTHA **N=1,504**		MEGA **N=1,248**		META-ANALYSIS **N=2,752**	
	236 recurrences		**233 recurrences**		**469 recurrences**	
Variables	HR (95%CI)	P	HR (95%CI)	P	HR (95%CI)	P
Gender (Men)	1.50 (1.12-2.01)	0.007	1.81 (1.35-2.43)	$8.6x10^{-5}$	1.64 (1.33-2.03)	$3.2x10^{-6}$
Age at the first VT (10 years increase)	1.13 (1.02-1.24)	0.018	0.95 (0.85-1.06)	0.385	1.05 (0.97-1.13)	0.265
Type of the first VT (DVT)	1.45 (1.01-2.09)	0.043	1.26 (0.96-1.65)	0.101	1.33 (1.09-1.61)	0.004
Characteristic of the first VT (Provoked)	0.79 (0.58-1.08)	0.144	0.69 (0.52-0.92)	0.012	0.74 (0.60-0.90)	0.003
ABO haplotypes						
A1	1.03 (0.83-1.27)	0.815	1.23 (0.99-1.53)	0.059	1.12 (0.96-1.31)	0.145
A2	1.22 (0.85-1.74)	0.283	1.18 (0.84-1.67)	0.334	1.20 (0.94-1.53)	0.146
O1	Reference		Reference		Reference	
O2	1.34 (0.72-2.49)	0.359	0.66 (0.27-1.62)	0.366	1.06 (0.57-2.00)	0.849
B	0.83 (0.59-1.17)	0.287	1.01 (0.71-1.43)	0.977	0.91 (0.71-1.17)	0.475
	Censoring after 4 years of follow-up					
	MARTHA **N=1,504**		MEGA **N=1,248**		META-ANALYSIS **N=2,752**	
	270 recurrences		**286 recurrences**		**556 recurrences**	
Variables	HR (95%CI)	P	HR (95%CI)	P	HR (95%CI)	P
Gender (Men)	1.58 (1.20-2.07)	0.001	1.80 (1.37-2.34)	$1.7x10^{-5}$	1.69 (1.40-2.04)	$5.9x10^{-8}$
Age at the first VT (10 years increase)	1.12 (1.03-1.23)	0.013	0.96 (0.87-1.06)	0.417	1.05 (0.97-1.12)	0.215
Type of the first VT (DVT)	1.44 (1.03-2.02)	0.033	1.34 (1.04-1.71)	0.021	1.37 (1.15-1.64)	$3.7x10^{-4}$
Characteristic of the first VT (Provoked)	0.84 (0.62-1.12)	0.228	0.66 (0.51-0.86)	0.002	0.73 (0.61-0.88)	0.001
ABO haplotypes						
A1	0.95 (0.78-1.16)	0.627	1.19 (0.98-1.45)	0.078	1.07 (0.93-1.23)	0.357
A2	1.17 (0.84-1.64)	0.353	1.18 (0.86-1.61)	0.304	1.18 (0.94-1.46)	0.151
O1	Reference		Reference		Reference	
O2	1.14 (0.60-2.14)	0.691	0.73 (0.34-1.56)	0.421	0.95 (0.56-1.62)	0.849
B	0.80 (0.58-1.10)	0.166	1.03 (0.75-1.41)	0.851	0.91 (0.73-1.14)	0.400

	Censoring after 5 years of follow-up					
	MARTHA **N=1,504** **301 recurrences**		MEGA **N=1,248** **334 recurrences**		META-ANALYSIS **N=2,752** **635 recurrences**	
Variables	HR (95%CI)	P	HR (95%CI)	P	HR (95%CI)	P
Gender (Men)	1.49 (1.15-1.93)	0.002	1.83 (1.43-2.34)	$1.7x10^{-6}$	1.66 (1.39-1.98)	$1.3x10^{-8}$
Age at the first VT (10 years increase)	1.13 (1.04-1.23)	0.005	0.95 (0.87-1.04)	0.294	1.05 (0.98-1.12)	0.192
Type of the first VT (DVT)	1.40 (1.02-1.91)	0.036	1.19 (0.95-1.49)	0.126	1.26 (1.07-1.47)	0.005
Characteristic of the first VT (Provoked)	0.93 (0.70-1.23)	0.606	0.62 (0.49-0.79)	$9.8x10^{-5}$	0.74 (0.62-0.87)	$4.4x10^{-4}$
ABO haplotypes						
A1	0.98 (0.81-1.19)	0.866	1.25 (1.04-1.50)	0.017	1.11 (0.98-1.27)	0.103
A2	1.13 (0.81-1.57)	0.470	1.18 (0.88-1.57)	0.268	1.16 (0.94-1.42)	0.166
O1	Reference		Reference		Reference	
O2	1.13 (0.62-2.08)	0.689	0.90 (0.47-1.69)	0.734	1.01 (0.65-1.59)	0.957
B	0.91 (0.68-1.22)	0.532	1.06 (0.79-1.42)	0.700	0.98 (0.80-1.21)	0.868

This table illustrates the variations over time of the associations between clinical variables/ABO haplotypes and VT recurrence. Several times of censoring were considered: from one year to five years since the first VT. For each delay of follow-up (i.e. since the first VT), individuals with no VT recurrence before the dedicated delay were right censored and so considered as controls for VT recurrence at this time point. In the various time windows discussed above, the number of VT recurrences that occured in the prospective phase were respectively: 9 (8%), 24 (13%), 39 (17%), 47 (17%) and 54 (18%).

Supplementary Table S5. Definition of the provoked character in MARTHA and MEGA

MARTHA study	MEGA study
• Surgery within 3 months before VT	• Surgery within 3 months before VT
• Pregnancy/ puerperium within 3 months before VT	• Pregnancy/ puerperium within 3 months before VT
• Oral contraceptive use within 3 months before VT	• Hormone use at the time of VT, including: hormone replacement therapy and hormonal contraceptives
	• Plaster cast within 3 months before VT
• Immobilization for 3 days or more within 3 months before VT	• Immobility in bed, in hospital: Confinement to bed ≥ 3 days in hospital, confinement to bed ≥ 3 days at home, within 3 months before VT
• Long travel (by car >10 hours ; by plane > 5 hours) within 3 months before VT	• Prolonged travel >4 hours within 2 months before VT
• Trauma of the lower limb within 3 months before VT	• Leg injury in 3 months before VT
• Pneumonia at time of VT	• Pneumonia in year before VT
• Infection at time of VT (urinary tract infection, pyelonephritis, arthritis, bursitis, sinusitis, pulpitis, inflammation elsewhere, hepatitis A, B or C)	• Infection in year before VT (urinary tract infection, pyelonephritis, arthritis, bursitis, sinusitis, pulpitis, inflammation elsewhere, hepatitis A, B or C)

Supplementary Table S6. Association of *ABO* haplotypes with first VT recurrence in MARTHA *ambispective* and MEGA stratified on the provoked character of the first VT

Variables	MARTHA *Ambispective* N=1,504 Nb recurrences=565		MEGA N=1,248 Nb recurrences=428		Meta-Analysis Fixed-effects	
	HR (95% CI)	**P**	**HR (95% CI)**	**P**	**HR (95% CI)**	**P**
ABO haplotypes – First VT provoked	**N=993 ; 368 recurrences**		**N=847 ; 237 recurrences**			
A1	1.16 (0.98-1.37)	0.072	1.22 (0.98-1.51)	0.070	1.18 (1.02-1.38)	0.029
A2	1.32 (0.94-1.86)	0.105	1.30 (0.93-1.82)	0.120	1.31 (1.04-1.66)	0.025
O1	Reference		Reference		Reference	
O2	0.97 (0.45-2.11)	0.940	0.66 (0.25-1.79)	0.420	0.84 (0.42-1.70)	0.627
B	1.15 (0.89-1.48)	0.289	0.93 (0.65-1.32)	0.670	1.07 (0.83-1.37)	0.615
ABO haplotypes – First VT unprovoked	**N=511 ; 197 recurrences**		**N=401 ; 191 recurrences**			
A1	1.17 (0.91-1.51)	0.213	1.23 (0.97-1.57)	0.094	1.20 (1.01-1.43)	0.037
A2	1.24 (0.86-1.78)	0.248	0.91 (0.60-1.37)	0.642	1.08 (0.81-1.44)	0.602
O1	Reference		Reference		Reference	
O2	1.41 (0.74-2.68)	0.291	1.09 (0.74-1.62)	0.667	1.22 (0.75-1.98)	0.431
B	0.82 (0.55-1.22)	0.325	1.02 (0.52-2.03)	0.947	0.95 (0.72-1.25)	0.698

HR: Hazard Ratio
CI: Confidence Interval

Supplementary Text. Sensitivity analysis on the weights estimation for MARTHA participants

Methods: To investigate the variability of the weights estimated from the MARTHA study and their impact on the weighted Cox model, we used a Monte Carlo method. From the death risk model, we estimated the survival function $\hat{S}(t_i|Z_i) = \exp(-\hat{A}(t_i|Z_i))$ of each individual i up to the time t_i (which corresponds to the time of collection of the information on VT recurrence). Assuming that the cumulative risk $\hat{A}(t_i|Z_i)$ follows a normal distribution, for each individual we randomly draw 1,000 values of the his/her cumulative risk from the distributions $N(\hat{A}(t_i|Z_i), SE(\hat{A}(t_i|Z_i)))$ and computed the corresponding survival probabilities $\hat{S}_k(t_i|Z_i) = \exp(-\hat{A}_k(t_i|Z_i))$ to obtain the set of individual weights w_{ik} for k=1,…,1000. Then 1,000 weighted Cox model for the VT recurrence were estimated.

Results: The distributions of the HR for the ABO blood groups from the 1,000 models for VT recurrence are shown in Supplementary Figure 6 where the value of the HR estimated in the initial model is presented as a dashed line. The empirical distributions are well centred at the initial estimated HRs and, for both A1 and A2, all estimated HRs are above 1 supporting our conclusions.

5.2 Analyses supplémentaires

Après la première soumission de cet article, j'ai eu accès aux données génétiques de l'étude EDITH (*voir section **4.2***) et pour affiner les résultats obtenus, j'ai analysé l'effet du groupe sanguin dans cette étude. Dans une seconde analyse exploratoire, je me suis intéressée aux variants génétiques identifiés par la plus récente méta-analyse de GWAS sur la MTEV, pour regarder si ces variants pouvaient également avoir un effet sur le risque de récidive de MTEV dans les études MARTHA, MEGA et EDITH.

5.2.1 Analyse du groupe sanguin dans EDITH

Le groupe sanguin a été génétiquement déterminé pour 1 716 individus de l'étude EDITH, à l'aide des polymorphismes rs8176719-delG (O1), rs41302905-T (O2), rs2519093-T (A1), rs1053878-A (A2) et rs8176743-T (B) comme indiqué dans *Goumidi et al.*. Ces individus ont été inclus pour leur première MTEV avec des critères d'inclusion similaires à ceux de l'étude MARTHA (sans homozygotie FV, FII, cancer, maladie auto-immune) et parmi lesquels 375 premières récidives de MTEV ont été prospectivement observées. Les mêmes analyses que celles présentées dans le Tableau 3 de l'article en page 62 ont été conduites et les résultats des trois études MARTHA, MEGA et EDITH ont été méta-analysés. Les résultats des analyses sur chacune des trois cohortes et de la méta-analyse sont présentés dans le **Tableau 2**.

Concernant les variables cliniques, l'ajout de l'étude EDITH a permis de confirmer l'association avec le sexe masculin et le caractère provoqué de la première MTEV et d'identifier de nouvelles associations : l'augmentation de l'âge de la première MTEV serait un facteur de risque de récidive de MTEV, HR = 1,06 ; pvaleur = $3{,}1\text{x}10^{-3}$ pour l'augmentation de 10 ans d'âge. Par contre, cette analyse ne confirme pas la tendance observée précédemment concernant le type de la première MTEV.

Concernant les haplotypes du groupe sanguin, les associations sont restées homogènes entre les études puisque les tests d'hétérogénéité de Cochran n'étaient pas significatifs (pvaleur > 0,5) (Cochran, 1954). De plus, l'haplotype A1 est toujours un facteur de risque de MTEV, seule une tendance est observée pour l'haplotype A2 et l'haplotype B n'est pas du tout associé à la récidive de MTEV. Néanmoins pour l'haplotype O2 qui a une fréquence relativement faible (environ 2 %), d'autres études seraient nécessaires pour affiner son effet sur la récidive de MTEV. En effet, même si dans l'étude MEGA l'effet de cet haplotype allait dans le sens inverse de celui observé dans MARTHA et EDITH, les tailles d'effet estimées dans ces deux études sont tout de même assez élevées (HR = 1,19 et HR = 1,28, respectivement, ce qui

donnerait HR = 1,24 [0,91 ; 1,69] en méta-analyse) ce qui souligne l'importance d'étudier séparément O1 et O2.

Tableau 2 : Associations entre les variables cliniques et les haplotypes du groupe sanguin avec le risque de récidive de MTEV dans les études MARTHA, MEGA et EDITH

Variables	MARTHA N = 1 504 Récidives = 565		MEGA N = 1 248 Récidives = 428		EDITH N = 1 716 Récidives = 375		Meta-analyse N = 4 468 Récidives = 1 368	
	HR (IC 95 %)	**P**	**HR (IC 95 %)**	**P**	**HR (IC 95 %)**	**P**	**HR (IC 95 %)**	**P**
Sexe								
Masculin	1,65 (1,36-2,01)	$4{,}0\times10^{-7}$	1,81 (1,46-2,25)	$5{,}9\times10^{-8}$	1,18 (0,95-1,46)	0,14	1,53 (1,36-1,73)	$3{,}1\times10^{-12}$
Age à la 1ère MTEV (augmentation de 10 ans)	1,08 (1,02-1,15)	0,02	0,99 (0,92-1,07)	0,81	1,10 (1,03-1,18)	$7{,}2\times10^{-3}$	1,06 (1,02-1,11)	$3{,}1\times10^{-3}$
Type de la 1ère MTEV								
TVP	1,17 (0,96-1,42)	0,14	1,15 (0,95-1,40)	0,16	0,87 (0,70-1,07)	0,19	1,06 (0,94-1,20)	0,32
Caractère de la 1ère MTEV								
Provoquée	0,99 (0,80-1,23)	0,92	0,61 (0,49-0,76)	$6{,}7\times10^{-6}$	0,45 (0,33-0,61)	$4{,}2\times10^{-7}$	0,70 (0,62-0,81)	$3{,}3\times10^{-7}$
Haplotypes du groupe sanguin								
A1	1,15 (1,00-1,32)	0,04	1,21 (1,03-1,42)	0,02	1,09 (0,92-1,31)	0,32	1,15 (1,06-1,26)	$1{,}7\times10^{-3}$
A2	1,27 (0,98-1,64)	0,06	1,11 (0,86-1,43)	0,41	1,03 (0,75-1,41)	0,86	1,15 (0,98-1,34)	0,08
O1	Référence		Référence		Référence		Référence	
O2	1,19 (0,73-1,94)	0,48	0,86 (0,50-1,49)	0,58	1,28 (0,83-1,98)	0,26	1,13 (0,85-1,50)	0,39
B	1,02 (0,82-1,27)	0,87	1,00 (0,78-1,29)	0,99	1,01 (0,77-1,33)	0,94	1,01 (0,88-1,17)	0,88

HR: Hazard Ratio

IC: Intervalle de Confiance

5.2.2 Analyse des facteurs génétiques de MTEV avec le risque de récidive

En parallèle du projet sur les haplotypes du groupe sanguin et suite à la publication de la dernière méta-analyse de GWAS sur la MTEV, j'ai regardé si les variants associés à la première MTEV le sont également avec la récidive (Thibord et al., 2022). Les associations entre ces variants génétiques et le risque de récidive de MTEV ont été étudiées dans MARTHA, MEGA et EDITH, puis méta-analysées. Parmi les 101 SNPs indépendants identifiés par *Thibord et al.* et ayant été correctement imputés dans les trois études, seulement 11 variants avec une fréquence supérieure à 1 % étaient significativement associés, au seuil nominal de 0,05, au risque de récidive de MTEV dans la méta-analyse. Les associations correspondantes sont présentées dans le **Tableau 3**. Il est à noter que plusieurs de ces SNPs sont localisés dans des gènes établis de la cascade de la coagulation, comme *F2* et *F11* qui codent pour des facteurs de la coagulation ou encore *FGG* qui code pour une composante de la molécule du fibrinogène. Malheureusement, aucun d'entre eux ne passe le seuil corrigé pour le nombre de polymorphismes testés de $0{,}05/101 = 5{,}0\times10^{-4}$.

Par la suite, les 101 polymorphismes ont été agrégés en un score de risque génétique (GRS). Le GRS de l'individu i se définit comme la somme des allèles à risque de MTEV des $P = 101$ SNPs, pondérés par leurs coefficients de régression respectifs (β) estimés dans la méta-analyse (Thibord et al., 2022) :

$$GRS_i = \sum_{j=1}^{P} SNP_{ij} \times \beta_j$$

Dans chacune des trois cohortes, l'association entre le GRS de MTEV et la récidive de MTEV a été estimée et les résultats sont présentés dans le **Tableau 4**. L'augmentation du GRS (représentant l'augmentation du risque génétique de MTEV) était significativement associée au risque de récidive de MTEV dans les études EDITH et MEGA (Hazard Ratio (HR) = 1,30 ; pvaleur = 0,002 et HR = 1,74 ; pvaleur = $1{,}5\text{x}10^{-11}$, respectivement), mais il ne l'était pas dans l'étude MARTHA (HR = 1,00 ; pvaleur = 0,98). Les associations du GRS de MTEV avec le risque de récidive de MTEV dans les études EDITH, MEGA et MARTHA ont été méta-analysées, permettant ainsi de mettre en évidence que les variants génétiques associés à la MTEV agrégés en score sont également associés à la récidive de MTEV (HR = 1,27 [1,16 ; 1,38] ; p = $1{,}19\text{x}10^{-7}$) malgré une hétérogénéité importante ($I^2 = 0{,}92$).

Tableau 3 : Résultats des analyses d'association entre les variants associés à la MTEV et le risque de récidive dans les études MARTHA, MEGA et EDITH

				MTEV (*Thibord et al. 2022*)		Récidive de MTEV							
						EDITH		MARTHA		MEGA		Méta-analyse	
SNP	Allèle testé	Localisation	Gène	RC*	Pvaleur	HR**	Pvaleur	HR	Pvaleur	HR	Pvaleur	HR	Pvaleur
rs3756011	A	intron	*F11*	1,23	$7{,}48\times10^{-198}$	1,28	$1{,}46\times10^{-3}$	1,03	0,68	1,17	0,03	1,14	$1{,}12\times10^{-3}$
rs2066864	A	UTR3	*FGG*	1,23	$1{,}98\times10^{-172}$	0,99	0,93	1,13	0,07	1,32	$1{,}55\times10^{-4}$	1,15	$1{,}21\times10^{-3}$
rs2274224	C	exon	*PLCE1*	1,04	$2{,}55\times10^{-9}$	1,11	0,18	1,20	$2{,}54\times10^{-3}$	1,07	0,36	1,14	$1{,}22\times10^{-3}$
rs505922	C	intron	*ABO*	1,35	$1{,}11\times10^{-425}$	1,08	0,37	1,15	0,02	1,14	0,07	1,13	$3{,}04\times10^{-3}$
rs710446	T	exon	*KNG1*	0,96	$5{,}92\times10^{-11}$	1,05	0,49	0,78	$3{,}25\times10^{-5}$	0,97	0,69	0,90	$6{,}91\times10^{-3}$
rs8110055	A	intron	*SLC44A2*	0,89	$5{,}36\times10^{-44}$	0,91	0,34	0,84	0,05	0,85	0,12	0,86	$7{,}79\times10^{-3}$
rs6503222	A	intron	*SMG6*	1,05	$1{,}59\times10^{-12}$	0,92	0,32	0,86	0,02	0,94	0,35	0,90	0,01
rs62282204	T	inter génique	*PIK3CB;LINC01391*	0,96	$1{,}87\times10^{-8}$	0,93	0,43	0,90	0,12	0,90	0,15	0,91	0,02
rs6993770	A	intron	*ZFPM2*	1,08	$4{,}48\times10^{-25}$	1,05	0,55	1,08	0,31	1,22	0,02	1,11	0,03
rs9611844	C	intron	*A4GALT*	1,10	$2{,}09\times10^{-21}$	1,01	0,91	1,09	0,34	1,24	0,02	1,12	0,04
rs2842700	A	intron	*C4BPA*	1,11	$5{,}95\times10^{-17}$	1,23	0,12	1,20	0,09	0,99	0,93	1,16	0,05

* Rapport de Cotes

** Hazard Ratio ajustés sur le sexe, les caractéristiques de la première MTEV (âge, EP/TVP, provoquée/non provoquée) ainsi que sur les quatre premières composantes principales de la stratification de la population dérivées des données génétiques

Tableau 4 : Résultats des analyses d'association entre les variants associés à la MTEV sous forme de scores génétiques et le risque de récidive dans les études MARTHA, MEGA et EDITH

	MARTHA		**MEGA**		**EDITH**		**Meta-analyse**	
	N = 1 518 Récidives = 571		N = 1 254 Récidives = 431		N=1 727 Récidives = 377		N=4 499 Récidives = 1 379	
	HR (IC 95 %)	**Pvaleur**	**HR (IC 95 %)**	**Pvaleur**	**HR (IC 95 %)**	**Pvaleur**	**HR (IC 95 %)**	**Pvaleur**
GRS 101 SNPs*	0,99 (0,86-1,13)	0,840	1,74 (1,48-2,04)	$2{,}10\times10^{-11}$	1,30 (1,08-1,55)	$4{,}40\times10^{-3}$	1,27 (1,16-1,38)	$1{,}19\times10^{-7}$
GRS 100 SNPs**	1,21 (1,00-1,45)	0,047	1,81 (1,44-2,27)	$3{,}70\times10^{-7}$	1,38 (1,09-1,75)	$8{,}30\times10^{-3}$	1,41 (1,25-1,59)	$5{,}11\times10^{-8}$
Score de 5 SNPs †	1,04 (0,98-1,11)	0,240	1,24 (1,12-1,34)	$8{,}60\times10^{-8}$	1,12 (1,03-1,21)	$7{,}20\times10^{-3}$	1,12 (1,07-1,16)	$4{,}16\times10^{-7}$
Score de 4 SNPs ††	1,07 (1,00-1,15)	0,043	1,20 (1,10-1,30)	$1{,}70\times10^{-5}$	1,11 (1,02-1,21)	0,015	1,20 (1,10-1,30)	$8{,}23\times10^{-7}$

* Score de risque génétique construit à partir des SNPs identifiés dans *(Thibord et al., 2022)* ** sans utiliser le rs6025 (*F5*)

† Score représentant la somme des allèles à risque de 5 SNPs très associés à la MTEV *(van Hylckama Vlieg et al., 2014)* †† sans utiliser le rs6025 (*F5*)

Néanmoins dans les analyses du risque de récidive de MTEV, l'effet de l'allèle à risque de MTEV du variant du gène *F5* (rs6025-T) n'était pas homogène puisqu'il semblait être protecteur dans l'étude MARTHA mais à risque dans les autres études (HR = 0,75 ; pvaleur = 0,01 , HR = 1,61 ; pvaleur = $4{,}34\text{x}10^{-5}$ et HR = 1,19 ; pvaleur = 0,23 dans les études MARTHA, MEGA et EDITH, respectivement). Ce variant est également celui avec l'effet le plus fort dans la méta-analyse de *Thibord et al.* (RC = 3,05) et il a donc un poids important dans le GRS de MTEV qui en découle. C'est pour cela que dans un second temps, j'ai construit un autre GRS sur $P = 100$ SNPs en enlevant le variant rs6025 (**Tableau 4**). Les résultats des études EDITH et MEGA sont restés similaires (HR = 1,38 ; pvaleur = 0,004 et HR = 1,81 ; pvaleur = $3{,}7\text{x}10^{-7}$, respectivement) mais pour l'étude MARTHA le GRS est devenu significativement associé au risque de récidive (HR = 1,21 ; pvaleur = 0,047), permettant ainsi de renforcer l'effet méta-analysé (HR = 1,41 ; pvaleur = $5{,}11\times10^{-8}$) avec cependant une hétérogénéité qui reste élevée ($I^2 = 0{,}72$).

Dans une autre analyse, je me suis intéressée à l'effet d'un autre score génétique qui avait été précédemment proposé par *van Hylckama Vlieg et al.* et qui représente la somme des allèles à risque de MTEV des polymorphismes situés dans seulement cinq gènes *F2*, *F5*, *F11*, *FGG* et *ABO* (van Hylckama Vlieg et al., 2014). Dans l'étude MARTHA, on retrouve un effet plus modéré du score sur le risque de récidive de MTEV, par rapport aux études EDITH et MEGA mais cette différence s'atténue lorsque le variant du gène du *F5* n'est pas utilisé (**Tableau 4**). Dans cette analyse, les HR représentent l'augmentation du risque instantané de récidive de MTEV pour l'augmentation d'un allèle à risque de MTEV, avec ajustement sur les covariables.

Les tailles d'effet de ces scores sont relativement comparables et des analyses comparatives des *Area Under the Curve* – AUC seraient nécessaires pour déterminer si le score composé des 100 polymorphismes pourrait permettre une meilleure discrimination du risque de récidive de MTEV que celui composé de seulement cinq polymorphismes.

5.3 Discussion

L'identification des déterminants environnementaux et génétiques de la récidive de la MTEV est un enjeu majeur de la recherche sur la MTEV. La stratégie de modélisation statistique proposée dans ce travail a permis d'intégrer à la fois des évènements rétrospectifs et prospectifs pour analyser l'effet de variables fixes dans le temps, dont les haplotypes du groupe sanguin *ABO*, sur le risque de récidive de MTEV dans l'étude MARTHA.

La modélisation proposée repose sur l'utilisation de poids permettant de limiter le biais de sélection dû à la mortalité qui est induit par l'analyse d'évènements pré-inclusion. La principale limite de la stratégie d'analyse des données ambispectives de MARTHA repose sur l'estimation du risque de décès et des poids qui en découlent, puisque le nombre de décès était assez faible (N = 73). De plus, cette estimation aurait pu être plus précise si d'autres variables, comme par exemple le nombre d'antécédents personnels de MTEV ainsi que le temps calendaire, avaient été pris en compte dans la modélisation. Néanmoins, cette stratégie d'analyse de données ambispectives peut s'étendre aux analyses GWAS mais également s'appliquer à d'autres phénotypes sous réserve qu'il soit possible d'avoir une estimation assez fiable du risque de décès dans la population étudiée, ou une population similaire.

La méta-analyse des études MARTHA et MEGA (totalisant 2 752 cas de MTEV dont 993 récidives) a permis de mettre en évidence que les haplotypes du groupe sanguin A1 et A2 seraient tous les deux des facteurs de risque de récidive de MTEV par rapport aux O1, ce qui est cohérent avec d'autres travaux montrant que les individus non OO seraient plus à risque de récidive (Dentali & Franchini, 2013; Gándara et al., 2013; Limperger et al., 2021). Cependant, contrairement au premier évènement, l'haplotype B ne semble pas être associé au risque de récidive de MTEV, ce qui laisse à penser qu'il existerait des différences mécanistiques entre la survenue d'une MTEV et de sa récidive. Par ailleurs, ce travail a également permis d'apporter de nouveaux éléments sur les facteurs de risque cliniques de la récidive de MTEV, confirmant ainsi que le sexe masculin est un facteur de risque de récidive de MTEV et que le caractère provoqué de la première MTEV est protecteur contre la récidive de MTEV. En effet, même si ce dernier n'était pas associé à la récidive dans l'étude MARTHA, ce phénomène pouvait s'expliquer par la longue durée de suivi (rétrospectif et prospectif) des individus de MARTHA et l'hypothèse de proportionnalité des risques qui n'était pas respectée. En restreignant l'analyse aux récidives survenant dans les deux années suivant le premier évènement (et en censurant les autres individus à deux ans de suivi), le caractère provoqué était bien protecteur de récidive de MTEV à la fois dans MARTHA et MEGA (*voir Supplementary Table S4 de l'article page 69*).

Ces éléments suggèrent que l'effet protecteur du caractère provoqué serait surtout présent dans les premières années suivant la MTEV et qu'il s'atténuerait par la suite.

Dans ce projet, pour analyser l'effet des haplotypes du groupe sanguin dans l'étude MEGA, un modèle de Cox classique a été utilisé. Cependant, le suivi des patients dans cette étude s'effectue en un seul recontact qui a lieu près de 10 ans après l'inclusion des sujets, ce qui pourrait nécessiter l'utilisation de poids pour limiter le biais de sélection dû à la mortalité post-inclusion. En effet, parmi les patients de l'étude MEGA, neuf patients ont été exclus car ils étaient décédés avant la phase de recontact et, de ce fait, aucune information sur la possible présence d'une récidive de MTEV post-inclusion n'était disponible. Le nombre de décès dans MEGA était trop petit mais il serait tout de même envisageable d'inférer des probabilités de survie jusqu'à la date du recontact à partir du modèle de décès estimé dans l'étude MARTHA, sous l'hypothèse que ces deux études soient similaires et que cette extrapolation puisse être faite. Néanmoins cela nécessiterait de re-générer des composantes principales à partir des données génétiques des échantillons de MARTHA et MEGA combinés, puis de ré-estimer le risque de décès dans MARTHA afin de pouvoir inférer les poids pour les individus de MEGA.

Pour affiner les associations observées dans la méta-analyse des études MARTHA et MEGA entre les haplotypes du groupe sanguin et la récidive de MTEV, les données prospectives de l'étude EDITH ont été étudiées. La puissance statistique pourrait être augmentée pour affiner les estimations des effets des haplotypes A2 et O2 sur le risque de récidive de MTEV puisque l'étude EDITH présente un schéma d'étude ambispectif. En effet, les individus ont été inclus pour un évènement de MTEV et certains d'entre eux avaient déjà présentés une récidive de MTEV avant leur inclusion. Les patients sont suivis régulièrement et les dates de décès ont été recueillies. Parmi les 2 122 individus ayant été suivis, 455 décès ont été observés et j'ai donc estimé le risque de décès dans cet échantillon en utilisant les mêmes covariables que dans l'étude MARTHA (*voir section « Risk of death estimation» de l'article en page 50*), afin d'inférer les poids pour pouvoir étudier les récidives prospectives et rétrospectives (N = 2 024 dont 673 récidives de MTEV). Cependant dans l'étude EDITH, l'âge de la première MTEV est très associé au risque de décès (HR = 1,36 pour une augmentation de 10 ans ; pvaleur = $6{,}8\times10^{-5}$) et cette variable a donc une importance plus grande lors du calcul des poids ce qui peut engendrer des poids très élevés ou très faibles. Etant donné que les individus de l'étude EDITH étaient assez âgés au moment de leur première MTEV (premier quartile = 40 ans ; troisième quartile = 72 ans), les poids inférés ont une très grande étendue (minimum : 0,2 ; maximum 87,7). Néanmoins, ces poids se stabilisent lorsqu'on enlève de l'analyse les individus ayant eu une première MTEV après 75 ans (minimum : 0,7 ;

maximum : 6,7) ou lorsque les estimations de l'étude MARTHA sont utilisées (sans tenir compte des composantes principales) (minimum : 0,9 ; maximum : 5,9) car l'âge de la première MTEV était moins associé au risque de décès et qu'il était protecteur (HR = 0,84 ; pvaleur = 0,029). L'extension de cette approche par un modèle de Cox pondéré pour analyser les données ambispectives dans l'étude EDITH nécessiterait donc plus de travail pour mieux comprendre les spécificités de cet échantillon et obtenir une estimation du risque de décès qui soit satisfaisante pour inférer des poids stables.

Outre l'application de cette stratégie de modélisation dans des études au schéma similaire ou encore l'utilisation de l'estimation du risque de décès pour générer des poids dans d'autres études de patients de MTEV, la perspective immédiate de ce projet est son implémentation en analyse GWAS pour étudier les facteurs de risque génétiques de la récidive de MTEV dans l'étude MARTHA. Je coordonne actuellement la plus grande méta-analyse internationale des facteurs de risque génétiques de la récidive de MTEV comptabilisant 6 504 cas de MTEV dont 1 797 récidives et constituée des études MARTHA, EDITH et MEGA ainsi que REVERSE (coordonnée par les Professeurs France Gagnon et Marc A. Rodgers au Canada), HVH (coordonnée par le Professeur Nicholas L. Smith aux Etats-Unis) et FHS (coordonnée par le Professeur Andrew D. Johnson aux Etats-Unis). Pour l'utilisation de cette modélisation dans l'étude MARTHA, dans un premier temps les poids étaient dépendants des SNPs étudiés car ils avaient été intégrés dans l'estimation du risque de décès. Cependant, les poids étaient trop sensibles aux variants basse fréquence et ce même en utilisant les méthodes classiques de standardisation des poids, notamment car le nombre de décès était trop faible. Pour analyser les données ambispectives, le modèle de Cox pondéré permet de gagner en puissance statistique mais les limites concernant le calcul des poids, énoncées ci-dessus, pourraient contrebalancer l'utilisation de cette approche. En effet, si le nombre d'évènements rétrospectifs est négligeable, il serait préférable de se restreindre à une analyse prospective qui est plus rigoureuse.

Dans ce projet, le traitement anticoagulant n'a pas pu être pris en compte, notamment à cause du recueil rétrospectif des données de MARTHA. Il aurait été intéressant de considérer cette variable comme dépendante du temps car le risque de récidive est différent selon que le patient est ou non sous traitement, ou alors de considérer le début de la période à laquelle le patient est à risque de récidive au moment de l'arrêt de ce traitement. Grâce aux bases de données françaises comme le Système National des Données de Santé il pourrait être envisageable de récupérer les informations concernant les délivrances des traitements

anticoagulants et de les croiser avec les études françaises (MARTHA, EDITH) que j'ai à ma disposition.

Dans une autre perspective, cette stratégie de modélisation pourrait être utilisée pour analyser les facteurs de risque génétiques de la récidive de MTEV dans l'étude MARTHA12. L'étude MARTHA12 est une extension de l'étude MARTHA puisque les mêmes critères de sélection ont été appliqués pour recruter environ 1 000 cas de MTEV sur la période 2008-2011. Les individus ont été génotypés avec une puce ADN classique ainsi qu'avec une puce d'exome permettant de mesurer uniquement des variants rares principalement localisés dans des exons. Pour ces patients qui n'ont pas été suivis prospectivement, les évènements de MTEV ont été collectés uniquement de façon rétrospective et il n'est donc pas possible d'estimer le risque de décès dans cette étude. Les composantes principales issues des données génétiques étant propres à chaque échantillon, il faudrait les générer en regroupant les études MARTHA et MARTHA12, puis ré-estimer le risque de décès dans MARTHA afin de pouvoir inférer les poids pour les individus de MARTHA12. Le modèle de Cox pondéré pourrait par la suite être utilisé en analyse GWAS mais également implémenté pour étudier les variants rares avec le risque de récidive de MTEV.

En utilisant ces cohortes de cas de MTEV, il serait intéressant d'évaluer la capacité prédictive des scores cliniques (Herdoo2, Dash, Vienna) de récidive de MTEV. Cependant le point commun de ces scores est qu'ils nécessitent tous la mesure des D-dimères, or cette mesure a été réalisée seulement chez quelques sujets de l'étude MARTHA et n'a pas toujours été faite à l'arrêt du traitement anticoagulant. A ma connaissance les D-dimères n'ont pas été mesurés dans les autres études.

6 Analyse des facteurs génétiques des NETs

Le deuxième travail méthodologique auquel je me suis intéressée dans ma thèse s'inscrivait dans un projet visant à étudier les déterminants génétiques des taux plasmatiques de NETs, un marqueur biologique ayant un rôle clé dans l'immunothrombose et suscitant de ce fait de plus en plus d'intérêt dans le domaine des pathologies thrombotiques (*voir section **1.4.2***). La mesure de ce biomarqueur est complexe et son protocole n'est pas encore bien établi. Cependant, l'identification de ses déterminants génétiques pourrait permettre de mieux comprendre les mécanismes biologiques impliqués dans la formation des NETs et permettre de réaliser des explorations en lien avec d'autres phénotypes, notamment via des analyses de corrélation génétique et de randomisation mendélienne. Ce biomarqueur a été mesuré dans un sous-échantillon de l'étude FARIVE pour lequel les données génétiques sont également disponibles (*voir section **4.3***). Comme illustré en **Figure 14**, la distribution des NETs dans l'étude FARIVE est semicontinue, c'est-à-dire qu'elle est caractérisée par un excès de valeurs en zéro suivi d'une distribution positive continue avec asymétrie à droite.

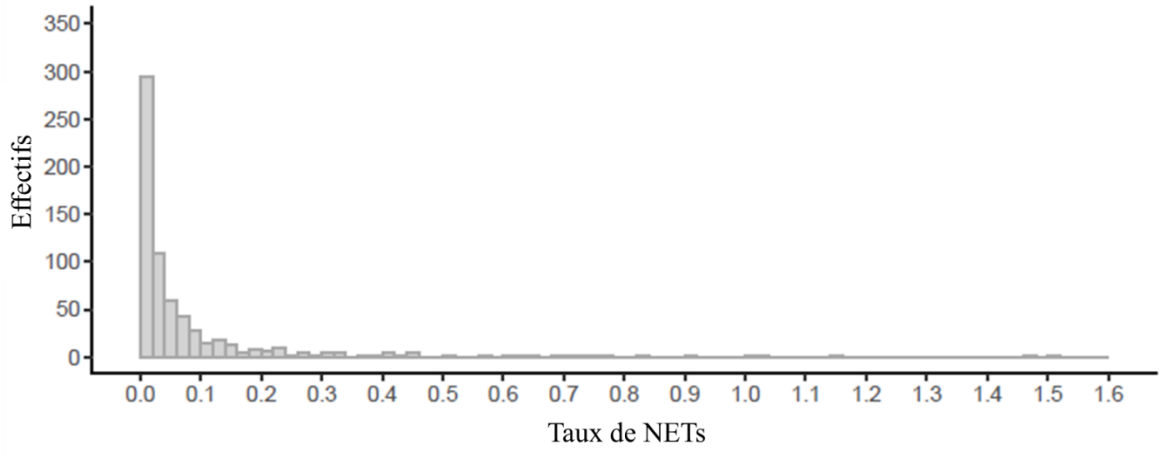


Figure 14 : Distribution des taux de NETs dans l'étude FARIVE

Comme indiqué précédemment, les études GWAS reposent généralement sur des modèles de régression logistique pour l'étude d'un trait binaire ou sur des modèles de régression linéaire pour l'étude d'un trait quantitatif dont la distribution est Gaussienne (ou normale). Au moment où je débutais cette thèse, aucune étude GWAS n'avait été réalisée sur un trait semicontinu, probablement parce que ce type de méthodologie n'est pas implémentée dans les logiciels couramment utilisés pour des études GWAS. Suite à une étude de la littérature, j'ai identifié les modèles qui me paraissaient adaptés à la modélisation d'une variable semicontinue.

Trois modèles ont été considérés dans ce travail et sont décrits plus en détail dans la section « *Statistical modeling for GWAS analysis of NETs plasma levels »* de l'article en page 93 : les modèles Tobit, Négatif Binomial et Composé Poisson-Gamma (Ehrenberg, 1959; Tobin, 1958; Tweedie, 1984). Ces trois modèles permettent d'obtenir un seul paramètre de régression pour mesurer l'association entre une variable explicative et la moyenne de la variable semicontinue à expliquer. D'autres modèles comme le modèle Two-part n'ont pas été étudiés car ils font l'hypothèse que la variable à expliquer est issue de deux processus de génération qui sont modélisés séparément, l'un pour la présence d'un zéro et l'autre pour la partie strictement continue. Le modèle Tobit permet de modéliser des distributions latentes gaussiennes qui sont censurées à un certain seuil qui peut être dû à une limite de détection par exemple. Le modèle Négatif Binomial est une extension du modèle de Poisson utilisé pour modéliser des données de comptage présentant une sur-dispersion. Le modèle Composé Poisson-Gamma permet de modéliser une somme de plusieurs variables suivant une loi Gamma dont le nombre est défini par une loi de Poisson.

Dans un premier temps, l'adéquation des modèles à la distribution des NETs dans FARIVE a été étudiée graphiquement ainsi qu'en utilisant le critère statistique du RMSE (*Root Mean Square Error*). Cette étude a permis de montrer que les modèles Négatif Binomial et Composé Poisson-Gamma étaient plus adaptés à la modélisation des NETs que le modèle Tobit (*voir Figures 1 & 2 et Table 2 de l'article*).

Pour autant, ces trois modèles ont été utilisés en GWAS ce qui a notamment permis d'identifier une inflation des p-valeurs des modèles Tobit et Négatif Binomial. Concernant le modèle Tobit, l'inflation des p-valeurs était principalement attribuable aux variants basse fréquence (fréquence < 5 %) car en restreignant l'analyse aux variants communs l'inflation était diminuée. Concernant le modèle Négatif Binomial, l'inflation des p-valeurs était due à la présence de valeurs positives extrêmes et, de façon plus modérée, aux variants basse fréquence. L'inflation des p-valeurs du modèle Négatif Binomial a été démontrée par simulation, et comparée au comportement du modèle Composé Poisson-Gamma.

Dans un premier temps, 10 000 échantillons *bootstrap* ont été simulés en tirant aléatoirement (avec remise) 657 valeurs depuis les taux de NETs observés. Pour chacun des 10 000 échantillons, quatre génotypes indépendants du taux de NETs ont été aléatoirement attribués aux individus, en respectant la loi d'Hardy-Weinberg et avec des fréquences alléliques de 1 %, 5 %, 10 % et 20 %. Les associations entre chacun des quatre génotypes générés aléatoirement et les taux de NETs des échantillons *bootstrap* ont été testées avec les modèles Négatif Binomial et Composé Poisson-Gamma sur les 10 000 échantillons. Le nombre de

résultats significatifs à différents seuils ($\alpha = 0,05$, $\alpha = 0,01$, et $\alpha = 0,001$) a permis de calculer les erreurs empiriques de type I.

Cette analyse a permis de confirmer que le modèle Négatif Binomial ne contrôlait pas bien l'erreur de type I (nombre important de faux positifs), indépendamment de la fréquence allélique du génotype étudié. A l'inverse, le modèle Composé Poisson-Gamma présentait un bon contrôle de l'erreur de type I même s'il pourrait être considéré comme étant un peu trop conservateur pour l'analyse des variants basse fréquence.

Dans un second temps pour évaluer la sensibilité du contrôle de l'erreur de type I de ces modèles face aux valeurs extrêmes, les individus simulés présentant des taux de NETs supérieurs à 0,5 (correspondant aux 3 % des valeurs les plus élevées dans FARIVE) ont été enlevés et les mêmes analyses ont été répétées. Dans cette situation, les deux modèles ont montré un bon contrôle de l'erreur de type I, même si le Négatif Binomial présentait une légère inflation pour les variants basse fréquence.

Le modèle Composé Poisson-Gamma s'est révélé être un modèle adéquat pour l'analyse des NETs dans FARIVE, en présentant un bon contrôle de l'erreur de type I et en étant robuste à la présence de valeurs extrêmes et de variants basse fréquence. De plus, les hypothèses de distribution de ce modèle sont assez plausibles concernant la production des NETs puisqu'on pourrait faire l'hypothèse que le nombre de neutrophiles activés suit une loi de Poisson et que la quantité de NETs relâchée par chacun d'eux suit une loi Gamma. L'utilisation de ce modèle en GWAS a permis d'identifier un locus significatif au seuil pangénomique localisé dans un ARN non codant hébergeant le miR-155. De façon intéressante, de récents travaux de biologie chez la souris ont montré que le miR-155 serait impliqué dans la formation des NETs et qu'il régulerait également l'expression de *PAD4*, un gène codant pour une protéine indispensable à la formation de NETs (Hawez et al., 2019, 2022).

Ce travail est la première GWAS des taux de NETs dans un échantillon de cas de MTEV et de témoins, réalisée avec un modèle adapté à la distribution semicontinue de ce biomarqueur et n'ayant pas nécessité de transformation (logarithmique, inverse normale). Il a permis d'apporter de nouveaux éléments confirmant des observations faites chez la souris et de mettre en avant les bonnes propriétés statistiques du modèle Composé Poisson-Gamma dont l'utilisation reste encore marginale. Par ailleurs, la sensibilité du modèle Négatif Binomial a déjà été mise en avant dans le cadre des analyses de RNA-seq et le Composé Poisson-Gamma pourrait être une bonne alternative qui mériterait d'être explorée (Gauthier et al., 2020).

6.1 Article 2 : Genome-wide association study of a semicontinuous trait: Illustration of the impact of the modeling strategy through the study of Neutrophil Extracellular Traps levels

Article 2 soumis (1er auteur) :

Munsch G, Proust C, Labrouche-Colomer S, […], Smadja M D, Jacqmin-Gadda H*, Trégouët D-A*.

Contribution :

- Identification des modèles pertinents
- Analyses statistiques
- Ecriture du manuscrit

La version soumise du manuscrit est présentée dans le section suivante.

Genome-wide association study of a semicontinuous trait: Illustration of the impact of the modeling strategy through the study of Neutrophil Extracellular Traps levels

Gaëlle Munsch[1+], Carole Proust[1], Sylvie Labrouche-Colomer[2,3], Dylan Aïssi[1], Anne Boland[4], Pierre-Emmanuel Morange[5,6], Anne Roche[7], Luc de Chaisemartin[8,9], Annie Harroche[10], Robert Olaso[4,11], Jean-François Deleuze[4,11], Chloé James[2,3], Joseph Emmerich[12,13], David M Smadja[14,15], Hélène Jacqmin-Gadda[1*], David-Alexandre Trégouët[1*]

[1] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France

[2] UMR1034, Inserm, Biology of Cardiovascular Diseases, University of Bordeaux, Pessac, France

[3] Laboratoire d'Hématologie, CHU de Bordeaux, Pessac, France

[4] Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), 91057, Evry, France

[5] INSERM UMR_S 1263, Nutrition Obesity and Risk of Thrombosis, Center for CardioVascular and Nutrition research (C2VN), Aix-Marseille University, Marseille 13385, France

[6] Laboratory of Haematology, La Timone Hospital, Marseille 13385, France

[7] Service pneumologie hôpital Bicêtre, France

[8] Service Auto-immunité, Hypersensibilité et Biothérapies, Hôpital Bichat, Assistance Publique-Hôpitaux de Paris, Paris, France

[9] Université Paris-Saclay, INSERM, Inflammation, Microbiome, Immunosurveillance, Orsay, France

[10] Service d'Hématologie Clinique Centre de Traitement de l'Hémophilie Hôpital Necker Enfants Malades, France

[11] Centre d'Etude du Polymorphisme Humain, Fondation Jean Dausset, Paris, France

[12] Department of vascular medicine, Paris Saint-Joseph Hospital Group, University of Paris, 185145 rue Raymond Losserand, Paris, 75674, France

[13] UMR1153, INSERM CRESS, 185 rue Raymond Losserand, Paris, 75674, France

[14] Innovative Therapies in Hemostasis, Université de Paris, INSERM, F-75006, Paris, France

[15] Hematology Department and Biosurgical Research Lab (Carpentier Foundation), Assistance Publique Hôpitaux de Paris, Centre-Université de Paris (APHP-CUP), F-75015, Paris, France

* These authors equally contributed to the work.

+ Corresponding author: Gaëlle Munsch, Université de Bordeaux, Case 11, 146 rue Léo Saignat CS61292, 33076 Bordeaux cedex. (gaelle.munsch@u-bordeaux.fr)

ABSTRACT

Over the last years there has been a considerable expansion of genome-wide association studies (GWAS) for discovering biological pathways underlying pathological conditions or disease biomarkers. These GWAS are often limited to binary or quantitative traits analyzed through linear or logistic models, respectively. In some situations, the distribution of the outcome may require more complex modeling, such as when the outcome exhibits a semicontinuous distribution characterized by an excess of zero values followed by a non-negative and right-skewed distribution. We here investigate three different modeling for semicontinuous data: Tobit, Negative Binomial and Compound Poisson-Gamma. Using both simulated data and a real GWAS on Neutrophil Extracellular Traps (NETs), an emerging biomarker in immuno-thrombosis, we demonstrate that Compound Poisson-Gamma was the most robust model with respect to low allele frequencies and outliers. This model further identified the MIR155HG locus as significantly ($p=1.4x10^{-8}$) associated with NETs plasma levels in a sample of 657 participants, a locus recently highlighted to be involved in NETs formation in mice. This work highlights the importance of the modeling strategy for GWAS of a semicontinuous outcome and suggests Compound Poisson-Gamma as an elegant but neglected alternative to Negative Binomial for modeling semicontinuous outcome in the context of genomic investigations.

Key-words: Semicontinuous outcome / Compound Poisson-Gamma / Negative Binomial / Genome Wide Association Study / Neutrophil Extracellular Traps

INTRODUCTION

Semicontinuous data, characterized by an excess of zeros followed by a non-negative and right-skewed distribution, are frequently observed in biomedical research (1). When the study aims at identifying determinants of such a semicontinuous biomarker, it must be handled as the outcome variable and due to the inflation of zeros, classical models such as linear regression cannot be applied without violating the Gaussian assumptions, even with a logarithmic or rank-based inverse-normal transformation. For instance, when the interest specifically lies in the identification of molecular determinants associated with a disease semicontinuous biomarker, as it is encountered in the omics era in order to identify/characterize new biological pathways, inform about drug discovery and help in individual risk prediction (2), the problem of how to model its distribution arises. Besides, in the context of Genome Wide Association Studies (GWAS), linear and logistic regression are often the only statistical models implemented in popular software. Users are then encouraged to transform their outcome of interest into a binary or a Gaussian variable at the cost of a loss of information and/or of complex interpretation of genetic association parameters.

Over the past decades several statistical models have been developed to model semicontinuous data by taking into account the mass of zeros. Among the most commonly used models are the Tobit and the two-part models (3, 4).

The two-part model and its extensions (5, 6) rely on the use of a logistic regression model to predict the probabilities of occurrence of zero values and of a linear regression model for the analysis of the strictly continuous outcome. The main assumption of this model is that the values of the outcome are derived from two different generating processes. This model has been used in various applications including the modeling of tumor size in cancer, food intake, microbiome abundance or individual costs of chronic kidney disease (7–11). However, the two-part model does not make possible the estimation of a single parameter that represents the association of an explanatory variable on the outcome. In contrast to the two-part model, Tobit models consider a single distribution of the outcome. In the case of zero-inflated data, the Tobit model assumes that the semicontinuous variable is a truncated observation of a Gaussian variable. This modeling is mainly used to account for floor or ceiling effect of the outcome variable that could be due to technical measurement limits (12–15).

Another possibility is to consider the outcome variable as quantitative discrete, which can be done in some cases by changing the unit of measurement through the use of a multiplicative factor, without losing precision. In this case, models for count data such as the Poisson model or the Negative Binomial model in presence of overdispersion can be used. These models are relevant as long as the proportion of zeros is not too high (16, 17). As the Tobit model, these models allow for a simple interpretation of the results since only one coefficient is estimated per explanatory variable. Extensions of these models have been developed to account for the zero mass (also known as ZIP for Zero-Inflated Poisson and ZINB

for Zero-Inflated Negative Binomial) but they make the assumption that the distribution of the outcome is composed of two generating processes, like the two-part models.

New models based on so-called Tweedie distributions (18–20) have recently emerged for the analysis of semicontinuous data but their use remains marginal (21). The Compound Poisson-Gamma model belongs to this Tweedie family. It assumes that the semicontinuous outcome is defined as a Poisson sum of gamma random variables. Semicontinuous data are then modeled through the use of a single distribution.

The choice between these different models is not obvious as each semicontinuous outcome has its own properties. As there is no established decision tool, the model to be applied should be chosen according to the distribution of the outcome and the clinical context (22).

In this work, we show the impact of the adopted model on the results of a GWAS that aimed at identifying genetic factors associated with plasma levels of Neutrophil Extracellular Traps (NETs), a semicontinuous biomarker involved in thrombosis. We highlight the differences between the models with respect to the flexibility of the underlying assumptions, the robustness to outliers and low allele frequency that can help to select the most appropriate model for future studies.

NETs are one of the emerging biomarkers with a key role in thrombosis that often present an excess of zero values (23–25). In the event of a vascular breach, neutrophils and platelets are the first cells to be recruited and activated (26). When neutrophils are activated by platelets, they have pro-inflammatory properties that can enhance tissue damage and induce thrombus formation in particular when they evolve towards a certain form of cell death leading to the release of their decondensed chromatin as a network of fibres also called NETs. NETs are composed of DNA fibres comprised of antimicrobial proteins and histones which promote coagulation, platelets activation and thus thrombus formation (27, 28). NETs are involved in many other biological mechanisms such as immune response to viruses, diabetes, cystic fibrosis, cancer tumor growth, progression and metastasis (29–33).

NETs plasma levels were here measured in 657 participants of the « FActeurs de RIsque et de récidives de la maladie thromboembolique Veineuse » (FARIVE) study (34). Genome wide genotype data were also available for these participants and then used to conduct a GWAS on NETs levels. We illustrate how the results of this GWAS are impacted by the statistical approach adopted to model NETs plasma levels.

SUBJECTS AND METHODS

The FARIVE study

The FARIVE study is a multicenter case-control study conducted between 2003 and 2007. The sample includes 607 patients with a documented episode of deep vein thrombosis and/or pulmonary embolism and 607 healthy individuals. A detailed description of the study can be found elsewhere (34). Briefly,

patients were not eligible if they were younger than 18 years, had previous venous thrombosis (VT) event, active cancer or recent history of malignancy (within 5 years). Controls were recruited over the same period and matched to cases according to age and sex. They did not have any history of venous and arterial thrombotic disease as well as cancer, liver or kidney failure.

NETs measurements

Neutrophil Extracellular Traps (NETs) were quantified by measuring myeloperoxidase (MPO)-DNA complexes using an in-house capture ELISA already described (35) in a subsample of 410 VT patients (7 months after their inclusion in the study once the anticoagulant treatment has stopped) and 327 controls (at their time of inclusion in the study). Briefly, microtiter plates were coated with anti-human MPO antibody. After blocking, serum samples were added together with a peroxidase-labeled anti-DNA antibody. After incubation, the peroxidase substrate was added and absorbance measured at 405 nm in a spectrophotometer.

Genotyping and Imputation

FARIVE participants were genotyped using the Illumina Infinium Global Screening Array v3.0 (GSAv3.0) microarray at the Centre National de Recherche en Génomique Humaine (CNRGH). Individuals with at least one of the following criteria were excluded: discordant sex information, relatedness individuals identified by pairwise clustering of identity by state distances (IBS), genotyping call rate lower than 99%, heterozygosity rate higher/lower than the average rate ± 3 standard deviation or of non-European ancestry. After applying these criteria, the final sample was composed of 1,077 individuals. Among the 730,059 variants genotyped, 145,238 variants without a valid annotation were excluded as well as 656 variants deviating from Hardy-Weinberg equilibrium in controls at $P<10\text{-}6$, 47,286 variants with a Minor Allele Count (MAC) lower than 20 and 1,774 variants with a call rate lower than 95%. This quality controls procedure was conducted using Plink v1.9 (36) and the R software v3.6.2. Finally, there were 535,105 markers left for imputation which was then performed with Minimac4 using the 1000 Genomes phase 3 version 5 reference panel (37).

Genome Wide Association Study of NETs plasma levels

The present study relies on a subsample of 657 individuals (372 VT cases and 285 controls) with both NETs measurements and imputed genetic data. All single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) greater than 0.01 and imputation quality score greater than 0.3 were tested for association with NETs plasma levels. As shown in the next section, 3 different statistical models were deployed. In all, associations were tested on imputed allelic doses and adjusted for potential confounders that is age, sex, smoking, case-control status and the four first principal components derived from genome wide genotype data (38–40). The standard genome-wide statistical threshold of 5x10-8 was used to consider SNPs as significantly associated with NETs plasma levels.

Statistical modeling for GWAS analysis of NETs plasma levels

Since we were interested in identifying genetic factors that influence mean NETs plasma levels, any statistical approach that treats independently the zeros mass and the distribution of positive values, such as the two-part, ZIP and ZINB models, was deemed not adapted to our application. As a consequence, only three models were compared in this study: Tobit, Compound Poisson-Gamma, and Negative Binomial models. The Poisson model was not investigated because NETs plasma data presented a large overdispersion (see Results section), a situation where Negative Binomial model is preferable. In this study, we aimed to identify the most suitable model for semicontinuous data in order to conduct a GWAS on NETs and highlight the difference between the three models.

Tobit model

In the Tobit model, the observed variable Y is assumed to be a right or left truncated observation of an underlying Gaussian latent variable (Y^*). Let c the constant threshold for truncation which needs to be known and is equal to zero in the context of zero-inflated data. Therefore, the Tobit model assumes that zero values are due to censoring or measurement limits and so they do not represent the true absence of the variable. In the case of a left truncation at 0 the values of the observed variable are:

$$Y = \begin{cases} Y^* & if\ Y > 0 \\ 0 & otherwise \end{cases}$$

The subsequent regression model is:

$$\mathbb{E}(Y^* | X) = \beta X$$

where $\mathbb{E}(Y^* | X)$ is the expected value of the underlying Gaussian variable Y^* conditioned on the explanatory variables X, and where β represents the regression parameters associated to X. The Tobit model is available in the *VGAM* R package (41).

Compound Poisson-Gamma model

An Exponential Dispersion Model (EDM) is a two-parameter family of distributions composed of a linear exponential family with an additional dispersion parameter (42). EDMs are characterized by their variance function $\mathbb{V}(.)$ that is an exponential function used to describe the relationship between the mean and the variance. If Y follows an EDM, then $\mathbb{E}(Y) = \mu$ and $Var(Y) = \Phi\mathbb{V}(\mu)$ with Φ a dispersion parameter. Tweedie models are a class of EDMs characterized by a power variance function: $\mathbb{V}(\mu) = \mu^p$ with p the index parameter (43, 44). Most of the usual distributions are included in the class of Tweedie models such as the normal $(p = 0)$, Poisson $(p = 1)$, gamma $(p = 2)$ and the inverse Gaussian $(p = 3)$ (45).

The probability density function of a Tweedie model is defined as (42):

$$f(y|\mu, \Phi, p) = a(y, \Phi, p) \exp\left(\frac{1}{\phi}\left(y\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right)\right)$$

where $a(.)$ is a given function.

The Compound Poisson-Gamma model belongs to the family of Tweedie models with $p \in]1;2[$. It simultaneously models the occurrence and the intensity of the semicontinuous outcome (46). The distribution of a variable Y following a Compound Poisson-Gamma model may be defined as a Poisson sum of M Gamma distributions:

$$Y = \begin{cases} 0, & \text{if } M = 0 \\ K_1 + K_2 + \cdots + K_M, & \text{if } M > 0 \end{cases}$$

where $M \sim Pois(\lambda)$, $K_i \sim Gamma(\alpha, \gamma)$ with α the shape parameter and γ the scale parameter, and where the values of K_i are iid and independent on M.

The Compound Poisson-Gamma model is a Tweedie model with the following parametrisation:

$\mu = \lambda\alpha\gamma$; $\Phi = \frac{\lambda^{1-p} * (\alpha\gamma)^{2-p}}{2-p}$; $p = \frac{\alpha+2}{\alpha+1} \in]1;2[$;

$$\mathbb{E}(Y) = \mu = \lambda\alpha\gamma$$

$$Var(Y) = \Phi\mu^p = \lambda\gamma^2\alpha * (1+\alpha)$$

Thus, direct modeling of the global expectation $\mathbb{E}(Y)$ is possible using a generalized linear model with a logarithmic link function to insure positivity of the means:

$$log(\mathbb{E}(Y \mid X)) = \beta X$$

We used the *cplm* R package to implement Compound Poisson-Gamma models (47).

Negative Binomial models

We also attempted to use a model for count data by multiplying NETs' values by 1,000 to ensure discreteness without creating new ex-aequos. Let Y be a random variable following a Poisson distribution which depends on a single parameter $\lambda > 0$:

$$\mathbb{E}(Y) = Var(Y) = \lambda$$

The Poisson model is adapted to model the expectation of a count variable using a generalized linear model with a logarithmic link function:

$$log(\mathbb{E}(Y \mid X)) = \beta X$$

The Negative Binomial model is an extension to the Poisson model in the presence of over-dispersion of the outcome: $Var(Y) > \mathbb{E}(Y)$ (48, 49). In that case, the variance of Y is linked to its expectation through the following relationship: $Var(Y) = \mathbb{E}(Y) + k * \mathbb{E}(Y)^2$ where $k > 0$ is a dispersion parameter. This model is also part of generalized linear models and its link function is the logarithm. Negative Binomial model can also be represented as Poisson distributions with a Gamma distributed means where $Y \sim Pois(\lambda)$ and $\lambda \sim Gamma(\alpha, \gamma)$ (50). However, unlike the Compound Poisson-Gamma presented above, the two variables Y and λ are not independent from each other.

Models' comparison

The three aforementioned tested models were applied to NETs data while adjusting for age, sex, smoking and case-control status. The fit of these models to NETs data were assessed in two ways. First, we computed the Root Mean Square Error (RMSE) of each tested model defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

where N is the sample size, $\hat{y}_i$ the prediction of the i^{th} individual according to its covariates provided by a given modeling approach and y_i the observed value. Instead of predicting $\hat{y}_i$ by $\mathbb{E}(Y|X,\hat{\beta})$ that cannot be equal to zero, we used simulated predictions. For each studied model, a random value was generated for each individual according to its explanatory variables and the estimated model parameters. This process was repeated 1,000 times and the mean of RMSEs over the 1000 replicates was reported. As the Tobit model predicts negative values that are not observed in our semicontinuous outcome, these were imputed at zero to calculate the corresponding RMSE.

Second, we graphically assessed the fit of models predictions using Quantile-Quantile plot (QQplot) of the observations and predictions for each tested model.

Simulation Study

A simulation study was conducted to evaluate the control of the type I error (α) of the Negative Binomial and Compound Poisson-Gamma models in the context of genetic association studies as well as their sensibility to outliers. From the observed NETs data distribution, we randomly generated $S = 1, \ldots, 10000$ bootstrapped samples of size $N = 657$. For each bootstrapped sample, all individuals were randomly assigned 4 independent genotypes under the assumption of Hardy-Weinberg and corresponding to 4 SNPs with allele frequencies 0.01, 0.05, 0.10 and 0.20. The association of SNPs with the outcome was tested under the assumption of additive allele effects. This procedure was used to simulate semicontinuous data that mimic the NETs data observed in FARIVE and to allow the evaluation of the robustness of the two studied models (Negative Binomial and Compound Poisson-Gamma models) to a deviation from their underlying distribution. To assess the robustness to outliers, each simulated dataset was also analysed after the exclusion of individuals with NET level higher than 0.5, a threshold corresponding on average to the exclusion of 3% of individuals.

For each tested model, the number of times a SNP was found statistically significant at α=0.05, α=0.01, and α=0.001 was used to compute its empirical type I error.

RESULTS

Population characteristics

The main characteristics of the FARIVE participants used in this work are presented in Table 1. There is approximately 40% of men, 20% of current smokers and individuals are on average 53 years old. The distribution of NETs plasma levels observed in FARIVE is shown in Figure 1A. Approximately 16% of exact zeros were observed with a higher proportion among VT cases compared to controls (20% and 10% respectively). To analyse NETs as count data, observed values were multiplied by 1,000. This induced a large overdispersion (mean=78; variance=26,064) leading to the adoption of a Negative Binomial model for analysing such data.

Clinical Variables & Goodness of fit

Table 2 reports the association of clinical covariates with NETs plasma levels in each of the three studied models, Tobit Compound Poisson-Gamma and Negative Binomial. The Tobit model assumes a linear association of the covariates on the expected mean of the latent variable, i.e. the true value of NETs. For example, each 10-year increase in age is associated with an increase of 0.005 on the expected mean of the latent variable of NETs plasma levels, given the other covariates are held constant. Regarding the two other models, as a logarithmic link function is used, the association of covariates on the expected mean of NETs is multiplicative. As a consequence, for the Compound Poisson-Gamma model, each 10-year increase of age is associated with an expected mean of NETs plasma levels multiplied by 1.05 (= $e^{0.05}$). Similar interpretation holds for the Negative Binomial model that yielded regression parameters very close to those obtained via the Compound Poisson-Gamma models.

RMSEs provided by the three models are shown in **Table 2**. The lowest RMSE was observed for the Compound Poisson-Gamma model while the Negative Binomial model exhibited the highest one. Graphically, the Compound Poisson-Gamma (**Figure 1C**) and the Negative Binomial (**Figure 1D**) showed similar distributions of their predicted values even if the right skewedness was slightly less pronounced for the Compound Poisson-Gamma distribution. These distributions were rather close to that observed for the original NETs data (**Figure 1A**). By contrast, the Tobit distribution (**Figure 1B**) substantially deviated from the original data and looked like a left-truncated Gaussian distribution.

Quantile-Quantile plots of the observed vs predicted values did not visually show obvious deviation from the bisection line, except for high values (above 0.5 in the original NET scale), for the Compound Poisson-Gamma (**Figure 2B**) and the Negative Binomial (**Figure 2C**) models. Conversely, for the Tobit model (**Figure 2A**), the QQplot line deviated from the bisection line from the lower values.

Altogether, these observations suggest that the Compound Poisson-Gamma model seems the most adequate to analyze FARIVE NETs data. Nevertheless, we conducted a GWAS on NETs plasma levels

using each of the three models discussed above in order to get additional elements of comparison between these models.

GWAS analysis on NETs plasma levels

A total of 9,670,724 autosomal SNPs with imputation criterion $r^2 > 0.3$ and minor allele frequencies (MAF) > 0.01 were tested for association with NETs plasma levels using the Tobit, Negative Binomial and Compound Poisson-Gamma models. Quantile-Quantile plots for the observed and expected p-values summarizing the GWAS results for each model are shown in **Figure 3**. While the whole set of association results was compatible with what was expected under the null hypothesis of no genetic association for the Compound Poisson-Gamma model **(Figure 3B)**, strong deviations were observed for the Tobit (**Figure 3A**) and Negative Binomial (**Figure 3C**) models. By restricting the GWAS results to SNPs with MAF greater than 5%, inflation was no longer observed for the Tobit model (**Supplementary Figure S1A**, genomic inflation factor λ=0.96) while the Negative Binomial model remained strongly inflated (**Supplementary Figure S1C**, λ =1.46).

To further explore the remaining inflation, we re-ran the GWAS under the Compound Poisson-Gamma and Negative Binomial models after excluding 19 FARIVE participants (~3%) with NETs plasma levels higher than 0.5. Inflation in the Negative Binomial model was considerably decreased (**Supplementary Figure S2B**) and completely vanished when we additionally restricted the GWAS analysis on SNP with MAF >5% (**Supplementary Figure S3B**, λ=1.03). Finally, we conducted simulation studies (see Methods) that demonstrated that the association test in the Negative Binomial model exhibited inflated type I error (α) for the three nominal values of α considered (**Supplementary Table S1**) when data distribution fit the one observed for NETs plasma levels in the FARIVE study. Type I error was rather well controlled in absence of extreme values (**Supplementary Table S2**). Of note, these simulations also show that the Compound Poisson-Gamma model generally well controls the nominal type I error. All these observations add support for the use of the Compound Poisson-Gamma model for the GWAS analysis of NETs plasma levels.

The corresponding Manhattan plot shown in **Supplementary Figure S4** revealed one genome-wide significant locus. The lead SNP rs57502213 is a deletion of two nucleotides (TC), mapping to the miR-155 hosting gene (*MIR155HG*). This variant had a MAF of ~7%, exhibited a good imputation quality (r^2=0.92) and its minor allele was associated with a 2.53 fold increase (95% confidence interval [1.85 – 3.47], *P*=1.42x10^{-8}) in NETs plasma levels. The average NETs plasma levels were higher in carriers of the deletion of the TC allele than in non-carriers (0.15 vs 0.07). This pattern of association was consistent in VT cases and in controls (**Table 3**). Full GWAS summary statistics are available on GWAS catalog under the accession number GCP000431(51).

DISCUSSION

This work was motivated by the search of genetic factors associated with NETs plasma levels exhibiting a semicontinuous distribution. We compared three different modeling strategies, Tobit, Negative Binomial and Compound Poisson-Gamma models, that handle the excess of zero and the asymmetric distribution while allowing the estimation of a single regression parameter for characterizing the association between an explanatory variable and the global mean of the semicontinuous outcome.

Visual inspection showed that both the Negative Binomial and Compound Poisson-Gamma models fit better the observed NETs distribution than the Tobit model. Indeed, the underlying hypothesis of a left-truncated Gaussian distribution with only two parameters makes the Tobit model less-flexible than Compound Poisson-Gamma and Negative Binomial models. RMSE analysis provided further support for the use of Compound Poisson-Gamma model. Of note, the definition of this model matches quite well the biological mechanisms underlying NETs production as it is intuitively reasonable to assume that the number of dead neutrophils follows a Poisson distribution, and that each of these rejects a certain quantity of NETs that would follow a Gamma distribution.

Our GWAS and simulation studies revealed that the Tobit and Negative Binomial models were prone to strong inflation of p-values. While this inflation could be attributable to SNPs with low allele frequency (MAF<5%) for the Tobit model, this inflation was due to both low allele frequency SNPs and extreme positive values of the outcome for the Negative Binomial model. The poor control of type I error by the Negative Binomial model has already been highlighted in previous work in the context of RNA-Seq analysis (52). The Compound Poisson-Gamma model was much more robust to these two phenomena. Using this model, we identified one significant locus on chr21q21.3 associated with NETs plasma levels. This locus maps to a long non coding RNA that hosts miR-155 (and as such is referred to as miR155HG, for Hosting Genes) and the lead SNP was rs57502213, an intronic deletion in *miR155HG*. While several recent publications have highlighted the role of miR-155 in the NETs formation (53, 54), little information is available in public resources about the possible functional impact of rs57502213. This SNP is in moderate linkage disequilibrium (pairwise D' > 0.50) with several other nearby variants located in *MRPL39*, *GABPA* and *APP*, the latter having also been reported to be involved in NETs formation (55). Note that another GWAS on NETs plasma levels has recently been conducted in the Rotterdam study (56). Despite the large sample size of this study (~5600 individuals), no significant genome-wide association was detected and the association of our lead SNP did not replicate there (P=0.14). However, different kits were used to measure NETs levels in the two studies and recent works have emphasized the need for standardized methods for NETs measurements (57, 58). Of note, in the Rotterdam study, NETs were analyzed using a log-transformed model suggesting that no zero values (or few) were observed. This contrasts with FARIVE data and could contribute to the heterogeneity of findings between studies. Nevertheless, because of the increasingly recognized role of

the chr21q21.3 locus in NETs biology, further works deserve to be conducted to clarify the genetic signal observed in the present study.

To conclude, our work indicates that the modeling strategy for a semicontinuous outcome is crucial, but not straightforward. The choice of the model should take into account the nature of the (biological) process generating zero values, the distribution of the outcome and, especially, the presence of extreme values. The Tobit model with only two parameters is less flexible than Compound Poisson-Gamma and Negative Binomial models and our work shows that the Compound Poisson-Gamma model, while still marginally used, is more robust than the Negative Binomial model to outliers and low allele frequency making. This make it well suitable for GWAS analysis on semicontinuous trait and its use as an alternative to Negative Binomial model would deserve to be explored in the context of RNA-Seq analysis.

DATA AND CODE AVAILABILITY

Full GWAS summary statistics are available on GWAS catalog under the accession number GCP000431. The code supporting the current study is available from the corresponding author on request.

ACKNOWLEDGEMENT

GM benefited from the EUR DPH, a PhD program supported within the framework of the PIA3 (Investment for the future). Project reference 17-EURE-0019.

Statistical analyses benefited from the CBiB computing centre of the University of Bordeaux.

This project was carried out in the framework of the INSERM GOLD Cross-Cutting program (P-EM, D-AT).

FUNDING

GM and D-AT are supported by the EPIDEMIOM-VT Senior Chair from the University of Bordeaux initiative of excellence IdEX.

The FARIVE study was supported by grants from the Fondation pour la Recherche Médicale, the Programme Hospitalier de recherche Clinique (PHRC 20 002; PHRC2009 RENOVA-TV), the Fondation de France, and the Leducq Foundation. FARIVE genetic data were funded by the GENMED Laboratory of Excellence on Medical Genomics [ANR-10-LABX-0013], a research program managed by the National Research Agency (ANR) as part of the French Investment for the Future.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

ETHICS APPROVAL

Research have been performed in accordance with the Declaration of Helsinki. The FARIVE study was approved by the “Comité consultatif de protection des personnes dans la recherche biomedicale” (Project n° 2002-034). Written informed consent to participate was obtained from all FARIVE participants.

REFERENCES

1. Min,Y. and Agresti,A. Modeling Nonnegative Data with Clumping at Zero: A Survey.

2. Folkersen,L., Gustafsson,S., Wang,Q., Hansen,D.H., Hedman,Å.K., Schork,A., Page,K., Zhernakova,D.V., Wu,Y., Peters,J., *et al.* (2020) Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.*, **2**, 1135–1148.

3. Cragg,J.G. (1971) Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica*, **39**, 829–844.

4. Tobin,J. (1958) Estimation of Relationships for Limited Dependent Variables. *Econometrica*, **26**, 24–36.

5. Farewell,V.T., Long,D.L., Tom,B.D.M., Yiu,S. and Su,L. (2017) Two-Part and Related Regression Models for Longitudinal Data. *Annu. Rev. Stat. Its Appl.*, **4**, 283–315.

6. Feng,X., Lu,B., Song,X. and Ma,S. (2019) Financial literacy and household finances: A Bayesian two-part latent variable modeling approach. *J. Empir. Finance*, **51**, 119–137.

7. Chen,E.Z. and Li,H. (2016) A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, **32**, 2611–2617.

8. Garbutt,D.J., Stern,R.D., Dennett,M.D. and Elston,J. (1981) A comparison of the rainfall climate of eleven places in West Africa using a two-part model for daily rainfall. *Arch. Meteorol. Geophys. Bioclimatol. Ser. B*, **29**, 137–155.

9. Hartman,B., Larson,C., Kunkel,C., Wight,C. and Warr,R.L. A Two-Part Model of the Individual Costs of Chronic Kidney Disease.

10. Rustand,D., Briollais,L., Tournigand,C. and Rondeau,V. (2022) Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data. *Biostatistics*, **23**, 50–68.

11. Tooze,J.A., Midthune,D., Dodd,K.W., Freedman,L.S., Krebs-Smith,S.M., Subar,A.F., Guenther,P.M., Carroll,R.J. and Kipnis,V. (2006) A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *J. Am. Diet. Assoc.*, **106**, 1575–1587.

12. Amore,M.D. and Murtinu,S. (2021) Tobit models in strategy research: Critical issues and applications. *Glob. Strategy J.*, **11**, 331–355.

13. Chen,T., Ma,S., Kobie,J., Rosenberg,A., Sanz,I. and Liang,H. (2016) Identification of significant B cell associations with undetected observations using a Tobit model. *Stat. Interface*, **9**, 79–91.

14. Debnath,A.K., Blackman,R. and Haworth,N. (2014) A Tobit model for analyzing speed limit compliance in work zones. *Saf. Sci.*, **70**, 367–377.

15. McBee,M. (2010) Modeling Outcomes With Floor or Ceiling Effects: An Introduction to the Tobit Model. *Gift. Child Q.*, **54**, 314–320.

16. van den Broek,J. (1995) A Score Test for Zero Inflation in a Poisson Distribution. *Biometrics*, **51**, 738–743.

17. Allison,P.D. (2012) Logistic Regression Using SAS: Theory and Application, Second Edition SAS Institute.

18. Tweedie,M.C. (1984) An index which distinguishes between some important exponential families. In.Vol. 579, pp. 579–604.

19. Gilchrist,R. and Drinkwater,D. (2000) The use of the Tweedie distribution in statistical modelling. In Bethlehem,J.G., van der Heijden,P.G.M. (eds), *COMPSTAT*. Physica-Verlag HD, Heidelberg, pp. 313–318.

20. Jørgensen,B., Martínez,J.R. and Vinogradov,V. (2009) Domains of attraction to Tweedie distributions. *Lith. Math. J.*, **49**, 399–425.

21. Kurz,C.F. (2017) Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Med. Res. Methodol.*, **17**, 171.

22. Brown,J.E. and Dunn,P.K. (2011) Comparisons of Tobit, Linear, and Poisson-Gamma Regression Models: An Application of Time Use Data. *Sociol. Methods Res.*, **40**, 511–535.

23. Kimball,A.S., Obi,A.T., Diaz,J.A. and Henke,P.K. (2016) The Emerging Role of NETs in Venous Thrombosis and Immunothrombosis. *Front. Immunol.*, **7**, 236.

24. de Boer,O., Li,X., Teeling,P., Mackaay,C., Ploegmakers,H., van der Loos,C., Daemen,M., de Winter,R. and van der Wal,A. (2013) Neutrophils, neutrophil extracellular traps and interleukin-17 associate with the organisation of thrombi in acute myocardial infarction. *Thromb. Haemost.*, **109**, 290–297.

25. Hisada,Y., Grover,S.P., Maqsood,A., Houston,R., Ay,C., Noubouossie,D.F., Cooley,B.C., Wallén,H., Key,N.S., Thålin,C., *et al.* (2020) Neutrophils and neutrophil extracellular traps enhance venous thrombosis in mice bearing human pancreatic tumors. *Haematologica*, **105**, 218–225.

26. Ruhnau,J., Schulze,J., Dressel,A. and Vogelgesang,A. (2017) Thrombosis, Neuroinflammation, and Poststroke Infection: The Multifaceted Role of Neutrophils in Stroke. *J. Immunol. Res.*, **2017**, 5140679.

27. Laridan,E., Martinod,K. and De Meyer,S.F. (2019) Neutrophil Extracellular Traps in Arterial and Venous Thrombosis. *Semin. Thromb. Hemost.*, **45**, 86–93.

28. Diamond,S.L. (2016) Systems Analysis of Thrombus Formation. *Circ. Res.*, **118**, 1348–1362.

29. Ng,H., Havervall,S., Rosell,A., Aguilera,K., Parv,K., von Meijenfeldt,F.A., Lisman,T., Mackman,N., Thålin,C. and Phillipson,M. (2021) Circulating Markers of Neutrophil Extracellular Traps Are of Prognostic Value in Patients With COVID-19. *Arterioscler. Thromb. Vasc. Biol.*, **41**, 988–994.

30. Wang,L., Zhou,X., Yin,Y., Mai,Y., Wang,D. and Zhang,X. (2019) Hyperglycemia Induces Neutrophil Extracellular Traps Formation Through an NADPH Oxidase-Dependent Pathway in Diabetic Retinopathy. *Front. Immunol.*, **9**, 3076.

31. Zhu,L., Liu,L., Zhang,Y., Pu,L., Liu,J., Li,X., Chen,Z., Hao,Y., Wang,B., Han,J., *et al.* (2018) High Level of Neutrophil Extracellular Traps Correlates With Poor Prognosis of Severe Influenza A Infection. *J. Infect. Dis.*, **217**, 428–437.

32. Martínez-Alemán,S.R., Campos-García,L., Palma-Nicolas,J.P., Hernández-Bello,R., González,G.M. and Sánchez-González,A. (2017) Understanding the Entanglement: Neutrophil Extracellular Traps (NETs) in Cystic Fibrosis. *Front. Cell. Infect. Microbiol.*, **7**, 104.

33. Masucci,M.T., Minopoli,M., Del Vecchio,S. and Carriero,M.V. (2020) The Emerging Role of Neutrophil Extracellular Traps (NETs) in Tumor Progression and Metastasis. *Front. Immunol.*, **11**, 1749.

34. Trégouët,D.-A., Heath,S., Saut,N., Biron-Andreani,C., Schved,J.-F., Pernod,G., Galan,P., Drouet,L., Zelenika,D., Juhan-Vague,I., *et al.* (2009) Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood*, **113**, 5298–5303.

35. Granger,V., Peyneau,M., Chollet-Martin,S. and de Chaisemartin,L. (2019) Neutrophil Extracellular Traps in Autoimmunity and Allergy: Immune Complexes at Work. *Front. Immunol.*, **10**, 2824.

36. Chang,C.C., Chow,C.C., Tellier,L.C., Vattikuti,S., Purcell,S.M. and Lee,J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

37. Das,S., Forer,L., Schönherr,S., Sidore,C., Locke,A.E., Kwong,A., Vrieze,S.I., Chew,E.Y., Levy,S., McGue,M., *et al.* (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.

38. White,P.C., Hirschfeld,J., Milward,M.R., Cooper,P.R., Wright,H.J., Matthews,J.B. and Chapple,I.L. c. (2018) Cigarette smoke modifies neutrophil chemotaxis, neutrophil extracellular trap formation and inflammatory response-related gene expression. *J. Periodontal Res.*, **53**, 525–535.

39. Ortmann,W. and Kolaczkowska,E. (2018) Age is the work of art? Impact of neutrophil and organism age on neutrophil extracellular trap formation. *Cell Tissue Res.*, **371**, 473–488.

40. Yuan,X.Z. Sex differences in neutrophil extracellular trap formation.

41. Yee,T.W. (2015) Vector generalized linear and additive models: with an implementation in R Springer, New York, NY.

42. Zhou,H., Qian,W. and Yang,Y. (2020) Tweedie gradient boosting for extremely unbalanced zero-inflated data. *Commun. Stat. - Simul. Comput.*, 10.1080/03610918.2020.1772302.

43. Jørgensen,B. (1987) Exponential Dispersion Models. *J. R. Stat. Soc. Ser. B Methodol.*, **49**, 127–145.

44. Fox,J. (2016) Applied regression analysis and generalized linear models Third edition. SAGE, Los Angeles.

45. Dunn,P.K. and Smyth,G.K. (2008) Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Stat. Comput.*, **18**, 73–86.

46. Dzupire,N.C., Ngare,P. and Odongo,L. (2018) A Poisson-Gamma Model for Zero Inflated Rainfall Data. *J. Probab. Stat.*, **2018**, 1–12.

47. Zhang,Y. (2013) Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Stat. Comput.*, **23**, 743–757.

48. Hoef,J.M.V. and Boveng,P.L. (2007) Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data? *Ecology*, **88**, 2766–2772.

49. Gardner,W., Mulvey,E.P. and Shaw,E.C. (1995) Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychol. Bull.*, **118**, 392–404.

50. Gorshenin,A.K. and Korolev,V.Yu. (2018) Scale Mixtures of Frechet Distributions as Asymptotic Approximations of Extreme Precipitation. *J. Math. Sci.*, **234**, 886–903.

51. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E., *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

52. Gauthier,M., Agniel,D., Thiébaut,R. and Hejblum,B.P. (2020) dearseq: a variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics Bioinforma.*, **2**, lqaa093.

53. Hawez,A., Taha,D., Algaber,A., Madhi,R., Rahman,M. and Thorlacius,H. (2022) MiR-155 regulates neutrophil extracellular trap formation and lung injury in abdominal sepsis. *J. Leukoc. Biol.*, **111**, 391–400.

54. Hawez,A., Al-Haidari,A., Madhi,R., Rahman,M. and Thorlacius,H. (2019) MiR-155 Regulates PAD4-Dependent Formation of Neutrophil Extracellular Traps. *Front. Immunol.*, **10**.

55. Canobbio,I., Visconte,C., Momi,S., Guidetti,G.F., Zarà,M., Canino,J., Falcinelli,E., Gresele,P. and Torti,M. (2017) Platelet amyloid precursor protein is a modulator of venous thromboembolism in mice. *Blood*, **130**, 527–536.

56. Donkel,S.J., Portilla Fernández,E., Ahmad,S., Rivadeneira,F., van Rooij,F.J.A., Ikram,M.A., Leebeek,F.W.G., de Maat,M.P.M. and Ghanbari,M. (2021) Common and Rare Variants Genetic Association Analysis of Circulating Neutrophil Extracellular Traps. *Front. Immunol.*, **12**, 615527.

57. Prével,R., Dupont,A., Labrouche-Colomer,S., Garcia,G., Dewitte,A., Rauch,A., Goutay,J., Caplan,M., Jozefowicz,E., Lanoix,J.-P., *et al.* (2022) Plasma Markers of Neutrophil Extracellular Trap Are Linked to Survival but Not to Pulmonary Embolism in COVID-19-Related ARDS Patients. *Front. Immunol.*, **13**, 851497.

58. Rada,B. (2019) Neutrophil Extracellular Traps. *Methods Mol. Biol. Clifton NJ*, **1982**, 517–528.

TABLES AND FIGURES LEGENDS

Figure 1: Distribution of NETs plasma levels

This figure presents the distribution of the observed NETs plasma levels (A), predictions from the Tobit model (B), the Compound Poisson-Gamma model (C) and the Negative Binomial model (D).

Figure 2: QQplots of observations and predictions from the three models

This figure presents the Quantile-Quantile plots of observations and predictions from Tobit (A), Compound Poisson-Gamma (B) and Negative Binomial (C) models. The red line represents the perfect match between observations and predictions.

Figure 3: Quantile-Quantile plots from the GWAS results on NETs plasma levels

This figure presents the Quantile-Quantile plots of the p-values distributions from the GWAS with Tobit (A), Compound Poisson-Gamma (B) and Negative Binomial (C) models.

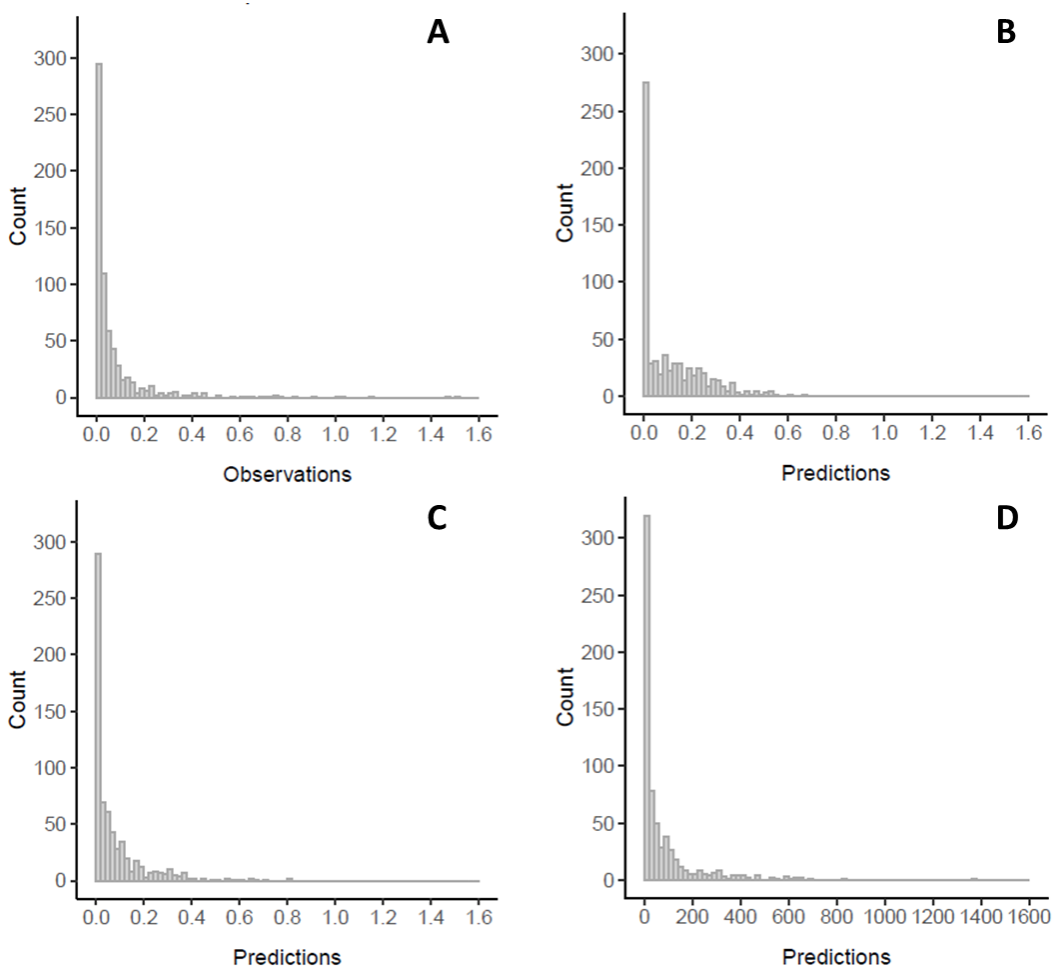


Figure 1: Distribution of NETs plasma levels

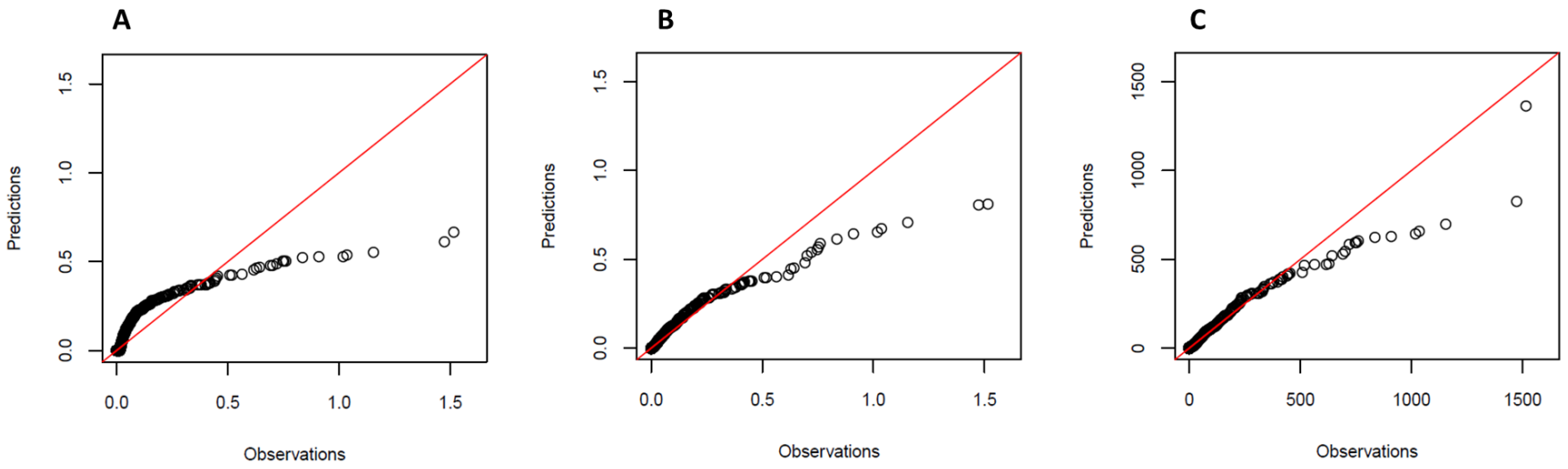


Figure 2: QQplots of observations and predictions from the three models

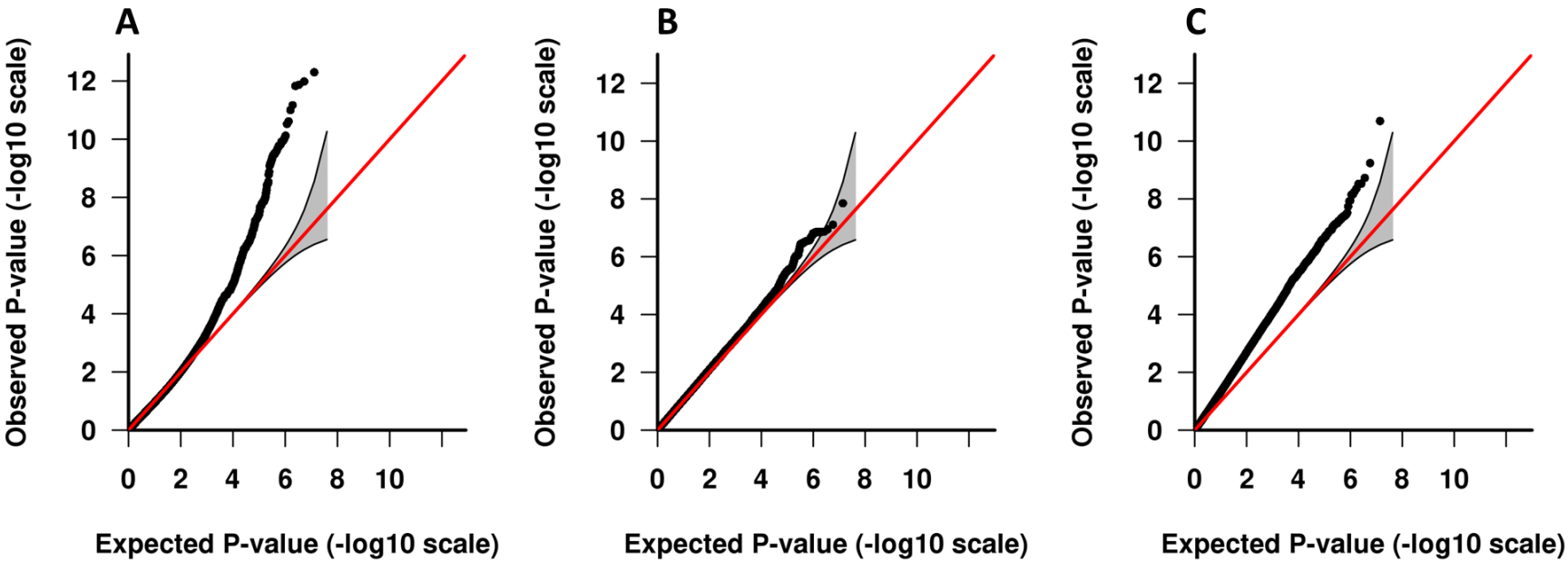


Figure 3: Quantile-Quantile plots from the GWAS results on NETs plasma levels

TABLES

Table 1: **Main characteristics of the FARIVE study**

	Total N=657	**VT‡ Cases N=372**	**Controls N=285**
	N (%)	**N (%)**	**N (%)**
Sex - Men	256 (39.0%)	141 (37.9%)	115 (40.4%)
Age at sampling *(Mean ± SD§)*	53.0 ± 18.8	53.3 ± 19.3	52.8 ± 18.2
Smoking status			
Current smoker	128 (19.5%)	62 (16.7%)	66 (23.2%)
Former smoker / Never	529 (80.5%)	310 (83.3%)	219 (76.8%)
Neutrophil Extracellular Traps levels			
All values			
Mean ± SD	0.08 ± 0.16	0.05 ± 0.11	0.12 ± 0.20
Median [Q1;Q3]	0.03 [$4x10^{-3}$;0.07]	0.03 [$2x10^{-3}$;0.05]	0.04 [0.01;0.12]
Exact zero	104 (15.8%)	75 (20.2%)	29 (10.2%)

§ Standard Deviation

‡ Venous Thrombosis

Table 2: Comparison of regression parameter estimates on the FARIVE data according to the three models

	Tobit	**Compound Poisson-Gamma**	**Negative Binomial***
	Beta (SD)	Beta (SD)	Beta (SD)
Covariates			
Age (10y)	0.005 (0.004)	0.05 (0.04)	0.05 (0.04)
Sex (Males)	-0.01 (0.02)	-0.08 (0.16)	-0.04 (0.13)
Smoking (Non smokers)	0.05 (0.02)	0.50 (0.19)	0.52 (0.17)
Status (Controls)	-0.08 (0.01)	-0.87 (0.15)	-0.87 (0.13)
RMSE*	198.7† [145.5 ; 251.9]	193.9 [180.8 ; 207.1]	211.5 [187.0 ; 236.0]

*: For the distribution of NETs multiplied by 1,000. Mean [min-max] over 1,000 bootstrapped samples

†: Negative predictions were censored at zero

Table 3: Association of rs57502213 with NETs plasma levels in the FARIVE study

	Genotype for rs57502213		
	TC/TC	**TC/-**	**-/-**
All individuals (N=657)			
N	568	88	1
Mean ± SD	0.07 ± 0.13	0.15 ± 0.28	0.07
Exact zero	91 (16.0%)	13 (14.8%)	-
Cases (N=372)			
N	323	48	1
Mean ± SD	0.04 ± 0.06	0.12 ± 0.26	0.07
Exact zero	66 (20.4%)	9 (18.7%)	-
Controls (N=285)			
N	245	40	-
Mean ± SD	0.10 ± 0.18	0.20 ± 0.30	-
Exact zero	25 (10.2%)	4 (10.0%)	-

Supplementary data

Supplementary Table S1 : Evaluation of the control of the type I error in 10 000 bootstrap samples according to the frequency of the SNP tested with Compound Poisson-Gamma and Negative Binomial models

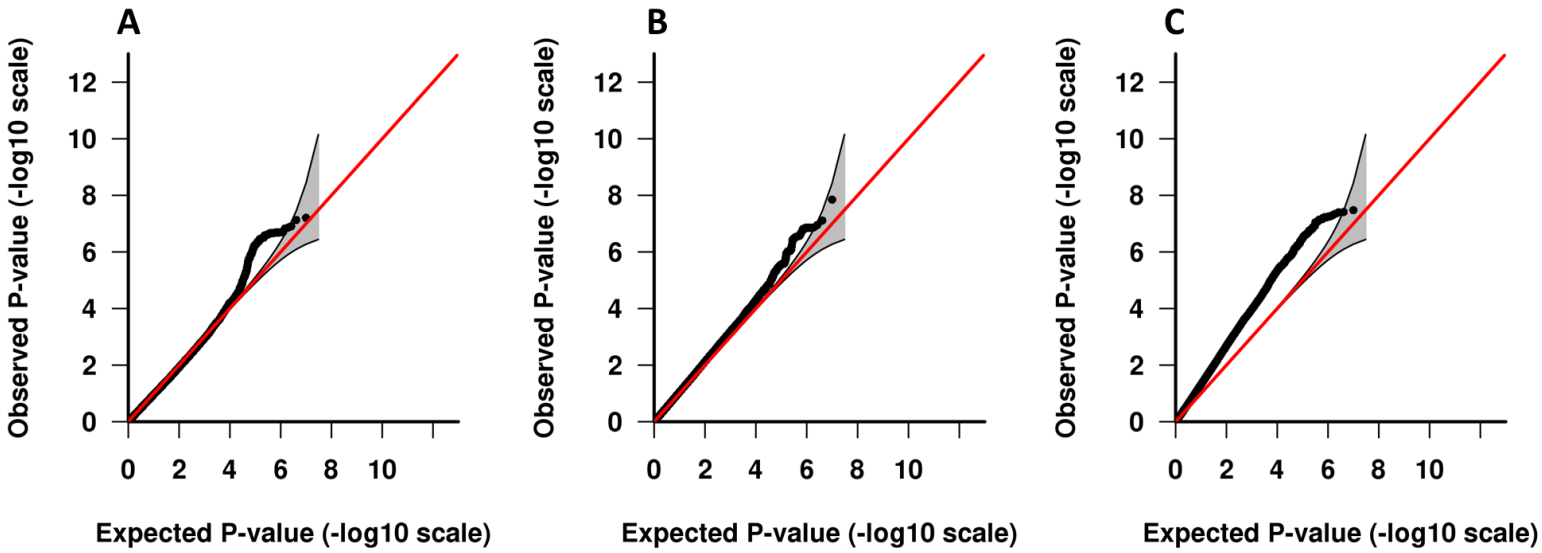
	Compound Poisson-Gamma				Negative Binomial			
α	**MAF 1%**	**MAF 5%**	**MAF 10%**	**MAF 20%**	**MAF 1%**	**MAF 5%**	**MAF 10%**	**MAF 20%**
0.05	0.043	0.053	0.054	0.059	0.119	0.112	0.117	0.114
0.01	0.007	0.010	0.013	0.013	0.039	0.036	0.038	0.038
0.001	0.0004	0.0014	0.0017	0.0011	0.0099	0.0077	0.0084	0.0072

Supplementary Table S2: Evaluation of the control of the type I error in 10 000 bootstrap samples when outliers are removed, according to the frequency of the SNP tested with Compound Poisson-Gamma and Negative Binomial models

	Compound Poisson-Gamma				Negative Binomial			
α	**MAF 1%**	**MAF 5%**	**MAF 10%**	**MAF 20%**	**MAF 1%**	**MAF 5%**	**MAF 10%**	**MAF 20%**
0.05	0.053	0.054	0.053	0.052	0.059	0.048	0.046	0.043
0.01	0.010	0.012	0.012	0.012	0.019	0.011	0.010	0.009
0.001	0.0005	0.0007	0.0013	0.0017	0.0047	0.0008	0.0010	0.0014

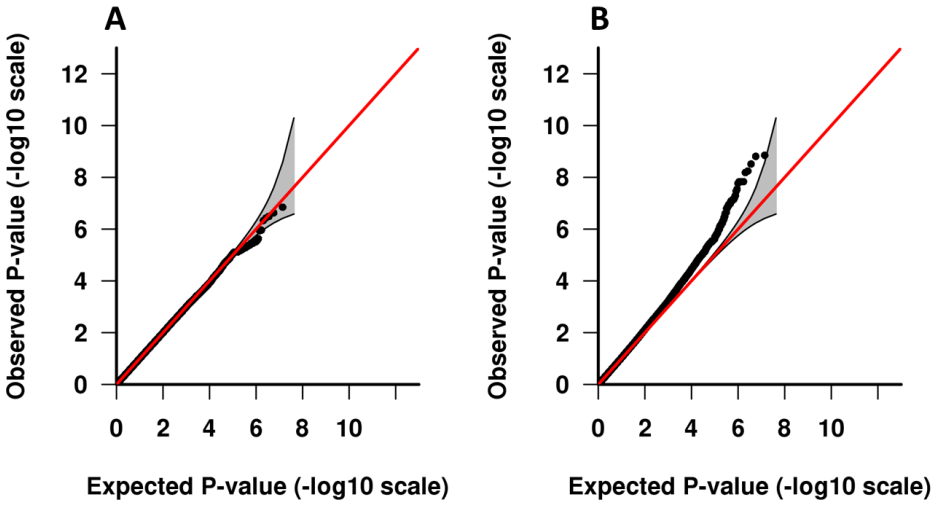
Supplementary Figure S1: Quantile-Quantile plots from the GWAS results with an allele frequency higher than 5% on NETs plasma levels

Tobit (A), Compound Poisson-Gamma (B) and Negative Binomial (C).



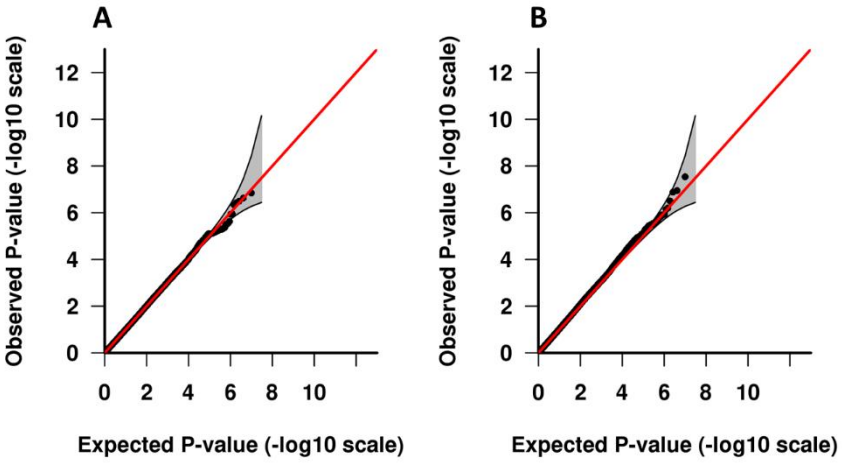
Supplementary Figure S2: Quantile-Quantile plots from the GWAS results on NETs plasma levels without outliers

Compound Poisson-Gamma (A) and Negative Binomial (B).

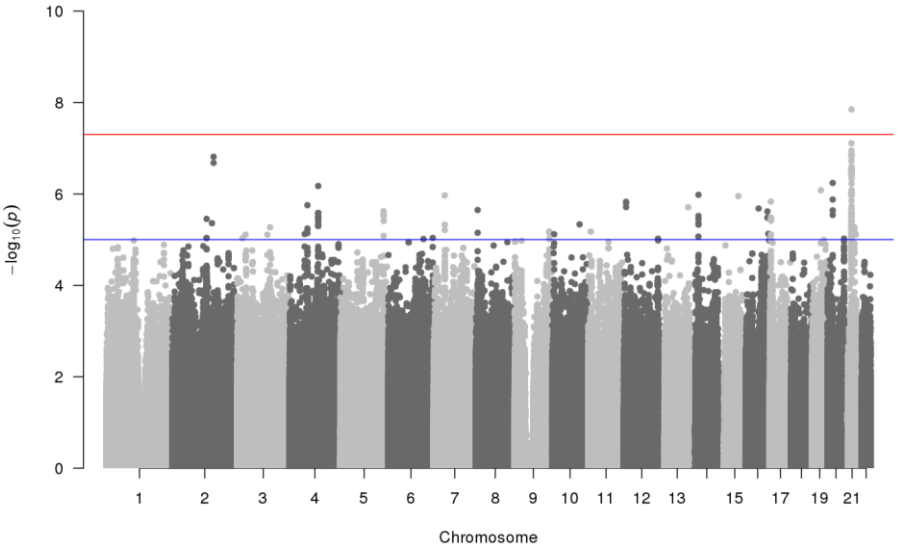


Supplementary Figure S3: Quantile-Quantile plots from the GWAS results with an allele frequency higher than 5% on NETs plasma levels without outliers

Compound Poisson-Gamma (A) and Negative Binomial (B).



Supplementary Figure S4: Manhattan plots from the GWAS results of NETs conducted with Compound Poisson-Gamma model



The $-log_{10}$ of the p-values are presented according to the position of the associated tested SNP across the genome. The genome wide significant threshold ($5x10^{-8}$) is represented with a red line.

6.2 Analyses supplémentaires

J'ai par la suite mené des analyses complémentaires sur ces données visant à 1) affiner l'origine du signal statistique obtenu sur le chromosome 21 et 2) étudier l'association entre les taux de NETs et les différentes protéines plasmatiques mesurées dans FARIVE.

6.2.1 Analyse haplotypique

La GWAS réalisée avec le modèle Composé Poisson-Gamma a permis d'identifier un signal dans la région de l'ARN non codant du miR-155 (*MIR155HG*) et le graphique d'association de la région est présenté en **Figure 15**.

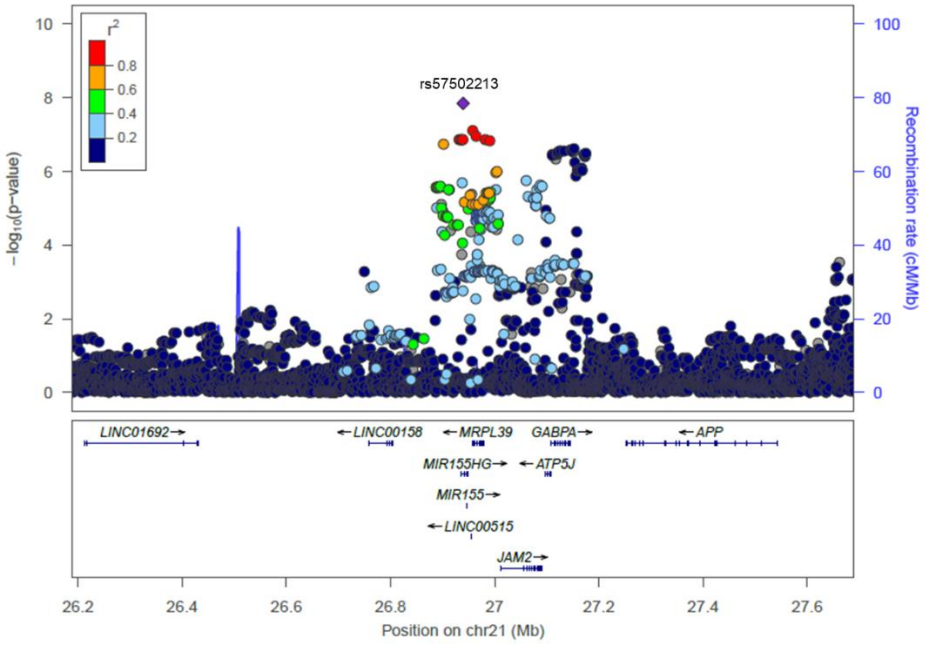


Figure 15 : Graphique d'association autour de la région identifiée par la GWAS des NETs dans l'étude FARIVE

Comme ce graphique l'illustre, d'autres variants en plus ou moins fort déséquilibre de liaison (r^2) avec le top SNP (rs57502213, délétion de deux nucléotides) étaient également très associés aux taux de NETs. Afin d'explorer plus en détail le signal obtenu dans cette région et tout particulièrement les variants proches des gènes *GABPA* et *APP*, de bons candidats d'après la littérature et par ailleurs impliqués dans le développement de pathologies neurodégénératives, j'ai mené une analyse haplotypique (Canobbio et al., 2017; Perdomo-Sabogal et al., 2016; Schmechel et al., 1988). A partir du déséquilibre de liaison et des haplotypes existants entre les SNPs les plus associés (pvaleur < 10^{-6}) aux taux de NETs dans la GWAS, j'ai pu identifier cinq

haplotypes fréquents à l'aide de l'outil en ligne *LDlink*, pouvant être inférés avec trois SNPs présentés en **Figure 16** (Machiela & Chanock, 2015).

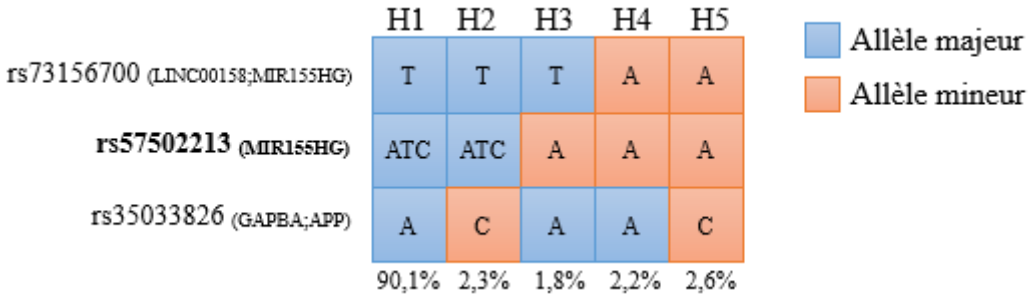


Figure 16 : Représentation des cinq haplotypes issus de la région identifiée par la GWAS des NETs dans l'étude FARIVE

L'association entre ces haplotypes (sous l'hypothèse d'effets additifs de chacun des cinq haplotypes) et les taux de NETs a été étudiée en utilisant le modèle Composé Poisson-Gamma ainsi que les mêmes covariables que pour la GWAS et les résultats sont présentés dans le **Tableau 5**. Cette analyse a permis de mettre en avant l'haplotype H5, composé des allèles mineurs des trois variants génétiques, avec des taux de NETs chez les porteurs de cet haplotype près de quatre fois supérieurs à ceux des porteurs de l'haplotype H1 (composé des allèles majeurs des trois variants génétiques). Les taux de NETs sont donc augmentés en présence de l'allèle mineur du rs57502213 (*MIR155HG)* mais ils le sont encore plus si l'individu est également porteur des allèles mineurs de rs73156700 (*LINC00158 ; MIR155HG*) et rs35033826 (*GAPBA ; APP*).

Tableau 5 : Associations entre les haplotypes du locus 21q21.3 du chromosome 21 et les taux de NETs dans l'étude FARIVE

	Risque Relatif (RR) [IC 95 %]	**P**
Haplotypes		
H1	Référence	
H2	1,04 [0,59 ; 1,83]	0,8945
H3	1,37 [0,72 ; 2,59]	0,3327
H4	1,37 [0,77 ; 2,45]	0,2817
H5	3,91 [2,53 ; 6,04]	$1{,}3\times10^{-9}$

Les cinq haplotypes du locus 21q21.3 ont été inférés à partir de rs57502213 (*MIR155HG),* rs73156700 (*LINC00158 ; MIR155HG*) et rs35033826 (*GAPBA ; APP*).

6.2.2 Analyse des taux plasmatiques de protéines

La technique de mesure des NETs étant très controversée, cette analyse supplémentaire avait pour objectif d'identifier d'autres partenaires moléculaires qui pourraient être impliqués dans les mécanismes de la production de NETs.

Dans l'étude FARIVE, les taux plasmatiques de 374 protéines ont été mesurés à l'inclusion des patients dans l'étude. J'ai donc testé leur association avec les taux de NETs avec le modèle Composé Poisson-Gamma en utilisant les mêmes variables d'ajustement que dans la GWAS ainsi que d'autres variables classiques de contrôle de ces données. Les taux plasmatiques de ces protéines présentant des corrélations plus ou moins fortes entre eux au sein de l'étude FARIVE, j'ai d'abord utilisé la méthode de Li & Ji qui se base sur les valeurs propres de la matrice de corrélation des protéines pour déterminer le nombre de protéines « indépendantes » (Li & Ji, 2005). Ce nombre a été estimé à 176 et j'ai ensuite utilisé la correction de Bonferroni pour 176 tests indépendants afin d'identifier les protéines dont les taux plasmatiques étaient significativement associés aux taux de NETs (pvaleur < 0,05/176 = $2{,}84\times10^{-4}$).

Les 374 protéines ont été testées mais seulement deux passaient le seuil de significativité (Risque Relatif (RR) = 0,60 ; pvaleur = $5{,}6\times10^{-6}$ et RR = 0,69 ; pvaleur = $1{,}5\times10^{-4}$, respectivement) : la première codée par le gène *CLEC3B* (*C-type lectin domain family 3 member B*) et la seconde par le gène *GPR183* (*G protein-coupled receptor 183*). Les associations de ces deux protéines étaient relativement homogènes entre les cas et les témoins même si elles sembleraient être plus associées chez les témoins, comme illustré dans le **Tableau 6**.

Tableau 6 : Associations entre les protéines plasmatiques et les taux de NETs dans l'étude FARIVE

		Global N = 657		Cas de MTEV N = 372		Témoins N = 285	
HPA	**Gène**	**RR**	**Pvaleur**	**RR**	**Pvaleur**	**RR**	**Pvaleur**
HPA034793	*CLEC3B*	0,60	$5{,}60\text{x}10^{-6}$	0,75	0,059	0,60	$8{,}45\text{x}10^{-6}$
HPA013784	*GPR183*	0,69	$1{,}51\text{x}10^{-4}$	0,75	0,043	0,65	$6{,}34\text{x}10^{-4}$
CABDY1707	*CRP*	1,27	$3{,}66\text{x}10^{-3}$	1,13	0,294	1,44	$9{,}48\text{x}10^{-4}$

De récents travaux sur *CLEC3B* ont montré que ce gène serait impliqué dans les mécanismes inflammatoires notamment via l'altération de la dégranulation des plaquettes et des neutrophiles (Rajasekaran et al., 2022). De plus, *CLEC3B* est également associé aux taux plasmatiques de la tétranectine, un régulateur de l'activation du plasminogène et des études ont

montré que les NETs contribueraient à la résistance de l'activation du plasminogène (Zhang et al., 2021). Le gène *GPR183* est impliqué dans les mécanismes inflammatoires puisqu'il serait exprimé principalement dans les lymphocytes et serait associé à la maladie inflammatoire de l'intestin (Bohrer et al., 2022; de Lange et al., 2017). A noter que l'association entre les taux de *CRP* (*C-reactive protein*) et les taux de NETs était proche du seuil de significativité (pvaleur = $3{,}7\times10^{-3}$), la CRP étant impliquée dans les mécanismes inflammatoires et son association avec les taux de NETs ayant déjà été montrée dans plusieurs travaux antérieurs (Kim et al., 2017; Zuo et al., 2020).

Les associations entre le rs57502213 (*MIR155HG)* identifié par la GWAS des taux de NETs, et les taux de *CLEC3B*, *GPR183* et *CRP* ont été évaluées chez 1 051 sujets de FARIVE (**Tableau 7**). Cependant, les résultats n'étaient pas significatifs et de la même façon les SNPs les plus associés aux taux de ces protéines ne ressortaient pas du tout dans la GWAS des NETs.

Tableau 7 : Résultats des analyses d'associations entre le rs57502213-A et les protéines des gènes *CLEC3B*, *GPR183* et *CRP* dans l'étude FARIVE (N = 1 051)

HPA	Gène	Beta (ET*)	Pvaleur
HPA034793	*CLEC3B*	-4,49 (11,9)	0,70
HPA013784	*GPR183*	-10,71 (11,7)	0,36
CABDY1707	*CRP*	1,15 (11,5)	0,92

* Ecart-type

Ces analyses complémentaires ont permis d'apporter de nouveaux éléments qui confortent le fait que la mesure des NETs utilisée dans FARIVE semble être un marqueur pertinent, ce qui améliore l'impact biologique des résultats de la GWAS sur les taux de NETs. De plus, ce travail a permis d'identifier d'autres biomarqueurs qu'il pourrait être pertinent de mesurer pour étudier leur lien avec les NETs.

6.3 Discussion

Les NETs sont un marqueur émergeant de l'inflammation qui pourrait être impliqué dans de nombreuses pathologies. La distribution semicontinue de ce biomarqueur nécessite une modélisation statistique adaptée pour l'analyse de ses déterminants. Ce travail est la première GWAS réalisée avec le modèle Composé Poisson-Gamma, un modèle adapté aux distributions semicontinues et présentant de bonnes propriétés statistiques avec un bon contrôle de l'erreur de type I, en étant robuste à la fois aux valeurs extrêmes et aux variants de faible fréquence allélique. L'utilisation de ce modèle a permis d'identifier plusieurs variants dans la région du chromosome 21 associés aux taux de NETs et localisés dans des gènes candidats préalablement identifiés dans des travaux expérimentaux chez la souris (Canobbio et al., 2017; Hawez et al., 2022, p. 155). Ce travail permet d'apporter de nouveaux éléments confirmant les observations faites chez la souris qui nécessiteraient une analyse plus approfondie de la région génomique identifiée et d'être répliqués dans un échantillon indépendant. En outre, un séquençage de cette région pourrait permettre d'identifier un variant rare sur l'haplotype H5 ou un autre haplotype avec une implication fonctionnelle (*voir section **6.2.1***). Par ailleurs, des travaux réalisés chez la souris ont mis en évidence que l'expression du gène *APP* dans les plaquettes serait impliquée dans la MTEV notamment via la régulation de la formation de la fibrine et de la quantité de NETs relâchée par les neutrophiles, renforçant ainsi l'importance de cette région dans la formation de NETs et l'intérêt qu'ils peuvent susciter pour d'autres pathologies (Canobbio et al., 2017).

Une limite majeure de ce travail concerne la mesure des NETs qui ne fait pas consensus puisque l'élaboration de recommandations internationales est toujours en cours. Néanmoins, les deux protéines significativement associées aux taux de NETs dans FARIVE sont impliquées dans des mécanismes inflammatoires ce qui suggère que la mesure réalisée reflète un mécanisme physiologique pertinent au regard de la physiologie des NETs.

De plus, si la taille de l'échantillon avait été plus importante, il aurait été intéressant d'étudier séparément les cas et les témoins de l'étude FARIVE puisque pour les témoins, la mesure des NETs a été réalisée à l'inclusion, en même temps que les mesures des protéines, alors que pour les cas de MTEV, les protéines ont aussi été mesurées à l'inclusion mais les taux de NETs ont été mesurés environ sept mois plus tard, ce qui correspond à l'arrêt du traitement anticoagulant.

A ce jour seule la GWAS sur les taux de NETs a été conduite mais d'autres analyses, comme des randomisations mendéliennes ou des corrélations génétiques, pourraient être

menées en utilisant les résultats de la GWAS pour identifier d'autres acteurs moléculaires. Ces analyses ont pour objectif d'identifier des relations causales ou des effets génétiques partagés entre deux phénotypes qui n'ont pas forcément été mesurés dans le même échantillon. Ces analyses seraient particulièrement intéressantes pour étudier le lien entre les NETs et la récidive de MTEV dont le caractère récurrent pourrait être expliqué par un dysfonctionnement des mécanismes de l'immunité acquise. De la même façon ces analyses pourraient être conduites avec d'autres phénotypes tels que la maladie d'Alzheimer ou les accidents vasculaires cérébraux, pour lesquels des travaux expérimentaux ont déjà mis en avant leur lien avec les NETs (Denorme et al., 2022; Laridan et al., 2017).

7 Projets annexes

Dans le cadre de ma thèse, j'ai également conduit une méta-analyse européenne sur les taux du récepteur du facteur H du complément 5 (CFHR5), une protéine impliquée dans la voie du complément même si elle n'apparait généralement pas sur les figures classiques de la cascade du complément comme celle que j'ai utilisée dans mon introduction (**Figure 7**) (Ferluga et al., 2017). Il m'a été proposé au cours de ma thèse de réaliser l'analyse GWAS de cette variable dans les études MARTHA et FARIVE où elle avait été mesurée. La distribution de cette variable étant Gaussienne, j'ai utilisé pour ces analyses GWAS le modèle linéaire classique implémenté dans Plink. J'ai ensuite réalisé une méta-analyse de ces résultats avec ceux obtenus dans une étude espagnole, l'étude *Riesgo de Enfermedad TROmboembólica VEnosa* – RETROVE, coordonnée par les Professeurs José Manuel Soria et Juan Carlos Souto. Cette méta-analyse portait donc sur près de 3 000 individus. Je me suis également chargée d'étudier l'effet des taux du CFHR5 sur le risque de récidive dans l'étude MARTHA, en utilisant uniquement les données prospectives. L'article associé à ce travail est actuellement en révision dans le journal *Nature Communications*. Au cours de la phase de révision de l'article, j'ai identifié d'autres études qui avaient réalisé des GWAS sur les taux du CFHR5, dont les projets *Omicscience* et *Decode* (Ferkingstad et al., 2021; Pietzner et al., 2021). Afin d'augmenter la puissance de nos analyses visant à identifier des polymorphismes associés à la variabilité inter-individuelle des taux plasmatiques de CFHR5, j'ai ensuite réalisé une méta-analyse de l'ensemble des données GWAS disponibles pour le CFHR5 qui contenait finalement plus de 50 000 sujets. Les résultats de ce travail sont détaillés en **Annexe 1** page 131.

De plus, au début de ma thèse j'ai appliqué la méthodologie originale des réseaux bayésiens que j'avais étudiée au cours de mon stage de Master 2 réalisé auprès de David-Alexandre Trégouët, pour identifier des micro-ARN associés au risque de récidive de MTEV dans l'étude MARTHA. Les réseaux bayésiens sont des modèles graphistes probabilistes représentés par des graphes acycliques orientés permettant de modéliser les dépendances et indépendances conditionnelles entre les variables et pouvant être modélisés sans qu'aucune relation *a priori* entre les variables ne soit connue. Des réseaux bayésiens ont été construits à partir de plus de 150 micro-ARN mesurés par une technique de séquençage dans le plasma de 344 patients. J'ai ensuite identifié les micro-ARN qui étaient trouvés comme étant des nœuds terminaux des réseaux pour tester leur effet sur le risque de récidive de MTEV dans l'étude MARTHA, en se basant sur l'hypothèse que ces nœuds terminaux, intégrant un mécanisme plus

global représenté par les relations au sein du réseau bayésien, pouvaient présenter les propriétés discriminantes les plus fortes vis-à-vis du risque de récidive. L'article associé à ce travail est présenté en **Annexe 2** page 218.

8 Discussion

Ces dernières années, les études d'associations pangénomiques sont devenues une stratégie couramment utilisée pour identifier des déterminants génétiques impliqués dans la physiopathologie des maladies humaines. Leur démocratisation a été facilitée notamment par la diminution importante des coûts des puces à ADN, le prix unitaire ayant été divisé par 10 pour atteindre désormais environ 30 € par échantillon. Le développement des techniques d'imputation ainsi que l'implémentation de méthodologies statistiques relativement standardisées dans des logiciels dédiés à l'analyse des données génétiques, ont également fortement contribué à la généralisation des approches dites GWAS. Cependant, ces logiciels se voulant faciles et rapides à utiliser n'implémentent généralement pas de méthodes statistiques plus complexes permettant d'analyser rigoureusement des données présentant des distributions moins conventionnelles ou issues de schémas d'études peu utilisés. C'est dans ce contexte que mes travaux de thèse ont porté sur le développement de deux approches méthodologiques permettant 1) d'analyser des données issues d'un schéma d'étude de type ambispectif et 2) de réaliser une analyse pangénomique d'une variable d'intérêt présentant une distribution semicontinue.

Dans mon premier travail de thèse, j'ai proposé une modélisation reposant sur un modèle de Cox pondéré permettant d'analyser des facteurs fixes dans le temps (comme les facteurs génétiques qui sont fixes dès la naissance) du risque de récidive de MTEV dans l'étude MARTHA qui repose sur un schéma d'étude ambispectif. En attribuant des poids aux individus à partir d'une estimation du risque de décès dans la population de MARTHA, cette approche a permis de maximiser la puissance statistique des analyses en intégrant à la fois les récidives pré- et post-inclusion, tout en tenant compte du biais de sélection lié à la survie des individus jusqu'au moment où leurs informations sur la récidive de MTEV ont été collectées. La principale limite de ce projet concerne l'estimation du risque de décès puisque le nombre de décès était faible (N = 73) et des variables d'intérêt comme le nombre d'antécédents personnels de MTEV ainsi que le temps calendaire n'ont pas pu être prises en compte dans la modélisation. De plus, l'intégration des données ambispectives a induit un non-respect de l'hypothèse de proportionnalité comme observé pour le caractère provoqué de la première MTEV.

En extension de ce projet, il serait intéressant d'évaluer la pertinence et les conditions qui pourraient permettre d'intégrer les polymorphismes lors de l'estimation du risque de décès afin que les poids soient dépendants du polymorphisme étudié. Dans un premier temps, cette

stratégie avait été adoptée mais les poids se sont avérés être trop sensibles aux variants basse fréquence et l'étendue des poids était trop importante malgré l'utilisation de méthodes de standardisation. Ce problème de valeurs extrêmes des poids a également été observé lors de l'estimation du risque du décès dans l'étude EDITH qui présente ce même schéma d'étude ambispectif et d'autres recherches sont nécessaires pour maximiser la puissance statistique pour l'analyse des récidives dans cet échantillon. De plus, dans l'étude EDITH toutes les caractéristiques des évènements antérieurs (date, localisation, …) ont été récoltées et il serait intéressant d'étudier non seulement les facteurs de risque de la première récidive mais également ceux des récidives suivantes à l'aide de modélisations de type multi-états si les effectifs sont suffisamment importants.

La stratégie de modélisation adoptée dans ce travail peut s'étendre aux analyses GWAS bien qu'il n'existe pas encore de logiciel optimisé pour estimer plusieurs millions de modèles de Cox pondérés. De plus, les paramètres obtenus s'interprètent comme ceux d'un modèle de Cox classique qui considère le même temps de base et il est donc possible de méta-analyser les résultats obtenus entre plusieurs études qui utilisent ces deux méthodes d'estimation. C'est d'ailleurs dans ce cadre que je coordonne la méta-analyse internationale des déterminants génétiques de la récidive de MTEV avec le consortium INVENT. Dans ce projet, des analyses par sous-groupes (selon le sexe, l'âge de la première MTEV, le type de la première MTEV à savoir TVP ou EP, le caractère provoqué de la première MTEV) sont aussi prévues. Les résultats de ces différentes méta-analyses sont attendus pour l'été prochain. Des analyses de type randomisation mendélienne (méthode permettant d'estimer l'effet causal d'une exposition sur un phénotype d'intérêt en se basant sur leurs déterminants génétiques) avec des phénotypes thrombotiques pertinents comme les plaquettes ou les taux des facteurs de la coagulation pourront être conduites à partir de ces résultats (Burgess et al., 2015; Didelez & Sheehan, 2007). Ces travaux permettront d'améliorer la compréhension des mécanismes biologiques sous-jacents de la récidive de MTEV, qui semblent être différents de ceux de la MTEV comme illustré lors de l'analyse des haplotypes du groupe sanguin ABO (*voir section **5.3***).

La réalisation des GWAS et les méta-analyses qui en découleront pourrait permettre, à terme, d'identifier un score pour prédire le risque de récidive de MTEV qui pourrait par la suite être utilisé en pratique clinique afin d'adapter la durée du traitement anticoagulant. En effet, si le patient a un risque très faible de récidive, il serait préférable de diminuer la durée du traitement afin de lui éviter de s'exposer à un risque accru d'hémorragie. A l'inverse, pour les patients avec un risque élevé de récidive, la durée de leur traitement anticoagulant pourrait être étendue.

Dans la seconde partie de ma thèse, des modèles adaptés aux distributions semicontinues (caractérisées par un excès de valeurs en zéro suivi d'une distribution continue avec asymétrie à droite) ont été comparés puis utilisés pour réaliser une analyse pangénomique d'un biomarqueur d'intérêt dans le champ des pathologies cardiovasculaires, les taux de NETs. Le modèle Composé Poisson-Gamma s'est avéré être un bon modèle dans le cadre de l'analyse des taux de NETs dans l'étude FARIVE. Il a montré de bonnes propriétés statistiques en étant à la fois robuste aux variants rares et à la présence d'individus avec des valeurs extrêmes. Le modèle Composé Poisson-Gamma est à l'heure actuelle principalement utilisé en économétrie ou en écologie mais son application dans les données de santé mériterait d'être davantage explorée, notamment dans le contexte des données de séquençage (ARN, micro-ARN, méthylation) qui sont principalement analysées avec le modèle Négatif Binomial malgré sa sensibilité aux valeurs extrêmes (Foster & Bravington, 2013; Gandy & Veraart, 2021; Gauthier et al., 2020; Jørgensen & Paes de Souza, 1994). Néanmoins, ce modèle n'est pas forcément adapté pour toutes les modélisations de données semicontinues et d'autres modélisations sont préférables lorsque la présence de zéros doit être étudiée séparément des valeurs strictement continues.

L'application de ce modèle à l'analyse pangénomique des taux de NETs dans l'étude FARIVE a permis d'identifier un bon candidat biologique situé dans le gène *MIR155HG*. Pour confirmer et affiner le résultat obtenu, il faudrait dans un premier temps répliquer cette association dans une étude indépendante et séquencer le locus d'intérêt pour identifier les variants ou l'haplotype fonctionnel sur lesquels des travaux expérimentaux pourraient être conduits.

Dans la continuité de ce projet, il pourrait être envisageable (malgré le faible effectif dans l'étude FARIVE) de réaliser des analyses de corrélation génétique ou de randomisation mendélienne pour identifier des variables biologiques et pathologies dans lesquelles les NETs pourraient être impliqués. De plus, l'activation de l'immunité acquise pourrait expliquer le caractère récurrent de la MTEV chez certains patients et il serait donc intéressant d'étudier l'effet des taux de NETs sur le risque de récidive de MTEV, ce que je ferai lorsque j'aurai finalisé la méta-analyse de GWAS de la récidive de MTEV.

En analyse de données, il est primordial d'adopter une modélisation statistique qui utilise toute l'information disponible afin de maximiser la puissance statistique et qui représente au mieux la variable d'intérêt, c'est-à-dire sans la transformer. Ce projet de thèse montre que malgré le nombre important de tests réalisés, cette philosophie peut également s'appliquer à l'analyse des données génétiques, même si le développement de logiciels serait nécessaire pour

optimiser les temps de calcul. En effet, le grand nombre de tests statistiques réalisés peut rendre certaines modélisations plus complexes à mettre en œuvre et il est de ce fait important d'utiliser des méthodes suffisamment rigoureuses dont le temps de calcul reste raisonnable. Les modélisations statistiques mises en œuvre dans ma thèse ouvrent de nouvelles perspectives, à la fois sur le plan clinique et sur le plan méthodologique, auxquelles j'espère pouvoir contribuer dans les prochaines années.

Annexes

Annexe 1 : *Elevated plasma Complement Factor H Regulating Protein 5 is associated with venous thromboembolism and COVID-19 severity*

Sanchez-Rivera L†, Iglesias MJ†, Ibrahim-Kosta M, […], **Munsch G**, […], Butler LM*, Trégouët D-A*, Odeberg J*. *Elevated plasma Complement Factor H Regulating Protein 5 is associated with venous thromboembolism and COVID-19 severity*. **medRxiv**. 2022 Jan 1;2022.04.20.22274046

Cet article a été mis à disposition sur le serveur de preprint de medRxiv (https://doi.org/10.1101/2022.04.20.22274046). La version actuellement en révision dans *Nature Communications* est présentée dans la section suivante. Pour une question de place, seuls les tableaux supplémentaires concernant la partie des analyses génétiques (à laquelle j'ai contribué) sont présentés.

Elevated plasma Complement Factor H Regulating Protein 5 is associated with venous thromboembolism

Maria Jesus Iglesias[1,2,3]†, Laura Sanchez-Rivera[1]†, Manal Ibrahim-Kosta[4], Clément Naudin[1,3], Gaëlle Munsch[5], Louisa Goumidi[4], Maria Farm[6,7], Philip M. Smith[8,9], Florian Thibord[10,11], Julia Barbara Kral-Pointner[12], Mun-Gwan Hong[1], Pierre Suchon[4], Marine Germain[5,13], Waltraud Schottmaier[12], Philip Dusart[1,3], Anne Boland[14,15], David Kotol[1], Fredrik Edfors[1], Mine Koprulu[16], Maik Pietzner[16,17], Claudia Langenberg[16,17,18], Scott M Damrauer[19,20], Andrew D Johnson[10,11], Derek M. Klarin[21], Nicholas L Smith[22,23,24], David M Smadja[25,26], Margareta Holmström[27], Maria Magnusson[6,27,28], Angela Silveira[8], Mathias Uhlén[1], Thomas Renné[29,30,31], Angel Martinez-Perez[32], Joseph Emmerich[33], Jean-Francois Deleuze[14,15,34], , Jovan Antovic[6,7], Jose Manuel Soria Fernandez[32], Alice Assinger[12], Jochen M Schwenk[1], Juan Carlos Souto Andres[35], Pierre-Emmanuel Morange[4,††], Lynn Marie Butler[1,3,6,7, ††], David-Alexandre Trégouët[5,12, ††, *], Jacob Odeberg[1,2,3,8,22, ††, *]

† These authors contributed equally:
Maria Jesus Iglesias, Laura Sanchez-Rivera.

†† These authors jointly supervised this work:
Pierre-Emmanuel Morange, Lynn Marie Butler, David-Alexandre Trégouët and Jacob Odeberg

*** Address correspondence to:**
David-Alexandre Trégouët: david-alexandre.tregouet@u-bordeaux.fr
Jacob Odeberg: Jacob.odeberg@scilifelab.se

Author affiliations:

1. Science for Life Laboratory, Department of Protein Science, CBH, KTH Royal Institute of Technology, SE-171 21 Stockholm, Sweden
2. Division of Internal Medicine, University Hospital of North Norway (UNN), PB100, 9038 Tromsø, Norway
3. Translational Vascular Research, Department of Clinical Medicine, UiT The Arctic University of Norway, 9019 Tromsø, Norway
4. Aix-Marseille Univ, INSERM, INRAE, C2VN, Laboratory of Haematology, CRB Assistance Publique - Hôpitaux de Marseille, HemoVasc (CRB AP-HM HemoVasc), Marseille, France
5. University of Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, team ELEANOR, Bordeaux, France
6. Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden
7. Department of Clinical Chemistry, Karolinska University Hospital, Stockholm, Sweden
8. Department of Medicine Solna, Karolinska Institute and Karolinska University Hospital, Stockholm. Sweden
9. Theme of Emergency and Reparative Medicine, Karolinska University Hospital, Stockholm, Sweden
10. Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Framingham, MA, USA
11. The Framingham Heart Study, Boston University, Framingham, MA, USA
12. Center for Physiology and Pharmacology, Institute of Vascular Biology and Thrombosis Research, Medical University of Vienna, Austria
13. Laboratory of Excellence GENMED (Medical Genomics), Bordeaux, France
14. Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), 91057, Evry, France
15. Laboratory of Excellence GENMED (Medical Genomics), Evry, France

16. MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge, CB2 0QQ, UK
17. Computational Medicine, Berlin Institute of Health at Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany
18. Precision Healthcare University Research Institute, Queen Mary University of London, UK
19. Corporal Michael Crescenz VA Medical Center, Philadelphia, Pennsylvania, USA
20. Department of Surgery and Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
21. Division of Vascular Surgery, Stanford University School of Medicine, Palo Alto, CA, USA
22. Department of Epidemiology, University of Washington, Seattle,WA, USA
23. Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA
24. Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, WA, USA
25. Hematology Department and Biosurgical Research Lab (Carpentier Foundation), European Georges Pompidou Hospital, Assistance Publique Hôpitaux de Paris, 20 rue Leblanc, Paris, 75015, France
26. Innovative Therapies in Haemostasis, INSERM, Université de Paris, 4 avenue de l'Observatoire, Paris, 75270, France
27. Coagulation Unit, Department of Haematology, Karolinska University Hospital, SE-171 76 Stockholm, Sweden
28. Department of Clinical Science, Intervention and Technology, Karolinska Institute, 171 77, Stockholm, Sweden
29. Institute for Clinical Chemistry and Laboratory Medicine, University Medical Centre Hamburg-Eppendorf, D-20246 Hamburg, Germany
30. Center for Thrombosis and Hemostasis (CTH), Johannes Gutenberg University Medical Center, D-55131 Mainz, Germany

31. Irish Centre for Vascular Biology, School of Pharmacy and Biomolecular Sciences, Royal College of Surgeons in Ireland, Dublin 2, D02 YN77, Ireland
32. Genomics of Complex Diseases Group. Research Institute Hospital de la Santa Creu i Sant Pau. IIB Sant Pau, Barcelona, Spain
33. Department of vascular medicine, Paris Saint-Joseph Hospital Group, INSERM 1153-CRESS, University of Paris Cité, 185 rue Raymond Losserand, Paris, 75674, France
34. Centre D'Etude du Polymorphisme Humain, Fondation Jean Dausset, Paris, France
35. Unitat d'Hemostàsia i Trombosi. Hospital de la Santa Creu i Sant Pau and IIB-Sant Pau, Barcelona, Spain

ABSTRACT

Venous thromboembolism (VTE) is a common, multi-causal disease with potentially serious short- and long-term complications. In clinical practice, there is a need for improved plasma biomarker-based tools for VTE diagnosis and risk prediction. We used multiplex proteomics profiling to screen plasma from patients with suspected acute VTE, and several case-control studies for VTE, to identify Complement Factor H Related Protein (CFHR5), a regulator of the alternative pathway of complement activation, as a novel VTE associated plasma biomarker. In plasma, higher CFHR5 levels were associated with increased thrombin generation potential and recombinant CFHR5 enhanced platelet activation *in vitro*. GWAS analysis of ~52,000 participants identified six loci associated with CFHR5 plasma levels, but Mendelian randomization did not demonstrate causality between CFHR5 and VTE. Our results indicate an important role for the regulation of the alternative pathway of complement activation in VTE and that CFHR5 represents a potential diagnostic and/or risk predictive plasma biomarker.

INTRODUCTION

Venous thromboembolism (VTE), comprising both pulmonary embolism (PE) and deep vein thrombosis (DVT) is a common, multi-causal disease with serious short and long-term complications. VTE has a high mortality rate in the first year, especially within the first 30 days (~30% for PE) and a high risk of recurrence, with a cumulative incidence rate of 25% within 10 years [1-4]. VTE diagnosis is challenging, as the predisposing common risk factors and clinical presentation can be consistent with multiple other conditions, particularly in the case of PE. Current VTE diagnostic work-up includes assessment of clinical probability, using clinical decision rules, e.g., the Well's score, in combination with elevated plasma D-dimer levels [5, 6]. D-dimer, a clot breakdown product, can be elevated in several other non-VTE conditions (e.g., inflammation, surgery, cancer), and its usefulness is limited to ruling out VTE in low probability cases. In medium and high probability cases, diagnostic imaging, is necessary to exclude or confirm diagnosis. However, despite progress in imaging techniques there is an urgent need for more precise biomarker-based tools to confirm or exclude VTE in a hospital setting. Several studies have proposed VTE biomarker candidates (e.g., p-selectin, microvesicles) [7], but none have reached clinical implementation. There is also a need for improved tools and plasma biomarkers for risk prediction in VTE. Risk scores based on clinical risk factors and D-dimer levels have been developed for recurrence prediction [8-14], but none are routinely integrated into clinical practice. In addition to the genetic variants currently used in clinical assessment of hereditary thrombophilia (e.g., Factor V Leiden, prothrombin mutation) there are other more recently discovered common gene variants that contribute to VTE risk [15-17]. However, even when these are also incorporated into risk scores, they still lack sufficient precision for VTE prediction on an individual basis [18, 19]. This likely reflects the interplay between transient and sustained risk factors in disease development, including acquired risk factors, genetics, and environmental exposures [20, 21].

VTE is a disease of the intravascular compartment and thus analysis of the blood proteome has the potential to capture the resulting effects of combined genetic, epi-genetic, and environmental contributors to risk variation. So far, a handful of plasma proteomics studies of VTE have been reported, presenting novel candidates associated with increased risk of VTE [22-27].

In this work, we aimed to identify novel biomarkers associated with acute VTE with a potential link to underlying VTE pathogenesis. We identify complement factor H-related protein 5 (CFHR5), a regulator of the alternative pathway of complement activation, as a novel VTE associated plasma biomarker. Our study suggests that CFHR5 could be involved in the underlying pathogenesis of VTE, and that it is a potential clinical biomarker for thrombotic disease diagnosis and/or risk prediction.

RESULTS

Affinity plasma proteomics identifies candidates associated with acute VTE

To identify plasma biomarkers for VTE we analysed samples collected as part of our *Venous thromboEmbolism BIOmarker Study* (*VEBIOS*) [23]. *VEBIOS* comprises two study arms; a prospective cohort of patients sampled at the Emergency Room (ER), Karolinska University Hospital, Sweden (*VEBIOS ER*) and a case/control study with patients sampled at an outpatient coagulation clinic after discontinuation of anticoagulant treatment after a first VTE event (*VEBIOS Coagulation*). The discovery cohort, *VEBIOS ER* (Figure 1A), consisted of patients (n=147) admitted to the ER with the suspicion of DVT in the lower limbs and/or PE. Following admission, both citrate and EDTA whole blood samples were collected from participants. Patient samples were classified as controls (n=96), when a VTE diagnosis was excluded by diagnostic imaging, and/or Well's clinical criteria with a normal D-dimer test, or cases (n=51) when VTE was confirmed by diagnostic imaging and anticoagulant treatment was initiated. A nested case/control sample set of 48 cases and 48 matched controls were selected for plasma protein analysis (Table 1). Target candidates for measurement were selected as previously described [23], based on: (i) indications from the literature, in house data or public repositories of a probable or plausible link to arterial or venous thrombosis (e.g., prior evidence of association with thrombosis or intermediate traits, or known involvement in biological pathways of relevance), including 124 that we predicted to have endothelial enriched expression [28], and (ii) the availability of target specific antibodies in the Human Protein Atlas (HPA) (see Methods). A total of 756 HPA antibodies, targeting 408 candidate proteins, were selected for incorporation into a single-binder suspension bead array (Figure S1A and Table S1, Tab 1), which was used to analyse plasma generated from the blood samples collected into citrate anticoagulant. The signal generated by antibody HPA059937, raised against the protein target sulfatase 1 (SULF1), was most strongly associated with VTE ($p<8.34E-06$) (Figure 1B, green point), with higher relative plasma levels in cases vs. controls (Figure 1D.i). Signals generated by a further seven antibodies were also associated with

VTE ($p<0.01$) (Table S1, Tab 1). Protein signatures in plasma can be differently affected by the sample matrix; anticoagulants can inhibit specific proteases, influence soluble protein interactions, and modify analyte stability. Thus, the anticoagulant type used has potential consequences for biomarker identification [29]. We therefore replicated the *VEBIOS ER* discovery screen in EDTA samples drawn in parallel from the same patients (Figure 1C). Of the eight antibodies that produced signals associated ($p<0.01$) with VTE in the citrate samples (Table S1, Tab 1), four were replicated in the EDTA samples (Figure 1B and C): HPA059937 (predicted target SULF1, green point), HPA044659 (predicted target Leukocyte surface antigen CD47 [CD47], blue point), HPA003042 (predicted target Adenosine receptor A2a [ADORA2A], orange point) and HPA002655 (predicted target P-Selectin [SELP], red point). In both anticoagulants, all four target candidates were elevated in cases vs. controls (Figure 1D and E, i-iv, Table S1, Tab 2). In all subsequent experiments, citrated blood was used in the analysis.

Previously, in the *VEBIOS Coagulation* study (n=177) [23], we identified 29 protein candidates in plasma that were associated with prior VTE. This study was composed of patients sampled 1-6 months after discontinuation of anticoagulation treatment (duration 6-12 months) following a first time VTE, or matched controls. Of the four antibodies that generated signals associated with acute VTE in citrate *and* EDTA plasma in *VEBIOS ER* (Figure 1B and C, marked with coloured circles), only HPA059937 (predicted target SULF1), produced a signal associated with prior VTE in the *VEBIOS Coagulation* study [23]. As our aim was to identify biomarkers associated with acute VTE that were potentially linked to the underlying disease risk, we prioritised this candidate.

Complement Factor H Related protein 5 (CFRH5) is associated with VTE

The antibodies used in the single-binder suspension bead arrays passed quality control for antigen binding specificity (see www.proteinatlas.org/), but selective binding to the target protein in context of the complex matrix of plasma requires verification, as antibody specificity and reliability can be a problematic issue [30-32] (Figure S1 C-F). To verify which protein(s) were captured by HPA059937 (predicted target SULF1) we performed immunocapture-mass

spectrometry (IC-MS). Two proteins were bound to HPA059937 with a z-score>3 in triplicate experiments; the predicted target, SULF1 (z-score = 4.02, with 1 Peptide Spectrum Match [PSM]), and Complement Factor H Related protein 5 (CFHR5) (z-score=5.09, with ≥21 PSM) (Figure 1F and Table S1, Tab 3). These data indicate that CFHR5 was the predominant protein captured by HPA059937 in plasma, whilst not ruling out concurrent binding of SULF1. High levels of CFHR5 have been detected in plasma by mass spectrometry (MS) [33], but SULF1 is below the MS detection threshold [34]. To further verify CFHR5 binding specificity of HPA059937, we developed three dual binder assays (Figure S1B and Figure 1G), all with a commercial monoclonal antibody against CFHR5 (MAB3845) as a detection antibody, combined with either: original antibody HPA059937 (Figure 1G.i), or one of two independent antibodies raised in house against CFHR5; HPA073894 (Figure 1G.ii) or HPA072446 (Figure 1G.iii), as bead coupled capture antibodies. We confirmed that detection antibody (MAB3845) specifically bound CFHR5 in plasma, using IC-MS analysis (Figure S2A). Western blot analysis showed that both MAB3845 and HPA073894 bound mono and homodimer of recombinant CFHR5, and that HPA073894 detected a band corresponding to CFHR5 size in plasma (Figure S2B).

When used to re-analyse the *VEBIOS ER* samples; all three assays consistently detected a higher level of target protein in cases vs. controls (Figure 1G.iv-vi, p=6E-04, 2.1E-03, 1E-04, respectively). MFI values from all three strongly correlated with those generated by HPA059937 in the *VEBIOS ER* discovery screen (Figure 1G.vii-ix) (Spearman´s ρ=0.83, 0.75 and 0.82, respectively [all p<1E-04]). We made five dual binder assays that targeted SULF1 using HPA059937 as capture antibody, together with different anti-SULF1 detection antibodies, but none gave a quantitative signal in a plasma dilution series, or buffer containing a dilution series of recombinant SULF1 protein. Western blot analysis showed that the anti-SULF1 HPA059937 detected the monomeric form of recombinant CFHR5 (rCFHR5) under non-reducing, but not reducing, conditions (Figure S2B), suggesting an off-target binding to an epitope created by a tertiary folded structure of CFHR5. Together, these data are consistent with CFHR5, as opposed

to SULF1, being the target protein of HPA059937 associated with VTE in the *VEBIOS ER* discovery screen.

CFHR5 is associated with VTE independent of D-dimer or CRP

We used data independent acquisition mass spectrometry (DIA-MS) to perform orthogonal validation of the results obtained from the analysis of CFHR5 plasma levels in *VEBIOS ER* using the dual binder assay with capture antibody HPA072446 (Figure 1G.iii). Data from these two independent assays correlated well (ρ=0.75, p<2.2E-16), and so the dual binder assay was used for quantification of CFHR5 in *VEBIOS ER* and an extended sample set of the *VEBIOS Coagulation* study (n=284) (Table S2, Tab 1 for cohort descriptive data).

In the *VEBIOS ER* and *Coagulation* study, mean CFHR5 concentrations in control plasma samples were 2840 ± 756 and 2470 ± 523ng/ml, respectively; levels in the range previously estimated by mass spectrometry (~1900ng/ml) [33]. In the *VEBIOS ER* discovery study, the mean absolute CFHR5 concentration was confirmed as higher in patients with confirmed VTE, compared to patients where VTE was ruled out (3430 ng/ml ± 782 [cases] vs. 2840 ng/ml ± 756 [controls], p=1.05E-03 [age and sex adjusted]); the odds ratio (OR) for diagnosis of acute VTE associated with one standard deviation (SD) increase of CFHR5 concentration was 2.54 [confidence interval (CI) 1.52-4.66], p=1.05E-03 (Figure 1H.i and Table 2). Consistent with the relative quantification results in our previous study [23], absolute CFHR5 concentration was associated with prior diagnosis of VTE in the extended *VEBIOS Coagulation* cohort, compared to controls (mean concentration 2680 ng/ml ± 556 [cases] vs. 2470 ng/ml ± 523 [controls], p= 8.42E-04 [age and sex adjusted]); the OR for first time VTE was 1.55 [CI 1.20-2.01], p=8.85E-04) (Figure 1J.i and Table 2).

We next investigated if CFHR5 levels were associated with VTE associated risk factors, such as age, body mass index [BMI] and routine clinical laboratory tests for blood markers associated with thrombosis risk (e.g., D-dimer, c-reactive protein [CRP], thrombocyte count) (Table S1, Tab 4). In

VEBIOS ER, CRP levels correlated with plasma CFHR5 concentration, in cases (ρ=0.57, p=4.58E-05) and controls (ρ=0.52, p=2.39E-04) (Figure 1H.ii and Table S1, Tab 4, Table A), but there was no strong correlation between CFHR5 and the other parameters measured, in cases or controls, including D-dimer (Figure 1H.iii and Table S1, Tab 4, Table A). Adjusting for CRP, CFHR5 remain significantly associated with acute VTE (OR=3.31 [CI 1.60-7.73] p=3.15E-03) (Table S1, Tab 10). In the *VEBIOS Coagulation* study, CFHR5 levels in cases correlated with both CRP (ρ=0.49, p=4.48E-10) and D-dimer (ρ=0.42, p=2.44E-07) (Figure 1J.ii and iii, right panel), but in controls these correlations were weak (CRP ρ=0.24, p=4.95E-03) or absent, respectively (Figure 1J.ii, and iii, left panel and Table S1, Tab 4, Table B). The association between CFHR5 and VTE remained significant in *VEBIOS Coagulation* when adjusted for CRP (OR=1.55 [CI 1.18-2.03] p=1.50E-03) or D-Dimer (OR=1.435 [CI 1.05-1.88] p=7.72E-03) (Table S1, Tab 10 Table A).

CFHR5 measurement can increase diagnostic accuracy in patients with likely VTE

To explore the potential usefulness of CFHR5 as a biomarker to be included in the diagnostic workup of suspected acute VTE, we assessed the discriminatory power of CFHR5 in *VEBIOS ER* using logistic regression in different models together, with D-dimer dichotomised using current Clinical Decision Rules (CDR) as 'positive' or 'negative' (below age adjusted cut-off [35]) and Wells score (VTE likely (≥2 for DVT and ≥4 for PE) or unlikely) (Table S1, Tab 5). In *VEBIOS ER*, D-dimer had negative predictive value (NPV) of 100% (0 false negatives) for VTE, while the specificity and positive predictive value (PPV) was only 62.8% and 74% respectively, with 16 false positive cases. Adding CFHR5 to the base model of D-dimer alone resulted in a non-significant improvement in AUC (0.88 versus 0.82, p=0.110), as did adding Wells score to the base model (AUC 0.85, p=0.33) (Table S1, Tab 5). D-dimer alone performed better than CFHR5 alone (AUC 0.73 versus 0.82, p=0.128). When stratifying patients based on Wells score, in the group where VTE was considered unlikely based on Wells score (n=43), adding CFHR5 to the base model resulted in a non-significant increased accuracy compared to the base model (AUC 0.84 versus

0.81, p=0.61). However, in the group where VTE was considered likely (n=41), the addition of CFHR5 to the base model resulted in a significantly increased accuracy compared to D-dimer alone (AUC 0.92 vs 0.83; p=0.035) (Table S1, Tab 5).

CFHR5 is specifically expressed in liver hepatocytes with other complement related genes

To further understand the expression characteristics of *CFHR5*, and to identify possible co-expressed or co-regulated proteins we used a whole transcriptome analysis approach. In a consensus dataset, consisting of normalized mRNA transcript (nTPM) levels across 55 different human tissue types, generated from Human protein Atlas (HPA) (https://www.proteinatlas.org) [36] and Genotype-Tissue Expression Project (GTEx) (www.gtexportal.org) [37] datasets, *CFHR5* was highly and specifically expressed in the liver (Figure 2A). Single cell analysis of liver tissue [38], showed that *CFHR5* was specifically expressed in the hepatocyte cellular compartment (Figure 2B). To identify transcripts potentially co-expressed or co-regulated with *CFHR5* in the liver, we analysed bulk RNAseq data (n=226) retrieved from GTEx portal V8. We generated Pearson correlation coefficients between *CFHR5* and all other expressed protein coding transcripts (19,525) (Table S1, Tab 6, Table A) and used gene ontology (GO) and reactome analysis [39, 40] to identify over-represented classes and pathways in the top 50 most highly correlated genes (Table S1, Tab 6, Table B and C). Results were consistent with known CFHR5 function; significant GO terms included *´complement activation´* (FDR adjusted p-value [PFDR]=2.4E-16) and *´humoral immune response´* (PFDR=1.2E-12) and reactome pathways included *´regulation of complement cascade´* (PFDR=3.8E-24). We performed an unbiased weighted network correlation analysis (WGCNA) [41] on the same dataset, where correlation coefficients between all transcripts, excluding those classified as non-coding, were calculated and subsequently clustered into related groups, based on expression similarity (Table S1, Tab 6, Table A [column F]). Transcripts that fulfilled the criteria of: (i) Pearson correlation with *CFHR5* >0.65 ($p < 0.001$) and (ii) annotation to the *CFHR5*-containing gene cluster in WGNCA (group 68), were identified as those most likely to be co-expressed or co-regulated. Of these, 13/18 [72%]

were other members of the complement cascade and 15/18 [83%] were also specifically expressed in liver [36] (Table S1, Tab 6, Table A, column F) (Table S1, Tab 6 Table A, column B, bold text). While these data indicate a degree of co-expression with *CFHR5* at the transcriptional level, plasma concentrations of the encoded proteins are subject to several post-transcriptional variables, such as translation efficiency, cellular release dynamics, protein stability and clearance. When these proteins were interrogated using the protein-protein interaction database STRING, v11 [42], 13/18 had high confidence functional and physical associations (Figure 2C) (Table S1, Tab 7, Table A). CFHR5 was most strongly linked to complement 3 (C3), the central hub of the largest of the three linked cluster groups identified (Figure 2C, clusters represented by green, red and cyan) (Table S1, Tab 7, Table B).

CFHR5 is associated with acute VTE independent of C3

Plasma levels of complement component C3 have previously been reported as associated with incident VTE [43]. To determine if the association between CFHR5 and VTE we observed is dependent on the concentration of C3, we developed an in-house dual binder quantitative assay to measure C3 in the *VEBIOS ER* and *VEBIOS Coagulation* cohorts. In *VEBIOS ER*, plasma C3 was not elevated in cases, compared to controls (Figure 2D.i and S1 Tab 2, panel B), CFHR5 and C3 did not significantly correlate in either group (Figure 2D.ii) and C3 was not associated with VTE (OR 1.04 [CI 0.68-1.58], p=0.86) (Table S1, Tab 8). Furthermore, the association with acute VTE for one SD increase in CFHR5 level remained unchanged (OR 2.65 [CI 1.53-5.01], p=1.26E-03) when including and adjusting for C3 concentration (together with age and sex), compared to when only adjusting for age and sex (OR 2.54 [1.52-4.66], p=0.001), Table 2, Table S1, Tab 8), demonstrating that CFHR5 is independently associated with acute VTE. In *VEBIOS Coagulation,* C3 levels were higher in plasma from cases, compared to controls (Figure 2E.i), and CFHR5 and C3 correlated with each other in both ([controls ρ = 0.46 p=<0.0001], [cases ρ = 0.47 p=<0.0001]). After adjusting for age and sex, one SD increase in C3 concentration was significantly associated with previous VTE (OR 1.52 [CI 1.18-2.01], p=1.93E-03). When adjusting for CFHR5 levels

(together with age and sex), this no longer reached significance (OR 1.31 [CI 0.99-1.78], p=0.064). The association with previous VTE for one SD increase in CFHR5 level in *VEBIOS Coagulation* was still nominally significant when adjusting for C3 levels (OR 1.36 [CI 1.03-1.82], p=0.032), although weaker compared to adjusting only for age and sex (OR 1.55 [1.2-2.01], p=8.85E-04, Table 2, Table S1, Tab 8).

The CFHR5 association with VTE replicates in additional cohorts

The identification of biomarkers associated with VTE diagnosis, or risk profiling, requires replication in independent cohorts, from different settings with different demographic profiles, to determine feasibility for potential translation to clinical practice. We sourced three independent replication cohorts to test the association of CFHR5 with VTE; the Swedish Karolinska Age Adjusted D-Dimer study (*DFW-VTE*) VTE study (n=200) consisting of patients with suspected VTE (cases; n=54, controls; n=146) [44] (Figure 3B), the French *FARIVE* study (n=1158) consisting of patients sampled during the week following a diagnosis of acute VTE (n=582), with hospital-based controls (n=576) [45] (Figure 3C) and the Spanish *Riesgo de Enfermedad TROmboembólica VEnosa* (*RETROVE*) study (n=668) of patients sampled post anticoagulant treatment (n=308), with population based controls (n=360) [46] (Figure 3E) (for all cohort details see Table S2, Tabs 2-4 and 6). The OR of VTE associated with CFHR5 per 1 SD increase in CFHR5 concentration was significant in all 3 replication cohorts: *DFW-VTE* (OR 1.80 [CI 1.29-2.58], p=7.65E-04) (Figure 3B), *FARIVE* (OR 1.24 [CI 1.10-1.40], p=3.98E-04) (Figure 3C), *RETROVE* (OR 1.29 [CI 1.09-1.53], p=2.4E-03) (Figure 3E) (Table 2). When samples from cases and controls were stratified according to CFHR5 concentration, the association with VTE was most pronounced in the third tertile, in all 5 cohorts analysed individually and in a meta-analysis (Table 2). These associations remain significant in subgroup meta-analyses when stratified by thrombosis type (DVT or PE), sex, or cause (provoked/unprovoked) (Table S1, Tab 9, Table A-C). In subgroup analyses in the individual cohorts, the association of CFHR5 with VTE did not reach significance in females in *VEBIOS Coagulation* and *FARIVE* and in males in *DFW-VTE*.

Furthermore, the association with provoked VTE in *RETROVE* and with unprovoked VTE in *FARIVE* were not significant. The results were consistent when further adjusting for BMI and/or CRP when this information was available (Table S1 Tab 10 Table A-D).

CFHR5 and risk of recurrent VTE

We measured plasma CFHR5 concentration in a sample of 669 VTE patients from the *MARseille THrombosis Association* Study (*MARTHA*) study that have been followed for VTE recurrence, among which 124 experienced a recurrent event (Table S2, Tab 5) [47]. After adjusting for sex, familial history of VTE, provoked or unprovoked status of the first VTE, age at first VTE, and BMI, the Hazard Ratio (HR) associated of 1 SD increase in CFHR5 levels was HR=1.13 [0.96-1.32], p=0.134. The association was consistent between females (HR=1.1 [0.90 -1.38]; p=0.320) and males (HR=1.14 [0.91-1.44]; p=0.260) and between patients with DVT (HR =1.18 [0.98-1.42]; p=0.080) or PE as first event (HR=1.13 [0.80-1.61]; p=0.489). This trend for association was strongest in the subgroup of patients with unprovoked first VTE (HR=1.32 [0.99–1.77], p=0.056), as no association was observed when the first event was provoked (HR=1.01 [0.83–1.23], p=0.900). However, the test for heterogeneity between these two HRs did not reach 0.05 significance (p=0.23).

Genome Wide Association Study on CFHR5 plasma levels

To explore if CFHR5 concentration in plasma was influenced by genetic variants, we first performed a meta-analysis of GWAS for dual binder assay based CFHR5 concentrations in individuals from the *FARIVE* (n=1,033), *RETROVE* (n=668) and *MARTHA* (n=1,266) studies. The results from the association results are summarised in Figure S3A. Of N= 7,135,343 SNPs tested in a total sample of 2967 individuals, one genome-wide significant (p<5E-08) signal was observed on chr1q31.3. The lead SNP at this locus was rs10737681, mapping to *CFHR1/CFHR4* (Figure S3B), and the G allele was associated with a one SD increase in CFHR5 levels of β= +0.25 ± 0.03 (p=6.49E-21). In the latest GWAS for VTE risk built on ~72K cases and >1M controls of European ancestry [17], the rs10737681–G allele was associated with a marginal (p=0.016)

371 decreased risk of VTE (Table S1, Tab 11). This pattern of association is not consistent with the
372 relationship between increased CFHR5 plasma levels and increased VTE risk observed in the
373 present studies, since the rs10737681-G CFHR5 increasing allele would have been expected to
374 be associated with increased VTE risk (Table S1, Tab 11). A second round of meta-analysis,
375 integrating GWAS summary statistics from 3 additional proteogenomic resources where CFHR5
376 was measured with different assays [see methods], totalling ~50,200 individuals, confirmed the
377 association of this locus with CFHR5 levels. Interestingly, the rs10737681 was also identified as
378 the lead SNP at this locus in the extended meta-analysis (β =0.28 ± 0.01, p=2.94E-396) (Table
379 S1, Tab 12). Strong linkage disequilibrium holds at the locus covering more than 10Mb and
380 extending from *CFHR1* to *CFHR5* (Figure 4B). The extended meta-analysis identified 5 additional
381 independent loci associated with CFHR5 levels: *HNF1A* (rs2393776, p=1.48E-21) on 12q24.31,
382 *JMJD1C* (rs7916868, p=4.61E-12) on 10q21.3, *TRIB1* (rs28601761, p=4.39E-09) on 8q24.13,
383 *DNAH10* (rs7133378, p=2.43E-08) also on 12q24.31 and *HNF4A* (rs1800961, p=4.97E-08) on
384 20q13.12 (Figure 4A and Table S1 Tab 12). All of the lead SNPs at these loci, except *HNF1A*
385 rs2393776 (p=0.17), demonstrated marginal (p<0.05) association with VTE risk (Table S1, Tab
386 11). However, only two, *JMJD1C*_ rs7916868 and *DNAH10*_rs7133378, showed patterns of
387 association with VTE that are compatible with the association of increased CFHR5 levels with
388 VTE risk. This explains why MR analyses are not supportive for a causal association between
389 increased CFHR5 levels and VTE (Table S1, Tab 13).
390 Of note, 1230 *MARTHA* participants with CFHR5 plasma levels have also been typed with an
391 Illumina exome12v1.2 DNA array [48] dedicated to the genotyping of coding polymorphisms,
392 mainly of low frequency. Capitalizing on this additional genetic resource, we investigated whether
393 low-frequency coding variants at the *CFHR5* locus (including the nearby
394 *CFHR1/CFHR2/CFHR3/CFHR4* genes) could contribute to the inter-individual variability of
395 CFHR5 plasma levels. Twelve rare variants were found polymorphic at this locus in *MARTHA*
396 participants (Table S1, Tab 14). Three of these variants showed evidence for association with

CFHR5 plasma (in bold). These were three rare non-synonymous *CFHR5* variants: rs139017763 (G278S) p=4.75E-05, rs41299613 (C208R) p=1.6E-03 and rs35662416 (R356H) p=7.1E-03, where rare minor alleles were associated with decreased CFHR5 plasma levels. It is important to emphasize that these 3 *CFHR5* non-synonymous variants were carried by 16 distinct individuals. Figure 4C illustrates the difference in CFHR5 plasma levels between the 16 carriers of rare CFHR5-associated variants and non-carriers. This difference remained significant (p=2.45E-07) after adjusting for the common rs10737681 variant identified in the GWAS analysis.

CFHR5 is associated with thrombin generation potential in patients with previous VTE

As thrombin generation has been associated with increased risk of VTE [49], we tested the association between CFHR5 plasma concentration and thrombin generation as measured by thrombinoscope in *MARTHA* (n=774 VTE cases, see Table S2, Tab 5 for details) with replication in *RETROVE* (308 cases/360 controls, see Table S2 Tab 4 for details). In both *MARTHA* and *RETROVE cases*, we find significant association between CFHR5 and lag time (ρ=0.181, p= p<0.0001 and ρ=0.176, p<0.0001, respectively), Endogenous Thrombin Potential (ETP) (ρ=0.105, p= 0.0036, and ρ=0.130, p<0.0001, respectively), peak (ρ=0.117, p= 0.0012, and ρ=0.132, p<0.0001), and ttPeak (ρ=0.116, p=0.0012, and ρ=0.086, p=0.0274 (see Table S1, Tab 15).

CFHR5 enhances platelet activation and degranulation in plasma

C3a, generated by cleavage of C3, can increase platelet activation [50-52]. As CFHR5 has a regulatory role upstream of C3/C3a activation, we investigated the effect of recombinant CFHR5 (rCFHR5) on platelet activation *in vitro.* Functionality of the rCFHR5 was validated by its capacity to form a homodimer complex and its ability to bind known interaction partners C-reactive protein (CRP) [53] and properdin [54] (see methods and Figure S2B and S4). Platelet rich plasma was pre-incubated with 6µg/ml recombinant CFHR5, a concentration corresponding to the upper range of that detected in the plasma of the VTE case group in *VEBIOS ER*, and in agreement with what has been reported in previous literature [53]. Platelet activation was measured by

surface expression of P-selectin, activated GP IIb/IIIa or CD63 (Figure 5A, B and C, respectively) in response to adenosine diphosphate (ADP), convulxin or TRAP6 (Figure 5A-C.i, ii and iii, respectively). Following stimulation with ADP, a higher percentage of platelets pre-incubated with CFHR5 expressed P-selectin (Figure 5A.i) (p=0.0056), activated GP IIb/IIIa (Figure 5B.i) (p=0.031) and CD63 (Figure 5C.i) (p=0.009), compared to the control. Pre-incubation with CHFR5 also potentiated platelet activation in response to convulxin or TRAP6 stimulation (Figure 5A-C.ii and iii), although the effect appeared to be more strongly linked to stimulus concentration, than that observed for ADP. Although CFHR5 potentiated the expression of platelet activation markers in response to ADP, it did not modify ADP-induced platelet aggregation (Figure 5D.i-iii). Washed platelet response to ADP, convulxin or TRAP6 was not modified by preincubation with CFHR5 (Figure S5) (ANOVA all p>0.05), indicating that additional components in plasma were required for the observed response, and that they are not a direct effect of CFHR5 on platelets.

A proposed function of CHFR5 is that it augments complement activation by antagonising complement factor H (CFH), the main negative regulator of alternative pathway (AP) activation in plasma. CFH inhibits C3 convertase, preventing formation of C3a [55], which has been suggested to have role in platelet activation and subsequent thrombosis formation [50]. To determine if CFHR5-induced augmentation of platelet activation was dependent on C3 cleavage and generation of C3a, we pre-treated platelets with an inhibitor of C3 cleavage and activation of C3a (compstatin), or an anti-C3a antibody, prior to CHFR5 pre-incubation and subsequent ADP stimulation. The potentiation effect of CHFR5 on baseline and ADP-induced activated GP IIb/IIIa expression (Figure 5E.i and 5F.i) was abolished following compstatin (Figure 5E.ii) or anti-C3a antibody (Figure 5F.ii) pre-treatment; data consistent with a complement dependent effect of CFHR5 on platelet activation.

Complement fragment 3c concentration correlates with CFHR5 in *VEBIOS ER* subset

As a marker for C3 cleavage and activation in plasma, we measured complement fragment 3c (C3c) in a subset of plasma samples from *VEBIOS ER*, selected based on a low (<2500 ng/ml,

10 samples) or high (>3800 ng/ml, 10 samples) plasma CFHR5 concentration. Mean C3c concentration was greatest in the high CFHR5 group (C3c (ng/ml) ± std dev: CFHR5 low: 0.91±0.2 vs CFHR5 high: 1.08±0.2), although this difference failed to reach statistical significance ($p<0.086$) (Figure S6A). However, C3c and CFHR5 concentrations were positively correlated across this sample set ($\rho=0.51$, $p<0.02$) (Figure S6B).

DISCUSSION

Here, we aimed to identify biomarkers associated with acute VTE that are linked to disease pathogenesis and risk. Using a nested case-control study, derived from a cohort of patients presenting to the ER with suspected acute VTE, and from a case-control study with patients that had suffered a previous first VTE, we identify CFHR5, a regulator of the alternative complement activation pathway, as such a biomarker. The association of CFHR5 with current or previous VTE was replicated in three additional cohorts or case-control studies, and we also found a trend for association with risk for recurrence of unprovoked VTE. We identify 6 independent loci with CFHR5 levels including the *CFHR1-5* gene cluster loci. We further provide evidence of a direct role of CFHR5 in the induction of a pro-thrombotic phenotype, through its effect on platelet activation. Our findings indicate CFHR5 has potential application as a clinical biomarker for VTE diagnosis and risk prediction, providing further support to the idea that complement regulation is a key element of VTE pathogenesis.

Currently, D-dimer is the only plasma biomarker used in VTE diagnostic work-up, but its clinical utility is limited to ruling out VTE in low-risk patients. Several studies have attempted to identify novel biomarkers with potential clinically usefulness for the confirmation of VTE diagnosis, and although a number have been identified [7], none have yet been implemented in clinical practice. For many, like D-dimer, elevated levels are a consequence of thrombosis formation, e.g. biomarkers of fibrinolysis, clot re-modelling or resolution (e.g. MMPs), inflammation secondary to local vascular and tissue injury (e.g., CRP, IL-6, IL-10, fibrinogen), or of endothelial and/or platelet activation (e.g. vWF, P-selectin) [7, 56, 57]. We found no correlation or association between D-dimer and CFHR5 in the acute VTE setting, supportive of that increased CFHR5 concentration at diagnosis is not secondary to thrombus formation. In contrast, we found a strong correlation between D-dimer and CFHR5 levels in patients followed up after ending treatment for a first VTE, but not in controls. D-dimer has been associated with increased risk of first and recurrent VTE [8-

14, 58] and thus, our results are consistent with a link between CFHR5 and subclinical coagulability in these patients at follow-up, possibly due to persistent risk factors.

The CFHR5 locus maps to chromosome 1q31.3 at one end, a gene cluster that spans approximately 350 kb including (in order from CFHR5) the CFHR2, CFHR4, CFHR1, CFHR3, and CFH loci. The rs10737681 we identified with genome wide association with plasma CFHR5 level maps between the CFHR4 and CFHR1 genes. Of note, the CFHR2 locus has just been identified as a novel susceptibility locus for VTE in the recently published international effort on VTE genetics [17]. The lead SNP at this locus is rs143410348, which is in moderate LD with the rs10737681 (r^2~0.40 in European population [59]), which here we found associated with CFHR5 plasma levels. In a combined meta-analysis of 37,770 individuals from 3 cohorts (EPIC/FARIVE/Omicscience) where it was imputed, rs143410348 was less strongly associated with CFHR5 levels than the lead rs10737681 (β=-0.15, p=1.9E-32 vs β =-0.27, p=2.1E-95). Altogether, the observations from the GWAS analyses emphasize the need for a deeper investigation of the genetic architecture of the CFHR1/CFHR4/CFHR2/CFHR5 locus with respect to CFHR5 levels and VTE risk. Five additional candidate loci were identified as participating to the genetic regulation of CFHR5 plasma levels: *DNAH10, HNF1A, HNF4A, JMJD1C* and *TRIB1*. All 5 loci have been reported to associate with various lipids traits (see GWAS catalogue [60]) and 3 (*HNF1A, HNF4A and TRIB1*) have been reported to also associate with liver enzymes. *DNAH10* is also known to be a locus involved in white & red blood cell biology [61] while the *JMJD1C* is a locus linked to platelet biology [62, 63]. Most of these traits are well known risk factors for VTE, and this may suggest that the association of CFHR5 levels with VTE risk implies many additional biological players with pleiotropic effects. This may explain why the MR analyses did not provide causal evidence for a link between CFHR5 levels and VTE risk.

The complement and haemostatic systems interact at several points during initiation, propagation, and regulation of complement activation and coagulation [64]. Studies have indicated a role of complement in VTE pathogenesis [43, 51], but underlying mechanisms are not well understood.

CFHR5 shares sequence and structural homology with Complement Factor H (CFH) [65], the main negative regulator of alternative pathway (AP) activation in plasma [55]. Under normal conditions, the AP is constitutively active through spontaneous hydrolysis of the thioester bond in C3 and the formation of the initial fluid phase C3 convertase, C3(H_2O)Bb, which cleaves C3 into C3a and C3b [55]. CFH promotes decay of the alternative and classical pathway convertases and is a cofactor in the cleavage of C3b, hereby regulating excess activation of C3 [66]. CFH inhibits C3 convertase, preventing formation of C3a. CFHR5 antagonizes CFH function, through competitive binding to C3b and its fragment C3d [67], thus deregulating AP activation. CFHR5 also promotes complement activation by interfering with CFH binding to CRP, pentraxin 3 (PTX3), and extracellular matrix (ECM) [68].

Elevated plasma C3 in baseline samples has been shown to be associated with increased risk of future VTE [43]. Consistent with these findings, C3 was associated with prior VTE in the *VEBIOS coagulation* study, but not with acute VTE in the *VEBIOS ER* study. In both cases, and in the previous study by Nordgaard *et. al.* [43], total C3 level, rather than the active form (C3a) is measured; it is possible that in acute VTE, regulation of C3 convertase (by CFHR5) is important, rather than absolute C3. Consistent with this, we observe a trend for higher plasma levels of complement C3c fragment, a marker of C3 activation, in samples with higher CFHR5 concentrations at VTE diagnosis. It could be speculated that the association of C3 with VTE in individuals sampled pre-VTE [43] or following treatment for a prior VTE, reflects co-regulation of CFHR5 and C3 under basal conditions, which would be consistent with our finding that in *VEBIOS coagulation*, the association with VTE for CFHR5 was weaker when adjusting for levels of C3, and *vice versa*.

The mechanisms underlying venous and arterial thrombosis development differ; venous thrombi contain an abundance of red blood cells trapped in a fibrin clot together with platelets, a structure quite distinct from the vast platelet aggregates found in arterial thrombi [69]. Thus, arterial thrombosis is treated with therapies that target platelet activation and/or aggregation while VTE

is traditionally treated with drugs targeting the coagulation system. Historically, platelet function has attracted attention primarily in arterial thrombosis, however more recently the role of platelets in VTE has been recognised [70]. Elevated levels of markers of platelet activation, such as P-selectin, are associated with acute VTE [7]; a protein we also identified as one of four candidates associated with VTE in the discovery screen of *VEBIOS ER*. Furthermore, anti-platelet therapy with acetylic salicylic acid had a protective effect against VTE [71], and reduced the size of venous thrombus linked to inhibition of platelet activation in mice [72]. Our results indicates that CFHR5 has a possible role in platelet activation, which could provide a mechanistic link to the observed association between CFHR5 plasma levels and VTE. Subramaniam *et. al.* showed that C3 and C5 affected platelet activation and tissue factor procoagulant activity by different mechanisms, independent of formation of the terminal complement C5b-C9 complex [52]. C3, but not C5, deficient mice had reduced platelet activation *ex vivo*, reduced platelet deposition *in vivo*, and reduced thrombosis incidence (<30 vs. 80% in wild type). These data indicate that in VTE C3 has an important role in initial haemostasis, independent of downstream complement proteins [51]. C3a, acting through platelet receptor C3aR, is suggested to have role in the activation of the glycoprotein IIb/IIIa fibrinogen receptor via intraplatelet signalling, and subsequent thrombosis formation [50]. The presence of a C3a receptor on human platelets has been controversial, with several contradictory studies [73, 74]. However, recent studies have confirmed the presence of C3aR on human platelets using several independent techniques [50]. Similar to the study of Subraminam *et. al.* [52], Sauter *et. al*. showed C3 deficient mice had prolonged bleeding time, that could be reversed by intravenous administration of C3a peptide [50]. The C3a-C3aR induced intracellular signaling was mediated through the Rap1b activation, where co-stimulation of platelets with C3a-ADP resulted in increased Rap1b activity on top of the platelet stimulation by only ADP. In our study, we observe a similar co-stimulatory effect of CFHR5 on ADP- (and convulxin- or TRAP6-) induced platelet activation. This effect was observed on platelets in plasma, but not on those that were pre-washed, consistent with the effect of CFHR5 on platelet

559 activation being due to its interaction with other complement factors (i.e.,C3) in plasma.
560 Furthermore, in the presence of compstatin, an inhibitor of C3 cleavage and formation of C3a, or
561 anti-C3a antibody, the co-stimulatory effect of CFHR5 was not observed. On basis of these
562 recently published mechanistic findings, our *in vitro* results indicate that CFHR5 regulation of the
563 alternative pathway of complement activation has a role in C3a mediated platelet activation in
564 thrombosis, providing a potential functional link to its association with acute VTE. The co-
565 stimulatory effect of CFHR5 on platelet activation did not translate into an effect on ADP-induced
566 platelet aggregation. Activated platelets express and secrete proinflammatory and procoagulant
567 factors that could directly drive VTE, independent of platelet aggregation [75]. In the *RETROVE*
568 study (where we found CFHR5 is associated with VTE and increased thrombin generation)
569 previous studies found no association between VTE and platelet aggregation in response to ADP
570 or epinephrine [46]. Thus, one could speculate that CFHR5 has a role in VTE-linked platelet
571 activation that is independent of platelet aggregation.
572 Our study has various strengths and limitations; *VEBIOS ER*, the discovery cohort, was derived
573 from a single centre, where blood sampling for plasma biobanking was performed in parallel to
574 that for routine tests after initial evaluation (before diagnostic imaging or anticoagulant treatment),
575 thus avoiding bias in inclusion or biobanking. Samples were handled according to standard clinical
576 chemistry lab routine, thus variations in needle-to-spin-to-freeze time were equivalent between
577 case and control samples. As biobanking was based on the routine sample flow, this increases
578 the feasibility that identified biomarker candidates are suitable for clinical translation into a routine
579 setting. Importantly, we demonstrate an association of CFHR5 with VTE in several independent
580 studies, that include patients in the acute setting, at follow up, and prior to recurrence. One
581 limitation of our study is that we have not analysed a cohort of individuals that were sampled prior
582 to VTE event. Our proteomics and GWAS analyses were mainly conducted in European ancestry
583 populations and should be further investigated in populations of other ancestry origin.

584 Our screening panel included many novel candidate proteins (e.g., selected based on GWAS
585 and/or transcriptomics studies) that are poorly understood or uncharacterized, and thus not
586 included in larger commercial panels, such as those available on Olink and Somascan screening
587 platforms. However, using a relatively small custom panel for screening meant we likely failed to
588 comprehensively identify all plasma proteins with currently unknown links to VTE. As our aim was
589 to identify biomarkers associated with acute VTE that were potentially linked to the underlying
590 disease pathogenesis and risk, we prioritised the antibody target HPA059937 (raised against
591 SULF1) for further work on the basis that higher plasma concentrations observed in individuals
592 with a documented increased risk of VTE (e.g., previous VTE in *VEBIOS Coagulation*) indicated
593 that it could also represent a constitutive and/or persistent risk factor. It is possible that any of the
594 3 other candidates that were associated with diagnosis of acute VTE (Figure 1D and E, i-iv) could
595 be more informative value in a clinical diagnostic tool for acute VTE than CFHR5. Further studies
596 are needed to investigate this. Some established procoagulant VTE associated proteins included
597 in the screening panel did not pass the significance threshold as VTE associated in *VEBIOS ER*
598 (Table S1, Tab 1). The control group in this cohort were patients seeking acute medical care with
599 symptoms that initially prompted a diagnostic workup for VTE, and both cases and controls had
600 elevated CRP levels, with no significant difference between them (Table 1), indicating an
601 inflammatory status in both. As F8 and vWF are acute phase reactants, the plasma levels of which
602 increase during inflammation, the lack of association of these proteins with VTE was likely due to
603 elevated levels in both cases and controls. Indeed, using the same assay, we previously reported
604 that both F8 and vWF had a strong association with VTE in *VEBIOS Coagulation* [23], a study
605 where healthy population-based controls were used.
606 From a technological perspective, our study demonstrates the need for orthogonal verification of
607 any potential biomarker identified using antibody-based proteomics screening [30, 31]. The same
608 caution should be extended to findings generated using other high throughput affinity proteomics
609 technologies vulnerable to non-specific protein binding, such as aptamer-based [76], where

missense single nucleotide polymorphisms can affect binding in a manner where a genetic difference drive associations, rather than protein levels (Figure S1 F) [77, 78]. Studies comparing different affinity proteomics technologies have found correlations of proteins assayed with two or more platforms to range from highly concordant (Spearman´s $\rho = 0.95$) to inversely correlated ($\rho = -0.48$) [77], highlighting further the need for orthogonal validation of any potential biomarker identified.

The next steps towards the translation of our findings into a clinical setting is to develop and establish standardised methods for quantification, to establish reference intervals and define cut off values with respect to specificity and sensitivity. Current clinical decision rule (CDR) in diagnostic workup of suspected acute VTE is based on age adjusted D-dimer and Wells score. In *VEBIOS ER* we found that adding CFHR5 to D-dimer increased diagnostic accuracy of acute VTE in the VTE-likely group (Wells score ≥2 for DVT and ≥4 in PE). This group represents the major diagnostic challenge, as an elevated D-dimer is common in several of the conditions associated with increased risk for VTE, e.g., cancer and surgery, both of which are included in Wells score. Therefore, according to current CDR, patients with high clinical probability based on Wells score proceed to diagnostic imaging without prior D-dimer testing [79, 80]. Thus, adding CFHR5 concentration to D-dimer in the diagnostic work-up could potentially reduce number of negative imaging procedures, to the benefit of patients and health care system. It remains to be established if the incorporation of CFHR5 measurements into clinical decision rules or other scores can improve predictive power. The inclusion of CFHR5 measurements as a diagnostic and/or risk predictive marker in randomized clinical trials of acute VTE and VTE recurrence would be particularly informative, as these are two areas of high clinical relevance in need of improved tools for clinical decision making. Furthermore, while our study indicates that CFHR5 has a functional role in VTE development, further studies are needed to understand the mechanism.

MATERIALS AND METHODS

PATIENTS AND SAMPLES

Discovery study

Venous thromboembolism biomarker study (VEBIOS)

VEBIOS is part of a collaboration between Karolinska University Hospital, Karolinska Institute and Royal Institute of Technology (KTH) designed to identify new plasma biomarkers for VTE [23]. *VEBIOS* comprises two different studies: (**i**) *VEBIOS ER study* is a prospective cohort study carried out at the Emergency Room (ER) at the Karolinska University Hospital in Solna, Sweden, between December 2010 and September 2013. All patients admitted with the suspicion of deep vein thrombosis (DVT) in the lower limbs and/or pulmonary embolism (PE), over 18 years old were eligible for the study. Exclusion criteria were patients with on-going anticoagulant treatment, pregnancy, active cancer, short life expectancy or lack of capacity to leave approved consent. A case was defined if a) VTE was confirmed by diagnostic imaging - compression venous ultrasonography (CUS) in patients with suspected DVT in the lower limbs, and computed tomography pulmonary angiography (CTPA) in patients with suspected PE, and b) anticoagulant treatment was initiated based on the VTE diagnosis. Patients with no evidence of an acute VTE, (neither by diagnostic imaging nor by Well's clinical criteria) that had a normal D-dimer test [5], were referred as controls in the study. All participants were sampled before any anticoagulant treatment. Whole blood was collected at the same timepoint in citrate or EDTA anticoagulant at the ER and sent within 30 minutes to the Karolinska University Laboratory. After centrifugation at 2000*g* for 15 minutes, plasma aliquots were snap frozen and stored at -80°C. *Data collection*: For each patient, doctors filled in a questionnaire detailing (1) any provoking factors within one month preceding the visit to the ER (2) current health situation, alcohol consumption and smoking habits; (3) family history of VTE (4) ongoing antithrombotic (antiplatelet) treatment and (5) estrogen containing contraceptives and hormone replacement therapy (women only). Information from the ER visit on patient sex, weight and height (when available) along with results from routine

laboratory tests e.g, blood count, D-dimer, C-reactive protein (CRP), creatinine, international normalized ratio (INR) and activated partial thromboplastin time (aPTT) were extracted from the medical records. In total, 158 patients were included (52 cases). For the present study, 48 cases were available for analysis and 48 controls were matched by sex, and as closely as possible by age (mean age difference [years] cases vs. controls, women: 0.95, men: 3.65). Clinical characteristics of the sample set is given in Table 1. (**ii**) *VEBIOS Coagulation study* is an on-going case-control study established January 2011 of patients sampled at an outpatient coagulation clinic sampled 1-6 months after discontinuation of 6-12 months anticoagulant treatment after a verified first VTE (DVT to the lower limbs and/or PE), sex and age matched with healthy controls from the population. Patients were between 18 to 70 years of age, free from cancer, severe thrombophilia and pregnancy at inclusion [23]. In the present study, we analysed an extended sample set of *VEBIOS Coagulation* comprising all available samples; 144 cases and 140 controls (Table S2, Tab 1). Approval for *VEBIOS* was granted by the regional research ethics committee in Stockholm, Sweden (KI 2010/636-31/4) and all participants gave informed written consent, in accordance with the Declaration of Helsinki.

Replication cohorts

The Swedish Karolinska Age Adjusted D-Dimer study (*DFW-VTE study*) includes patients with clinically suspected acute VTE, prospectively recruited from the ER of Karolinska University Hospital in Huddinge, Stockholm, as previously described [44]. The patients were out-patients with low-to-high probability of acute PE or DVT in a lower limb. The study was approved by the regional ethics review board in Stockholm (DNR 2013-2143-31-2), and all participants gave informed written consent, in accordance with the Declaration of Helsinki. For the current study, biobanked plasma aliquots collected at the ER visit were available for a subset of subjects comprising 15 patients with PE, 39 with DVT, and 146 controls where VTE was excluded. Controls were identified based on negative diagnostic imaging, or a low Well´s score together with negative D-dimer. Clinical characteristics are described in Table S2, Tab 2.

688 The *FARIVE study* is a French multicentre case-control study carried out between 2003-2009, as
689 previously described [45]. The study consists of patients with first confirmed VTE (DVT to the
690 lower limbs and/or PE) from 18 years of age, matched to hospital controls with no previous
691 thrombotic event. All patients were free of known or recently discovered cancer at the time of VTE
692 diagnosis. Patients treated for cancer >5 years before the episode without recurrence could be
693 included. The study was approved by the Paris Broussais-HEGP ethics committee in Paris (2002-
694 034) and all participants gave informed written consent, in accordance with the Declaration of
695 Helsinki. In the current study we used a subset of *FARIVE* samples (n=1158), as previously
696 described [23, 81]. From most cases, blood was collected in the first week after diagnosis and
697 during anticoagulant treatment initiation. Clinical characteristics are described in Table S2, Tab
698 3. Information of sex was obtained from medical records and population registries.
699 The *Riesgo de Enfermedad TROmboembólica VEnosa (RETROVE) study* is a prospective case–
700 control study of 400 consecutive patients with VTE (cancer associated thrombosis excluded) and
701 400 healthy control volunteers. Individuals were recruited at the Hospital de la Santa Creu i Sant
702 Pau of Barcelona (Spain) between 2012 and 2016. Controls were selected according to the age
703 and sex distribution of the Spanish population (2001 census) [82]. All individuals were ≥ 18 years.
704 All procedures were approved by the Institutional Review Board of the Hospital de la Santa Creu
705 i Sant Pau, and all participants gave informed written consent, in accordance with the Declaration
706 of Helsinki. In the current study, samples from 308 cases and 360 controls were used. Clinical
707 characteristics are described in Table S2, Tab 4.
708 The *Marseille Thrombosis Association study* (*MARTHA*) is a population based single centre study,
709 as previously described [81]. Recruitment in *MARTHA* started in 1994 at Timone Hospital in
710 Marseille (France) and is still ongoing. The cohort from 1994 and 2008, includes a total of 1542
711 VTE-cases (66% women) that donated blood for further analysis. All patients had a history of a
712 first VTE documented by venography, Doppler ultrasound, angiography and/or
713 ventilation/perfusion lung scan [47]. Ethical approval was granted from the Department of Health

and Science, France (2008-880 & 09.576) and all participants gave informed written consent, in accordance with the Declaration of Helsinki. In the current study, proteomics data generated for 1322 sampled *MARTHA* cases was used. For 669 of the *MARTHA* cases, follow up data up to 12 years post-event was available and used to analyse risk of recurrent VTE. For a subset of 774 cases data for thrombin generation potential (TGP) was available for the same samples used for CFHR5 measurement [83], which was used to analyse the association between CFHR5 and blood coagulability. Clinical characteristics are described in Table S2, Tab 5.

ANALYSIS OF PLASMA BY TARGETED AFFINITY PROTEOMICS

The candidate target selection was based on the following, as previously described [23]: (1) proteins with previous support, or hypothesis of association with VTE and/or intermediate traits, and, (2) availability of corresponding antibodies assessed for target specificity in the Human Protein Atlas (HPA) antibody resource. Based on type of prior support and/or rationale for inclusion, selected candidate targets were grouped into four categories, ranging from ´known/probable´ (A) to ´plausible/hypothetical´ (D):

- Category A: Proteins with an established VTE association, including support from functional analysis, e.g., von Willebrand factor (VWF) [84, 85].
- Category B: Targets with: (a) a reported genetic association with VTE, such as single nucleotide polymorphism (SNP) in the gene/locus e.g. Adhesion G protein-coupled receptor B3 (ADGRB3) [86], or (b) an associated with cardiovascular events and/or arterial thrombosis, based on genetic and/or functional data e.g., class IA phosphoinositide 3-kinase β (PI3Kβ) [87]**.**
- Category C: (a) protein encoded by genes we previously identified as having body-wide endothelial cell enriched expression e.g., Myc target 1 (MYCT1) [28], or (b) proteins involved in intermediate traits related to thrombosis, e.g., protein disulphide isomerase A4 (PDIA4) [88], or (c) plasma proteins we previously identified as associated with myocardial infarction or stroke [89].

- Category D: proteins with functions in pathways of relevance to thrombosis or intermediate traits, in the absence of evidence for a direct role e.g., integrin alpha 4 subunit (ITGA4) [90].

Following assessment of available target specific antibodies in the HPA resource, a final panel of 408 target proteins were selected for the discovery screen (from 586 proposed candidates). Target categories for each candidate are given in Table S1, Tab 1.

Plasma proteomic profiles in *VEBIOS ER* were generated using multiplexed suspension bead arrays (SBA) with 756 individual HPA antibodies targeting the 408 proteins (Table S1, Tab 1), using identical design, procedures and methods as previously described [23]. Briefly, paired samples were randomly distributed within the same 96-well area. Two suspension bead arrays composed of 380 antibodies and 4 controls were used to sequentially generate profiles of the 96 samples in parallel. Median fluorescence intensity (MFI) values were obtained from the suspension bead array assay by detecting at least 32 beads per ID and sample with the FlexMap 3D instrument (Luminex® Corp). Proteomics profiling was performed in both EDTA and Citrate plasma. The selected target proteins and categories are given in Table S1, Tab 1. A significance threshold of $p<0.01$ in both EDTA and Citrate plasma was used as selection criteria.

Immunocapture mass spectrometry (IC-MS)

IC-MS was performed in triplicate of pooled plasma, as previously described [23] using the HPA059937 antibody (Atlas Antibodies) or MAB3845 (RnD Biosystems) and rabbit or mouse immunoglobulin G (rIgG, AB-105-C [RnD] and PMP01X [Biorad], respectively) as respective negative controls. In brief, samples were treated in 10 mM dithiothreitol followed by 50 mM chloroacetamide. Overnight sample digestion at 37 °C using Trypsin, was quenched with 0.5% (v/v) trifluoroacetic acid. Digested samples were analyzed using an Ultimate 3000 RSLC nanosystem (Dionex) coupled to a Q-Exactive HF (Thermo). Resulting raw files were searched using the engine Sequest and Proteome Discoverer platform (PD, v1.4.0.339, Thermo Scientific and Uniprot whole human proteteome [20180131, for HPA058337], or MaxQuant [91] (v. 2.1.4.0) against whole human proteome (UniProt,20210811, for MAB3845) using default settings and

label-free quantification. An internal database containing the most common proteins detected by IC-MS in plasma was used to calculate Z-scores [32]. A z-score of ≥3, corresponding to a p-value <0.01 [Confidence level 99%], was used as cut-off.

Mass Spectrometry analysis

Sample preparation

Blood plasma was diluted 10 times with 1x PBS, 1% sodium deoxycholate and processed as described in [92] and above in the IC-MS section. The digested samples were desalted using in-house prepared StageTips packed with Empore C18 Bonded Silica matrix (CDS Analytical, CN: 98-0604-0217-3) as described in [93]. Briefly, three layers of octadecyl membrane were placed in 200µl pipette tips, activated by addition of 100% acetonitrile (ACN) and subsequently equilibrated with 0.1% TFA. Approximately 15µg of peptides was added to the StageTip membrane and washed twice with 0.1% TFA. The peptides were eluted in two-step elution with 30µl of solvent containing 80% ACN, 0.1% formic acid (FA). Each desalting step required an in between centrifugation for 2 min at 1000x*g*. Desalted peptides were vacuum-dried and stored at -20 °C. Prior to LC-MS/MS analysis samples were dissolved in Solvent A (3% ACN, 0.1% FA) and amount corresponding to approximately 3 µg of raw plasma subjected to LC-MS/MS analysis.

DIA-MS analysis

The LC-MS/MS analysis was performed using an online system of Ultimate 3000 LC (Thermo Scientific) connected to Q Exactive HF (Thermo Scientific) mass spectrometer. First, the amount corresponding to 3µg of raw plasma was loaded onto a trap column (CN: 164535, Thermo Scientific) and washed for 3 minutes at 7µl/minute with solvent A. Peptides were separated by a 25 cm analytical column (CN: ES802A, Thermo Scientific) following a linear 40-minute gradient ranging from 1% to 32% Solvent B (95% ACN, 0.1% FA) at 0.7µl/ minute. The washout of analytical column was performed with 99% B for 1 minute followed by two seesaw gradients from 1% to 99% Solvent B over 4 minutes. Column was then equilibrated for 9 minutes with 99% Solvent A. The MS was operated in DIA mode with each cycle comprising of one full MS scan

791 performed at 30,000 resolution (AGC target $3e^6$, mass range 300-1,200 m/z and injection time
792 105 ms) followed by 30 DIA MS/MS scans with 10 m/z windows with 1 m/z margin ranging 350-
793 1,000 m/z at 30,000 resolution (AGC target $1e^6$, NCE 26, isolation window 12 m/z, injection time
794 55 ms).
795 *Data processing*
796 Resulting raw files were converted to mzML format using peak picking filter within ProteoWizard
797 provided software tool msConvert [94]. Resulting mzML files were searched using EncyclopeDIA
798 [95] against a spectral library generated with a deep learning network Prosit, which is integrated
799 into ProteomicsDB [96]. A whole human proteome (Homo Sapiens UniProt ID: #UP000009606,
800 20,205 entries, accessed 20170918) was used as a background proteome. Finally, the
801 quantification reports were saved, and the protein quantities calculated using top3 method from
802 the peptide intensities [97].
803 **Western blotting**
804 Recombinant CFHR5 (rCFHR5, 100 ng, R&D), normal human plasma (NP, 1 µl, George king)
805 and plasma depleted of the 14 most abundant proteins by depletion spin column (Thermo
806 scientific) (DP, 10 µl) were loaded on SDS PAGE 4-12% (Invitrogen) in non-reducing (without
807 dithiothreitol [DTT], NR) or reducing conditions (with DTT, R). After electrophoresis and transfer
808 onto PVDF membrane, protein was detected using antibodies HPA059937 (original target
809 SULF1), HPA072446 and MAB3845 (both targeting CFHR5). After incubation with horseradish
810 peroxidase (HRP)-coupled goat anti-rabbit or anti-mouse antibodies (1:2000, Dako), bands were
811 detected using chemiluminescence (ECL, Biorad). Molecular weight attributed using PageRuler™
812 prestained protein ladder (Thermo scientific). WB analysis verified that HPA072446 and
813 MAB3845 bind monomeric and homodimeric form of recombinant CFRHR5, and that HPA072446
814 detects a band in plasma corresponding to CFHR5 (Figure S2B).
815 **In-house developed bead based dual binder immunoassays**

A Suspension Bead Array (SBA) was built with the capture antibodies raised against human extracellular sulfatase 1-SULF1 (rabbit polyclonal HPA059937) and human CFHR5 (rabbit polyclonal HPA072446 and HPA073894) covalently coupled to color-coded magnetic beads as previously described [98, 99]. Bead-coupled rabbit IgG and mouse-IgG and bare beads were included as negative controls. Mouse anti-human SULF1 antibodies ABIN525031 (Abnova), ab172404 (Abcam), PA5-113112 (Thermofisher) HPA054728 and HPA051204 (Atlas antibodies), and mouse monoclonal anti-human CFHR5 (R&D systems, MAB3845) antibody were labelled with biotin and used as detection antibodies in combination with their respective capture antibodies. Citrate plasma samples were thawed on ice and centrifuged for 1 min at 2000 rpm and diluted in buffer polyvinyl casein 10% rIgG (PVXcas 10% rIgG; polyvinyl alcohol, Sigma Aldrich P8136; polyvinylpyrrolidone, Sigma Aldrich PVP360; Blocker Casein, Thermo 37528), heated at 56°C for 30 min and incubated with the SBA overnight. The detection antibody was used at 1ug/mL for 90 min, and streptavidin- R-phycoerythrin (R-PE) conjugate (Life Technologies; SA10044) was used for the fluorescence read out in Luminex platform. The dual binder assays based on HPA059937, HPA072446 or HPA073894 as capture antibodies together with the monoclonal anti-human CFHR5 were used to measure samples in the *VEBIOS ER* cohort. Median fluorescence intensity (MFI) values were acquired by FlexMap 3D instrument (Luminex® Corp). The polyclonal anti-human CFHR5 HPA072446 capture antibody and the anti-human CFHR5 MAB3845 detection antibody were selected for development of a quantitative assay together with human recombinant CFHR5 (R&D systems; 3845-F5).

Absolute quantification of CFHR5, C3, C3c and D-dimer in plasma

For CFHR5 quantification, rabbit polyclonal anti-human CFHR5 HPA072446 (Atlas Antibodies) and mouse monoclonal anti-human CFHR5 (R&D systems; MAB3845) antibodies were used in a dual binder assay. Human recombinant CFHR5 (R&D systems; 3845-F5) spiked into chicken plasma (Sigma Aldrich, P3266) was used as a standard. All samples were diluted 1:300 in PVXcas 10% rIgG. For C3 quantification, mouse anti-human C3 and mouse monoclonal anti-

842 human C3 antibodies (Bsi0263, Bsi0190, respectively, Biosystems International) were used in a
843 dual binder assay. Human recombinant C3 (Sigma Aldrich, C2910) spiked into C3 depleted serum
844 (Merck, 234403) was used as a standard. All samples were diluted 1:5000 and analysed as
845 described above. Concentration of complement fragment 3c (C3c) in 20 *VEBIOS ER* cohort
846 samples, selected based on low (<2500 ng/ml) or high (3800 ng/ml) plasma concentrations of
847 CFHR5 were measured using a commercial sandwich C3c ELISA kit (Nordic BioSite, Sweden).
848 Samples were measured in duplicate and C3c concentration calculated using a standard curve.
849 D-dimer was quantified by ELISA (D-Di 96 test, product #00947, Asserachrom) following the
850 manufacturer´s instructions. In the *DFW-VTE* study, D-dimer values were analysed at the
851 Karolinska University Hospital Laboratory in fresh samples sent for routine clinical chemistry
852 analysis, as part of the work up in the ER.

853 **Statistical analysis**

854 **Plasma protein profiling and quantification**

855 Median fluorescence intensity (MFI) values were obtained from the suspension bead array assay
856 or dual binder assay were processed as follow: (1) probabilistic quotient normalization as
857 accounting for any potential sample dilution effects [100], and (2) multidimensional MA (M=log
858 ratio; A=mean average, scales) normalization to minimize the difference amount the subgroups
859 of the samples generated by experimental factor as multiple batches [101]. Log-transformation
860 was applied to reduce any skewness in the proteomic data distribution. Quantitative plasma levels
861 for CFHR5 and C3 resulted from extrapolating processed MFI values to standard curves
862 generated from 4 or 5 parametric logistic models [102]. Association of target proteins with VTE
863 was tested using linear regression analysis while adjusting for age and sex, unless stated
864 otherwise. BMI data was lacking for 33 of the 96 patients (7 cases and 26 controls), so we did not
865 adjust for BMI in the discovery analysis, however adjustment with BMI, CRP or the combination
866 of both were applied to CFHR5 quantitative data (Table S1, Tab 10) when data was available.
867 Correlation estimations among protein and/or clinical variables were calculated by Spearman´s

rank method. Odds ratio analyses were performed scaling the data to one standard deviation, and significance was obtained from applying generalized linear models. All tests were two-tailed. Analyses were performed using the R statistical computing software (versions R3.2.0-R4.0.5) [103] and data visualizations were created using GraphPad Prism (version 9.1.2).

Association of CFHR5 levels with VTE recurrence in the *MARTHA* cohort was assessed using a Cox survival model with left truncature at age at sampling. Analysis was adjusted for sex, familial history of VTE, provoked or unprovoked status of the first VTE, age at first VTE, and BMI, and were conducted using the Survival R package. The heterogeneity of the association between CFHR5 and VTE (recurrence) according to specific subgroups was assessed using the Cochran-Mantel-Haenszel statistical test [104].

Diagnostic Prediction model

Discriminatory accuracy of plasma concentrations of CFHR5 and of D-dimer categorized as 'positive´ or ´negative' using age adjusted D-dimer cutoff [35] in the different models was assessed using logistic regress ion analysis and presented as Area Under the Receiver Operating Characteristics curve (AUC). Statistical analyses were performed using R version 4.0.3. ROC curves for the different biomarker-based risk models based on plasma concentration CFHR5, dichotomized data on D-dimer (positive or negative) and Wells score (´VTE likely´ (≥2 for DVT and ≥4 for PE) or ´VTE unlikely´) were compared using the function roc.test (Delonge's test) in the RStudio attachment. All tests were two-tailed.

CFHR5 mRNA expression across human organs

As part of the Human Protein Atlas (HPA, www.proteinatlas.org/), the average TPM value of all individual samples for each human tissue in both the HPA and Genotype-Tissue Expression Project (GTEx) transcriptomics datasets were used to estimate the respective gene expression levels. To be able to combine the datasets into consensus transcript expression levels, a pipeline was set up to normalize the data for all samples. In brief, all TPM values per sample were scaled to a sum of 1 million TPM (denoted pTPM) to compensate for the non-coding transcripts that had

been previously removed. Next, all TPM values of all samples within each data source were normalized separately using Trimmed mean of M values to allow for between-sample comparisons. The resulting normalized transcript expression values (nTPM) were calculated for each gene in each sample. For further details see www.proteinatlas.org/about/assays+annotation - normalization rna. Analysis of liver single cell transcriptomes and visualisation was performed as part of the HPA single-cell transcriptomics map [38] from data generated in [105]. Gene ontology

CFHR5 mRNA co-expression analysis

Liver bulk RNAseq data analysed in this study was part of the Genotype-Tissue Expression (GTEx) Project (gtexportal.org) [106] (dbGaP Accession phs000424.v7.p2) (n=226). Pearson correlation coefficients were calculated between *CFHR5* expression values and those for all other mapped protein coding genes across the sample set. Weighted correlation network (WGCNA) analysis: The R package WGCNA was used to perform co-expression network analysis for gene clustering, on log2 expression values. The analysis was performed according to recommendations in the WGCNA manual. Genes with too many missing values were excluded using the goodSamplesGenes() function. The remaining genes were used to cluster the samples, and obvious outlier samples were excluded. Using these genes and samples a soft-thresholding power was selected and the networks were constructed using a minimum module size of 15 and merging threshold of 0.05. Eigengenes were calculates from the resulting clusters and eigengene dendrograms were constructed using the plotEigengeneNetworks() function. Gene ontology and reactome analysis was performed using (http://geneontology.org/docs/go-enrichment-analysis/) [39, 40], PFDR values for the top six over represented terms in each category are provided in Table S1, Tab 6).

Genome wide genotyping methods

RETROVE samples were typed with the Illumina Infinium Global Screening Array (GSA) v2.0 array at the Spanish National Cancer Research Centre in the Human Genotyping lab, a member

of CeGen. After genotyping, all monomorfic and unannotated variants were removed as well as polymorphisms with call rate <95% and those whose genotype distributions deviate from Hardy-Weinberg equilibrium at $p < 0.000001$. Remaining polymorphisms were then imputed using the TOPMed r2 reference panel using Eagle v2.4. *FARIVE* participants were genotyped using the Illumina Infinium Global Screening Array v3.0 (GSAv3.0) at the Centre National de Recherche en Génomique Humaine (CNRGH). A control quality has been performed on individuals and genetic variants using Plink v1.9 [107] and the R software v3.6.2 [103]. Individuals with at least one of the following criteria were excluded: discordant sex information (N=20), relatedness individuals (N=9) identified by pairwise clustering of identity by state distances (IBS), genotyping call rate lower than 99% (N=5), heterozygosity rate higher/lower than the average rate +/- 3 standard deviation (N=34). These criteria led to a final sample composed of 1,266 individuals. Among the 730,059 genotyped variants, we excluded 145,238 variants with incorrect annotation, 656 variants with deviation from Hardy-Weinberg equilibrium (HWE) in controls using the statistical threshold of $p<10^{-6}$, 47,286 variants with a Minor Allele Count (MAC) lower than 20, 1,774 variants with a call rate lower than 95%. Finally, 535,105 markers passed the control quality and were used for the imputation. The Imputation was performed with Minimac4 using the 1000 Genomes phase 3 version 5 reference panel [108]. *MARTHA* participants were genotyped with the Illumina bead arrays [109]. Quality control procedures were as previously described [110-112]. Briefly, SNPs showing genotyping call rate <99%, significant ($p<10^{-5}$) deviation from Hardy-Weinberg Equilibrium (HWE), with minor allele frequency (MAF) less than 1% in were filtered out. Individuals were excluded based on the following criteria: (i) genotyping success rates <95%, (ii) close relatedness as detected by pairwise clustering of identity by state distances (IBS) and multi-dimensional scaling (MDS) using PLINK, (iii) genetic outliers using principal components approach as calculated by EIGENSTRAT. After application of quality control filters, 1525 participants remained for association testing with CFHR5 plasma levels. We imputed genotypes

by using MaCH version 1.0.18.c and haplotypes from the 1000 Genomes Total European Ancestry (EUR) population (August 2010 release).

Genome Wide Association Study on CFHR5 plasma levels

All SNPs with imputation quality criterion (r^2/info score) greater than 0.30 and minor allele frequency (MAF) greater than 0.01 in each participating cohorts (*FARIVE, MARTHA, RETROVE*) were tested for association with CFHR5 plasma levels. Associations were assessed using a linear regression model adjusted for age, sex and study-specific principal components derived from genome-wide genotype data. Results obtained in the different contributing cohorts were then meta-analysed through a fixed effect model as implemented in the GWAMA software.[113].

A second round of meta-analysis was performed by integrating GWAS summary statistics from 3 additional proteogenomic resources where CFHR5 have been measured. These include 35,559 Icelander participants of the Decode project [114] and 10,708 participants from the Fenland study with CFHR5 plasma levels measured using the Somalogic platform[115] and an additional independent sample of 1,178 EPIC participants with CFHR5 measured using the Olink platform [116]. Of note, in a sample of 485 Fenland participants with both measurements CFHR5, the correlation between Somalogic and Olink-derived CFHR5 levels was 0.35 [Supplementary Data Set 2 in [77]]. To meta-analyse GWAS results for CFHR5 plasma levels measured with three different techniques, (dual binder assay, Somalogic and Olink), we used the Z-score fixed-effect model implemented in the METAL software [117]. In order to conduct the other genetic analyses, normalized regression coefficient and their standard error were derived from Z-scores using the method described in Chauhan et al [118].

Shared genetics between CFHR5 plasma levels and VTE risk

Using summary statistics of the latest GWAS on VTE risk [17], we deployed two complementary approaches to assess whether genetics could support a causal role for CFHR5 levels on VTE risk. First, we performed a colocalisation analysis [119] at each locus presenting with SNPs significantly ($p < 5E-08$) associated CFHR5 levels to estimate the posterior probability (PP4) of a

shared causal variant between CFHR5 and VTE risk. For this, all SNPs located at +/- 100kb from a lead variant were investigated through the use of the coloc R package. Second, we conducted a Mendelian Randomization (MR) analysis using, as instrumental variables, the lead SNPs at each locus found genome-wide significantly associated with CFHR5 plasma levels. MR estimates were computed with fixed effect Inverse Variance Weighted method as implemented in the MendelianRandomization R library. Additional MR methodologies expected to be more robust to the presence of pleiotropy including the Weighted Median [120] and Egger [121] methods, were also used.

Measurement of Thrombin Generation Potential

Thrombin generation potential (TGP) was measured in fresh frozen platelet poor plasma (PPP) using the Calibrated Automated Thrombogram (CAT®) method according to the manufacturer's instructions. Analyses in *MARTHA* and *RETROVE* are as previously described [83, 122]**.** Output parameters recorded were Lagtime (min), the time to the initial generation of thrombin after induction; Endogenous Thrombin Potential (ETP)(nmol/min), equal to the area under the Thrombogram curve; Peak (nmol/L), the maximum amount of thrombin produced after induction by 5pM tissue factor, time to peak (ttPeak, min), time from initiation of assay to peak thrombin generation.

Validation of structural functionality of recombinant CFHR5

In order to confirm the functionality of the commercial rCFHR5 reagent used, we evaluated its binding capacity to its known partners C-reactive protein (CRP) [53] and properdin [54], using in-house designed ELISA. Flat-bottom microtiter plates (Nunc) were coated with recombinant CRP (1 µg, AG723-M Sigma), properdin (5 µg, 341283 -Sigma) or BSA (5 µg, negative control). After blocking with TBS-T containing 3% BSA, serial dilutions of rCFHR5 (3845-F5, RnD systems) or heat denatured rCFHR5 (mock) were added in triplicate and incubated for 1 hour at 20°C. rCFHR5 binding to CRP or properdin was detected with monoclonal mouse anti-human CFHR-5 (MAB3845, RnD) followed by HRP-labeled rabbit anti-mouse Ig (Dako). Binding was visualized

997 using o-Phenylenediamine dihydrochloride (OPD) (P9187 Sigma) and absorbance measured at
998 492 nm following the addition of 3 M HCl as stop solution in microplate photometer (Multiskan
999 FC, Thermoscientific). rCFHR5 bound CRP and properdin in a dose-dependent manner (Figure
1000 S4A.i and S4B.i, respectively). Binding properties were impaired following rCFHR5 heat-
1001 denaturation, indicating structure-dependent function (Figure S4A.ii–B.ii). Immunostaining of
1002 rCFHR5 by Western blot in non-reducing conditions revealed two bands (Figure S2B, lane 1 and
1003 7), showing intact capacity for dimerization [123].

1004 **Effect of recombinant CFHR5 on platelet activation**

1005 Blood was drawn from healthy volunteers free from any anti-platelet therapy for at least 10 days
1006 and anticoagulated with sodium citrate. All donors signed informed consent, in accordance with
1007 approval of the Human Ethics Committee of the Medical University of Vienna (EK237/2004) and
1008 the Declaration of Helsinki. Whole blood was centrifuged (120*g*, 20 min, room temperature (RT))
1009 and platelet-rich plasma (PRP) harvested. To obtain isolated platelets, PRP was diluted with PBS
1010 and treated with PGI_2 (100ng/ml), centrifuged for 90 sec at 3000*g* and platelets were resuspended
1011 in PBS. This step was repeated twice. Platelet-rich plasma (PRP) or isolated platelets were
1012 incubated with recombinant CFHR5 (rCFHR5) in PBS (6µg/ml, 3845-F5, R&D systems) or PBS
1013 alone for 10 minutes before treatment with varying concentrations of ADP (1-5µM), TRAP-6 (3-
1014 15µM) or convluxin (1-6ng/ml) for 15 minutes. Platelets were subsequently incubated with primary
1015 antibodies: anti-human CD62P-AF647 (AK4), anti-human CD63-PE (H5C6) or anti-human
1016 CD41/CD61-FITC (PAC-1) (all Biolegend) for 20 minutes, washed (PBS then 500*g* for 10 min),
1017 then fixed with 1% paraformaldehyde and incubated with Alexa Fluor 647-streptavidin (Jackson
1018 Immuno Research, Ely, UK) for 20 minutes. Samples were analysed by flow cytometry (Cytoflex,
1019 Beckman Coulter GmbH, Krefeld, Germany) and data processed using Cytexpert (Beckman
1020 Coulter GmbH, Krefeld, Germany). In some experiments PRP was incubated for 20 minutes with
1021 0.25% DMSO, 100µM compstatin, PBS, 10µg/ml anti-C3a/C3a (desArg) (clone K13/16), prior to
1022 assay as described above.

Effect of recombinant CFHR5 on platelet aggregation

PRP was prepared from citrated blood of healthy volunteers by centrifugation (120*g*, 20min, RT). Platelet aggregation was determined by light transmission aggregometry using a Platelet Aggregation Profiler PAP-8 (möLab), defining 0% aggregation as naïve PRP as and 100% aggregation as platelet-poor plasma (PPP), which was obtained by centrifuging PRP in the presence of 0.1µg/ml PGI2 (1000*g*, 90sec, RT). To determine the effect of CFHR5 on platelet aggregation, PRP was monitored in the aggregometer for 1 minute before addition of 6µg/ml CFHR5 or PBS). After 10 minutes, PRP was stimulated with 7µM ADP and aggregation monitored for further 10 minutes.

Independent data access and analysis: Dr. Maria Jesus Iglesias had full access to all the data in the study and takes responsibility for its integrity and the data analysis.

Data availability: In order to minimize the possibility of unintentionally sharing information that can be used to re-identify private information, a subset of the data that support the findings of this study are available from the corresponding authors upon reasonable request. By contacting the corresponding authors (JO, DAT), procedures for sharing data, analytic methods, and study materials for reproducing the results or replicating the procedure can be arranged. When submitting an access request, please indicate: [name of PI and host organisation/contact details (including your name and email)/scientific purpose of data access request/commitment to inform when the data has been used in a publication/commitment not to host or share the data outside the requesting organisation/statement of non-commercial use of data].

Source data are provided with this paper.

The GWAS data is available through GWAS catalogue (GCP ID: GCP000508**).**

Availability of materials: Availability of human plasma samples in respective study are subject to limitations in local ethical permits and discretion of respective study PI.

ACKNOWLEDGEMENTS

We thank research nurses Anna Fahlén and Doris Näslin for assistance with inclusion in the *VEBIOS Coagulation* study. We are grateful to all the participants of the EPIC-Norfolk study who have been part of the project and to the many members of the study teams at the University of Cambridge who have enabled this research.

Data usage: We used data from the Genotype-Tissue Expression (GTEx) Project (gtexportal.org) [37]. The GTEx project was supported by the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Figures: Some parts of the figures were created with BioRender.com.

Funding: The study was supported by grants from Stockholm County Council (Region Stockolm) to JO (FoUI-948982, FoUI-949280, FoUI-950776, FoUI-952641, FoUI-952912, FoUI-954024, FoUI-955547**)**, from Familjen Erling Personssons Stiftelse to MU, Knut och Alice Wallenberg foundation to JO (2020.0182, 2020.0241), from Swedish Heart Lung Foundation to LB (20170759, 20170537), Swedish Research Council (VR) to LMB (2019-01493), HelseNord to JO (HNF1544-20). The Human Protein Atlas (HPA) is funded by The Knut and Alice Wallenberg Foundation. *MARTHA* and *FARIVE* related genetics research programs were funded by the GENMED Laboratory of Excellence on Medical Genomics [ANR-10-LABX-0013], a research program managed by the National Research Agency (ANR) as part of the French Investment for the Future and supported by the French INvestigation Network on Venous Thrombo-Embolism (*INNOVTE*). *MARTHA* and *FARIVE* genetic data analyses benefit from the technical support of the CBiB computing centre of the University of Bordeaux. GM and D-AT are supported by the EPIDEMIOM-VT Senior Chair from the University of Bordeaux initiative of excellence IdEX. GM benefited from the EUR DPH, a PhD program supported within the framework of the PIA3 (Investment for the future). Project reference 17-EURE-0019. The *RETROVE* study was supported by grants PI12/00612 and PI15/0026. Genotyping of the *RETROVE* samples was supported by grant PT17/0019, of the PE I+D+i 2013-2016, funded by ISCIII and ERDF. T.R.

acknowledges German Research Foundation (DFG) grants: 25440785 - SFB877, P6 - KFO306, 80750187 - SFB841. The EPIC-Norfolk study (DOI 10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1 MC-UU_12015/1 and MC_UU_00006/1) and Cancer Research UK (C864/A14136). The genetics work in the EPIC-Norfolk study was funded by the Medical Research Council (MC_PC_13048). SMD is supported by IK2-CX001780. This publication does not represent the views of the Department of Veterans Affairs or the United States Government. SMD receives research support from RenalytixAI and Novo Nordisk, all outside the scope of the current research.

AUTHOR CONTRIBUTIONS

Conceptualisation: JO, DAT, LMB, PEM.

Supervision: JO, DAT, LMB, PEM

Proteomics data and analyses: MJI, LSR, MGH, JMS, PMS, DK, FE, DAT, GM, LMB

Experimental data and analyses: MJI, LSR, CN, LMB, PD, WCS, AA, JBKP, MIK, PS

GWAS data and analyses: DAT, GM, MG, LG, FT, AB, JFD, MK, MP, CL, SMD, ADJ, NLS, DMK.

Resources and cohorts: MF, JA, AA, MU, TR, LG, MIK, MM, MH, DMS, AS, AMP, JE, JFD, JMSF, JCS, LMB, DAT, PEM, JO,

Writing – Original Draft: LMB, JO, MJI, DAT

Writing – Review & Editing: All

Visualisation: LMB, MJI, DAT, GM, CN, JBKP, JO.

Funding Acquisition: JO, DAT, LMB, PEM, MU, TR, JFD, JE, JMSF, SMD.

COMPETING INTERESTS

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1. Plasma proteomics profiling identifies CFHR5 associated with VTE. (A) Overview of the *VEBIOS ER* discovery cohort. 756 HPA antibodies, targeting 408 candidate proteins, were used to analyse plasma samples using affinity proteomics. Log fold changes in antibody MFI (mean fluorescent intensity) signal were calculated between VTE cases and controls in **(B)** citrate or **(C)** EDTA anticoagulated plasma; coloured circles indicate antibodies that generated signals significantly associated with VTE in both. MFI signals generated by these antibodies for controls and cases in **(D)** citrate plasma and **(E)** EDTA plasma. **(F)** Immunocapture-mass spectrometry identification of protein targets of HPA059937. **(G)** Dual binder assays were developed using an anti-CFHR5 detection antibody, combined with **(i)** HPA059937 (raised against SULF1) **(ii)**, anti-CFHR5 HPA073894 or **(iii)** anti-CFHR5 HPA072446 as capture antibodies. Monoclonal anti-CFHR5 (MAB3845) was in applied as detection in the three combination **(i-iii)**. CFHR5 levels in the citrated plasma samples were re-analysed, using the respective dual binder assays, to determine **(vii-ix)** levels (MFI) in controls vs. cases and **(vii-ix)** the correlation between the signal and those generated by the original single binder assay using HPA059937. Dual binder assay using capture antibody HPA072446 with a recombinant protein standard and MAB3845 as detection antibody, was used for absolute quantification of CFHR5 in samples from: **(H)** *VEBIOS ER* and **(J)** *VEBIOS Coagulation*. CFHR5 concentration was **(i)** measured in controls and cases, with associated OR (odds ratio per 1 standard deviation increase) or **(ii)** used to determine the correlation with C-reactive protein (CRP), or **(iii)** D-dimer concentration. ****p<0.00001, ***p<0.0001, **p<0.001, *p<0.01. For summary statistics see Table S1, Tab 2, Panel A.

Figure 2: *CFHR5* is expressed in hepatocytes and is VTE-associated independent of C3. **(A)** mRNA expression of *CFHR5* across 55 different human tissue types. **(B)** Expression of *CFHR5* in different liver cell types, analysed by ssRNAseq. **(C)** STRING protein-protein interaction analysis for genes identified as potentially co-expressed with *CFHR5* in liver by correlation-based

analysis of bulk mRNAseq. Coloured circles indicate closest network clusters. Complement component 3 (C3) concentration was measured in plasma from **(D)** *VEBIOS ER* or (E) *VEBIOS coagulation* to determine: **(i)** differences between controls and cases, or **(ii)** correlation with CFHR5 in controls (left) or cases (right). **p<0.01. Summary of statistical analysis can be found in Table S1, Tab 2, Panel B. C4BPA complement factor 4 binding protein, CFI complement factor I, CFB complement factor B, CFH; complement factor H, C1S; complement component 1, C1R; complement component 1, C2; complement component 2, C8a; complement component C8 alpha chain, C8b; complement component C8 beta chain, C5 complement component 5, C9; complement component 9.

Figure 3: CFHR5 concentration is associated with VTE in 5 independent studies. Plasma samples were generated as part of: **(A)** the Swedish *VEBIOS ER* or **(B)** the Swedish *DFW-VTE* study, both of which recruited patients presenting with suspected VTE. Samples were drawn pre-treatment, and cases and controls were identified based on confirmed or ruled out diagnosis. **(C)** The French *FARIVE* study recruited patients with confirmed acute VTE, with controls recruited from hospital patients treated for non-VTE causes. Samples were drawn within one week from diagnosis, during initiation of treatment. **(D)** The Swedish *VEBIOS Coagulation* or **(E)**, Spanish *RETROVE* study recruited cases from patients who had a prior first time VTE, sampled post-treatment (6-12 months anticoagulants), with healthy controls recruited from the general population. CFHR5 concentration was measured in the respective samples using a dual binder assay. **p<0.001, ***p<0.0001 adjusted for difference in sex and age. OR (1SD) = Odds ratio for 1 standard deviation elevation. CI= confidence interval.

Figure 4: GWAS analysis identifies a CFHR5 pQTL on Chromosome 1 q31.3. (A) Manhattan plot of the meta-analysis on INVENT-MVP consortium resources [17] showing six loci associated with CHFR5 plasma levels and VTE risk: *CFHR1,CFHR4* (rs10737681, p=2.94E-396), *HNF1A* (rs2393776, p=1.48E-21), *JMJD1C* (rs7916868, p=4.61E-12), *TRIB1* (rs28601761, p=4.39E-09),

DNAH10 (rs7133378, p=2.43E-08) and *HNF4A* (rs1800961, p=4.97E-08). Lead SNPs at *HNF1A* and *DNAH10* are rs2393776 and rs7133378, respectively. They are ~3Mb apart and do not show any linkage disequilibrium (pairwise r^2=0). **(B)** Regional association plot [124] at the Chromosome 1 locus covering more than 10Mb from *CFHR1* to *CFHR5* around the lead SNP associated with CFHR5 plasma levels. **(C)** CFHR5 plasma levels for 16 patients who are carriers of rare non-synonymous CFHR5-associated variants (rs139017763, rs41299613 or rs35662416) and non-carriers. ****p<0.00001. See also Table S1, Tabs 11-13.

Figure 5: Recombinant CFHR5 enhances platelet activation in platelet rich plasma.
Platelet activation was measured by surface expression of **(A)** P-selectin, **(B)** activated GP IIb/IIIa ($PAC1^+$) or **(C)** CD63, following treatment of platelet rich plasma with different concentrations of **(i)** adenosine diphosphate (ADP) **(ii)** convulxin or **(iii)** TRAP6, following pre-incubation with recombinant CFHR5, or PBS control. **(D)** Platelet aggregation was measured in response to ADP (2μm) following pre-incubation with recombinant CFHR5, or PBS control: **(i)** representative aggregation curve, **(ii)** maximum aggregation and **(iii)** slope. ADP-induced platelet activated GP IIb/IIIa ($PAC1^+$) was measured following preincubation with: (**E**) **(i)** DMSO (control) or **(ii)** compstatin, or (**F**) **(i)** PBS (control) or **(ii)** an anti-C3a antibody, followed by treatment with PBS or rCFHR5. US: unstimulated (PBS control). Each experiment is represented by an individual point and paired experiments connected by a dotted line. *p<0.05 **p<0.01 ***p<0.001 (p for trend bottom right).

TABLES

Table 1: Clinical Characteristic of the *VEBIOS ER* sample set

Table 2: Odds ratio of VTE associated with CHR5 concentration in five independent studies separated in tertiles

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY FILES

Figure S1: Antibody based suspension bead assay concepts.

Figure S2: Verification of antibody target and CFHR5 dual binder quantitative assay

Figure S3: GWAS analysis identifies a CFHR5 pQTL on Chromosome 1 q31.3.

Figure S4. Recombinant CFHR5 (rCFHR5) binding to CRP and Properdin

Figure S5: Recombinant CFHR5 does not potentiate platelet activation of washed platelets.

Figure S6. Relationship between CFHR5 and C3c levels in *VEBIOS ER* sample sub-set

Supplementary Table S1:

Tab 1: Selection of antibodies with p values, FC and log P values (Citrate & EDTA)

Tab 2: Summary of the statistic values for Figure 1 C-J and 2 D-E

Tab 3: IC-MS SULF1

Tab 4: Risk factors & clinical chemistry

Tab 5: UAC

Tab 6: Liver co-expression

Tab 7: STRING

Tab 8: C3 and CFHR5

Tab 9: Meta sub-analyses

Tab 10: Sub-analysis (BMI, CRP)

Tab 11: GWAS metanalysis A

Tab 12: GWAS metanalysis B

Tab 13: GWAS metanalysis C

Tab 14: Rare variants

Tab 15: TGP analysis

Supplementary Table S2:

Tab 1: VEBIOS Coagulation

Tab 2: DFW-VTE

Tab 3: FARIVE

Tab 4: RETROVE

Tab 5: MARTHA

Tab 6: ALL COHORTS

Source file

References

1. Johansson, M., L. Johansson, and M. Lind, *Incidence of venous thromboembolism in northern Sweden (VEINS): a population-based study.* Thromb J, 2014. **12**(1): p. 6.
2. Heit, J.A., *Epidemiology of venous thromboembolism.* Nature reviews. Cardiology, 2015. **12**(8): p. 464-74.
3. Sogaard, K.K., et al., *30-year mortality after venous thromboembolism: a population-based cohort study.* Circulation, 2014. **130**(10): p. 829-36.
4. Martinez, C., et al., *Epidemiology of first and recurrent venous thromboembolism: A population-based cohort study in patients without active cancer.* Thrombosis and Haemostasis, 2014. **112**(2): p. 255-263.
5. Wells, P.S., et al., *Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis.* N Engl J Med, 2003. **349**(13): p. 1227-35.
6. Wells, P.S., et al., *Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer.* Ann Intern Med, 2001. **135**(2): p. 98-107.
7. Jacobs, B., A. Obi, and T. Wakefield, *Diagnostic biomarkers in venous thromboembolic disease.* J Vasc Surg Venous Lymphat Disord, 2016. **4**(4): p. 508-17.
8. Eichinger, S., et al., *Risk assessment of recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism: the Vienna prediction model.* Circulation, 2010. **121**(14): p. 1630-6.
9. Tosetto, A., et al., *Predicting disease recurrence in patients with previous unprovoked venous thromboembolism: a proposed prediction score (DASH).* J Thromb Haemost, 2012. **10**(6): p. 1019-25.
10. Verhovsek, M., et al., *Systematic review: D-dimer to predict recurrent disease after stopping anticoagulant therapy for unprovoked venous thromboembolism.* Ann Intern Med, 2008. **149**(7): p. 481-90, W94.
11. Rodger, M.A., et al., *Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy.* Canadian Medical Association Journal, 2008. **179**(5): p. 417-26.
12. Baglin, T., et al., *Unprovoked recurrent venous thrombosis: prediction by D-dimer and clinical risk factors.* J Thromb Haemost, 2008. **6**(4): p. 577-82.
13. Bruinstroop, E., et al., *Elevated D-dimer levels predict recurrence in patients with idiopathic venous thromboembolism: a meta-analysis.* J Thromb Haemost, 2009. **7**(4): p. 611-8.
14. Douketis, J., et al., *Patient-level meta-analysis: effect of measurement timing, threshold, and patient age on ability of D-dimer testing to assess recurrence risk after unprovoked venous thromboembolism.* Ann Intern Med, 2010. **153**(8): p. 523-31.
15. Lindstrom, S., et al., *Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism.* Blood, 2019. **134**(19): p. 1645-1657.
16. Klarin, D., et al., *Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease.* Nat Genet, 2019. **51**(11): p. 1574-1579.
17. Thibord, F., et al., *Cross-Ancestry Investigation of Venous Thromboembolism Genomic Predictors.* Circulation, 2022. **146**(16): p. 1225-1242.

18. Morange, P.E. and D.A. Tregouet, *Current knowledge on the genetics of incident venous thrombosis.* J Thromb Haemost, 2013. **11 Suppl 1**: p. 111-21.
19. Martinelli, I., V. De Stefano, and P.M. Mannucci, *Inherited risk factors for venous thromboembolism.* Nat Rev Cardiol, 2014. **11**(3): p. 140-56.
20. Goldhaber, S.Z., *Risk factors for venous thromboembolism.* J Am Coll Cardiol, 2010. **56**(1): p. 1-7.
21. Crous-Bou, M., L.B. Harrington, and C. Kabrhel, *Environmental and Genetic Risk Factors Associated with Venous Thromboembolism.* Semin Thromb Hemost, 2016. **42**(8): p. 808-820.
22. Jensen, S.B., et al., *Discovery of novel plasma biomarkers for future incident venous thromboembolism by untargeted synchronous precursor selection mass spectrometry proteomics.* Journal of Thrombosis and Haemostasis, 2018. **16**: p. 1763-1774.
23. Bruzelius, M., et al., *PDGFB, a new candidate plasma biomarker for venous thromboembolism: results from the VEREMA affinity proteomics study.* Blood, 2016. **128**(23): p. 59-67.
24. Blostein, M.D., et al., *Elevated plasma gas6 levels are associated with venous thromboembolic disease.* J Thromb Thrombolysis, 2011. **32**(3): p. 272-8.
25. Song, Y., et al., *Increased expressions of integrin subunit beta1, beta2 and beta3 in patients with venous thromboembolism: new markers for venous thromboembolism.* Int J Clin Exp Med, 2014. **7**(9): p. 2578-84.
26. Memon, A.A., et al., *Identification of novel diagnostic biomarkers for deep venous thrombosis.* British Journal of Haematology, 2018. **181**: p. 378-385.
27. Ten Cate, V., et al., *Protein expression profiling suggests relevance of noncanonical pathways in isolated pulmonary embolism.* Blood, 2021. **137**(19): p. 2681-2693.
28. Butler, L.M., et al., *Analysis of Body-wide Unfractionated Tissue Data to Identify a Core Human Endothelial Transcriptome.* Cell Syst, 2016. **3**(3): p. 287-301 e3.
29. Mathews, J.A., et al., *Considerations for Soluble Protein Biomarker Blood Sample Matrix Selection.* AAPS J, 2020. **22**(2): p. 38.
30. Baker, M., *Reproducibility crisis: Blame it on the antibodies.* Nature, 2015. **521**(7552): p. 274-6.
31. Weller, M.G., *Quality Issues of Research Antibodies.* Anal Chem Insights, 2016. **11**: p. 21-7.
32. Fredolini, C., et al., *Systematic assessment of antibody selectivity in plasma based on a resource of enrichment profiles.* Sci Rep, 2019. **9**(1): p. 8324.
33. Schwenk, J.M., et al., *The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays.* J Proteome Res, 2017. **16**(12): p. 4299-4310.
34. Desiere, F., et al., *The PeptideAtlas project.* Nucleic Acids Res, 2006. **34**(Database issue): p. D655-8.
35. Douma, R.A., et al., *Potential of an age adjusted D-dimer cut-off value to improve the exclusion of pulmonary embolism in older patients: a retrospective analysis of three large cohorts.* BMJ, 2010. **340**: p. c1475.
36. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
37. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.

38. Karlsson, M., et al., *A single-cell type transcriptomics map of human tissues.* Sci Adv, 2021. **7**(31).
39. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
40. Gene Ontology, C., *The Gene Ontology resource: enriching a GOld mine.* Nucleic Acids Res, 2021. **49**(D1): p. D325-D334.
41. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
42. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Res, 2019. **47**(D1): p. D607-D613.
43. Norgaard, I., S.F. Nielsen, and B.G. Nordestgaard, *Complement C3 and High Risk of Venous Thromboembolism: 80517 Individuals from the Copenhagen General Population Study.* Clin Chem, 2016. **62**(3): p. 525-34.
44. Farm, M., et al., *Age-adjusted D-dimer cut-off leads to more efficient diagnosis of venous thromboembolism in the emergency department: a comparison of four assays.* J Thromb Haemost, 2018. **16**(5): p. 866-875.
45. Zhu, T., et al., *Association of influenza vaccination with reduced risk of venous thromboembolism.* Thromb Haemost, 2009. **102**(6): p. 1259-64.
46. Llobet, D., et al., *Platelet hyperaggregability and venous thrombosis risk: results from the RETROVE project.* Blood Coagul Fibrinolysis, 2021. **32**(2): p. 122-131.
47. 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, 2012. **157**(2): p. 230-9.
48. Lindström, S., et al., *A large-scale exome array analysis of venous thromboembolism.* Genet Epidemiol, 2019. **43**(4): p. 449-457.
49. Wang, H., et al., *D-dimer, thrombin generation, and risk of a first venous thrombosis in the elderly.* Res Pract Thromb Haemost, 2021. **5**(5): p. e12536.
50. Sauter, R.J., et al., *Functional Relevance of the Anaphylatoxin Receptor C3aR for Platelet Function and Arterial Thrombus Formation Marks an Intersection Point Between Innate Immunity and Thrombosis.* Circulation, 2018. **138**(16): p. 1720-1735.
51. Afshar-Kharghan, V., *Complement and clot.* Blood, 2017. **129**(16): p. 2214-2215.
52. Subramaniam, S., et al., *Distinct contributions of complement factors to platelet activation and fibrin formation in venous thrombus development.* Blood, 2017. **129**(16): p. 2291-2302.
53. McRae, J.L., et al., *Human factor H-related protein 5 has cofactor activity, inhibits C3 convertase activity, binds heparin and C-reactive protein, and associates with lipoprotein.* J Immunol, 2005. **174**(10): p. 6250-6.
54. Chen, Q., et al., *Complement Factor H-Related 5-Hybrid Proteins Anchor Properdin and Activate Complement at Self-Surfaces.* J Am Soc Nephrol, 2016. **27**(5): p. 1413-25.
55. Cserhalmi, M., et al., *Regulation of regulators: Role of the complement factor H-related proteins.* Semin Immunol, 2019. **45**: p. 101341.
56. Audu, C.O., et al., *Inflammatory biomarkers in deep venous thrombosis organization, resolution, and post-thrombotic syndrome.* J Vasc Surg Venous Lymphat Disord, 2020. **8**(2): p. 299-305.

57. Mosevoll, K.A., et al., *Altered plasma levels of cytokines, soluble adhesion molecules and matrix metalloproteases in venous thrombosis.* Thromb Res, 2015. **136**(1): p. 30-9.
58. van Hylckama Vlieg, A., et al., *The risk of a first and a recurrent venous thrombosis associated with an elevated D-dimer level and an elevated thrombin potential: results of the THE-VTE study.* J Thromb Haemost, 2015. **13**(9): p. 1642-52.
59. Machiela, M.J. and S.J. Chanock, *LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants.* Bioinformatics, 2015. **31**(21): p. 3555-7.
60. Sollis, E., et al., *The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource.* Nucleic Acids Res, 2022.
61. Chen, M.H., et al., *Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations.* Cell, 2020. **182**(5): p. 1198-1213.e14.
62. Johnson, A.D., et al., *Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists.* Nat Genet, 2010. **42**(7): p. 608-13.
63. Eicher, J.D., et al., *Replication and hematological characterization of human platelet reactivity genetic associations in men from the Caerphilly Prospective Study (CaPS).* J Thromb Thrombolysis, 2016. **41**(2): p. 343-50.
64. Markiewski, M.M., et al., *Complement and coagulation: strangers or partners in crime?* Trends Immunol, 2007. **28**(4): p. 184-92.
65. McRae, J.L., et al., *Human factor H-related protein 5 (FHR-5). A new complement-associated protein.* J Biol Chem, 2001. **276**(9): p. 6747-54.
66. Zipfel, P.F., et al., *Factor H and disease: a complement regulator affects vital body functions.* Mol Immunol, 1999. **36**(4-5): p. 241-8.
67. Murphy, B., et al., *Factor H-related protein-5: a novel component of human glomerular immune deposits.* Am J Kidney Dis, 2002. **39**(1): p. 24-7.
68. Csincsi, A.I., et al., *Factor H-related protein 5 interacts with pentraxin 3 and the extracellular matrix and modulates complement activation.* J Immunol, 2015. **194**(10): p. 4963-73.
69. Koupenova, M., et al., *Thrombosis and platelets: an update.* Eur Heart J, 2017. **38**(11): p. 785-791.
70. Montoro-Garcia, S., et al., *The Role of Platelets in Venous Thromboembolism.* Semin Thromb Hemost, 2016. **42**(3): p. 242-51.
71. Simes, J., et al., *Aspirin for the prevention of recurrent venous thromboembolism: the INSPIRE collaboration.* Circulation, 2014. **130**(13): p. 1062-71.
72. Tarantino, E., et al., *Role of thromboxane-dependent platelet activation in venous thrombosis: Aspirin effects in mouse model.* Pharmacol Res, 2016. **107**: p. 415-425.
73. Fukuoka, Y. and T.E. Hugli, *Demonstration of a specific C3a receptor on guinea pig platelets.* J Immunol, 1988. **140**(10): p. 3496-501.
74. Polley, M.J. and R.L. Nachman, *Human platelet activation by C3a and C3a des-arg.* J Exp Med, 1983. **158**(2): p. 603-15.
75. Heemskerk, J.W., N.J. Mattheij, and J.M. Cosemans, *Platelet-based coagulation: different populations, different functions.* J Thromb Haemost, 2013. **11**(1): p. 2-16.
76. Joshi, A. and M. Mayr, *In Aptamers They Trust: The Caveats of the SOMAscan Biomarker Discovery Platform from SomaLogic.* Circulation, 2018. **138**(22): p. 2482-2485.

77. Pietzner, M., et al., *Synergistic insights into human health from aptamer- and antibody-based proteomic profiling.* Nat Commun, 2021. **12**(1): p. 6822.
78. Olson, N.C., et al., *Soluble Urokinase Plasminogen Activator Receptor: Genetic Variation and Cardiovascular Disease Risk in Black Adults.* Circ Genom Precis Med, 2021. **14**(6): p. e003421.
79. Kline, J., *Response by Kline to Letter Regarding Article, "Over-Testing for Suspected Pulmonary Embolism in American Emergency Departments: The Continuing Epidemic".* Circ Cardiovasc Qual Outcomes, 2020. **13**(4): p. e006588.
80. Zarabi, S., et al., *Physician choices in pulmonary embolism testing.* CMAJ, 2021. **193**(2): p. E38-E46.
81. Trégouët, D.A., et al., *Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: Results from aGWAS approach.* Blood, 2009. **113**(21): p. 5298-5303.
82. Vazquez-Santiago, M., et al., *Platelet count and plateletcrit are associated with an increased risk of venous thrombosis in females. Results from the RETROVE study.* Thromb Res, 2017. **157**: p. 162-164.
83. Rocanin-Arjo, A., et al., *A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential.* Blood, 2014. **123**(5): p. 777-785.
84. Smith, N.L., et al., *Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium.* Circulation, 2010. **121**(12): p. 1382-92.
85. Tsai, A.W., et al., *Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE).* Am J Med, 2002. **113**(8): p. 636-42.
86. 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, 2010. **8**(12): p. 2671-9.
87. Laurent, P.A., et al., *Platelet PI3Kβ and GSK3 regulate thrombus stability at a high shear rate.* Blood, 2015. **125**(5): p. 881-8.
88. Cho, J., *Protein disulfide isomerase in thrombosis and vascular inflammation.* J Thromb Haemost, 2013. **11**(12): p. 2084-91.
89. Matic, L.P., et al., *Novel Multiomics Profiling of Human Carotid Atherosclerotic Plaques and Plasma Reveals Biliverdin Reductase B as a Marker of Intraplaque Hemorrhage.* JACC Basic Transl Sci, 2018. **3**(4): p. 464-480.
90. Nalls, M.A., et al., *Multiple loci are associated with white blood cell phenotypes.* PLoS Genet, 2011. **7**(6): p. e1002113.
91. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.
92. Kotol, D., et al., *Targeted proteomics analysis of plasma proteins using recombinant protein standards for addition only workflows.* Biotechniques, 2021. **71**(3): p. 473-483.
93. Rappsilber, J., Y. Ishihama, and M. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics.* Anal Chem, 2003. **75**(3): p. 663-70.
94. Chambers, M.C., et al., *A cross-platform toolkit for mass spectrometry and proteomics.* Nat Biotechnol, 2012. **30**(10): p. 918-20.

95. Searle, B.C., et al., *Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry.* Nat Commun, 2018. **9**(1): p. 5128.
96. Gessulat, S., et al., *Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning.* Nat Methods, 2019. **16**(6): p. 509-518.
97. Silva, J.C., et al., *Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition.* Mol Cell Proteomics, 2006. **5**(1): p. 144-56.
98. Drobin, K., P. Nilsson, and J.M. Schwenk, *Highly multiplexed antibody suspension bead arrays for plasma protein profiling.* Methods Mol Biol, 2013. **1023**: p. 137-45.
99. Neiman, M., et al., *Selectivity analysis of single binder assays used in plasma protein profiling.* Proteomics, 2013. **13**(23-24): p. 3406-10.
100. Dieterle, F., et al., *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.* Anal Chem, 2006. **78**(13): p. 4281-90.
101. Hong, M.G., et al., *Multidimensional Normalization to Minimize Plate Effects of Suspension Bead Array Data.* J Proteome Res, 2016. **15**(10): p. 3473-3480.
102. Häussler, R.S., et al., *Systematic Development of Sandwich Immunoassays for the Plasma Secretome.* Proteomics, 2019. **19**(15): p. e1900008.
103. *R Core Team, R A Language and Environment for Statistical Computing*. 2020, Fundation for Statistical Computing, Vienna, Austria.
104. MANTEL, N. and W. HAENSZEL, *Statistical aspects of the analysis of data from retrospective studies of disease.* J Natl Cancer Inst, 1959. **22**(4): p. 719-48.
105. MacParland, S.A., et al., *Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations.* Nat Commun, 2018. **9**(1): p. 4383.
106. Consortium, G.T., *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
107. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
108. Das, S., et al., *Next-generation genotype imputation service and methods.* Nat Genet, 2016. **48**(10): p. 1284-1287.
109. Sennblad, B., et al., *Genome-wide association study with additional genetic and post-transcriptional analyses reveals novel regulators of plasma factor XI levels.* Hum Mol Genet, 2017. **26**(3): p. 637-649.
110. Germain, M., et al., *Genetics of venous thrombosis: insights from a new genome wide association study.* PLoS One, 2011. **6**(9): p. e25581.
111. Antoni, G., et al., *Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels.* BMC Med Genet, 2011. **12**: p. 102.
112. Germain, M., et al., *Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism.* Am J Hum Genet, 2015. **96**(4): p. 532-42.
113. Magi, R. and A.P. Morris, *GWAMA: software for genome-wide association meta-analysis.* BMC Bioinformatics, 2010. **11**: p. 288.
114. Ferkingstad, E., et al., *Large-scale integration of the plasma proteome with genetics and disease.* Nat Genet, 2021. **53**(12): p. 1712-1721.
115. Pietzner, M., et al., *Mapping the proteo-genomic convergence of human diseases.* Science, 2021. **374**(6569): p. eabj1541.

116. Koprulu, M., et al., *From genome to phenome via the proteome: broad capture, antibody-based proteomics to explore disease mechanisms*. 2022 medRxiv 2022.08.19.22278984

117. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.

118. Chauhan, G., et al., *Association of Alzheimer's disease GWAS loci with MRI markers of brain aging.* Neurobiol Aging, 2015. **36**(4): p. 1765.e7-1765.e16.

119. Giambartolomei, C., et al., *Bayesian test for colocalisation between pairs of genetic association studies using summary statistics.* PLoS Genet, 2014. **10**(5): p. e1004383.

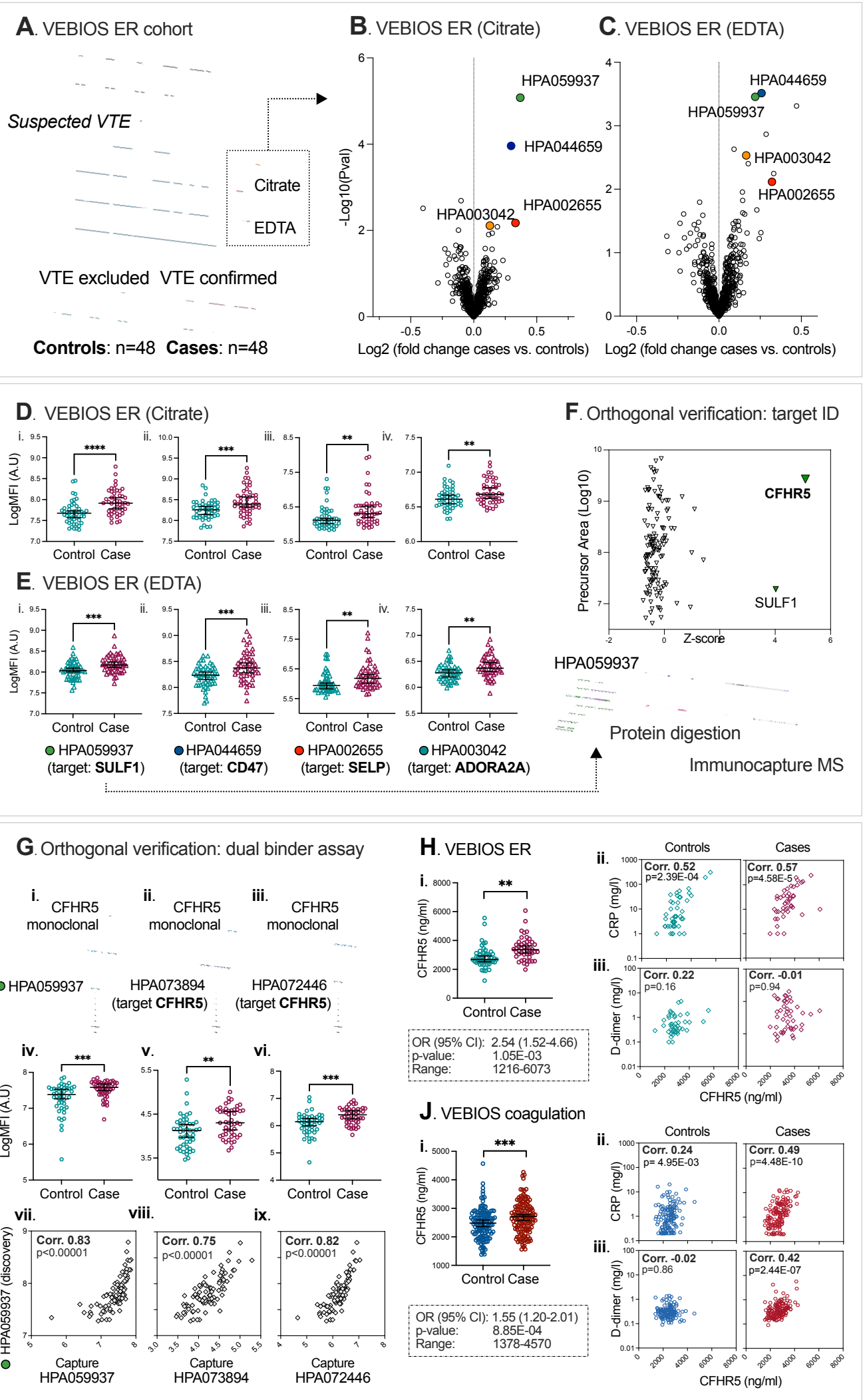
120. Hartwig, F.P., G. Davey Smith, and J. Bowden, *Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption.* Int J Epidemiol, 2017. **46**(6): p. 1985-1998.

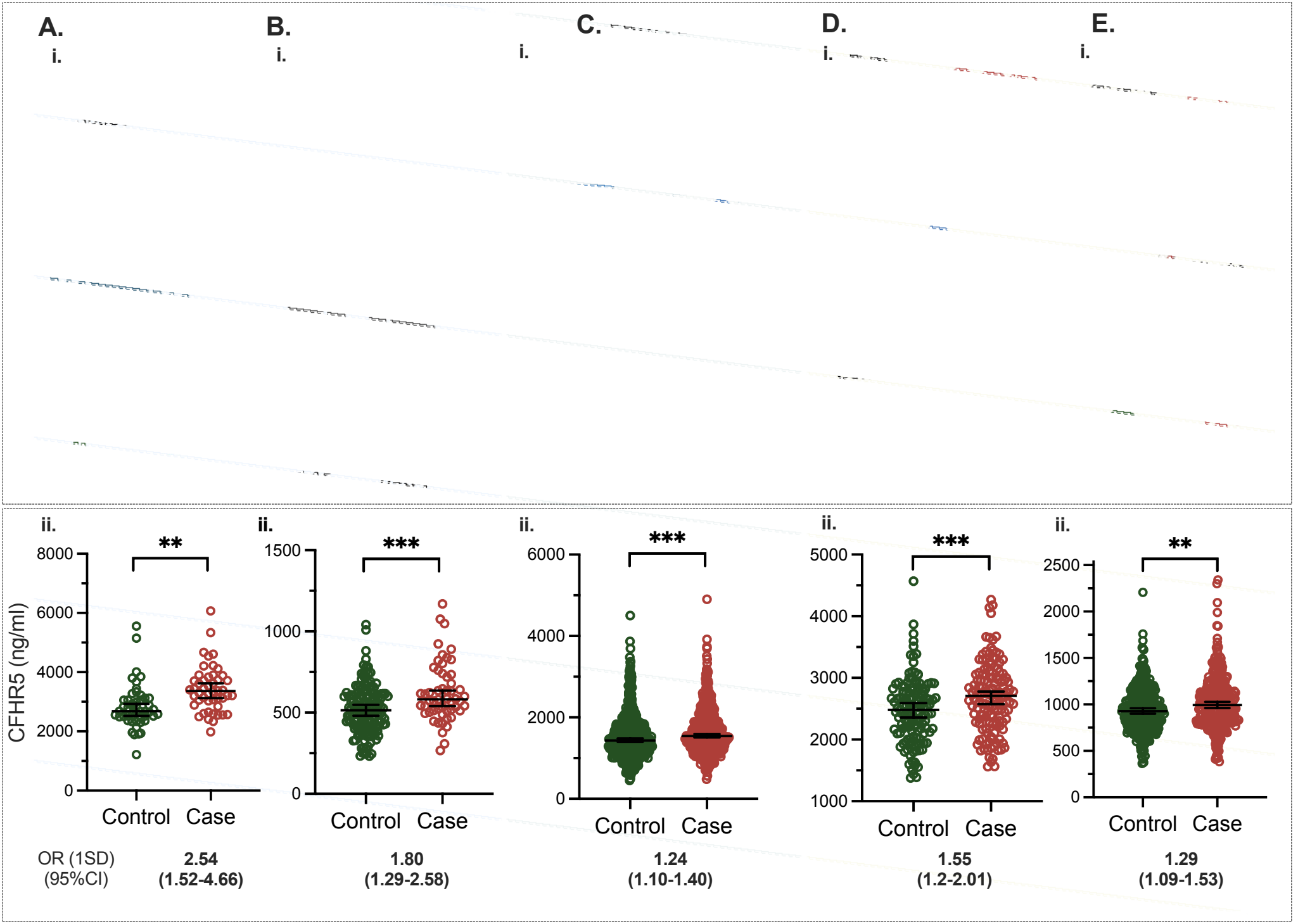
121. Bowden, J., G. Davey Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* Int J Epidemiol, 2015. **44**(2): p. 512-25.

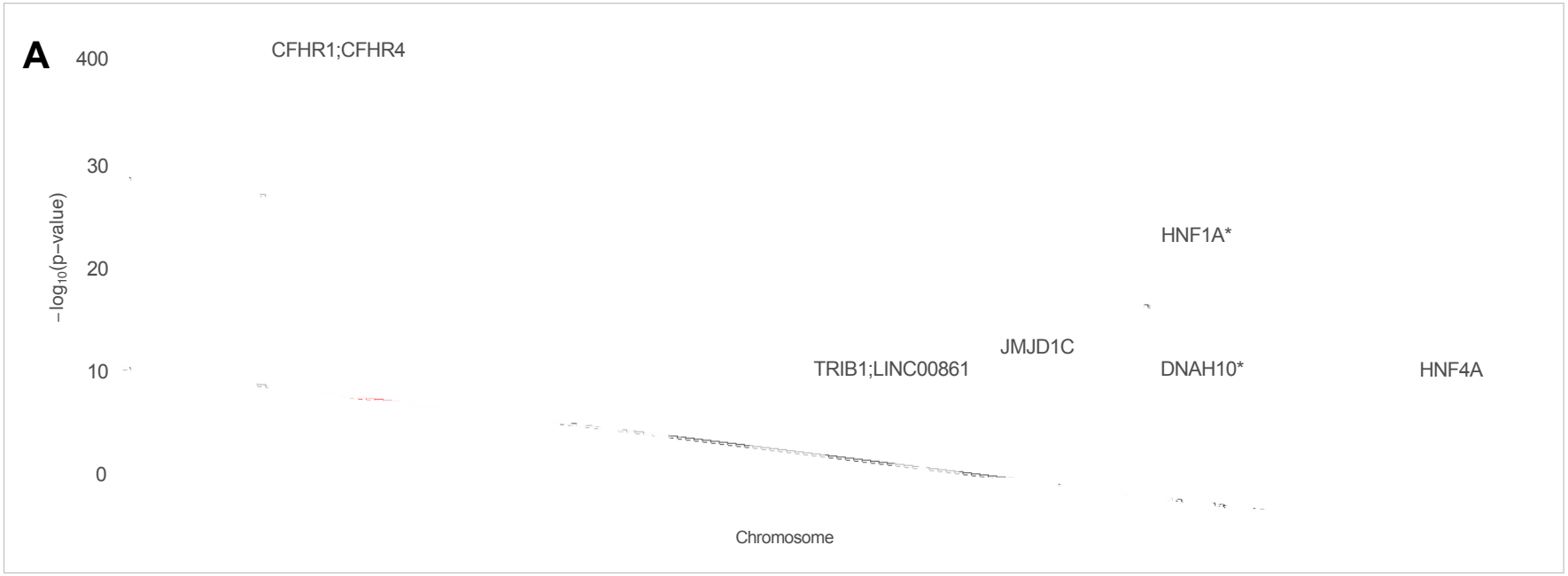
122. Martin-Fernandez, L., et al., *Genetic Determinants of Thrombin Generation and Their Relation to Venous Thrombosis: Results from the GAIT-2 Project.* PLoS One, 2016. **11**(1): p. e0146922.

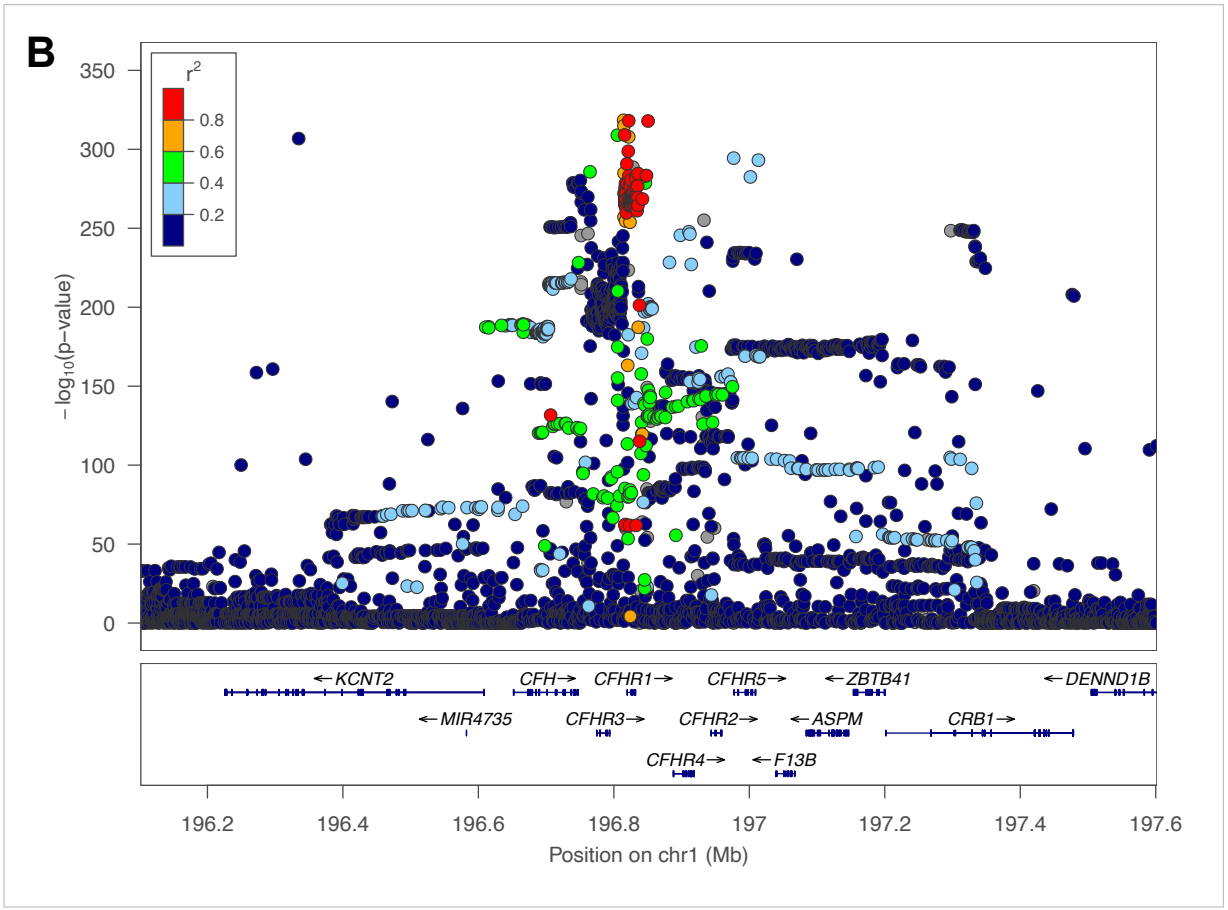
123. Goicoechea de Jorge, E., et al., *Dimerization of complement factor H-related proteins modulates complement activation in vivo.* Proc Natl Acad Sci U S A, 2013. **110**(12): p. 4685-90.

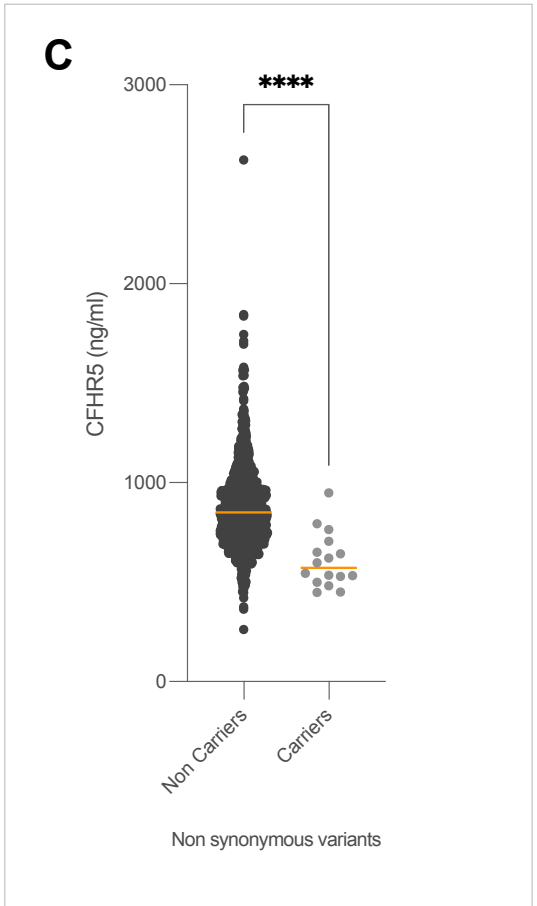
124. Pruim, R.J., et al., *LocusZoom: regional visualization of genome-wide association scan results.* Bioinformatics, 2010. **26**(18): p. 2336-7.

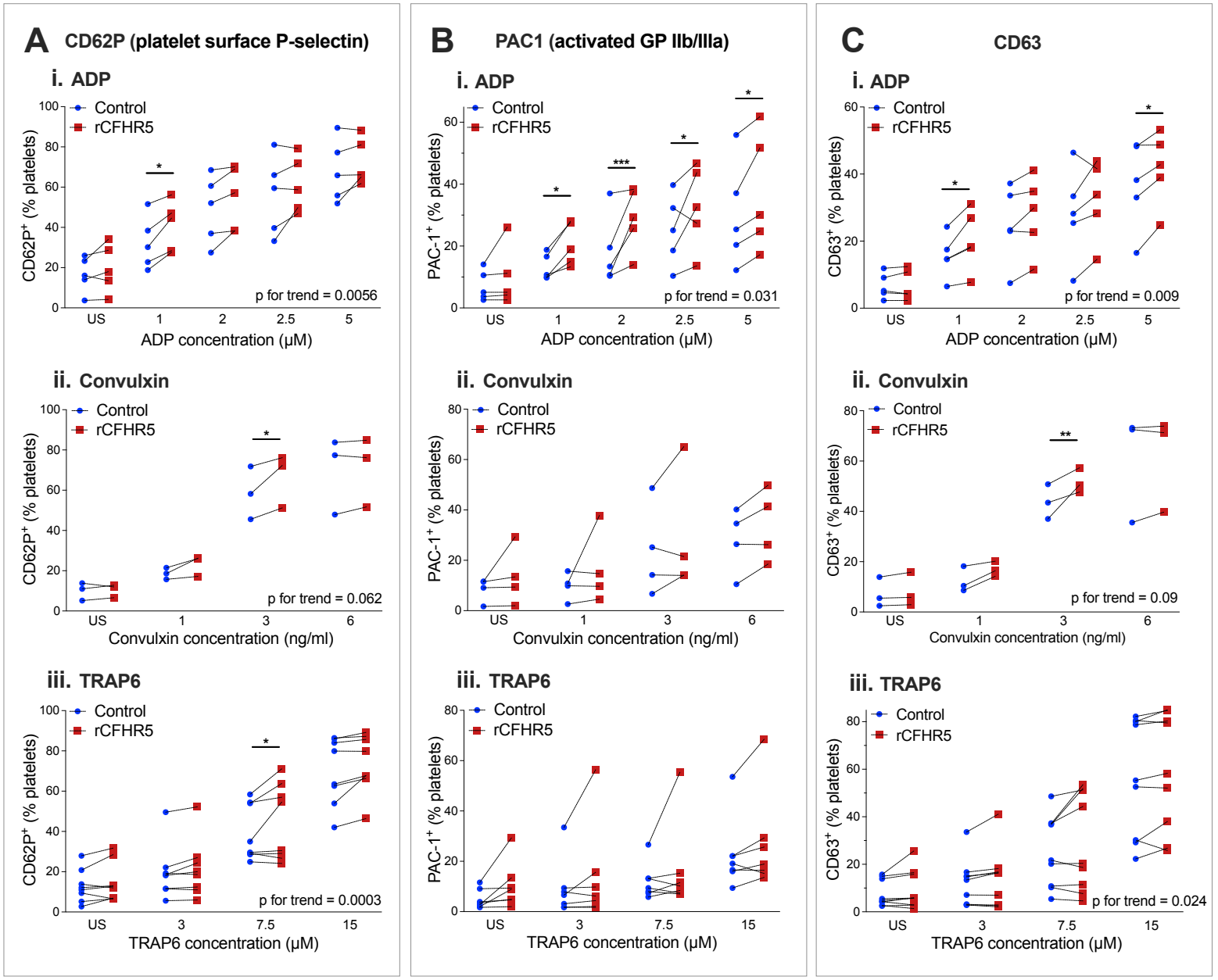

A. VEBIOS ER cohort
Suspected VTE
Citrate
EDTA
VTE excluded VTE confirmed
Controls: n=48 Cases: n=48
B. VEBIOS ER (Citrate)
HPA059937
HPA044659
HPA003042
HPA002655
-Log10(Pval)
Log2 (fold change cases vs. controls)
C. VEBIOS ER (EDTA)
HPA044659
HPA059937
HPA003042
HPA002655
Log2 (fold change cases vs. controls)
D. VEBIOS ER (Citrate)
LogMFI (A.U)
Control Case
E. VEBIOS ER (EDTA)
HPA059937 (target: SULF1)
HPA044659 (target: CD47)
HPA002655 (target: SELP)
HPA003042 (target: ADORA2A)
F. Orthogonal verification: target ID
Precursor Area (Log10)
Z-score
CFHR5
SULF1
HPA059937
Protein digestion
Immunocapture MS
G. Orthogonal verification: dual binder assay
CFHR5 monoclonal
HPA059937
HPA073894 (target CFHR5)
HPA072446 (target CFHR5)
LogMFI (A.U)
HPA059937 (discovery)
Capture HPA059937
Capture HPA073894
Capture HPA072446
Corr. 0.83 p<0.00001
Corr. 0.75 p<0.00001
Corr. 0.82 p<0.00001
H. VEBIOS ER
CFHR5 (ng/ml)
OR (95% CI): 2.54 (1.52-4.66)
p-value: 1.05E-03
Range: 1216-6073
J. VEBIOS coagulation
OR (95% CI): 1.55 (1.20-2.01)
p-value: 8.85E-04
Range: 1378-4570
Controls
Cases
CRP (mg/l)
D-dimer (mg/l)
Corr. 0.52 p=2.39E-04
Corr. 0.57 p=4.58E-5
Corr. 0.22 p=0.16
Corr. -0.01 p=0.94
Corr. 0.24 p= 4.95E-03
Corr. 0.49 p=4.48E-10
Corr. -0.02 p=0.86
Corr. 0.42 p=2.44E-07
CFHR5 (ng/ml)

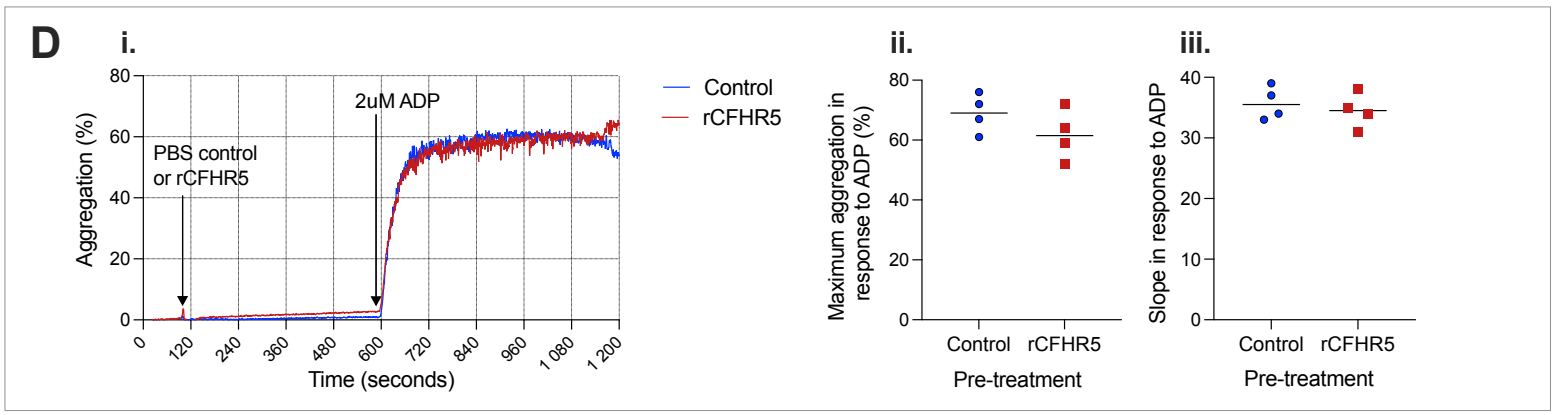

A.
B.
C.
D.
E.
i.
ii.
CFHR5 (ng/ml)
**
***
***
***
**
Control
Case
OR (1SD)
(95%CI)
2.54
(1.52-4.66)
1.80
(1.29-2.58)
1.24
(1.10-1.40)
1.55
(1.2-2.01)
1.29
(1.09-1.53)


A
400
30
20
10
0
−log10(p−value)
CFHR1;CFHR4
HNF1A*
JMJD1C
TRIB1;LINC00861
DNAH10*
HNF4A
Chromosome


B
r2
0.8
0.6
0.4
0.2
350
300
250
200
150
100
50
0
−log10(p−value)
←KCNT2
CFH→
CFHR1→
CFHR5→
←ZBTB41
←DENND1B
←MIR4735
CFHR3→
CFHR2→
←ASPM
CRB1→
CFHR4→
←F13B
196.2
196.4
196.6
196.8
197
197.2
197.4
197.6
Position on chr1 (Mb)


C
****
3000
2000
1000
0
CFHR5 (ng/ml)
Non Carriers
Carriers
Non synonymous variants


A
CD62P (platelet surface P-selectin)
i. ADP
Control
rCFHR5
CD62P+ (% platelets)
p for trend = 0.0056
US
1
2
2.5
5
ADP concentration (μM)
ii. Convulxin
p for trend = 0.062
Convulxin concentration (ng/ml)
iii. TRAP6
p for trend = 0.0003
TRAP6 concentration (μM)
B
PAC1 (activated GP IIb/IIIa)
PAC-1+ (% platelets)
p for trend = 0.031
p for trend = 0.09
p for trend = 0.024
C
CD63
CD63+ (% platelets)
p for trend = 0.009


D
i.
ii.
iii.
Aggregation (%)
PBS control
or rCFHR5
2uM ADP
Time (seconds)
Control
rCFHR5
Maximum aggregation in response to ADP (%)
Pre-treatment
Slope in response to ADP

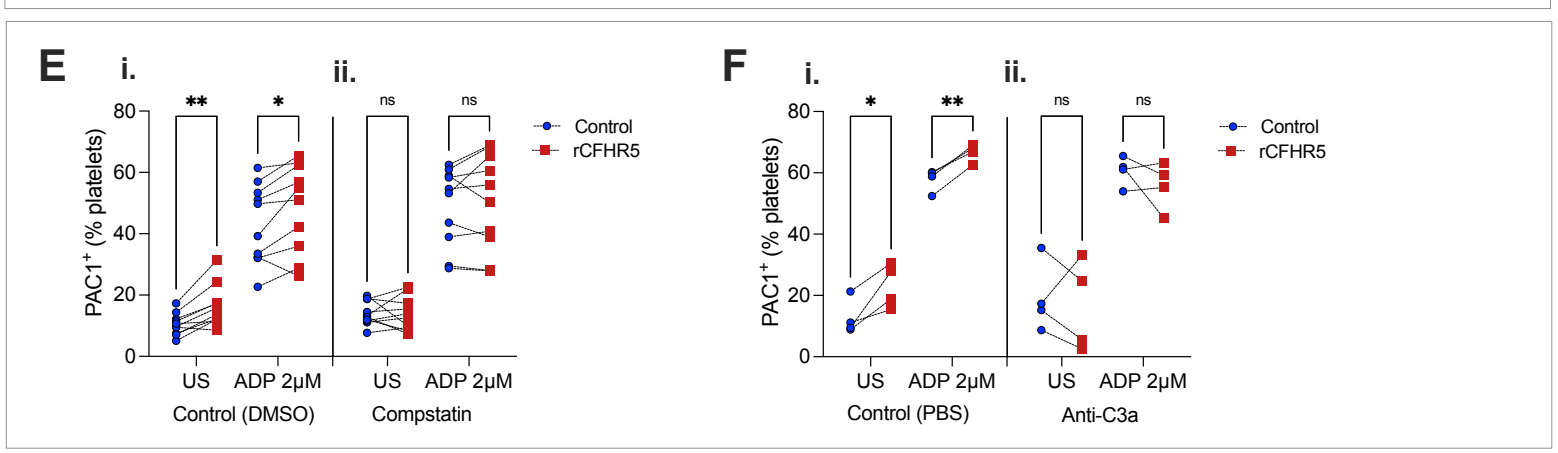

E
i.
ii.
PAC1+ (% platelets)
US
ADP 2μM
Control (DMSO)
Compstatin
ns
F
Control (PBS)
Anti-C3a

Table 1. Characteristics of VEBIOS ER case control discovery study

Variables	Cases (n=48)		Controls (n=48)		P value
Thrombosis localisation	n	%	n	%	
DVT, lower limbs	20	41.6	-	-	
- Proximal	17	85	-	-	
PE	21	43.7	-	-	
DVT and PE	7	14.5	-	-	
Sex and biometry					
Sex; women	22	45.8	22	45.8	1
Age (years) (mean ± SD and range)	56.6 ± 19	[19-89]	56.6 ± 17	[23-88]	0.982
BMI (kg/m2) (mean ± SD, range)	26.8 ± 5.89	[19.4-44.5]	28.9 ± 6.66	[19.8-44.9]	0.221
- Missing	7	14.6	26	54.2	
Obese (BMI ≥ 30 kg/m2) (n, %)	6	12.5	7	14.6	0.782
Current smoking ‡					
- No	38	79.2	37	77.1	0.309
- Yes	4	8.3	9	18.8	
- Missing	6	12.5	2	4.2	
Family history					
VTE, First degree relative < 60 years old					
- No	28	58.3	41	85.4	0.0388
- Yes	14	29.2	6	12.5	
- Missing	6	12.5	1	2.1	
Provoked risk factors (all)†	25	52.1	21	43.8	0.555
Estrogen containing contraceptives and hormone replacement therapy*	5	10.4	5	10.4	1
Concentration of markers measured					
CFHR5 (ng/mL)					
- Mean ± SD	3430 ± 782		2840 ± 756		<0.001
- Median [Min,Max]	3360 [1990-6070]		2680 [1220-5560]		
C3 (µg/mL)					
- Mean ± SD	668 ± 217		660 ± 206		0.865
- Median [Min,Max]	653 [63.7-1210]		656 [292-1190]		
- Missing values (n, %)	1 (2.1%)		4 (8.3%)		
D-dimer (ng/mL)					
- Mean ± SD	4620 ± 4630		880 ± 926		<0.001
- Median [Min,Max]	3600 [414, 26200]		521 [92.3, 4410]		
- Missing values (n, %)	2 (4.2%)		6 (12.5%)		
CRP (mg/L)					
- Mean ± SD	41.2 ± 55.2		23.0 ± 52.4		0.112
- Median [Min,Max]	18.0 [1.00, 246]		4.50 [1.00, 304]		
- Missing values (n, %)	4 (8.3%)		2 (4.2%)		
LPK ($x10^9$/L)					
- Mean ± SD	8.68 ± 2.25		9.10 ± 5.21		0.62
- Median [Min,Max]	9.10 [3.40, 12.4]		7.80 [4.10, 36.9]		
- Missing values (n, %)	5 (10.4%)		3 (6.3%)		
Hb (g/L)					
- Mean ± SD	134 ± 17.9		138 ± 15.2		0.349
- Median [Min,Max]	135 [53.0, 179]		138 [96.0, 173]		
- Missing values (n, %)	5 (10.4%)		3 (6.3%)		
TPK ($x10^9$/L)					
- Mean ± SD	221 ± 74.6		253 ± 91.4		0.073
- Median [Min,Max]	206 [108, 538]		240 [29.0, 530]		
- Missing values (n, %)	5 (10.4%)		3 (6.3%)		

DVT, deep vein thrombosis; **PE**, pulmonary embolism; **n**, numbers; **SD**, standard deviation; †, within one month from diagnosis or index date (immobilization with trauma, surgery, cast and/or orthosis and bedrest more than 3 days of sickness); **‡**, within the last year; *,on-going treatment; **CFHR5**, Complement Factor H-related protein 5; **C3**, Complement 3; **CRP**, C-reactive protein; **LPK**, leucocytes; **Hb**, hemoglobin; **TPK**, thrombocytes. P value was obtained by Student t.test and Pearson's Chi-squared Test for numerical and categorical variables, repectively.

Table 2

		VEBIOS ER	DFW-VTE	FARIVE	VEBIOS Coagulation	RETROVE	Meta-analysis
VTE ALL	OR (1SD)	2.54	1.80	1.24	1.55	1.29	1.35
	(95%CI)	(1.52-4.66)	(1.29-2.58)	(1.10-1.40)	(1.2-2.01)	(1.09-1.53)	(1.23-1.47)
	P-value	1.05E-03	7.65E-04	3.98E-04	8.85E-04	2.4E-03	1.94E-11
	Cases	48	54	582	142	308	1134
	Controls	48	146	576	135	360	1265
CFHR5 range (ng/mL)		1216-6073	232-1170	450-4904	1378-4570	364-2341	232 – 6073
Tertile 1	OR (1SD) (95%CI)	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	P-value	NA	NA	NA	NA	NA	
	Cases	10	9	165	39	91	314
	Controls	22	57	221	54	132	486
CFHR5 range (ng/mL)		1216-2653	232-476	450-1315	1378-2333	364-858	232-2653
Tertile 2	OR (1SD)	1.64	2.72	1.40	1.41	1.03	1.3
	(95%CI)	(0.58-4.80)	(1.15-6.86)	(1.05-1.86)	(0.78-2.55)	(0.69-1.52)	(1.10-1.66)
	P-value	0.36	0.03	0.02	0.25	0.88	4.38E-03
	Cases	13	21	198	45	94	371
	Controls	19	46	188	47	128	428
CFHR5 range (ng/mL)		2675-3338	478-601	1317-1716	2335-2779	859-1063	478-3338
Tertile 3	OR (1SD)	9.05	2.93	1.75	2.51	1.67	1.97
	(95%CI)	(2.93-31.61)	(1.24-7.37)	(1.31-2.33)	(1.39-4.63)	(1.13-2.47)	(1.60-2.42)
	P-value	2.51E-04	0.02	1.28E-04	2.68E-03	9.90E-03	1.43E-10
	Cases	25	24	219	58	123	449
	Controls	7	43	167	34	100	351
CFHR5 range (ng/mL)		3345-6073	603-1170	1717-4904	2782-4569	1064-2341	603-6073

Tertile 1= reference value 1; OR = odds ratio; 1SD= 1 standard deviation; CI= confidence interval. Data adjusted for age and sex.

Study design and study protocols

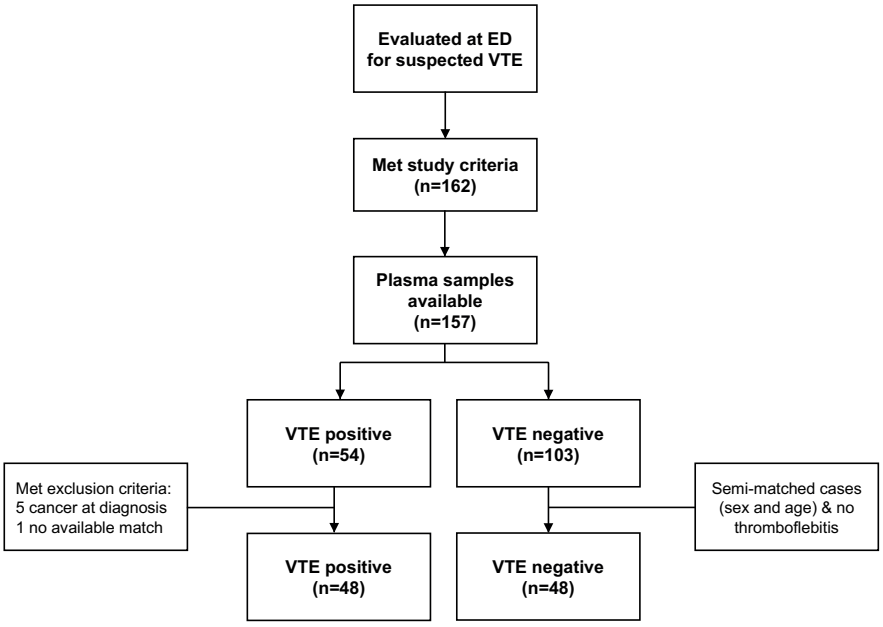
Overall design of proteomics discovery study

We performed an initial antibody based targeted plasma proteomics screening in a nested case/control study based on the prospective **VEBIOS ER cohort study** of patients with suspected acute VTE recruited in an emergency ward. As our aim was to identify biomarkers associated with acute VTE that were potentially linked to the underlying disease risk, rather than secondary to the acute thrombosis, the results were examined for overlap with published results from an identical proteomics screen in a case/control study, the **VEBIOS Coagulation study**, with patients recruited in an outpatient clinic at follow up after ending treatment for a prior VTE. The study protocols for the two studies are described below.

The **VEBIOS ER** *(Venous thromboEmbolism BIOmarker Study in Emergency Room)* study is a prospective cohort study of patients admitted to a hospital with suspected acute VTE. Patients were recruited at the Emergency Room at Karolinska University Hospital Stockholm between December 2010 and September 2013 Sweden. Inclusion was based on an initial suspicion of acute VTE on behalf of the attending physician. All patients admitted with the suspicion of deep vein thrombosis (DVT) in the lower limbs and/or pulmonary embolism (PE), over 18 years old were eligible for the study. Exclusion criteria were patients with on-going anticoagulant treatment, pregnancy, active cancer, short life expectancy or lack of capacity to leave approved consent. A case was defined if a) VTE was confirmed by diagnostic imaging - compression venous ultrasonography (CUS) in patients with suspected DVT in the lower limbs, and computed tomography pulmonary angiography (CTPA) in patients with suspected PE, and b) anticoagulant treatment was initiated based on the VTE diagnosis. Patients with no evidence of an acute VTE, (neither by diagnostic imaging nor by Well's clinical criteria) that had a normal d-dimer test were referred as controls in the study. All participants were sampled for biobanking before any anticoagulant treatment. Whole blood was collected in citrate or EDTA anticoagulant at the ER and sent within 30 minutes to the Karolinska University Laboratory. After centrifugation at 2000*g* for 15 minutes, plasma aliquots were snap frozen and stored at -80°C. *Data collection*: For

each patient, doctors filled in a questionnaire detailing (1) any provoking factors within one month preceding the visit to the ER (2) current health situation, alcohol consumption and smoking habits; (3) family history of VTE (4) ongoing antithrombotic (antiplatelet) treatment and (5) estrogen containing contraceptives and hormone replacement therapy (women only). Medical information from the ER visit (e.g. results of diagnostic workup and imaging procedures, medication, patient weight and height) along with results from routine laboratory tests (e.g. D-dimer, CRP, blood cell count) were extracted from the medical records. Informed written consent was obtained from all participants in accordance with the Declaration of Helsinki. VEBIOS was approved by the regional research ethics committee in Stockholm, Sweden (KI 2010/636-31/4). In total, 158 patients were included (52 cases) between December 2010 and September 2013. For the present study, 48 cases were available for analysis and 48 controls were matched as closely as possible by age and sex. Details of the discovery set is given in Table 1 in the manuscript. A flow diagram of the discovery cohort is shown below.

VEBIOS ER COHORT

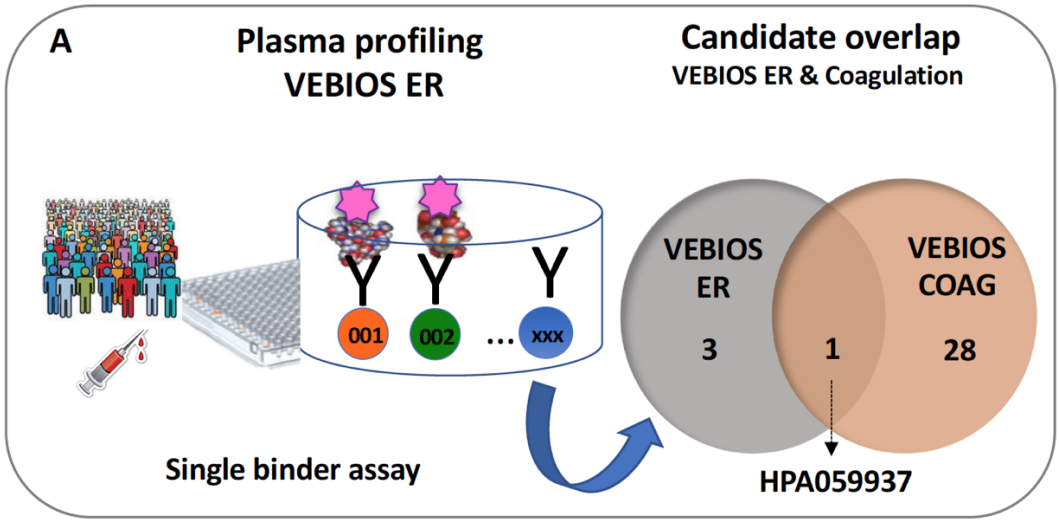


The **VEBIOS Coagulation** (Venous thromboEmbolism BIOmarker Study in Coagulation unit) is an on-going case-control study, initiated in January 2011, with recruitment at the out-patient Coagulation clinic at Karolinska University Hospital Solna, Sweden. The study protocol has been described previously by Bruzelius, M., *et al.* (Blood, 2016. 128(23): p. 59-67.) Eligible cases were between 18-70 years of age at the time of first VTE, confirmed by diagnostic imaging; venous ultrasonography in patients with DVT of the lower limbs, and computed tomography pulmonary angiography (CTPA) or ventilation perfusion scintigraphy (V/Q lung scan) in patients with pulmonary embolism (PE). Cases were recruited from three regional hospitals after referral from the emergency clinics following diagnosis. Cases had been treated with anticoagulants (vitamin K antagonist, direct oral anticoagulant or low molecular weight heparin) for 6-12 months and were sampled 1-6 months after discontinuation of treatment. Patients with severe thrombophilia, i.e. anti-thrombin, protein S and protein C deficiencies, antiphospholipid syndrome, homozygosity for either factor V Leiden (G1691A) or the G20210A polymorphism in prothrombin gene, or a combined heterozygosity, were excluded. Controls were randomly recruited from the general population of Stockholm County, using the Swedish Tax Agency register and matched to cases for age (± 2 years) and sex, and sampled within one year of the index date (when blood sampling took place) of the matched case. Exclusion criteria were history of VTE, pregnancy in the 3 months prior to index date or active cancer within the last 5 years. *Data collection:* Participants completed a questionnaire regarding (a) demography, (b) provoking factors within three months preceding VTE diagnosis, or the time of sampling for the controls, (c) co-existing cardiovascular risk factors and chronic co-morbidities, (d) current health situation; alcohol consumption, smoking habits, physical activity (e) family history of VTE and other CVD (f) ongoing medication (g) reproductive history and hormonal use (women only). *Blood sampling*: All participants were sampled in the morning after an overnight fast at the Coagulation Unit, Karolinska Hospital. Whole blood was collected in citrate or EDTA anticoagulant and centrifuged at 2000 g for 15 minutes at room temperature. Plasma aliquots were snap frozen and stored at −80°C. Blood and plasma samples were sent to the Karolinska University Laboratory for measurement of routine laboratory tests (CRP, D-dimer, blood glucose, creatinine, albumin, blood count, lipid profile and liver enzymes). *Biometry*: On the index date participant weight, height, blood pressure and pulse in left arm in sitting position and waist and hip circumference were measured.

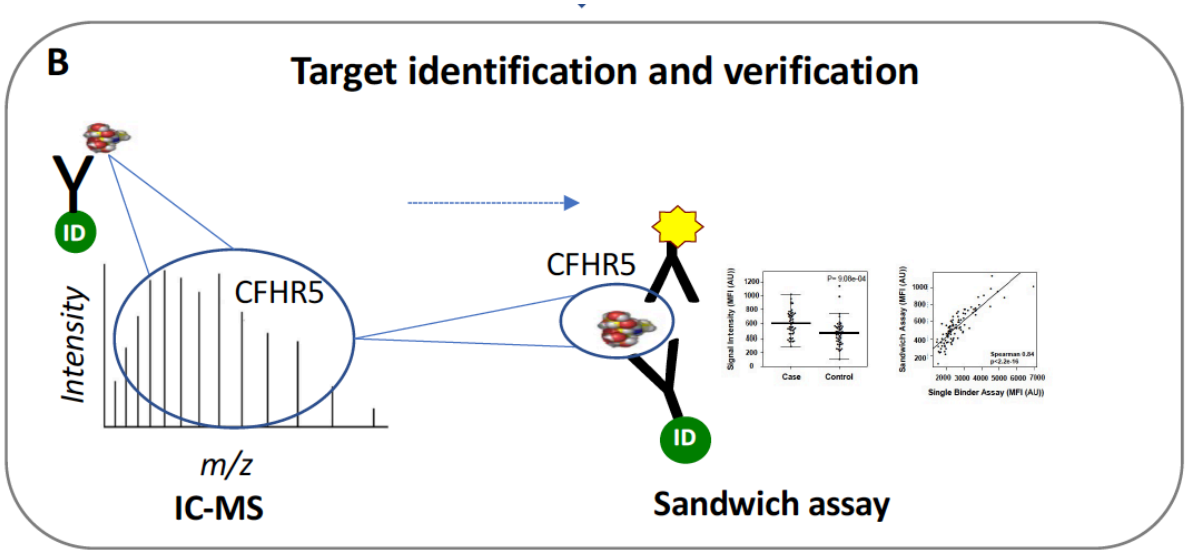
Informed written consent was obtained from all participants in accordance with the Declaration of Helsinki. VEBIOS was approved by the regional research ethics committee in Stockholm, Sweden (KI 2010/636-31/4). In the current study, the results from our previously published proteomics screening of 96 cases and 96 matched controls was used in the candidate biomarker selection stage. VEBIOS Coagulation is an ongoing study and in the current work, CFHR5 was quantified in an extended sample set of *VEBIOS Coagulation* comprising 144 cases and 140 controls available in 2017. Details of the extended sample set is given in Table S2_Tab 1.

Candidate biomarker identification and verification

A discovery affinity proteomics antibody screening array was designed to target a panel of 408 candidate proteins selected based on known or hypothesized involvement in pathways or intermediate traits of relevance for thrombosis, haemostasis and/or cardiovascular disease, and/or with endothelial enriched expression. This panel was used to screen citrate samples from VEBIOS ER followed by a technical replication in EDTA plasma sampled in parallel. Antibodies yielding significant signals of association with diagnosis of acute VTE in VEBIOS ER was examined for overlap with our previously published results generated with an identical proteomics panel in 96 cases and 96 controls in VEBIOS Coagulation study, see figure A below.



The one overlapping candidate antibody (HPA059937) was selected for target identification by ImmunoCapture Mass Spectrometry (IC-MS) followed by orthogonal verification with dual binder assays, identifying CFHR5 as the VTE-associated target protein of HPA059937, see figure B below.



Biomarker validation and replication in 5 independent studies.

A dual binder assay for quantification of CFHR5 concentration in plasma using recombinant CFHR5 protein as standard was developed. The assay was used to measure absolute CFHR5 concentrations in 5 independent cohorts and case/control studies, where patients had been sampled either at time of the acute event (**VEBIOS ER, DFW-VTE**), within days of diagnosis (**FARIVE**), or after ending anticoagulant treatment for a previous VTE (**VEBIOS Coagulation, RETROVE**). Study design and protocols for the DFV-VTE, FARIVE, and RETROVE studies are described below.

The DFW-VTE study (*D-dimer, Fibrin monomer and Wells score in VTE study*) is a prospective single-center study and has previously been described by Farm *et al* (J Thromb Haemost, 2018. 16(5): p. 866-875(11)). All patients were included at admittance to the ER of Karolinska University Hospital in Huddinge, Sweden, between April 2014 and May 2015. DVT was confirmed by either doppler ultrasonography or CUS whereas PE was confirmed by either CTPA or ventilation/perfusion lung scintigraphy. A total of 954 patients (125 VTE positive) were included. All relevant data

was extracted from electronic medical records (EMR). Blood was collected in 0.109 mol/L (3.2%) sodiumcitrate vacutainer tubes by direct venipuncture. After centrifugation at 3000 × *g* for 10 minutes the samples were frozen at -80° C within one hour. The samples were thawed once in 37° C water-bath and re-frozen at -80° C within 20 minutes after aliquotation. The study protocol was approved by the regional Ethical Review Agency in Stockholm, Sweden (Dnr 2013-2143-31-2). Informed written consent was obtained from all participants in accordance with the Declaration of Helsinki. For the current study, CFHR5 was measured in stored plasma samples for a subset of 200 patients (54 VTE). Clinical characteristics are described in Table S2, Tab_2.

FARIVE is a French multicentre case-control study carried out between 2003-2009 that recruited consecutive inpatients or outpatients treated for a first episode of DVT and/or PE, confirmed by diagnostic imaging, from 18 years of age. The study protocol has been previously described by Trégouët, D.A., et al. (Blood, 2009. **113**(21): p. 5298-5303). Controls were age- and sex matched and consisted of in- and outpatients, free of history of venous or arterial thrombotic disease. Exclusion criteria were a cancer diagnosis, short life expectancy owing to other causes and renal or liver failure. The study was approved by the Paris Broussais-HEGP ethics committee in Paris (2002-034) and all participants gave informed written consent, in accordance with the Declaration of Helsinki. Blood sampling was carried out at the respective centers. Blood in citrate anticoagulant was centrifuged at 2000 g for 20 minutes at room temperature and plasma aliquots were collected and snap frozen and stored at −80°C. Whole blood and plasma were collected to the biobank centralied at Hôpital Européen Georges Pompidou (HEGP) in Paris. A total of 587 cases were sampled at the study entry, within the first week after diagnosis and anticoagulant treatment initiation. For the current study, CFHR5 was measured in plasma samples from a subset of 582 cases and 576 controls with existing genotype data. Clinical characteristics are described in Table S2, Tab_3.

The **RETROVE** (*Riesgo de Enfermedad TROmboembólica VEnosa*) study is a prospective case–control study of 400 consecutive patients with VTE (cancer associated thrombosis excluded) and 400 healthy control volunteers. Individuals were

recruited at the Hospital de la Santa Creu i Sant Pau of Barcelona (Spain) between 2012 and 2016. Controls were selected according to the age and sex distribution of the Spanish population (2001 census). All individuals were ≥ 18 years. Blood samples from the patients were taken at least 6 months after thrombosis in order to minimize the influence of the acute phase. None of the participants was using oral anticoagulants, heparin or antiplatelet therapy at the time of blood collection. All procedures were approved by the Institutional Review Board of the Hospital de la Santa Creu i Sant Pau, and all participants gave informed written consent, in accordance with the Declaration of Helsinki. In the current study, CFHR5 was measured in aliqouted plasma samples from 308 cases and 360 controls. Clinical characteristics are described in Table S2, Tab_4.

Test for association with risk of VTE recurrence

The developed assay was further used for quantification of plasma concentration of CFHR5 in samples from a cohort of patients in the **MARTHA study** (described below) that were followed up for up to 12 years after previous VTE.

The *MARTHA* (*Marseille Thrombosis Association*) study is a population based single centre study. Recruitment in *MARTHA* started in 1994 at Timone Hospital in Marseille (France) and is still ongoing. The cohort from 1994 and 2008, includes a total of 1542 VTE-cases (66% women) that donated blood for further analysis. Details of study protocol is previously described by Trégouët, D.A., et al (Blood, 2009. 113(21): p. 5298-5303). The cases were unrelated consecutively recruited whites with a documented history of VTE and without strong known risk factors, including AT, PC, or PS deficiency, homozygosity for FV Leiden or FII 20210A, and lupus anticoagulant. In all cases, history of VTE was recorded before the laboratory diagnosis was made. The thrombotic events were documented by venography, Doppler ultrasound, spiral computed tomographic scanning angiography, and/or ventilation/perfusion lung scan. Study subjects were interviewed by a physician about their medical history, which emphasized manifestations of DVT and PE using a standardized questionnaire. The date of occurrence of every episode of VTE and the presence of precipitating factors (such as surgery, trauma, prolonged immobilization, pregnancy or puerperium, and

oral contraceptive intake) were collected. VTE was classified as secondary when occurring within 3 months after exposure to exogenous risk factors, including surgery, trauma, immobilization for 7 days or more, oral contraceptive use, pregnancy, and puerperium. In the absence of these risk factors, VTE was defined as primary. Ethical approval was granted from the Department of Health and Science, France (2008-880 & 09.576) and all participants gave informed written consent, in accordance with the Declaration of Helsinki. CFHR5 concentration was measured in 1,266 of the MARTHA cases. For 669 of these cases, follow up data up to 12 years post-event was available and used to analyse risk of recurrent VTE. For 774 of the cases, thrombinoscope data was available and used to test association of CFHR5 concentration and Thrombin Generation Potential (TGP). Clinical characteristics of these two datasets are described in Table S2, Tab 5

Figure S1

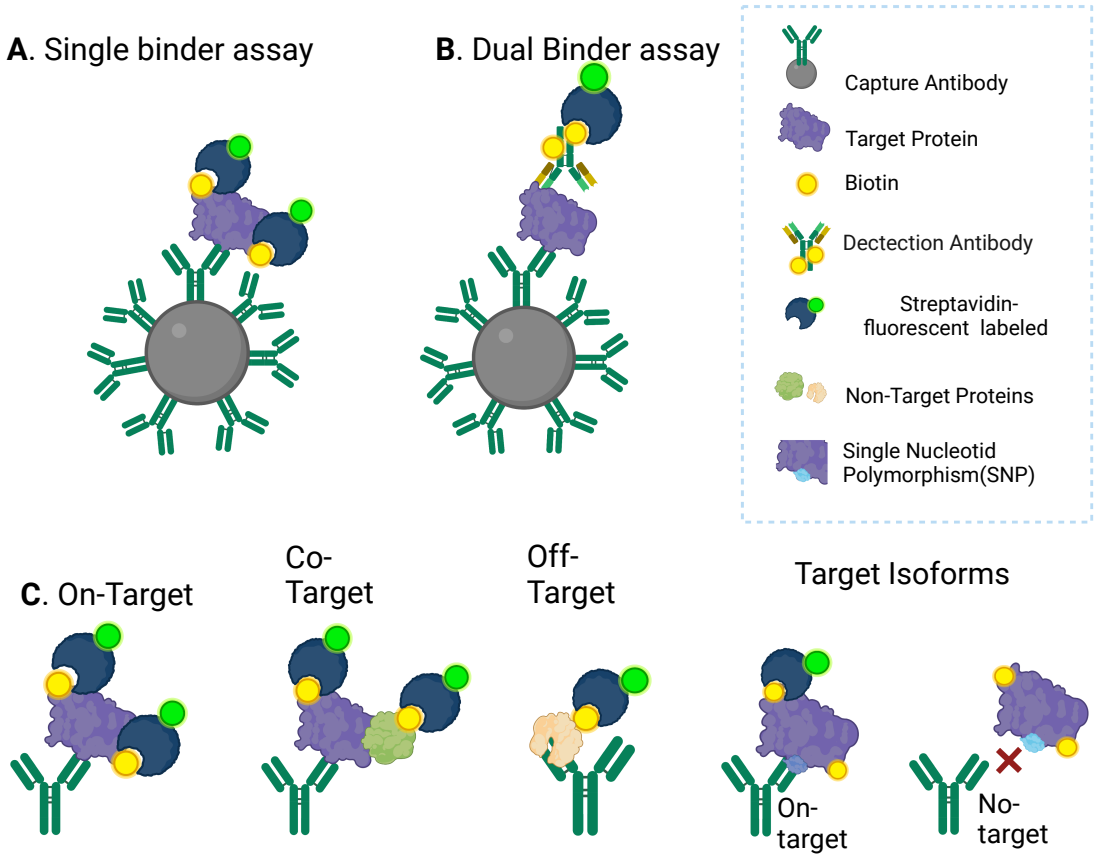


Figure S1: Antibody based suspension bead assay concepts. The antibody raised towards intended protein target is covalently coupled to colour-coded, micrometer-sized beads. (**A**) In the single binder assay, all proteins within the plasma sample are labelled with biotin before incubation with beads. Biotin-labelled proteins bound to the capture antibody are detected by fluorescent-labelled-streptavidin. The suspension is analysed by a cytometry-based instrument (Luminex), where the colour of the antibody-coupled bead provides the antibody ID, and the mean fluorescence intensity provides a relative measure of the bound target protein corresponding to plasma levels. (**B**) In the dual binder assay captured proteins are unlabelled, and detection of target protein bound to the capture antibody is through a secondary target-specific biotin-labelled detection antibody, followed by fluorescent-labelled-streptavidin addition and analysis on a Luminex instrument. As detection in the single binder assay is based on detection of biotin bound directly to target proteins, the signal can reflect either (**C**) the intended on-target binding, (**D**) co-target binding where the target protein is

complexed with non-target protein or **(E)** off-target binding, i.e., binding of a non-target protein**.** (**F**) An amino-acid substitution in the epitope of the target protein can lead to iso-form specific binding.

Figure S2

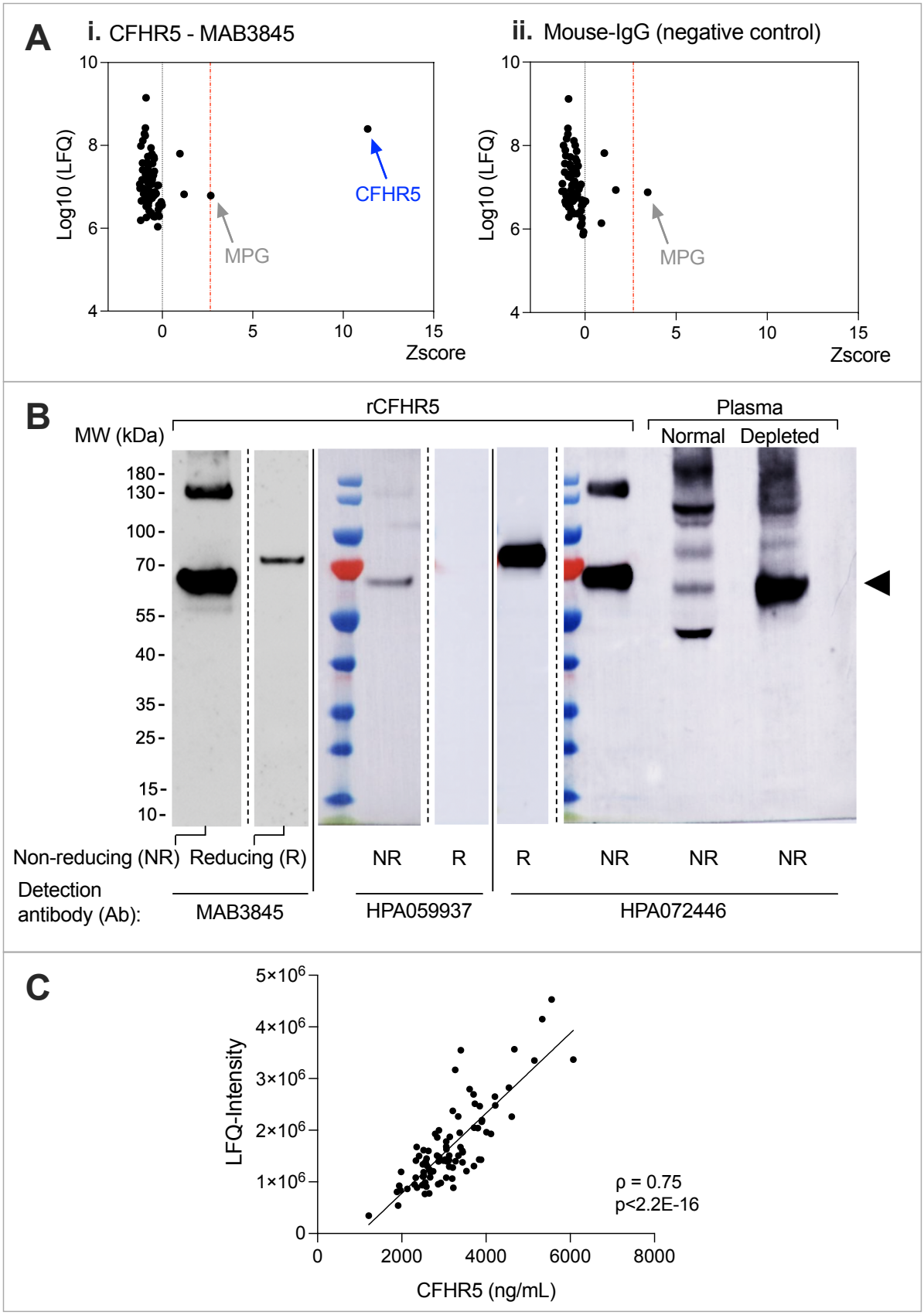

A
i. CFHR5 - MAB3845
ii. Mouse-IgG (negative control)
Log10 (LFQ)
Zscore
CFHR5
MPG
B
rCFHR5
Plasma
Normal
Depleted
MW (kDa)
180
130
100
70
55
40
35
25
15
10
Non-reducing (NR)
Reducing (R)
Detection antibody (Ab):
MAB3845
HPA059937
HPA072446
C
LFQ-Intensity
CFHR5 (ng/mL)
ρ = 0.75
p<2.2E-16

Figure S2: Verification of antibody target and CFHR5 dual binder quantitative assay

(A) Immunocapture mass spectrometry analysis of monoclonal anti-CFHR5 detection antibody (MAB3845) and negative control (Mouse IgG). **(B)** Immunodetection of CFHR5 by Western blot. Recombinant CFHR5 (rCFHR5, 100 ng), normal plasma (NP, 1 µl) and 14 most abundant proteins depleted plasma (DP, 10 µl) loaded on SDS PAGE in non-reducing (without DTT, **NR**) or reducing conditions (with DTT, **R**). After electrophoresis and transfer on PVDF membrane, protein was detected using antibodies against SULF1 (HPA059937) and CFHR5 (HPA072446 and MAB3845). The arrow shows the band corresponding to CFHR5 protein. **(C)** Spearman´s correlation analysis between quantitative dual binding assay and label-free quantitative data-independent acquisition mass spectrometry data (LFQ-DIA-MS) in *VEBIOS ER*.

Figure S3

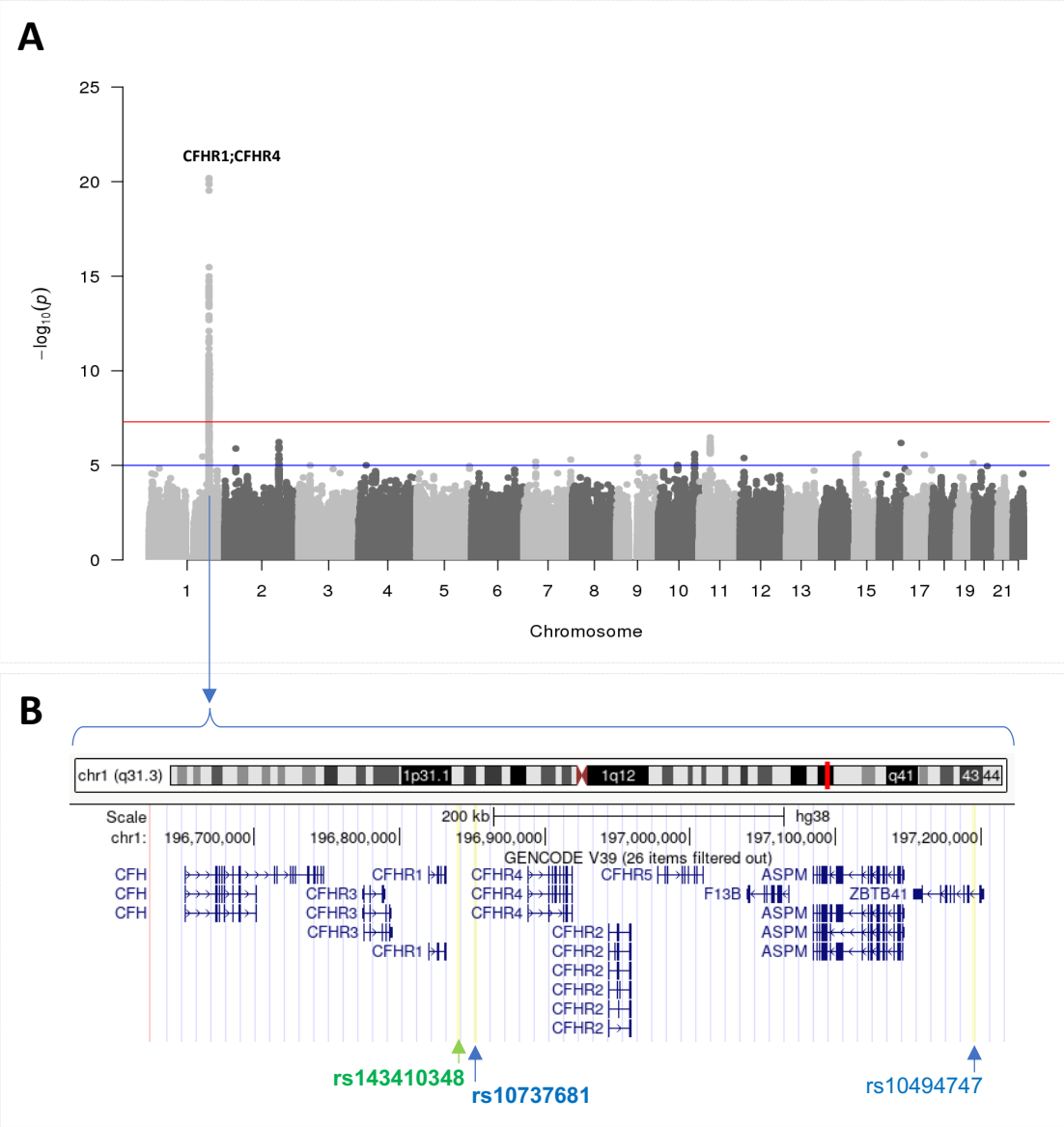


Figure S3: GWAS analysis identifies a CFHR5 pQTL on Chromosome 1 q31.3.A meta-analysis of GWAS data for 7,135,343 SNPs tested for association with CFHR5 concentrations in a total sample of 2967 individuals from the *FARIVE*, *MARTHA* and *RETROVE* studies: **(A)** identified one genome-wide significant (p<5E-08; red line) signal on chr1q31.3. **(B)** The lead SNP at this locus, rs10737681, maps between the *CFHR1* and *CFHR4* gene loci in the gene cluster of *CFHR1-5*. A borderline significant association (p=9.83E-08) with CFHR5 levels was observed at the rs10494747, mapping to the *ZBTB41* gene. Indicated in green is the rs143410348 recently identified with genome wide significance as associated with VTE risk [17].

Figure S4

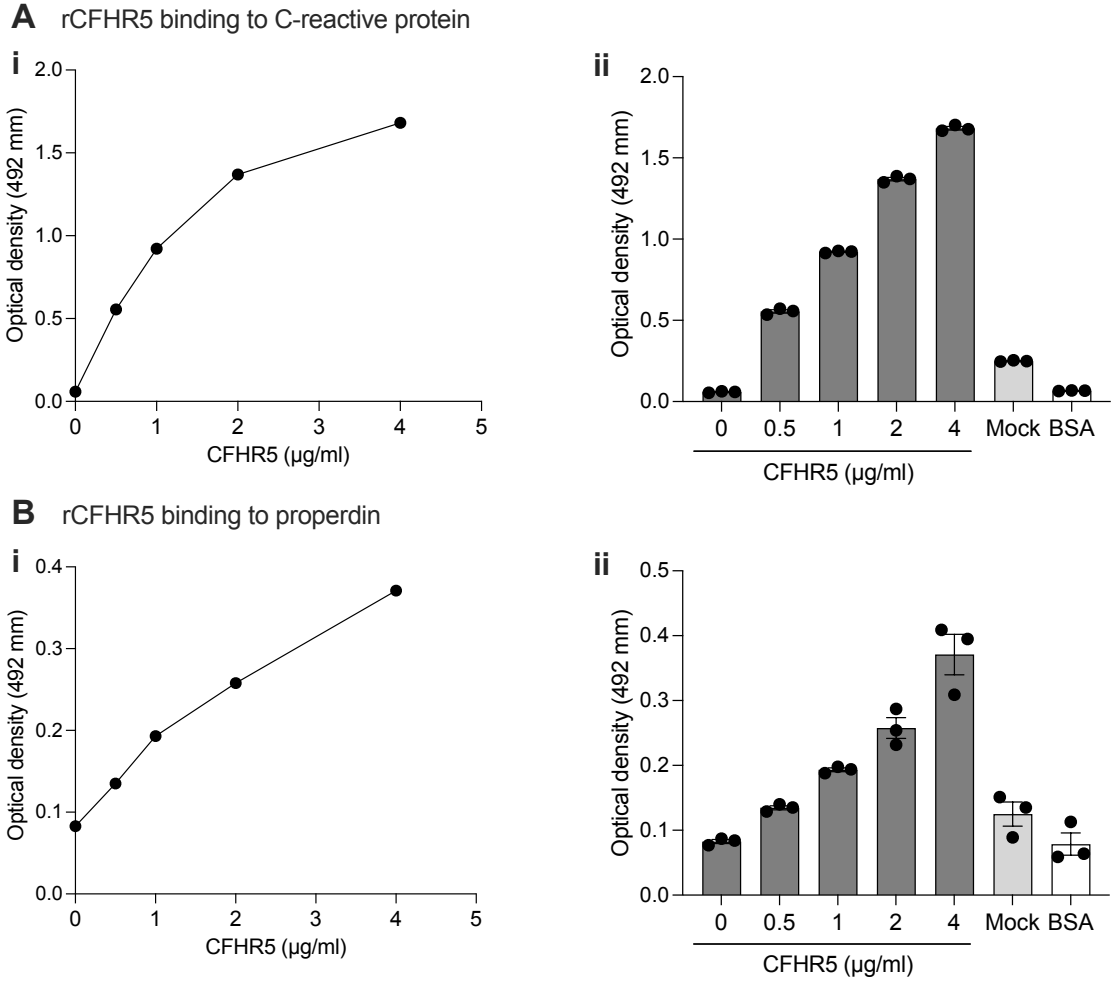


Figure S4. Recombinant CFHR5 (rCFHR5) binding to C-reactive protein (CRP) and Properdin. (**A**) CRP or (**B**) properdin, were immobilized on microtiter plates, before serial dilutions of rCFHR5 were added and binding assayed by ELISA. Data is displayed as mean signal across serial dilution or (ii) measurements for individual replicates and controls (mock; heat-denaturation rCFHR5, BSA; bovine serum albumin).

Figure S5

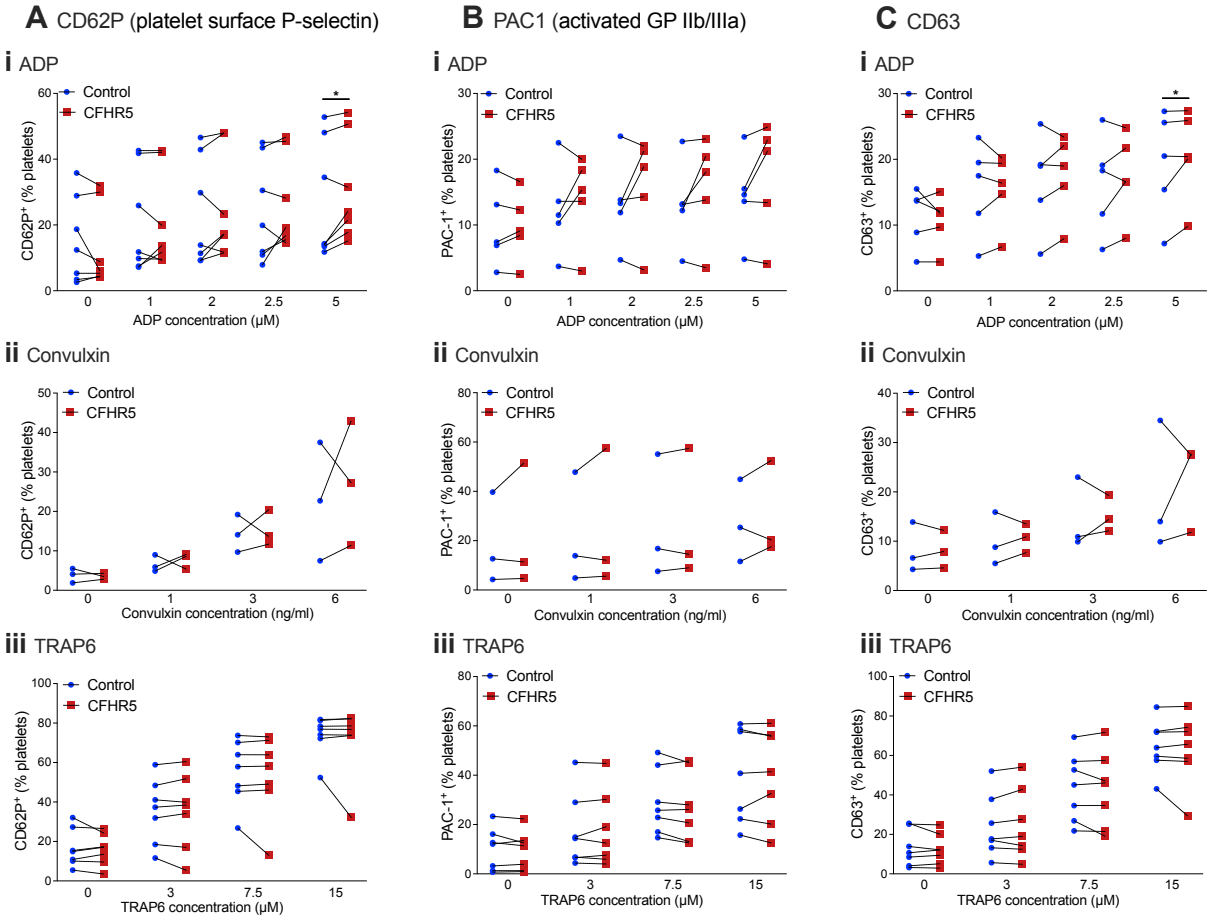


Figure S5: Recombinant CFHR5 does not potentiate platelet activation of washed platelets. Platelet activation was measured by surface expression of (A) P-selectin, (B) activated GP IIb/IIIa ($PAC1^+$) or (C) CD63, following treatment of washed platelets with different concentrations of: (i) adenosine diphosphate (ADP) (ii) convulxin or (iii) TRAP6, following preincubation (10 minutes) with 6 μg/ml recombinant CFHR5, or PBS control. Each experiment is represented by an individual point and paired experiments connected by a dotted line.

Figure S6

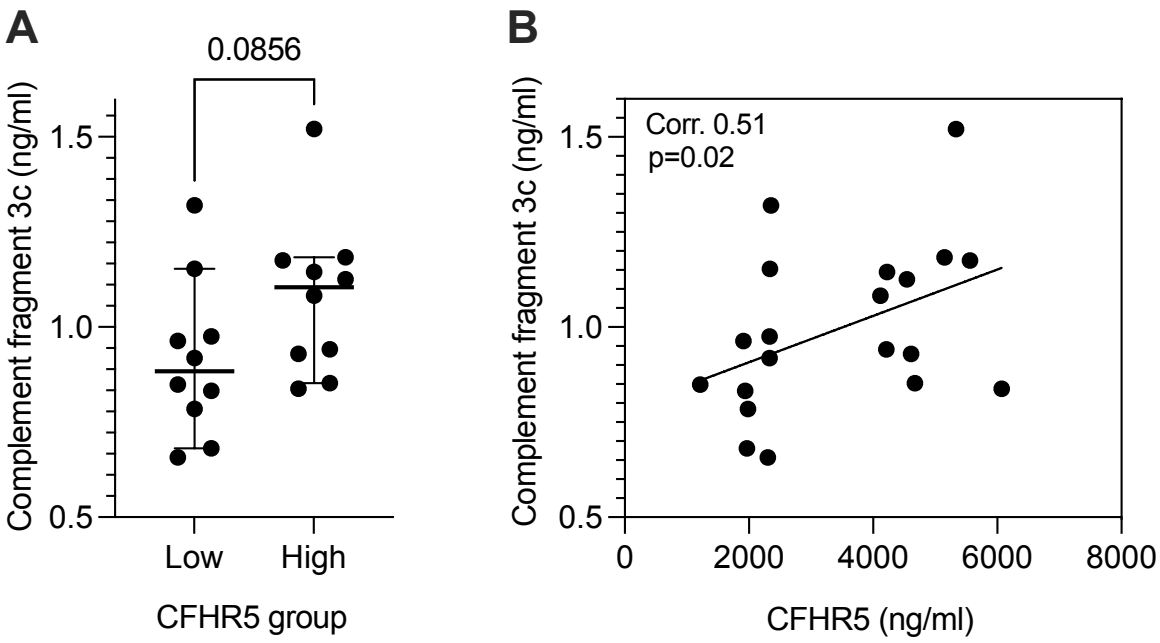


Figure S6. Relationship between CFHR5 and C3c levels in *VEBIOS ER* sample sub-set

Concentration of complement fragment 3c (C3c) was measured by ELISA in 20 VEBIOS ER cohort samples, which were selected based on high or low plasma concentration of CFHR5 (low CFHR5 group [<2500 ng/ml, n=10]; high CFHR5 group [>3800 ng/ml, n=10]). Data is presented by **(A)** CFHR5 group and **(B)** correlation between CFHR5 and C3c across all analysed samples.

Tab_9: Relative risk of VTE associated with CFHR5 by thrombosis type,sex, and causality catergory.

Table A: Relative risk of VTE associated with CFHR5 by thrombosis type

		VEBIOS ER	DFW-VTE	FARIVE	VEBIOS Coag.	RETROVE	Meta-analysis
DVT only	OR (95%CI)	1.81 (1.05-3.41)	1.61 (1.10-2.40)	1.25 (1.05-1.47)	1.54 (1.13-2.12)	1.31 (1.09-1.60)	1.35 (1.21-1.51)
	P-value	4.15E-02	1.50E-02	8.94E-03	6.78E-03	4,76E-03	6,74E-08
	Cases	20	39	166	66	173	464
	Controls	48	146	576	135	360	1265
	CFHR5 range (ng/mL)	1216-6073	232-1048	450-4504	1378-4570	364-2301	232-6073
PE and PE+DVT	OR (95%CI)	2.83 (1.59-5.61)	2.38 (1.42-4.20)	1.23 (1.08-1.40)	1.49 (1.11-2.04)	1.26 (1.02-1.56)	1.33 (1.20-1.47)
	P-value	1.07E-03	1.42E-03	1,98E-03	9.26E-03	3,49E-02	2,53E-08
	Cases	28	15	415	76	135	669
	Controls	48	146	576	135	360	746
	CFHR5 range (ng/mL)	1216-5560	232-1170	450-4904	1378-4570	364-2341	232-5560

OR:Odds ratio per 1 SD increase in CFHR5 concentration, adjusted for differences in age and sex, separated by thrombosis type

Table B. Relative risk of VTE associated with CFHR5 by sex

		VEBIOS ER	DFW-VTE	FARIVE	VEBIOS Coag.	RETROVE	Meta-analysis
VTE_Female only	OR (95%CI)	2.14 [1.13-4.92]	2.23 [1.36-3.93]	1.15 [0.98-1.36]	1.15 [0.79-1.68]	1.38 [1.08-1.78]	1.28 (1.13-1.44)
	P-value	0,04	2,71E-03	0,08	0,48	9,23E-03	6,57E-05
	Cases	22	23	348	56	142	591
	Controls	22	90	330	53	183	678
	CFHR5 range (ng/mL)	1216-6073	232-1075	450-4504	1388-4570	364-2341	232-6073
VTE_Male only	OR (95%CI)	3.09 [1.45-7.89]	1.51 [0.98-2.43]	1.43 [1.19-1.73]	2.00 [1.40-2.95]	1.23 [0.98-1.54]	1.46 (1.29-1.65)
	P-value	8,27E-03	0,072	1,75E-04	2,22E-04	7,26E-02	2,87E-09
	Cases	26	31	234	86	166	543
	Controls	26	56	246	82	177	587
	CFHR5 range (ng/mL)	1939-5335	246-1170	482-4904	1378-4268	375-2301	246-5335

OR: Odds ratio per 1 SD increase in CFHR5 concentration, adjusted for differences in age including all VTE patients, separated by sex

Table C. Relative risk of VTE associated with CFHR5 by causality categories: provoked/unprovoked VTE

		VEBIOS ER	DFW-VTE	FARIVE	VEBIOS Coag.	RETROVE	Meta-analysis
VTE-Unprovoked	OR (95%CI)	2.01 (1.14-3.87)	NA	1.19 (0.98-1-43)	1.52 (1.16-2.03)	1.33 (1.10-1.61)	1.33 (1.18-1.50)
	P-value	0,02		0,074	3,33E-03	3,10E-03	2,04E-06
	Cases	25		156	90	203	474
	Controls	48		576	140	360	1124
	CFHR5 range (ng/mL)	1216-5560		450-4504	1378-4570	364-2341	364-5560
VTE-Provoked	OR (95%CI)	2.65 (1.48-5.48)	NA	1.26 (1.11-1.43)	1.46 (1.04-2.09)	1.21 (0.95-1.53)	1.30 (1.17-1.44)
	P-value	3,14E-03		4,01E-04	0,03	0,125	1,00E-06
	Cases	23		425	54	105	607
	Controls	48		576	140	360	1124
	CFHR5 range (ng/mL)	1216-6073		450-4904	1378-4570	364-2206	364-6073

OR: Odds ratio per 1 SD increase in CFHR5 concentration, adjusted for differences in age and sex, separated by whether the VTE event was unprovoked or provoked

Tab_11: Association of CFHR5 associated lead SNPs with VTE risk in the INVENT-MVP consortium resources (Thibord et al, Circulation. 2022 Oct 18;146(16):1225-1242)

				Association with CFHR5 levels			Association with VTE risk			Colocalisation
CHR:POS:NEA:EA	**RS**	**Localisation**	**GENE**	**BETA(SE)**	**Z**	**P**	**Log(OR)**	**se(Log(OR))**	**P**	**PP4 probability**
1:196851676:T:G	rs10737681	Intergenic	CFHR1;CFHR4	0.28 (0.01)	42.59	$2.94x10^{-396}$	-0.0162	0.0068	**0.01679**	1.27%
8:125487789:C:G	rs28601761	Intergenic	TRIB1;LINC00861	-0.04 (0.01)	-5.87	$4.39x10^{-9}$	0.0176	0,0060	3,41E-03	15.7%
10:63229171:A:T	rs7916868	Intronic	JMJD1C	0.04 (0.01)	6.92	$4.61x10^{-12}$	0.0183	0.0059	2,02E-03	39.3%
12:120986603:A:G	rs2393776	Intronic	HNF1A	0.06 (0.01)	9.54	$1.48x10^{-21}$	-0.0081	0.0059	0,1718	2.9%
12:123924955:A:G	rs7133378	Intronic	DNAH10	-0.04(0.01)	-5.56	$2.43x10^{-8}$	-0.0223	0.0063	4,37E-04	42.4%
20:44413724:T:C	rs1800961	Exonic	HNF4A	0.10 (0.02)	5.45	$4.97x10^{-8}$	-0.0368	0.0166	0,0263	8.22%

Tab_12: Lead SNPs genome-wide significantly associated with CFHR5 levels in the extended meta-analysis of GWAS datasets

				FARIVE				MARTHA				RETROVE_cases				RETROVE_controls				OMICSCIENCE				EPIC				DECODE (N=35 559)				META-ANALYSIS			
CHR:POS:NEA:EA	RS	Localisation	GENE	BETA(SE)	Z	P	N	BETA(SE)	Z	P	N	BETA(SE)	Z	P	N	BETA(SE)	Z	P	N	BETA(SE)	Z	P	N	BETA(SE)	Z	P	N	BETA(SE)	Z	P	N	BETA(SE)	Z	P	N
1:196851676:T:G	rs10737681	Intergenic	CFHR1;CFHR4	0.16(0.05)	3.49	$5.06x10^{-4}$	1033	0.27(0.04)	6.51	$1.10x10^{-10}$	1266	0.29(0.09)	3.44	$5.86x10^{-4}$	308	0.38(0.08)	4.80	$1.62x10^{-6}$	360	0.28(0.01)	20.0	$1.80x10^{-89}$	10708	0.23(0.04)	5.54	$3.05x10^{-8}$	1178	0.30(0.01)	36.08	$4.64x10^{-285}$	35338	0.28 (0.01)	42.59	$2.94x10^{-396}$	50191
8:125487789:C:G	rs28601761	Intergenic	TRIB1;LINC00861	-0.07(0.04)	-1.46	0.144	1033	-0.02(0.04)	-0.46	0.646	1266	-0.14(0.08)	-1.74	0.083	308	-0.05(0.08)	-0.64	0.525	360	-0.01(0.01)	-0.96	0.333	10708	-0.12(0.04)	-2.84	$4.49x10^{-3}$	1178	-0.04(0.01)	-5.38	$7.48x10^{-8}$	35369	-0.04 (0.01)	-5.87	$4.39x10^{-9}$	50222
10:63229171:A:T	rs7916868	Intronic	JMJD1C	-0.01(0.04)	-0.23	0.820	1033	0.03(0.04)	0.73	0.467	1266	0.06(0.08)	0.79	0.430	308	-0.16(0.07)	-2.11	0.035	360	0.04(0.01)	2.64	$8.25x10^{-3}$	10708	0.07(0.04)	1.62	0.106	1178	0.05(0.01)	6.53	$6.39x10^{-11}$	35351	0.04 (0.01)	6.92	$4.61x10^{-12}$	50204
12:120986603:A:G	rs2393776	Intronic	HNF1A*	0.02(0.05)	0.41	0.682	1033	0.09(0.04)	2.11	0.035	1266	0.12(0.08)	1.41	0.159	308	0.03(0.08)	0.40	0.686	360	0.07(0.01)	5.09	$3.65x10^{-7}$	10708	0.04(0.04)	0.84	0.402	1178	0.06(0.01)	7.77	$7.77x10^{-15}$	35365	0.06 (0.01)	9.54	$1.48x10^{-21}$	50218
12:123924955:A:G	rs7133378	Intronic	DNAH10	-0.04(0.05)	-0.89	0.375	1033	-0.02(0.04)	-0.40	0.689	1266	-0.20(0.08)	-2.30	0.021	308	-0.08(0.08)	-1.02	0.308	360	-0.04(0.01)	-2.61	$9.16x10^{-3}$	10708	-0.04(0.04)	-0.91	0.366	1178	-0.04(0.01)	-4.50	$6.67x10^{-6}$	35369	-0.04(0.01)	-5.56	$2.43x10^{-8}$	50222
20:44413724:T:C	rs1800961	Exonic	HNF4A	0.06(0.13)	0.46	0.648	1033	0.21(0.12)	1.84	0.066	1266	0.07(0.24)	0.30	0.763	308	0.19(0.22)	0.89	0.374	360	0.09(0.04)	2.32	0.020	10708	0.04(0.12)	0.36	0.715	1178	0.09(0.02)	4.61	$4.00x10^{-6}$	35371	0.10 (0.02)	5.45	$4.97x10^{-8}$	50224

*top snp of this locus was rs7135337 (R^2=0,83 ; D'=0,96 with rs2393776), intergenic to XLOC_009911;HNF1A-AS1, but exhibiting heterogeneity across studies

Tab_13: Mendelian Randomization analysis investigating the causal link between CFHR5 levels and VTE risk

Methods	Estimate	Standard Error (Estimate)	P	$P_{heterogeneity}$	I^2	Intercept	$P_{intercept}$
IVW	-0.0502	0.0615	0.414	3.532 E-07	86.89		
Weighted Median	-0.0607	0.0233	0.009				
Egger	-0.0875	0.0918	0.3402			0.0065	0.5586

N.B In these MR analyses, the 6 independent lead SNPs for CFHR5 (Table A) were used as instrument variables

Tab_14: Association of coding polymorphisms at the CFHR1-CHFR5 gene cluster with CFHR5 plasma levels in the MARTHA Study

				Common Homozygotes	Heterozygotes	Rare Homozygotes	Association Pvalue*
CFHR2	rs138792300_T/C	exm133636	Mean	866.5501	859.32	-	
			SD	201.7235	NA	-	
			N	1229	1	-	NA
CFHR2	rs7417769_G/A	exm133651		Common Homozygotes	Heterozygotes	Rare Homozygotes	
			Mean	859.711	869.9894	879.3656	
			SD	210.0704	196.858	190.3602	
			N	526	541	160	
							p = 0.2606
CFHR2	rs181498339_A/T	exm133666		Common Homozygotes	Heterozygotes	Rare Homozygotes	
			Mean	866.3337	889.8745	-	
			SD	201.6064	214.1214	-	
			N	1219	11	-	
							p = 0.6127
CFHR2	rs41257904_G/T	exm133727		Common Homozygotes	Heterozygotes	Rare Homozygotes	
			Mean	865.7398	941.8477	-	
			SD	200.2927	304.5941	-	
			N	1217	13	-	
							p = 0.1919
CFHR2	rs41310132_A/G	exm133734		Common Homozygotes	Heterozygotes	Rare Homozygotes	
			Mean	867.3962	850.2161	-	
			SD	204.8161	126.2275	-	
			N	1169	61	-	
							p = 0.4688
CFHR5	rs57960694_G/A	exm133770		Common Homozygotes	Heterozygotes	Rare Homozygotes	
			Mean	866.3779	934.5767	-	
			SD	201.8326	82.39749	-	
			N	1227	3	-	
							p = 0.5158
CFHR5	rs151134004_T/G	exm133777		Common Homozygotes	Heterozygotes	Rare Homozygotes	
			Mean	866.6771	703.23	-	
			SD	201.6698	NA	-	
			N	1229	1	-	
							p = 0.4526
CFHR5	**rs41299613_T/C**	exm133780		Common Homozygotes	Heterozygotes[2]	Rare Homozygotes	
			Mean	867.3763	526.22	-	
			SD	201.1603	75.30122	-	
			N	1227	3	-	
							p = 0.0016
CFHR5	**rs139017763_G/A**	exm133790		Common Homozygotes	Heterozygotes[2]	Rare Homozygotes	
			Mean	868.76	620.9922	-	
			SD	200.7478	146.9869	-	
			N	1219	11	-	
							p = 4.75E^{-05}
CFHR5	**rs35662416_G/A**	exm133801		Common Homozygotes	Heterozygotes	Rare Homozygotes[3]	
			Mean	870.274	822.1695	669	
			SD	201.4904	198.5919	176.6211	
			N	1141	87	2	
							p = 0.0071
CFHR5	rs144438200_G/A	exm133821		Common Homozygotes	Heterozygotes	Rare Homozygotes	
			Mean	866.6696	828.1	-	
			SD	201.8976	100.1842	-	
			N	1226	4	-	
							p = 0.6935
CFHR5	rs141321678_T/G	exm133827		Common Homozygotes	Heterozygotes	Rare Homozygotes	
			Mean	866.7805	803.5633	-	
			SD	201.884	276.5459	-	
			N	1222	3	-	
							p = 0.4817

***p-values were obtained under the assumption of additive allele effects and were adjusted for age, sex and anticoagulant therapy**
1: CFHR5 values in the 3 heterozygotes were 597.97, 532.88 and 447.81
2: CFHR5 values in the 11 heterozygotes were 643.07, 650.10, 451.18, 621.44, 705.78, 534.39, 500.15, 764.95, 481.47, 528.89 and 949.49
3: CFHR5 values in rare homozygotes were 793.89 and 544.11

Annexe 2 : *Bayesian network analysis of plasma microARN sequencing data in patients with venous thrombosis*

Thibord F, **Munsch G**, Perret C, […], Deleuze J-F, Morange P-E*, Trégouët D-A*. *Bayesian network analysis of plasma microRNA sequencing data in patients with venous thrombosis.* **Eur Heart J Suppl**. 2020 Apr 1;22(Supplement_C):C34–45.

European Heart Journal Supplements (2020) **22** (Supplement C), C34–C45
The Heart of the Matter
doi:10.1093/eurheartj/suaa008

ESC European Society of Cardiology

Bayesian network analysis of plasma microRNA sequencing data in patients with venous thrombosis

Florian Thibord[1,2], Gaëlle Munsch[1], Claire Perret[3], Pierre Suchon[4], Maguelonne Roux[3], Manal Ibrahim-Kosta[4,5], Louisa Goumidi[5], Jean-François Deleuze[6,7], Pierre-Emmanuel Morange[4,5†], and David-Alexandre Trégouët[1*†]; on behalf of the GENMED Consortium

[1]*Institut National pour la Santé et la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1219, Bordeaux Population Health Research Center, University of Bordeaux, 146 rue Léo Saignat, Bordeaux 33076, France;*
[2]*Pierre Louis Doctoral School of Public Health, Sorbonne-Université, 15 rue de l'école de médecine, Paris 75006, France;*
[3]*Sorbonne Universités, Université Pierre et Marie Curie (UPMC Univ Paris 06), INSERM UMR_S 1166, 91 Boulevard de l'Hôpital, Paris 75013, France;*
[4]*Laboratory of Haematology, La Timone Hospital, 278 rue Saint Pierre, Marseille 13385, France;*
[5]*INSERM UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, Center for CardioVascular and Nutrition research (C2VN), Aix-Marseille University, 278 rue Saint Pierre, Marseille 13385, France;*
[6]*Centre National de Recherche en Génomique Humaine, Direction de la Recherche Fondamentale, CEA, 2 rue Gaston Crémieux, Evry 91057, France; and*
[7]*CEPH, Fondation Jean Dausset, 27 rue Juliette Dodu, Paris 75010, France*

KEYWORDS
Venous thrombosis;
plasma miRNA;
Next generation sequencing;
Biomarkers;
Genome Wide
Association Study

MicroRNAs (miRNAs) are small regulatory RNAs participating to several biological processes and known to be involved in various pathologies. Measurable in body fluids, miRNAs have been proposed to serve as efficient biomarkers for diseases and/or associated traits. Here, we performed a next-generation-sequencing based profiling of plasma miRNAs in 344 patients with venous thrombosis (VT) and assessed the association of plasma miRNA levels with several haemostatic traits and the risk of VT recurrence. Among the most significant findings, we detected an association between hsa-miR-199b-3p and haematocrit levels ($P = 0.0016$), these two markers having both been independently reported to associate with VT risk. We also observed suggestive evidence for association of hsa-miR-370-3p ($P = 0.019$), hsa-miR-27b-3p ($P = 0.016$) and hsa-miR-222-3p ($P = 0.049$) with VT recurrence, the observations at the latter two miRNAs confirming the recent findings of Wang *et al*. Besides, by conducting Genome-Wide Association Studies on miRNA levels and meta-analyzing our results with some publicly available, we identified 21 new associations of single nucleotide polymorphisms with plasma miRNA levels at the statistical significance threshold of $P < 5 \times 10^{-8}$, some of these associations pertaining to thrombosis associated mechanisms. In conclusion, this study provides novel data about the impact of miRNAs' variability in haemostasis and new arguments supporting the association of few miRNAs with the risk of recurrence in patients with venous thrombosis.

*Corresponding author. Tel: +33 5 4730 4254, Email: david-alexandre.tregouet@u-bordeaux.fr; david-alexandre.tregouet@inserm.fr
† These authors contributed equally to this study.

PALABRAS CLAVE
Trombosis venosa;
Plasma miRNA;
Secuenciación de última generación;
Biomarcadores;
Genoma completo;
Estudio de asociación

Los micro-ARN (miARN) son pequeñas moléculas de ARN reguladoras que participan en varios procesos biológicos y están implicados en diversas patologías. Mensurables en los líquidos corporales, se ha planteado que los miARN pueden ser biomarcadores eficaces para el diagnóstico de enfermedades y/o características asociadas. Aquí hemos llevado a cabo un análisis de miARN plasmático con tecnología de secuenciación de última generación en 344 pacientes con trombosis venosa (TV) y hemos evaluado la asociación de los niveles de miARN con distintas características hemostáticas y el riesgo de recidiva de TV. Entre los hallazgos más significativos, hemos detectado una asociación entre hsa-miR-199b-3p y los niveles de hematocritos ($p = 0{,}0016$); dos marcadores que se habían asociado de forma independiente con el riesgo de sufrir TV. Asimismo, hemos observado una evidencia indicativa de asociación entre hsa-miR-370-3p ($p = 0{,}019$), hsa-miR-27b-3p ($p = 0{,}016$) y hsa-miR-222-3p ($p = 0{,}049$) y la recidiva de TV; los resultados los dos últimos miARN confirman los hallazgos recientes de Wang *et al.* (Clin Epigenetics, 2019). Además, al efectuar estudios de asociación del genoma completo sobre los niveles de miARN y al metaanalizar nuestros resultados con otros disponibles públicamente, hemos identificado 21 asociaciones nuevas de polimorfismos de un solo nucleótido (PSN) con niveles de miARN plasmático con un umbral de significación estadística de $p < 5 \times 10^{-8}$; algunas de estas asociaciones pertenecen a los mecanismos patogénicos de la trombosis.
Como conclusión, en este estudio se proporcionan nuevos datos sobre el impacto de la variabilidad de miARN en la hemostasia y nuevos argumentos que apoyan la asociación de algunas secuencias de miARN con el riesgo de recidiva en pacientes con trombosis venosa.

关键词
静脉血栓形成;
血浆miRNA;
下一代测序;
生物标记;
基因组广泛;
关联研究

微小 RNA（miRNA）是参与多种生物学过程的小型调节性 RNA，已知参与各种病理过程。在体液中可测量的 miRNA 已被提议为疾病和/或相关性状的有效生物标记。我们在此对 344 名静脉血栓形成（VT）患者的血浆 miRNA 进行了基于下一代测序的分析，并评估了血浆 miRNA 水平与几种止血性状和 VT 复发风险之间的关系。我们的重大发现之一是，我们检测到 hsa-miR-199b-3p 与血细胞比容水平之间存在关联（$p = 0.0016$），这两个标志物均已被独立报道与 VT风险相关。我们还观察到了提示性的证据，表明 hsa-miR-370-3p（$p = 0.019$）、hsa-miR-27b-3p（$p = 0.016$）和 hsa-miR-222-3p（$p = 0.049$）与 VT 复发相关，后两个 miRNA 的观察结果证实了 Wang □□的最新发现。（临床表观遗传学（Clin Epigenetics），2019 年）。此外，通过对 miRNA 水平进行基因组广泛关联研究并通过一些可获得的公开结果对我们的结果进行荟萃分析，我们在 $p < 5 \times 10^{-8}$ 的统计学显著性阈值下发现了 21 个 SNP 与血浆 miRNA 水平的新关联，其中一些关联与血栓形成的相关机制有关。
总而言之，这项研究提供了有关 miRNA 变异性在止血方面的影响的新数据，并为少数 miRNA 与静脉血栓形成患者复发风险的相关性提供了新论据。

Introduction

Venous thrombosis (VT), including deep vein thrombosis (DVT) and pulmonary embolism (PE), affects about 1 200 000 individuals each year in Europe and is thus the third most common cardiovascular disease after coronary artery disease and stroke.[1] It is a severe disorder that leaves many patients (25–50%) with a debilitating post-thrombotic syndrome[2] and whose PE manifestation kills many of them (6% acute and 20% after 1 year).[3] About 50% of VT are unprovoked, i.e. they occur without clear external factors like surgery, trauma, immobilization, hormone use or cancer. The annual recurrent rate is ~6% and about 25% of patients with unprovoked VT will face a recurrent event after a 6-month course of anticoagulant treatment.[4] Thus, the secondary prevention of VT in this specific population group of patients with a first unprovoked VT is a major health issue.

There is an urgent need to better understand the pathophysiological mechanisms leading to VT in order to develop targeted therapeutic and preventative strategies to save lives, improve quality of life and reduce healthcare costs. Effective preventative options are available in the form of anticoagulant treatments, but these are associated with major bleeding complications. There are unmet needs to develop predictive biomarkers with high sensitivity and specificity for accurate identification of patients who will develop a recurrence, to avoid unacceptably high risk of bleeding complications in patients at low risk of recurrence. Indeed, preventing thrombosis without inducing bleeding is the holy grail of anticoagulant therapy. Currently, there are no commercially available anticoagulants that achieve this.

Predicting the risk of recurrence and discriminating between fatal (PE) and non-fatal (DVT) events in unprovoked VT patients remain challenging. There is so far no

established biomarkers that serve these aims, even if D-dimers measurement has been proposed[5] but lacks specificity. We here propose a comprehensive microRNA (miRNA) profiling from plasma samples of VT patients aimed at discovering miRNA-derived biomarkers discriminating between PE and DVT and associated with VT recurrence. MicroRNAs represent a class of small (~22 nucleotides) non-coding RNAs that participate in genes post-transcriptional regulation.[6] It is now well-established that miRNAs are involved in the development of human diseases, in particular, cardiovascular ones.[7] Several genes participating to thrombosis associated mechanisms have already been suspected to be subject to miRNA regulation.[8-11] So far, epidemiological studies looking for association of plasma miRNAs with VT outcomes are still sparse. Using plasma samples of 20 VT cases and 20 healthy individuals, Starikova *et al.*[12] assessed the association of 97 miRNAs with VT risk among which 9 were found significantly ($P < 0.05$) associated with the outcome. As for Wang *et al.*,[13] by looking for the association of 110 miRNAs with the risk of VT recurrence in plasma samples of 39 cases and 39 controls, 12 miRNAs were identified. None of these observations, which were obtained on miRNA data profiled using RT-qPCR techniques, have yet been replicated.

Briefly, we here performed plasma miRNA profiling in 391 VT patients using a next-generation sequencing technology and assessed the association of identified miRNAs with several haemostatic traits and VT associated clinical outcomes. Association analyses were conducted using an original Bayesian network (BN) inference strategy aimed at identifying miRNAs with the highest abilities to serve as relevant biomarkers. In addition, we integrated genome-wide genotype data with miRNA expression levels in order to identify miRNAs that are under strong genetic control.

Methods

The MARTHA microRNA sequencing study

The MARseille THrombosis Association project refers to a collection of VT patients recruited at the La Timone Hospital in Marseille, France, initially between 1994 and 2005 and further extended over the 2010–12 period. Detailed description of this collection has already been previously provided.[14]

The present study relies on a subsample of 391 VT patients that had been previously genotyped for genome-wide polymorphisms using dedicated genotyping array[15,16] and with available plasma samples. For each sample, total RNA was extracted from 400 μL citrate plasma sample using miRNeasy Serum/Plasma kit from Qiagen. From 6 μL of total RNA, plasma miRNA libraries were then prepared with NEBNext Multiplex Small RNA Library Prep Set for Illumina. The manufacturer's protocol was followed, with an optimized size selection method via Ampure XP beads, a specific dilution of adapters to 1/10, and 15 cycles of PCR amplification, using adapter sequences GATC GG AAGAGCACACGTCTGAACTCCAGTCAC and CGACAGGTTCAG AGTTCTACAGTCCGACGATC for 3' and 5' ends, respectively. Detailed characteristics of the experimental protocol for libraries preparation and sequencing have already been described.[17]

MicroRNA alignment and quantification processes

Sequenced data were processed with the bioinformatic OptimiR pipeline[17] in order to detect and quantify miRNAs. Briefly, OptimiR aligned miRNAs to a library composed of mature miRNA references sequences from miRBase 21.[18] For miRNA integrating genetic variants in their sequence (called polymiRs), the reference library was upgraded by OptimiR with sequences integrating alternative alleles. Ambiguous alignments were resolved using a scoring algorithm that keeps only the most likely alignment while considering the frequent post-transcriptional modifications that miRNAs can undergo.[19] Reads aligned on polymiRs were kept if they were consistent with the sample's genotype, otherwise, they were discarded.[17]

From the resulting miRNA abundances, we performed several quality assessments in order to discard unreliable data. First, samples that were poorly sequenced, i.e. with <100 000 reads aligned, were discarded ($n = 3$) as well as samples identified to be haemolyzed ($n = 34$). The degree of haemolysis was determined based on the optical density at 414 nm, and values exceeding 0.2 were defined as haemolyzed samples.[20] Finally, in order to retain only highly expressed miRNAs, we kept only those with at least five counts in at least 75% of the remaining samples.

Abundances were then normalized using the rlog method from the DESeq2 R library.[21] This normalization process takes into account differences in library sizes due to library preparation and sequencing protocols, and stabilize variance across miRNAs and samples to respect homoscedasticity constraints for further analysis. Principal component analysis (PCA) was applied to normalized abundances in order to identify individuals with outliers miRNA profiles. Individuals deviating by 3 SD from the centres of the first four PCAs ($n = 10$) were further excluded from downstream analyses, leaving 344 individuals for BN and association analyses.

Bayesian network analysis

A BN is a probabilistic directed acyclic graphical model that represents relationships among a large number of variables (here mainly miRNAs) with the aim of modelling the dependencies/interactions and conditional independencies between variables.[22,23] Generally, any BN is defined by a directed acyclic graph structure $G = (V,E)$ where V is the set of variables and E the set of edges representing the directional relationships between variables and P a joint probability distribution of the variables in the network. Three types of nodes can be identified in a given BN: the root nodes that are variables found to influence several other variables but are not themselves influenced by any other variables, the internal nodes that are both influenced by and modulate other variables, and finally terminal nodes that are variables that are not identified as influencing others (see *Figure 1*). Any variable influencing another variable in the network is referred to as a parental node for this later variable. In the following, we will mainly

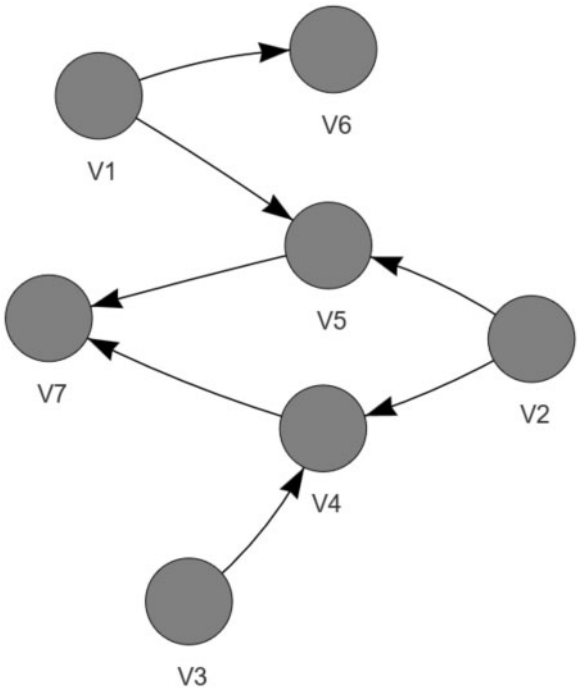


Figure 1 A Bayesian network example. In this illustrative BN example, variables V1, V2, and V3 are root nodes, V4 and V5 are internal nodes, and V6 and V7 are terminal nodes. V3 is also a parental node for V4 which is itself a parental node for V7.

focus on terminal nodes assuming that such nodes, as integrating the cumulative upstream effects of other variables, would serve as more relevant and powerful endophenotypes to be tested in relation to some outcomes of interest. In that context, BN analysis can also be viewed as a data reduction technique since, instead of testing the association of all initial variables with a given outcome, only the terminal nodes will be tested for association, reducing then the multiple testing burden. In this article, BNs will be constructed with the «bnlearn» package[24] that implements the relatively fast *tabu search* algorithm handling both discrete and continuous variables. In the current application, BNs will be created from all expressed miRNAs but also with the age and sex variables. These two latter variables have been shown to have strong influence on circulating miRNA levels[25,26] and their integration in the BN analysis can then add information to more efficiently model the dependencies and conditional independence between some miRNAs.

Because *tabu search* is a greedy search algorithm, it may end up into a local optimum. To overcome such situation and to assess the stability of the BN analysis in identifying robust terminal nodes, we generated 2000 bootstrapped datasets composed of 95% of the initial samples and for each boostrapped datasets, we randomly shuffled the way the input variables were ordered in the initial dataset. For each shuffled bootstrapped dataset, a BN was constructed and the terminal nodes identified. After 2000 bootstrap, we calculated the number of times a given variable was identified as terminal node.

In order to assess whether the observed distribution of the number of terminal node's occurrences deviates from the null hypothesis of no correlation structure between miRNAs, a permutation strategy was adopted. For each permutation, we randomly selected at least 40 variables whose values were permuted between individuals in order to break down the original data correlation structure. We generated 2000 of such permuted datasets and constructed a BN on each of them. From these permuted BNs, we counted the maximum number of times a given variable (that could be any miRNA, age, or sex) was identified as a terminal node and used this maximum value as a cut-off to identify robust terminal miRNAs in the unpermuted analysis above.

Association analysis with haemostatic traits and clinical outcomes

Identified terminal miRNAs were tested for association with several haemostatic traits available in MARTHA participants (see *Table 1*). Association analyses were performed using linear regression model and adjusted for age, sex, anticoagulant therapy, and combined plasma levels of hsa-let-7d-5p, hsa-let-7g-5p and let-7i-5p measured by qPCR, which serve as a control reference of miRNA levels.[27] Individuals under anticoagulant therapy at the time of blood sampling were excluded for the analysis on protein C, protein S, and prothrombin time. For association testing, log-transformation was applied to the following variables: Activated Thrombin Generation Potential biomarkers (Endogenous Thrombin Potential, Lagtime), Partial Thromboplastin Time, Factor VIII, Homocystein, Plasminogen Activator Inhibitor-1, Tissue Factor Principal Inhibitor, and von Willebrand Factor.

Terminal miRNAs were also tested for association with the DVT vs. PE outcome using a logistic regression model, while a Cox model was used to assess their association with VT recurrence whose information was available in 228 patients only. For the latter analysis, we applied the Cox survival model with left truncature[28] and adjusted for age, sex, body mass index (BMI), and smoking. To address the multiple testing issue associated with the number of terminal miRNAs that will be tested for association with the phenotypes, we applied a Bonferroni correction based on the effective number of independent variables.[29]

Genome-wide miR-eQTL analysis

As MARTHA participants have been typed for high-density genotyping arrays and imputed for common polymorphisms available in the 1000 G reference panel, we performed genome-wide association study (GWAS) on each expressed miRNA for identifying miRNA expression quantitative trait loci (miR-eQTL) using the mach2QTL programme.[30] Analyses were performed under the assumption of additive genetic effects and adjusting for the following covariates: sex, age of blood collection, anticoagulant prescription, RT-qPCR measured hsa-let-7 combination,[27] and the four first principal genetic components retrieved from PCA analysis as previously described.[15,16] GWAS results were filtered out for variants with minor allele frequency lower than 0.05 and with imputation criterion r^2 below 0.4. Finally, we combined the results of our miR-eQTL analysis with those previously described by Nikpay *et al.*[31] and

Table 1 Characteristics of the MARTHA miRNA cohort

Variables	*N*	Mean ± SD[a]
Gender (male/female)	344	144/200
Age (years)	344	52.1 ± 14.5
Smoking (yes/no)	343	94/249
BMI (kg/m^2)	331	25.86 ± 4.62
Deep vein thrombosis/pulmonary embolism	344	259/85
Anticoagulant therapy (yes/no)	344	122/222
Antithrombin (IU/mL)	313	102.41 ± 11.59
Activated partial thromboplastin time (s)	341	33.42 ± 6.02
D-dimers (μg/mL)	184	0.39 ± 0.33[b]
FV (IU/mL)	150	109.21 ± 22.26
FVIII (IU/dL)	294	135.07 ± 48.31
FXI (IU/mL)	336	130.78 ± 31.99
Fibrinogen (g/L)	342	3.42 ± 0.66
Haematocrit (L/L)	343	0.42 ± 0.03
Homocysteine (μmol/L)	304	12.26 ± 5.65
Platelet count (G/L)	344	254.62 ± 64.91
Mean platelet volume (fL)	344	7.90 ± 0.77
Haemoglobin (g/dL)	344	140.42 ± 13.19
PAI-1 (UI/mL)	272	12.25 ± 13.44
Protein C (IU/mL)	318	99.55 ± 40.56
Protein S (IU/mL)	322	81.3 ± 27.49
TAFI (μg/mL)	336	15.27 ± 4.72
TFPI (ng/mL)	336	14.17 ± 6.84
vWF (IU/dL)	308	154.34 ± 67.74
Prothrombin time (%)	344	87.63 ± 27.95
Thrombin generation	193	
Endogeneous thrombin potential (nM·min)		1761.44 ± 280.31
Peak (nM)		340.35 ± 57.51
Lagtime (min)		3.34 ± 1.17
VT recurrence during follow-up (yes/no)	228	41/187

[a]Count data are shown for categorical variables, other reported values were mean ± standard deviation.
[b]In about 50% participants, D-dimers values were below the detection limit (0.22) and thus discarded. Mean and SD were then computed over all D-dimer values >0.22.

available at https://zenodo.org/record/2560974 in order to identify additional single nucleotide polymorphism (SNP) × miRNA associations. For this, a random-effect model-based meta-analysis was adopted as implemented in the GWAMA software.[32] SNP × miRNA associations were considered as *cis* effects when the SNP maps ± 1 Mb from the mature miRNA position. Otherwise, they were considered as *trans*. Any association with *P*-value $<3.2\times10^{-10}$ corresponding to the Bonferroni threshold corrected for the number of tested SNP × miRNA associations was considered as genome-wide significant. We also used a miRNA-wide threshold of $P<5\times10^{-8}$, the standard statistical threshold generally advocated in the context of a single GWAS, to identify additional suggestive associations.

Results

The MARTHA microRNA cohort

Detailed description of the clinical and biological characteristics of the 344 participants is shown in *Table 1*. Of note, 228 patients have been followed for the risk of recurrence for a mean time period of 11.4 ± 4.3 years. During this period, 41 patients experienced a new VT event.

After the application of the OptimiR workflow, 162 miRNAs were found expressed in the 344 MARTHA participants. Full miRNA data are provided in Supplementary material online, *Table S1*. The most expressed miRNA was the hsa-miR-122-5p (Supplementary material online, *Figure S1*), a miRNA known to be mainly expressed in liver and that was previously shown to be amongst the most abundant plasma miRNAs.[33] Additional highly expressed miRNAs were hsa-miR-486-5p, hsa-miR-92a-3p, and hsa-miR-451a (Supplementary material online, *Figure S1*). Of note, the 25 most expressed miRNAs accounted for >90% of all sequenced reads that were aligned to miRNA mature sequences.

BN analysis of microRNA data

Under the null hypothesis of no specific structure in the miRNA data, all miRNAs were identified as a terminal node at least once and, on average, a miRNA was found as a

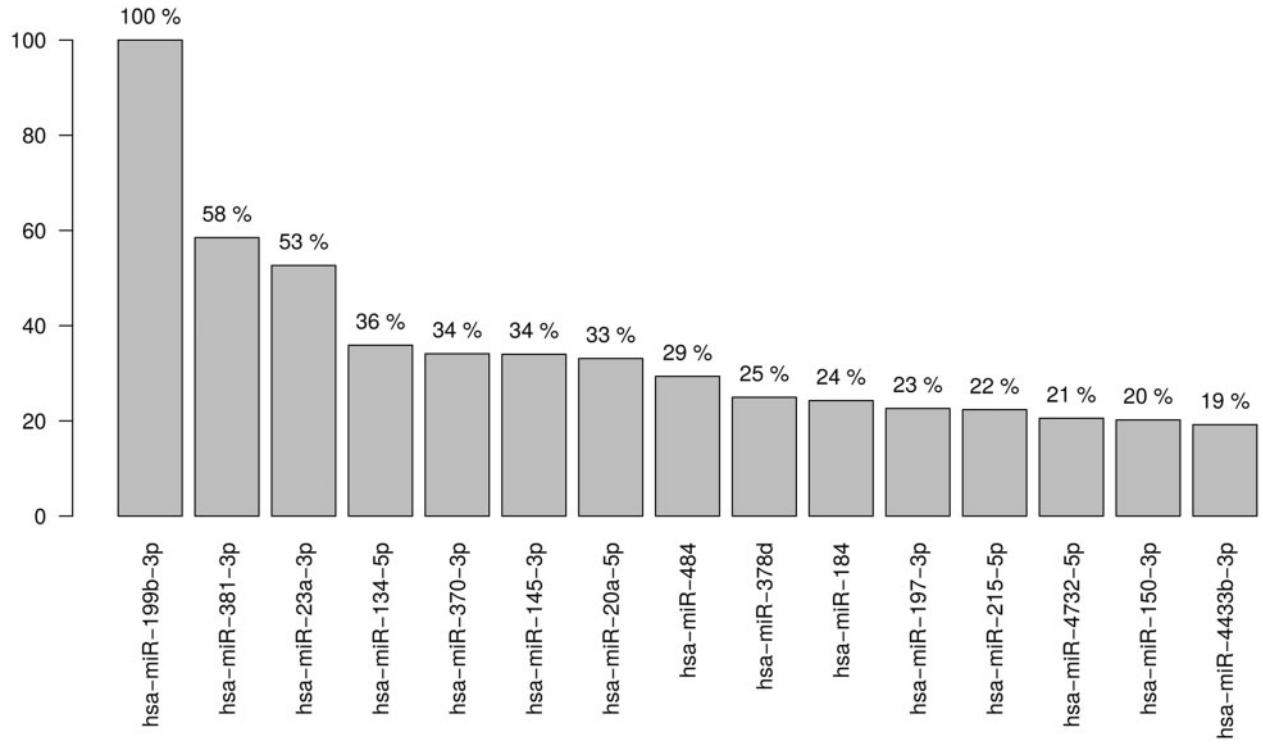


Figure 2 Percentage of significant terminal miRNAs found in 2000 bootstrapped Bayesian networks. The bootstrap BN analysis identified 15 terminal miRNAs with an occurrence percentage over the significance threshold (18.3%) determined by the permutation analysis.

terminal node in 6.3% ± 3.5 of the permuted BNs, with a maximum of 18.3%. Using the latter threshold, the bootstrap BN analysis identified 15 terminal miRNAs and the number of times each of them was found as a terminal node in boostrapped BNs is shown in *Figure 2*.

Association of microRNAs' levels with VT-associated biological and clinical traits

The application of the Li and Ji multiple testing procedure[29] estimated the number of effective independent terminal miRNAs as 14, leading to an adapted Bonferroni threshold of 3.6×10^{-3}. At this statistical level, only one association between terminal miRNAs and haemostatic traits was detected. Plasma levels of hsa-miR-199b-3p were negatively correlated ($\rho = -0.17$, $P = 0.0016$) with haematocrit levels. Interestingly, this miRNA has recently been reported to associate with VT risk[12] whose association with haematocrit levels have already been described.[34,35] The full results of the scan for association between miRNAs and haemostatic traits are given in Supplementary material online, *Table S2*.

Of note, the strongest association of terminal miRNAs with recurrence risk was observed for hsa-miR-370-3p [HR = 1.77 (1.09–2.88), $P = 0.019$], this miRNA being also the terminal miRNA that discriminated the most between DVT and PE [OR for PE = 0.72 (0.49–1.05), $P = 0.090$] (*Table 2*). Of interest, one of our terminal miRNAs, hsa-miR-197-3, was reported to associate with VT recurrence in Wang *et al.*[13] However, we did not observe here such trend for association [HR = 0.78 (0.35–1.76), $P = 0.55$]. Nevertheless, among the nine additional miRNAs reported in Wang *et al.* and also expressed in MARTHA, we found two with a suggestive association with VT recurrence: hsa-miR-27b-3p [HR = 0.4 (0.2–0.79), $P = 0.016$] and hsa-miR-222-3p [HR = 1.76 (1.01–3.08), $P = 0.049$] (Supplementary material online, *Table S3*).

miR-eQTL analyses

At the pre-specified genome-wide statistical level of 3.2×10^{-10}, three SNP × miRNA associations, all *cis*, were identified in the MARTHA study (*Table 3*). These were observed for rs12473206 with hsa-miR-4433b-3p ($P = 8.12 \times 10^{-35}$), rs2127870 with hsa-miR-625-3p ($P = 9.57 \times 10^{-26}$), and rs140930133 with hsa-miR-941 ($P = 5.07 \times 10^{-15}$). The latter two have already been observed in whole blood[36] and adipose tissue.[37] Using a more liberal miRNA-wide threshold of $P = 5 \times 10^{-8}$, 10 additional suggestive associations, 1 in *cis* and 9 in *trans*, were observed (*Table 3*). Regional association plots and boxplot summarizing the genotype × miRNA associations at these 13 main candidates are shown in Supplementary material online.

Of note, the most significant association was observed between hsa-miR-4433b-3p and rs12473206, a variant located within the mature miRNA sequence. It can be speculated that this variant impacts the maturation process of the miRNA or its target spectrum, and thus influences its plasma expression levels. In addition, two SNPs with *cis* effects on miRNA levels (thereafter referred to as *cis* miSNPs) have been previously found to associate with levels of the protein encoded by the miRNA host gene. In whole blood, the miSNP rs2127870 was reported to

Table 2 Association of terminal miRNAs with VT outcomes in the MARTHA miRNA study

miRNA	VT recurrence		Pulmonary embolism vs. deep vein thrombosis	
	HR (95% CI)	P^a	OR (95% CI)	P^b
hsa-miR-370-3p	1.77 (1.09-2.88)	0.019	0.72 (0.49-1.05)	0.090
hsa-miR-184	0.53 (0.30-0.95)	0.024	1.23 (0.92-1.66)	0.153
hsa-miR-4732-5p	0.41 (0.18-0.92)	0.024	0.70 (0.39-1.22)	0.218
hsa-miR-4433b-3p	1.54 (1.04-2.29)	0.033	1.01 (0.75-1.36)	0.930
hsa-miR-215-5p	0.63 (0.37-1.09)	0.091	1.11 (0.73-1.67)	0.633
hsa-miR-134-5p	1.58 (0.85-2.91)	0.142	0.89 (0.57-1.39)	0.601
hsa-miR-381-3p	1.45 (0.83-2.56)	0.194	0.81 (0.53-1.23)	0.327
hsa-miR-145-3p	0.51 (0.15-1.76)	0.278	0.62 (0.24-1.56)	0.311
hsa-miR-23a-3p	0.67 (0.26-1.70)	0.393	1.00 (0.51-1.93)	0.999
hsa-miR-197-3p	0.78 (0.35-1.76)	0.555	1.41 (0.79-2.56)	0.251
hsa-miR-150-3p	1.23 (0.53-2.83)	0.629	0.90 (0.49-1.66)	0.743
hsa-miR-484	1.20 (0.56-2.59)	0.637	1.27 (0.69-2.38)	0.447
hsa-miR-199a-3p	0.80 (0.22-2.86)	0.726	1.17 (0.46-2.97)	0.746
hsa-miR-378d	0.81 (0.15-4.56)	0.812	0.41 (0.10-1.46)	0.184
hsa-miR-20a-5p	1.09 (0.40-2.95)	0.863	0.74 (0.36-1.52)	0.411

[a]*P*-values were obtained from the Likelihood Ratio test statistic associated with a Cox survival model adjusted for age, sex, BMI, and smoking.
[b]*P*-values obtained from a logistic model adjusted for age, sex, BMI, and smoking.

Table 3 Significant associations at the $5{\cdot}10^{-8}$ statistical level between SNPs and plasma miRNA levels in the MARTHA miRNA study

miRNA	miRNA host gene	Top SNP Associated	MAF	r^2	Chr	Distance to 5' miRNA	Effect (SD)	*P*-value	SNP Genomic Context
Cis associations									
hsa-miR-4433b-3p	Intergenic	rs12473206	0.23	0.99	2	−13	0.979 (0.080)	$8.12{\cdot}10^{-35}$	exonic_ncRNA (hsa-miR-4433b)
hsa-miR-625-3p	FUT8	rs2127870	0.27	0.99	14	141 025	0.533 (0.051)	$9.57{\cdot}10^{-26}$	Intergenic
hsa-miR-941	DNAJC5	rs140930133	0.19	0.97	20	8822	−0.349 (0.045)	$5.07{\cdot}10^{-15}$	Intronic (DNAJC5)
hsa-miR-432-5p	RTL1	rs201969986	0.29	0.95	14	177 423	−0.346 (0.063)	$3.31{\cdot}10^{-8}$	Intergenic
Trans associations									
hsa-miR-184		rs144867605	0.07	0.82	11	75 957 983	0.804 (0.134)	$2.02{\cdot}10^{-9}$	Intergenic
hsa-miR-654-5p		rs11109171	0.44	0.99	12	98 098 091	−0.246 (0.042)	$3.28{\cdot}10^{-9}$	Intergenic
hsa-miR-320c		rs10151482	0.06	0.93	14	41 934 917	0.427 (0.074)	$6.47{\cdot}10^{-9}$	Intergenic
hsa-miR-184		rs143007764	0.06	0.65	3	142 899 139	0.916 (0.161)	$1.14{\cdot}10^{-8}$	Intergenic
hsa-miR-1-3p		rs73245753	0.12	0.79	4	26 292 392	0.589 (0.105)	$2.31\ 10^{-8}$	Intergenic
hsa-miR-330-3p		rs1554362	0.45	0.82	2	101 221 457	−0.227 (0.041)	$2.81{\cdot}10^{-8}$	Intronic (LINC01849)
hsa-miR-582-3p		rs4522365	0.13	0.83	15	29 964 742	0.314 (0.057)	$2.91{\cdot}10^{-8}$	Intergenic
hsa-miR-4446-3p		chr12:95274192:I	0.09	0.61	12	95 274 192	−0.492 (0.089)	$3.07{\cdot}10^{-8}$	Intergenic
hsa-miR-320d		rs12800249	0.05	0.63	11	21 240 436	0.481 (0.088)	$4.33{\cdot}10^{-8}$	Intronic (NELL1)

MAF, minor allele frequency; r^2, imputation quality criterion.

influence FUT8 levels,[38] *FUT8* being the host gene for hsa-miR-625-3p. Similarly, the *DNAJC5* rs2427555 that is in very strong linkage disequilibrium (LD) with the miSNP rs140930133, we here found associated with plasma hsa-miR-941 levels, has been reported to influence the expression of *DNAJC5* in lymphoblastoid cells.[39] These observations are supportive elements for the observed miSNP associations and would suggest a joint regulation of hsa-miR-625-3p and hsa-miR-941 expressions with those of their host genes as already documented for several miRNAs.[40]

One *trans*-eQTL located in the long non-coding RNA (lncRNA) LINC01849 was associated with hsa-miR-330-3p. The identified *trans* miSNP, rs1554362, is also an eQTL for

Table 4 Association of SNPs with plasma miRNA levels identified in Nikpay *et al.*[31] that nominally replicated ($P < 0.05$) in MARTHA miRNA study

					NIKPAY ($N = 710$)				MARTHA ($n = 344$)				
miRNA	SNP	Chr	Position(bp)	EA	EAF	β	SE	*P*	EAF	R^2	β	SE	P^a
Cis associations													
miR-941	rs2427550	20	62547575	A	0.23	−0.157	0.023	3.96×10^{-11}	0.19	0.99	−0.339	0.044	5.76×10^{-15}
miR-584-5p	rs17795259	5	148416952	C	0.15	0.268	0.018	1.35×10^{-45}	0.15	0.99	0.213	0.043	4.82×10^{-7}
miR-4433b-5p	rs2059631	2	64574682	A	0.43	0.289	0.017	1.57×10^{-56}	0.45	1.00	0.129	0.029	4.96×10^{-6}
miR-139-3p	rs4944563	11	72316881	C	0.17	0.169	0.026	1.18×10^{-10}	0.14	1.00	0.182	0.042	6.82×10^{-6}
miR-181a-5p	rs74746864	1	199023240	G	0.11	0.175	0.025	4.12×10^{-12}	0.13	0.95	0.221	0.066	4.27×10^{-4}
miR-425-5p	rs7623513	3	142100428	C	0.15	−0.044	0.007	7.48×10^{-10}	0.12	0.95	−0.166	0.054	1.04×10^{-3}
let-7e-5p	rs2198171	19	52174483	G	0.27	−0.089	0.014	3.10×10^{-10}	0.25	0.97	−0.124	0.043	1.83×10^{-3}
miR-197-3p	rs7355073	1	110129740	T	0.16	−0.078	0.011	1.23×10^{-12}	0.19	1.00	−0.118	0.041	2.10×10^{-3}
miR-26b-5p	rs12623740	2	219665715	A	0.49	−0.060	0.007	3.37×10^{-18}	0.51	0.99	−0.138	0.051	3.24×10^{-3}
miR-152-3p	rs9910516	17	46183160	A	0.23	0.093	0.016	1.52×10^{-08}	0.27	0.95	0.089	0.033	3.44×10^{-3}
miR-27b-3p	rs10993381	9	97639463	T	0.07	0.170	0.016	2.00×10^{-24}	0.06	0.99	0.148	0.055	3.86×10^{-3}
miR-182-5p	rs2693738	7	129431977	G	0.32	0.115	0.020	2.36×10^{-08}	0.37	0.82	0.166	0.063	4.30×10^{-3}
miR-181a-3p	rs1434282	1	199010721	C	0.27	0.211	0.022	9.03×10^{-21}	0.26	0.98	0.122	0.048	5.57×10^{-3}
miR-181a-5p	rs12125200	1	198992043	A	0.27	0.340	0.013	1.13×10^{-111}	0.24	0.96	0.124	0.049	5.79×10^{-3}
miR-584-5p	rs4147470	5	148528107	T	0.49	−0.131	0.014	7.71×10^{-20}	0.51	1.00	−0.081	0.032	6.15×10^{-3}
miR-26b-5p	rs833083	2	219336959	T	0.41	−0.076	0.006	3.96×10^{-30}	0.43	0.81	−0.137	0.057	7.96×10^{-3}
miR-181a-5p	rs878254	1	199257141	A	0.48	−0.122	0.015	3.54×10^{-15}	0.49	0.90	−0.104	0.045	0.010
miR-181a-5p	rs2360961	1	199000277	C	0.40	−0.151	0.016	4.39×10^{-20}	0.40	0.94	−0.095	0.043	0.014
miR-30d-5p	rs13282464	8	135707922	T	0.15	0.092	0.007	2.02×10^{-33}	0.17	1.00	0.047	0.023	0.020
miR-4433b-5p	rs6740438	2	64528086	C	0.13	0.163	0.029	1.78×10^{-08}	0.15	0.98	0.083	0.041	0.022
miR-30d-5p	rs13268530	8	135727196	T	0.15	0.095	0.007	1.68×10^{-35}	0.17	0.99	0.045	0.023	0.024
miR-21-5p	rs2665392	17	57809453	A	0.16	0.059	0.011	3.59×10^{-08}	0.16	0.88	0.078	0.041	0.027
miR-4433b-5p	rs35503140	2	64539015	C	0.21	−0.130	0.022	9.86×10^{-09}	0.19	0.95	−0.071	0.037	0.029
miR-584-5p	rs9325124	5	148248818	A	0.39	−0.085	0.015	7.62×10^{-09}	0.45	1.00	−0.056	0.031	0.036
miR-181a-5p	rs3861924	1	199121330	A	0.18	0.137	0.020	2.06×10^{-11}	0.20	0.96	0.097	0.054	0.037
miR-1908-5p	rs174561	11	61582708	C	0.30	0.151	0.012	4.76×10^{-31}	0.26	1.00	0.052	0.030	0.040
miR-151a-3p	rs11167012	8	141968408	A	0.42	0.059	0.006	3.79×10^{-24}	0.40	1.00	0.061	0.036	0.045
miR-139-3p	rs10898849	11	72269302	T	0.25	0.124	0.022	3.30×10^{-08}	0.27	1.00	0.054	0.032	0.046
let-7i-5p	rs6581454	12	62934442	G	0.47	0.039	0.006	3.04×10^{-11}	0.44	0.99	0.034	0.021	0.049
Trans associations													
miR-222-3p	rs11070216	15	39817245	T	0.19	−0.067	0.012	4.87×10^{-08}	0.19	0.97	−0.198	0.051	5.06×10^{-5}
miR-222-3p	rs970280	15	39864403	G	0.32	−0.064	0.010	8.79×10^{-10}	0.32	0.94	−0.113	0.042	3.57×10^{-3}
miR-143-3p	rs4734879	8	106583124	G	0.28	0.239	0.031	2.88×10^{-14}	0.24	0.96	0.098	0.038	5.60×10^{-3}
miR-1-3p	rs11906462	20	61158952	T	0.20	0.310	0.033	6.28×10^{-20}	0.23	0.42	0.262	0.116	0.012
miR-320a	rs1443651	2	68569316	G	0.45	−0.036	0.006	7.12×10^{-10}	0.44	1.00	−0.053	0.028	0.029
miR-16-5p	rs137214	22	35288857	T	0.28	0.041	0.007	1.76×10^{-08}	0.29	0.97	0.088	0.050	0.040
miR-126-3p	rs600038	9	136151806	C	0.21	0.055	0.009	5.95×10^{-09}	0.34	1.00	0.041	0.024	0.041
miR-320c	rs1443651	2	68569316	G	0.45	−0.031	0.005	2.77×10^{-10}	0.44	1.00	−0.066	0.039	0.045

[a]One-sided test *P*-value.
EA, effect allele; EAF, effect allele frequency.

the PDCL3 transcript levels in different tissues according to the GTeX database.[41] Another intronic miSNP located in the *NELL1* gene was associated with hsa-miR-320d levels. The seven other *trans* eQTL are located in intergenic regions.

We sought to *in silico* replicate these miSNP associations using the results from Nikpay *et al.*[31] who scanned for genetic polymorphisms associated with miRNA levels in 710 plasma samples. Unfortunately, as the Nikpay *et al.* study relied on a genotyping array focusing mainly on coding regions and used a very stringent imputation quality criterion ($r^2 > 0.9$), it was not possible to assess all our candidate associations. Only four were testable (hsa-miR-941 × rs140930133, hsa-miR-432-5p × rs201969986, hsa-miR-654-5p × rs11109171, hsa-miR-320c × rs10151482) among which only the association of rs140930133 with hsa-miR-941 levels replicated ($P = 6.3 \times 10^{-11}$).

Conversely, we looked into the MARTHA results to replicate the 223 miSNP associations that were significantly ($P < 5 \times 10^{-8}$) detected in the Nikpay *et al.* study. We were able to test 92 of them among which 37 replicated at the nominal level of $P = 0.05$ in MARTHA (*Table 4*). These involved 29 *cis* and 8 *trans* miSNP associations.

Among these eight *trans* miSNP associations, three deserve to be highlighted. First, plasma levels of hsa-miR-143-3p were influenced by the intronic *ZFPM2* rs4734879, *ZFPM2* being a locus reported to associate with venous thrombosis risk[42] and platelet function.[43] In MARTHA, plasma levels of hsa-miR-143-3p were negatively significantly correlated with BMI ($\rho = -0.24$, $P = 3.6 \times 10^{-4}$) and borderline significant with PAI-1 activity levels ($\rho = -0.21$, $P = 5.3 \times 10^{-3}$) (Supplementary material online, *Table S2*). Second, hsa-miR-126-3p plasma levels were associated with the rs600038 located in the promoter region of the *ABO* gene. This polymorphism is in strong LD with several other *ABO* polymorphisms that are known to associate with VT risk, including the rs579459 ($r^2 = 0.99$) tagging for the A1 ABO blood group. In MARTHA, plasma levels of hsa-miR-126-3p were strongly and positively correlated ($\rho \sim 0.20$) with red cells ($P = 1.73 \times 10^{-5}$), lymphocytes ($P = 2.5 \times 10^{-4}$), platelets ($P = 5.9 \times 10^{-4}$), and polynuclear ($P = 6.0 \times 10^{-4}$) (Supplementary material online, *Table S2*). Third, polymorphisms (rs970280, rs11070216) in the promoter region of the *THBS1* gene were found associated with plasma levels of hsa-miR-222-3p. This miRNA has been previously reported to associate with the risk of VT recurrence[13] and has a suggestive association ($P = 0.049$) in our study (Supplementary material online, *Table S3*), where it positively correlated with antithrombin levels ($\rho = 0.21$, $P = 8.8 \times 10^{-4}$) (Supplementary material online, *Table S2*). THBS1 encodes Thrombospondin-1 and is known to be involved in angiogenesis and platelet aggregation.[44,45]

Finally, we performed a random-effect meta-analysis of both datasets in order to discover additional miSNPs. At the 5×10^{-8} statistical threshold, we identified seven new *cis* and five new *trans* miSNP associations (*Table 5*). None of these miSNP associations appeared to involve loci with documented link with thrombosis related traits.

Discussion and conclusion

In this study, we reported the largest investigation to date of miRNA plasma profiling in a cohort of VT patients. Capitalizing on the application of a next-generation sequencing technology, known to be more efficient and sensitive to detect and quantify miRNAs compared with microarray or RT-qPCR techniques, we were able to detect 162 highly expressed miRNAs. These miRNAs were then tested for association with several VT-related phenotypes including 38 haematological traits and VT recurrence. In order to deal with the correlation between miRNA levels and reduce the multiple testing burden associated with the number of tested miRNAs, we deployed an original BN analysis aimed at identifying miRNAs that could serve as more powerful biomarkers for the investigated traits. In addition, as our studied VT patients had been previously typed for genome-wide genotypes, we were able to perform GWAS on each of the 162 miRNAs, and combined our results with some previously obtained in disease-free individuals in order to identify novel associations of common SNPs with plasma miRNA levels.

Several conclusions could be derived from this work. First, we did not identify any miRNA that significantly associated with the risk of VT recurrence. In our study, the miRNA that discriminated the most between patients with or without recurrence, but also between DVT vs. PE patients, was the hsa-miR-370-3p. Several works have already reported the involvement of has-miR-370-3p in lipids metabolism[46–49] and one of the most robust target gene for hsa-miR-370-3p is *CPT1A*[50] whose role in lipid metabolism is also very documented.[51–53] Hsa-miR-370-3p is also predicted to target drug-metabolism genes, such as *CYP2D6* and *VKORC1L1*,[50] that are related to the warfarin anticoagulant pharmacotherapy. Aside this miRNA, we observed a trend of association with VT recurrence for the hsa-mir-27b-3p and hsa-miR-222-3p that had been previously identified in Wang *et al.*[13] but these associations ($P = 0.016$ and $P = 0.0495$, respectively) did not survive any multiple testing correction (Supplementary material online, *Table S3*). Larger studies would be mandatory to confirm these observations and increase our chance to identify other miRNAs associated with the risk of recurrence in VT patients. Second, we observed several significant associations of miRNAs with haematological traits that deserve further replication in independent studies. One can highlight the significant correlation between haematocrit levels and plasma levels of hsa-miR-199b-3p, a miRNA that has been reported to be associated with VT risk.[12] Third, our miR-QTL study identified about 25 significant ($P < 5 \times 10^{-8}$) associations of SNPs with plasma miRNA levels, of which, to the best of our knowledge, 21 have never been reported, including a dozen of *trans* associations. These associations could help deciphering the genomic architecture of complex diseases where miRNAs are involved. For example, plasma levels of hsa-miR-143-3p were found to be associated with the rs4734879 mapping to *ZFPM2*, a gene known to associate with platelet function[43] and VT risk.[42] We also observed a strong association of rs12473206 with plasma levels of hsa-miR-4433b-3p, a miRNA whose serum levels have recently shown to be associated with stroke.[54] The impact of this SNP on stroke risk deserves to be further and deeply investigated. The results of our GWAS on miRNA levels were combined with those obtained by Nikpay *et al.*[31] and freely available at https:// zenodo. org/ . However, only SNPs with imputation quality greater than 0.90 are available at this resource, which has hampered our ability to replicate some of the main associations observed in the MARTHA miRNA study. To facilitate future studies aimed at disentangling the genetic regulation of miRNAs, the results of the 162 GWAS performed on miRNA levels in MARTHA will be available for download at https://zenodo.org/.

Altogether, this study produced a rich source of information relating to plasma miRNAs and biological/clinical traits associated with VT that could be of great use to generate and/or validate new hypothesis.

Supplementary material

Supplementary material is available at *European Heart Journal-Supplement* online.

Table 5 Significant ($P < 5 \times 10^{-8}$) associations of miSNP with miRNA plasma levels derived from the MARTHA miRNA and Nikpay *et al.*[31] meta-analysis

					MARTHA					Nikpay				Combined			
miRNA	chr	Position (bp)	SNP	EA	EAF	r^2	β	SE	*P*	EAF	β	SE	*P*	P^a	β	SE	P^b
Cis associations																	
miR-181b-5p	1	199257141	rs878254	A	0.485	0.90	−0.054	0.032	0.0916	0.480	−0.071	0.013	$1.64\ 10^{-7}$	0.61	−0.069	0.012	$3.18\ 10^{-8}$
miR-148a-3p	7	25991977	rs9639523	T	0.375	0.87	−0.081	0.034	0.0191	0.344	−0.072	0.013	$2.03\ 10^{-7}$	0.80	−0.073	0.013	$8.41\ 10^{-9}$
let-7a-5p	9	96916230	rs10512230	T	0.287	1.00	0.040	0.031	0.1934	0.315	0.026	0.004	$6.49\ 10^{-8}$	0.67	0.027	0.005	$2.19\ 10^{-8}$
let-7d-5p	9	97229465	rs4497033	T	0.492	0.99	−0.061	0.036	0.0895	0.463	−0.028	0.005	$1.50\ 10^{-7}$	0.36	−0.029	0.005	$3.85\ 10^{-8}$
miR-2110	10	115933905	rs17091403	T	0.091	1.00	−0.141	0.043	$1.13\ 10^{-3}$	0.074	−0.103	0.023	$9.90\ 10^{-6}$	0.44	−0.112	0.020	$4.34\ 10^{-8}$
miR-342-3p	14	100256449	rs8011282	C	0.474	0.99	0.095	0.030	$1.39\ 10^{-3}$	0.487	0.067	0.014	$5.65\ 10^{-6}$	0.41	0.073	0.013	$3.68\ 10^{-8}$
miR-99b-5p	19	52160843	rs11084100	C	0.392	1.00	−0.067	0.024	$5.17\ 10^{-3}$	0.419	−0.065	0.012	$1.12\ 10^{-7}$	0.94	−0.066	0.011	$1.50\ 10^{-9}$
Trans associations																	
miR-215-5p	2	171402733	rs724806	C	0.252	0.97	0.091	0.057	0.1123	0.326	0.143	0.027	$1.44\ 10^{-7}$	0.40	0.134	0.024	$4.09\ 10^{-8}$
miR-10b-5p	7	13236107	rs6948643	G	0.264	1.00	−0.071	0.040	0.0766	0.285	−0.09	0.017	$2.84\ 10^{-7}$	0.66	−0.087	0.016	$4.62\ 10^{-8}$
let-7d-3p	11	2611449	rs1024164	A	0.133	0.87	−0.083	0.034	0.0147	0.092	−0.065	0.013	$7.78\ 10^{-7}$	0.63	−0.068	0.012	$3.18\ 10^{-8}$
miR-378a-3p	11	133763476	rs10894759	A	0.317	0.99	0.066	0.028	0.0206	0.296	0.059	0.011	$7.86\ 10^{-7}$	0.82	0.060	0.011	$3.58\ 10^{-8}$
miR-7-5p	15	41614621	rs7163989	G	0.293	0.99	−0.112	0.041	$6.68\ 10^{-3}$	0.278	−0.089	0.016	$1.48\ 10^{-7}$	0.61	−0.093	0.016	$2.70\ 10^{-9}$

EAF, estimated allele frequency; r^2, imputation quality criterion; β, allele effect.
[a]*P*-value of the test for heterogeneity between the MARTHA and Nikpay studies.
[b]*P*-value of the combined effect obtained through a random-effect meta-analysis of the results of both studies.

Funding

F.T., G.M., and M.G. were financially supported by the GENMED Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013). D.A.T. was financially supported by the «EPIDEMIOM-VTE» Senior Chair from the Initiative of Excellence of the University of Bordeaux. MiRNA sequencing in the MARTHA study was performed on the iGenSeq platform (ICM Institute, Paris) and supported by a grant from the European Society of Cardiology for Medical Research Innovation. Bioinformatics and statistical analyses benefit from the CBiB computing centre of the University of Bordeaux. This paper was published as part of a supplement supported by an educational grant from Boehringer Ingelheim.

Conflict of interest: none declared.

References

1. Goldhaber SZ. Venous thromboembolism: epidemiology and magnitude of the problem. *Best Pract Res Clin Haematol* 2012;**25**:235–242.
2. Galanaud J-P, Monreal M, Kahn SR. Epidemiology of the post-thrombotic syndrome. *Thromb Res* 2018;**164**:100–109.
3. White RH. The epidemiology of venous thromboembolism. *Circulation* 2003;**107**:41–48.
4. Prandoni P, Bernardi E, Marchiori A, Lensing AWA, Prins MH, Villalta S, Bagatella P, Sartor D, Piccioli A, Simioni P, Pagnan A, Girolami A. The long term clinical course of acute deep vein thrombosis of the arm: prospective cohort study. *BMJ* 2004;**329**:484–485.
5. Kearon C, Parpia S, Spencer FA, Schulman S, Stevens SM, Shah V, Bauer KA, Douketis JD, Lentz SR, Kessler CM, Connors JM, Ginsberg JS, Spadafora L, Julian JA. Long-term risk of recurrence in patients with a first unprovoked venous thromboembolism managed according to d-dimer results; a cohort study. *J Thromb Haemost* 2019;**17**:1144–1152.
6. Bartel DP. Metazoan microRNAs. *Cell* 2018;**173**:20–51.
7. McManus DD, Freedman JE. MicroRNAs in platelet function and cardiovascular disease. *Nat Rev Cardiol* 2015;**12**:711–717.
8. Marchand A, Proust C, Morange P-E, Lompré A-M, Trégouët D-A. miR-421 and miR-30c inhibit SERPINE 1 gene expression in human endothelial cells. *PLoS One* 2012;**7**:e44532.
9. Arroyo AB, Los Reyes-García AM, de Teruel-Montoya R, Vicente V, González-Conejero R, Martínez C. microRNAs in the haemostatic system: more than witnesses of thromboembolic diseases? *Thromb Res* 2018;**166**:1–9.
10. Vossen CY, Hylckama Vlieg A, van Teruel-Montoya R, Salloum-Asfar S, Haan H, de Corral J, Reitsma P, Koeleman BPC, Martínez C. Identification of coagulation gene 3'UTR variants that are potentially regulated by microRNAs. *Br J Haematol* 2017;**177**:782–790.
11. Sennblad B, Basu S, Mazur J, Suchon P, Martinez-Perez A, Hylckama Vlieg A, van Truong V, Li Y, Gådin JR, Tang W, Grossman V, Haan HG, de Handin N, Silveira A, Souto JC, Franco-Cereceda A, Morange P-E, Gagnon F, Soria JM, Eriksson P, Hamsten A, Maegdefessel L, Rosendaal FR, Wild P, Folsom AR, Trégouët D-A, Sabater-Lleal M. Genome-wide association study with additional genetic and post-transcriptional analyses reveals novel regulators of plasma factor XI levels. *Hum Mol Genet* 2017;**26**:637–649.
12. Starikova I, Jamaly S, Sorrentino A, Blondal T, Latysheva N, Sovershaev M, Hansen J-B. Differential expression of plasma miRNAs in patients with unprovoked venous thromboembolism and healthy control individuals. *Thromb Res* 2015;**136**:566–572.
13. Wang X, Sundquist K, Svensson PJ, Rastkhani H, Palmér K, Memon AA, Sundquist J, Zöller B. Association of recurrent venous thromboembolism and circulating microRNAs. *Clin Epigenetics* 2019;**11**:28.
14. Oudot-Mellakh T, Cohen W, Germain M, Saut N, Kallel C, Zelenika D, Lathrop M, Trégouët D-A, Morange P-E. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. *Br J Haematol* 2012;**157**:230–239.
15. Germain M, Saut N, Oudot-Mellakh T, Letenneur L, Dupuy A-M, Bertrand M, Alessi M-C, Lambert J-C, Zelenika D, Emmerich J, Tiret L, Cambien F, Lathrop M, Amouyel P, Morange P-E, Trégouët D-A. 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**:e38538.
16. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng L-C, de Andrade M, de Visser MCH, Wiggins KL, Suchon P, Saut N, Smadja DM, Le Gal G, van Hylckama Vlieg A, Di Narzo A, Hao K, Nelson CP, Rocanin-Arjo A, Folkersen L, Monajemi R, Rose LM, Brody JA, Slagboom E, Aïssi D, Gagnon F, Deleuze J-F, Deloukas P, Tzourio C, Dartigues J-F, Berr C, Taylor KD, Civelek M, Eriksson P, Psaty BM, Houwing-Duitermaat J, Goodall AH, Cambien F, Kraft P, Amouyel P, Samani NJ, Basu S, Ridker PM, Rosendaal FR, Kabrhel C, Folsom AR, Heit J, Reitsma PH, Trégouët D-A, Smith NL, Morange P-E. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet* 2015;**96**:532–542.
17. Thibord F, Perret C, Roux M, Suchon P, Germain M, Deleuze J-F, Morange P-E, Trégouët D-A; GENMED Consortium. OPTIMIR, a novel algorithm for integrating available genome-wide genotype data into miRNA sequence alignment analysis. *RNA* 2019;**25**:657–668.
18. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;**42**:D68–D73.
19. Ameres SL, Zamore PD. Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol* 2013;**14**:475–488.
20. Kirschner MB, Edelman JJB, Kao S-H, Vallely MP, Van Zandwijk N, Reid G. The impact of hemolysis on cell-free microRNA. *Biomarkers. Front Genet* 2013;**4**:94.
21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
22. Ramachandran P, Sánchez-Taltavull D, Perkins TJ. Uncovering robust patterns of microRNA co-expression across cancers using Bayesian Relevance Networks. *PLoS One* 2017;**12**:e0183103.
23. Töpner K, Rosa GJM, Gianola D, Schön C-C. Bayesian networks illustrate genomic and residual trait connections in maize (Zea mays L.). *G3 GenesGenomesGenetics* 2017;**7**:2779–2789.
24. Scutari M. Learning Bayesian Networks with the bnlearn R Package. *J Stat Softw* 2010;**35**:1–22.
25. Florijn BW, Bijkerk R, van der Veer EP, van Zonneveld AJ. Gender and cardiovascular disease: are sex-biased microRNA networks a driving force behind heart failure with preserved ejection fraction in women? *Cardiovasc Res* 2018;**114**:210–225.
26. Huan T, Chen G, Liu C, Bhattacharya A, Rong J, Chen BH, Seshadri S, Tanriverdi K, Freedman JE, Larson MG, Murabito JM, Levy D. Age-associated microRNA expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell* 2018;**17**:e12687.
27. Chen X, Liang H, Guan D, Wang C, Hu X, Cui L, Chen S, Zhang C, Zhang J, Zen K, Zhang C-Y. A combination of Let-7d, Let-7g and Let-7i serves as a stable reference for normalization of serum microRNAs. *PLoS One* 2013;**8**:e79652.
28. Tsai W-Y, Jewell NP, Wang M-C. A note on the product-limit estimator under right censoring and left truncation. *Biometrika* 1987;**74**:883–886.
29. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 2005;**95**:221–227.
30. 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**:816–834.
31. Nikpay M, Beehler K, Valsesia A, Hager J, Harper M-E, Dent R, McPherson R. Genome-wide identification of circulating-miRNA expression quantitative trait loci reveals the role of several miRNAs in the regulation of Cardiometabolic phenotypes. *Cardiovasc Res* 2019;**115**:1629–1645.
32. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 2010;**11**:288.
33. Rubio M, Bustamante M, Hernandez-Ferrer C, Fernandez-Orth D, Pantano L, Sarria Y, Piqué-Borras M, Vellve K, Agramunt S, Carreras R, Estivill X, Gonzalez JR, Mayor A. Circulating miRNAs, isomiRs and

small RNA clusters in human plasma and breast milk. *PLoS One* 2018; **13**:e0193527.
34. Braekkan SK, Mathiesen EB, Njølstad I, Wilsgaard T, Hansen J-B. Hematocrit and risk of venous thromboembolism in a general population. The Tromso study. *Haematologica* 2010;**95**:270–275.
35. Rezende SM, Lijfering WM, Rosendaal FR, Cannegieter SC. Hematologic variables and venous thrombosis: red cell distribution width and blood monocyte count are associated with an increased risk. *Haematologica* 2014;**99**:194–200.
36. Huan T, Rong J, Liu C, Zhang X, Tanriverdi K, Joehanes R, Chen BH, Murabito JM, Yao C, Courchesne P, Munson PJ, O'Donnell CJ, Cox N, Johnson AD, Larson MG, Levy D, Freedman JE. Genome-wide identification of microRNA expression quantitative trait loci. *Nat Commun* 2015;**6**:6601.
37. Civelek M, Hagopian R, Pan C, Che N, Yang W, Kayne PS, Saleem NK, Cederberg H, Kuusisto J, Gargalovic PS, Kirchgessner TG, Laakso M, Lusis AJ. Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits. *Hum Mol Genet* 2013; **22**:3023–3037.
38. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK, Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, Bradley JR, Janjic N, Roberts DJ, Ouwehand WH, Todd JA, Soranzo N, Suhre K, Paul DS, Fox CS, Plenge RM, Danesh J, Runz H, Butterworth AS. Genomic atlas of the human plasma proteome. *Nature* 2018;**558**:73–79.
39. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET. Population genomics of human gene expression. *Nat Genet* 2007;**39**:1217–1224.
40. Wang Y-P, Li K-B. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics* 2009;**10**:218.
41. GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet* 2013;**45**:580–585.
42. Klarin D, Emdin CA, Natarajan P, Conrad MF, Kathiresan S. Genetic analysis of venous thromboembolism in UK Biobank identifies the ZFPM2 locus and implicates obesity as a causal risk factor. *Circ Cardiovasc Genet* 2017;**10**. pii: e001643.
43. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, Lambourne JJ, Sivapalaratnam S, Downes K, Kundu K, Bomba L, Berentsen K, Bradley JR, Daugherty LC, Delaneau O, Freson K, Garner SF, Grassi L, Guerrero J, Haimel M, Janssen-Megens EM, Kaan A, Kamat M, Kim B, Mandoli A, Marchini J, Martens JHA, Meacham S, Megy K, O'Connell J, Petersen R, Sharifi N, Sheard SM, Staley JR, Tuna S, van der Ent M, Walter K, Wang S-Y, Wheeler E, Wilder SP, Iotchkova V, Moore C, Sambrook J, Stunnenberg HG, Di Angelantonio E, Kaptoge S, Kuijpers TW, Carrillo-de-Santa-Pau E, Juan D, Rico D, Valencia A, Chen L, Ge B, Vasquez L, Kwan T, Garrido-Martín D, Watt S, Yang Y, Guigo R, Beck S, Paul DS, Pastinen T, Bujold D, Bourque G, Frontini M, Danesh J, Roberts DJ, Ouwehand WH, Butterworth AS, Soranzo N. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* 2016;**167**:1415–1429.e19.
44. Lawler PR, Lawler J. Molecular basis for the regulation of angiogenesis by thrombospondin-1 and -2. *Cold Spring Harb Perspect Med* 2012;**2**:a006627.
45. Trumel C, Plantavid M, Lévy-Tolédano S, Ragab A, Caen JP, Aguado E, Malissen B, Payrastre B. Platelet aggregation induced by the C-terminal peptide of thrombospondin-1 requires the docking protein LAT but is largely independent of alphaIIb/beta3. *J Thromb Haemost* 2003;**1**:320–329.
46. Iliopoulos D, Drosatos K, Hiyama Y, Goldberg IJ, Zannis VI. MicroRNA-370 controls the expression of microRNA-122 and Cpt1alpha and affects lipid metabolism. *J Lipid Res* 2010;**51**:1513–1523.
47. Gao W, He H-W, Wang Z-M, Zhao H, Lian X-Q, Wang Y-S, Zhu J, Yan J-J, Zhang D-G, Yang Z-J, Wang L-S. Plasma levels of lipometabolism-related miR-122 and miR-370 are increased in patients with hyperlipidemia and associated with coronary artery disease. *Lipids Health Dis* 2012;**11**:55.
48. Benatti RO, Melo AM, Borges FO, Ignacio-Souza LM, Simino L. A P, Milanski M, Velloso LA, Torsoni MA, Torsoni AS. Maternal high-fat diet consumption modulates hepatic lipid metabolism and microRNA-122 (miR-122) and microRNA-370 (miR-370) expression in offspring. *Br J Nutr* 2014;**111**:2112–2122.
49. Tian D, Sha Y, Lu J-M, Du X-J. MiR-370 inhibits vascular inflammation and oxidative stress triggered by oxidized low-density lipoprotein through targeting TLR4. *J Cell Biochem* 2018; **119**:6231–6237.
50. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H, Chiew M-Y, Tai C-S, Wei T-Y, Tsai T-R, Huang H-T, Wang C-Y, Wu H-Y, Ho S-Y, Chen P-R, Chuang C-H, Hsieh P-J, Wu Y-S, Chen W-L, Li M-J, Wu Y-C, Huang X-Y, Ng FL, Buddhakosai W, Huang P-C, Lan K-C, Huang C-Y, Weng S-L, Cheng Y-N, Liang C, Hsu W-L, Huang H-D. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* 2018;**46**:D296–D302.
51. Gagnon F, Aïssi D, Carrié A, Morange P-E, Trégouët D-A. Robust validation of methylation levels association at CPT1A locus with lipid plasma levels1. *J Lipid Res* 2014;**55**:1189–1191.
52. Frazier-Wood AC, Aslibekyan S, Absher DM, Hopkins PN, Sha J, Tsai MY, Tiwari HK, Waite LL, Zhi D, Arnett DK. Methylation at CPT1A locus is associated with lipoprotein subfraction profiles. *J Lipid Res* 2014;**55**:1324–1330.
53. Irvin MR, Zhi D, Joehanes R, Mendelson M, Aslibekyan S, Claas SA, Thibeault KS, Patel N, Day K, Jones LW, Liang L, Chen BH, Yao C, Tiwari HK, Ordovas JM, Levy D, Absher D, Arnett DK. Epigenome-wide association study of fasting blood lipids in the Genetics of Lipid-lowering Drugs and Diet Network study. *Circulation* 2014; **130**:565–572.
54. Sonoda T, Matsuzaki J, Yamamoto Y, Sakurai T, Aoki Y, Takizawa S, Niida S, Ochiya T. Serum microRNA-based risk prediction for stroke. *Stroke* 2019;**50**:1510–1518.

Références

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Abd ElHafeez, S., D'Arrigo, G., Leonardis, D., Fusaro, M., Tripepi, G., & Roumeliotis, S. (2021). Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. *Oxidative Medicine and Cellular Longevity*, *2021*, 1–6. https://doi.org/10.1155/2021/1302811

Alonso, A., Seguí-Gómez, M., de Irala, J., Sánchez-Villegas, A., Beunza, J. J., & Martínez-Gonzalez, M. Á. (2006). Predictors of follow-up and assessment of selection bias from dropouts using inverse probability weighting in a cohort of university graduates. *European Journal of Epidemiology*, *21*(5), 351–358. https://doi.org/10.1007/s10654-006-9008-y

Angelini, D. E., Kaatz, S., Rosovsky, R. P., Zon, R. L., Pillai, S., Robertson, W. E., Elavalakanar, P., Patell, R., & Khorana, A. (2022). COVID-19 and venous thromboembolism: A narrative review. *Research and Practice in Thrombosis and Haemostasis*, *6*(2), e12666. https://doi.org/10.1002/rth2.12666

Antoni, G., Oudot-Mellakh, T., Dimitromanolakis, A., Germain, M., Cohen, W., Wells, P., Lathrop, M., Gagnon, F., Morange, P.-E., & Tregouet, D.-A. (2011). Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. *BMC Medical Genetics*, *12*, 102. https://doi.org/10.1186/1471-2350-12-102

Arshad, N., Isaksen, T., Hansen, J.-B., & Brækkan, S. K. (2017). Time trends in incidence rates of venous thromboembolism in a large cohort recruited from the general population. *European Journal of Epidemiology*, *32*(4), 299–305. https://doi.org/10.1007/s10654-017-0238-y

Aulchenko, Y. S., Struchalin, M. V., & van Duijn, C. M. (2010). ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, *11*(1), 134. https://doi.org/10.1186/1471-2105-11-134

Authors/Task Force Members, Konstantinides, S. V., Torbicki, A., Agnelli, G., Danchin, N., Fitzmaurice, D., Galiè, N., Gibbs, J. S. R., Huisman, M. V., Humbert, M., Kucher, N., Lang, I., Lankeit, M., Lekakis, J., Maack, C., Mayer, E., Meneveau, N., Perrier, A., Pruszczyk, P., … Spyropoulos, A. C. (2014). 2014 ESC Guidelines on the diagnosis and management of acute pulmonary embolism. *European Heart Journal*, *35*(43), 3033–3080. https://doi.org/10.1093/eurheartj/ehu283

Avnery, O., Martin, M., Bura-Riviere, A., Barillari, G., Mazzolai, L., Mahé, I., Marchena, P. J., Verhamme, P., Monreal, M., Ellis, M. H., & RIETE Investigators. (2020). D-dimer levels and risk of recurrence following provoked venous thromboembolism: Findings from the RIETE registry. *Journal of Internal Medicine*, *287*(1), 32–41. https://doi.org/10.1111/joim.12969

Ayed Alanazi, O., Mohamed Abo El-Fetoh, N., Ali Mohammed, N., Mozil Aquab Alanizy, T., Wadi Alanazi, Y., Saleh Alanazi, M., Aiash Alrwaili, A., Hamoud Alruwaili, A., Hussain Alanazi, A., & Saad Alanazi, A. (2017). Deep Venous Thrombosis among hypertensive patients in King

Abdulaziz University (KAU) Hospital, Jeddah, Kingdom of Saudi Arabia. *Electronic Physician*, *9*(10), 5472–5477. https://doi.org/10.19082/5472

Basic Steps in Hemostasis—Nurseinfo. (2020, February 22). https://nurseinfo.in/basic-steps-in-hemostasis/

Bauer, K. A. (2003). Management of thrombophilia. *Journal of Thrombosis and Haemostasis: JTH*, *1*(7), 1429–1434. https://doi.org/10.1046/j.1538-7836.2003.00274.x

Baylis, R. A., Smith, N. L., Klarin, D., & Fukaya, E. (2021). Epidemiology and Genetics of Venous Thromboembolism and Chronic Venous Disease. *Circulation Research*, *128*(12), 1988–2002. https://doi.org/10.1161/CIRCRESAHA.121.318322

Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A. C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., Holmans, P. A., Boland, A., Damotte, V., van der Lee, S. J., Costa, M. R., Kuulasmaa, T., Yang, Q., de Rojas, I., Bis, J. C., … Lambert, J.-C. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics*, *54*(4), Article 4. https://doi.org/10.1038/s41588-022-01024-z

Bezemer, I. D., Bare, L. A., Doggen, C. J. M., Arellano, A. R., Tong, C., Rowland, C. M., Catanese, J., Young, B. A., Reitsma, P. H., Devlin, J. J., & Rosendaal, F. R. (2008). Gene Variants Associated With Deep Vein Thrombosis. *JAMA*, *299*(11), 1306–1314. https://doi.org/10.1001/jama.299.11.1306

Bezemer, I. D., van der Meer, F. J. M., Eikenboom, J. C. J., Rosendaal, F. R., & Doggen, C. J. M. (2009). The Value of Family History as a Risk Indicator for Venous Thrombosis. *Archives of Internal Medicine*, *169*(6), 610–615. https://doi.org/10.1001/archinternmed.2008.589

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ (Clinical Research Ed.)*, *310*(6973), 170. https://doi.org/10.1136/bmj.310.6973.170

Blom, J. W., Doggen, C. J. M., Osanto, S., & Rosendaal, F. R. (2005). Malignancies, prothrombotic mutations, and the risk of venous thrombosis. *JAMA*, *293*(6), 715–722. https://doi.org/10.1001/jama.293.6.715

Bohrer, A. C., Castro, E., Tocheny, C. E., Assmann, M., Schwarz, B., Bohrnsen, E., Makiya, M. A., Legrand, F., Hilligan, K. L., Baker, P. J., Torres-Juarez, F., Hu, Z., Ma, H., Wang, L., Niu, L., Wen, Z., Lee, S. H., Kamenyeva, O., Tuberculosis Imaging Program, … Mayer-Barber, K. D. (2022). Rapid GPR183-mediated recruitment of eosinophils to the lung after Mycobacterium tuberculosis infection. *Cell Reports*, *40*(4), 111144. https://doi.org/10.1016/j.celrep.2022.111144

Bonaventura, A., Vecchié, A., Dagna, L., Martinod, K., Dixon, D. L., Van Tassell, B. W., Dentali, F., Montecucco, F., Massberg, S., Levi, M., & Abbate, A. (2021). Endothelial dysfunction and immunothrombosis as key pathogenic mechanisms in COVID-19. *Nature Reviews Immunology*, *21*(5), Article 5. https://doi.org/10.1038/s41577-021-00536-9

Børsting, C., & Morling, N. (2013). Single-Nucleotide Polymorphisms. In *Encyclopedia of Forensic Sciences* (pp. 233–238). Elsevier. https://doi.org/10.1016/B978-0-12-382165-2.00042-8

Bruzelius, M., Iglesias, M. J., Hong, M.-G., Sanchez-Rivera, L., Gyorgy, B., Souto, J. C., Frånberg, M., Fredolini, C., Strawbridge, R. J., Holmström, M., Hamsten, A., Uhlén, M., Silveira, A., Soria, J. M., Smadja, D. M., Butler, L. M., Schwenk, J. M., Morange, P.-E., Trégouët, D.-A., &

Odeberg, J. (2016). PDGFB, a new candidate plasma biomarker for venous thromboembolism: Results from the VEREMA affinity proteomics study. *Blood*, *128*(23), e59–e66. https://doi.org/10.1182/blood-2016-05-711846

Bucciarelli, P., Passamonti, S. M., Biguzzi, E., Gianniello, F., Franchi, F., Mannucci, P. M., & Martinelli, I. (2012). Low borderline plasma levels of antithrombin, protein C and protein S are risk factors for venous thromboembolism. *Journal of Thrombosis and Haemostasis*, *10*(9), 1783–1791. https://doi.org/10.1111/j.1538-7836.2012.04858.x

Buil, A., Trégouët, D.-A., Souto, J. C., Saut, N., Germain, M., Rotival, M., Tiret, L., Cambien, F., Lathrop, M., Zeller, T., Alessi, M.-C., Rodriguez de Cordoba, S., Münzel, T., Wild, P., Fontcuberta, J., Gagnon, F., Emmerich, J., Almasy, L., Blankenberg, S., … Morange, P.-E. (2010). C4BPB/C4BPA is a new susceptibility locus for venous thrombosis with unknown protein S-independent mechanism: Results from genome-wide association and gene expression analyses followed by case-control studies. *Blood*, *115*(23), 4644–4650. https://doi.org/10.1182/blood-2010-01-263038

Bunce, P. E., High, S. M., Nadjafi, M., Stanley, K., Liles, W. C., & Christian, M. D. (2011). Pandemic H1N1 Influenza Infection and Vascular Thrombosis. *Clinical Infectious Diseases*, *52*(2), e14–e17. https://doi.org/10.1093/cid/ciq125

Burgess, S., Timpson, N. J., Ebrahim, S., & Davey Smith, G. (2015). Mendelian randomization: Where are we now and where are we going? *International Journal of Epidemiology*, *44*(2), 379–388. https://doi.org/10.1093/ije/dyv108

Canobbio, I., Visconte, C., Momi, S., Guidetti, G. F., Zarà, M., Canino, J., Falcinelli, E., Gresele, P., & Torti, M. (2017). Platelet amyloid precursor protein is a modulator of venous thromboembolism in mice. *Blood*, *130*(4), 527–536. https://doi.org/10.1182/blood-2017-01-764910

Chalal, N., & Demmouche, A. (2013). Venous thromboembolic disease in the region of Sidi Bel Abbes, Algeria: Frequency and risk factors. *The Pan African Medical Journal*, *16*, 45. https://doi.org/10.11604/pamj.2013.16.45.2620

Christiansen, C. F., Schmidt, M., Lamberg, A. L., Horváth-Puhó, E., Baron, J. A., Jespersen, B., & Sørensen, H. T. (2014). Kidney disease and risk of venous thromboembolism: A nationwide population-based case-control study. *Journal of Thrombosis and Haemostasis*, *12*(9), 1449–1454. https://doi.org/10.1111/jth.12652

Cochran, W. G. (1954). The Combination of Estimates from Different Experiments. *Biometrics*, *10*(1), 101–129. https://doi.org/10.2307/3001666

Comp, P. C., & Esmon, C. T. (1984). Recurrent venous thromboembolism in patients with a partial deficiency of protein S. *The New England Journal of Medicine*, *311*(24), 1525–1528. https://doi.org/10.1056/NEJM198412133112401

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., & Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science (New York, N.Y.)*, *261*(5123), 921–923. https://doi.org/10.1126/science.8346443

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Dahlbäck, B., Carlsson, M., & Svensson, P. J. (1993). Familial thrombophilia due to a previously unrecognized mechanism characterized by poor anticoagulant response to activated protein C: Prediction of a cofactor to activated protein C. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(3), 1004–1008.

Danckwardt, S., Hentze, M. W., & Kulozik, A. E. (2013). Pathologies at the nexus of blood coagulation and inflammation: Thrombin in hemostasis, cancer, and beyond. *Journal of Molecular Medicine*, *91*(11), 1257–1271. https://doi.org/10.1007/s00109-013-1074-5

de Haan, H. G., van Hylckama Vlieg, A., Germain, M., Baglin, T. P., Deleuze, J.-F., Trégouët, D.-A., & Rosendaal, F. R. (2018). Genome-Wide Association Study Identifies a Novel Genetic Risk Factor for Recurrent Venous Thrombosis. *Circulation: Genomic and Precision Medicine*, *11*(2). https://doi.org/10.1161/CIRCGEN.117.001827

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, *49*(2), 256–261. https://doi.org/10.1038/ng.3760

Deitelzweig, S. B., Johnson, B. H., Lin, J., & Schulman, K. L. (2011). Prevalence of clinical venous thromboembolism in the USA: Current trends and future projections. *American Journal of Hematology*, *86*(2), 217–220. https://doi.org/10.1002/ajh.21917

Delluc, A., Tromeur, C., Le Ven, F., Gouillou, M., Paleiron, N., Bressollette, L., Nonent, M., Salaun, P.-Y., Lacut, K., Leroyer, C., Le Gal, G., Couturaud, F., Mottier, D., & EPIGETBO study group. (2016). Current incidence of venous thromboembolism and comparison with 1998: A community-based study in Western France. *Thrombosis and Haemostasis*, *116*(5), 967–974. https://doi.org/10.1160/TH16-03-0205

Dennis, J., Johnson, C. Y., Adediran, A. S., de Andrade, M., Heit, J. A., Morange, P.-E., Trégouët, D.-A., & Gagnon, F. (2012). The endothelial protein C receptor (PROCR) Ser219Gly variant and risk of common thrombotic disorders: A HuGE review and meta-analysis of evidence from observational studies. *Blood*, *119*(10), 2392–2400. https://doi.org/10.1182/blood-2011-10-383448

Denorme, F., Portier, I., Rustad, J. L., Cody, M. J., Araujo, C. V. de, Hoki, C., Alexander, M. D., Grandhi, R., Dyer, M. R., Neal, M. D., Majersik, J. J., Yost, C. C., & Campbell, R. A. (2022). Neutrophil extracellular traps regulate ischemic stroke brain injury. *The Journal of Clinical Investigation*, *132*(10). https://doi.org/10.1172/JCI154225

Dentali, F., & Franchini, M. (2013). Recurrent venous thromboembolism: A role for ABO blood group? *Thrombosis and Haemostasis*, *110*(12), 1110–1111. https://doi.org/10.1160/TH13-09-0780

Didelez, V., & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, *16*(4), 309–330. https://doi.org/10.1177/0962280206077743

Domeij-Arverud, E., Labruto, F., Latifi, A., Nilsson, G., Edman, G., & Ackermann, P. W. (2015). Intermittent pneumatic compression reduces the risk of deep vein thrombosis during post-operative lower limb immobilisation. *The Bone & Joint Journal*, *97-B*(5), 675–680. https://doi.org/10.1302/0301-620X.97B5.34581

Duncan, L. E., Ostacher, M., & Ballon, J. (2019). How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology*, *44*(9), Article 9. https://doi.org/10.1038/s41386-019-0389-5

Duthé, G., Samuel, O., & Solaz, A. (2021). Dynamiques, enjeux démographiques et socioéconomiques du vieillissement dans les pays à longévité élevée. *Population*, *Vol. 76*(2), 223–224. https://doi.org/10.3917/popu.2102.0223

Egeberg, O. (1965). Inherited Antithrombin Deficiency Causing Thrombophilia. *Thrombosis and Haemostasis*, *13*(2), 516–530. https://doi.org/10.1055/s-0038-1656297

Ehrenberg, A. S. C. (1959). The Pattern of Consumer Purchases. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *8*(1), 26–41. https://doi.org/10.2307/2985810

Eichinger, S., Heinze, G., Jandeck, L. M., & Kyrle, P. A. (2010). Risk assessment of recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism: The Vienna prediction model. *Circulation*, *121*(14), 1630–1636. https://doi.org/10.1161/CIRCULATIONAHA.109.925214

Enga, K. F., Brækkan, S. K., Hansen-Krone, I. J., le CESSIE, S., Rosendaal, F. R., & Hansen, J.-B. (2012). Cigarette smoking and the risk of venous thromboembolism: The Tromsø Study. *Journal of Thrombosis and Haemostasis*, *10*(10), 2068–2074. https://doi.org/10.1111/j.1538-7836.2012.04880.x

Engelmann, B., & Massberg, S. (2013). Thrombosis as an intravascular effector of innate immunity. *Nature Reviews. Immunology*, *13*(1), 34–45. https://doi.org/10.1038/nri3345

Farmer-Boatwright, M. K., & Roubey, R. A. S. (2009). Venous Thrombosis in the Antiphospholipid Syndrome. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *29*(3), 321–325. https://doi.org/10.1161/ATVBAHA.108.182204

Ferkingstad, E., Sulem, P., Atlason, B. A., Sveinbjornsson, G., Magnusson, M. I., Styrmisdottir, E. L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B. V., Jensson, B. O., Zink, F., Halldorsson, G. H., Masson, G., Arnadottir, G. A., Katrinardottir, H., Juliusson, K., Magnusson, M. K., Magnusson, O. Th., … Stefansson, K. (2021). Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics*, *53*(12), 1712–1721. https://doi.org/10.1038/s41588-021-00978-w

Ferluga, J., Kouser, L., Murugaiah, V., Sim, R. B., & Kishore, U. (2017). Potential influences of complement factor H in autoimmune inflammatory and thrombotic disorders. *Molecular Immunology*, *84*, 84–106. https://doi.org/10.1016/j.molimm.2017.01.015

Foster, S. D., & Bravington, M. V. (2013). A Poisson–Gamma model for analysis of ecological non-negative continuous data. *Environmental and Ecological Statistics*, *20*(4), 533–552. https://doi.org/10.1007/s10651-012-0233-0

Gándara, E., Kovacs, M. J., Kahn, S. R., Wells, P. S., Anderson, D. A., Chagnon, I., Gal, G., Solymoss, S., Crowther, M., Carrier, M., Langlois, N., Kovacs, J., Ma, J., Carson, N., Ramsay, T., & Rodger, M. A. (2013). Non-OO blood type influences the risk of recurrent venous thromboembolism: A cohort study. *Thrombosis and Haemostasis*, *110*(12), 1172–1179. https://doi.org/10.1160/TH13-06-0488

Gandy, A., & Veraart, L. A. M. (2021). Compound Poisson models for weighted networks with applications in finance. *Mathematics and Financial Economics*, *15*(1), 131–153. https://doi.org/10.1007/s11579-020-00268-9

García Raso, A., Ene, G., Miranda, C., Vidal, R., Mata, R., & Llamas Sillero, M. P. (2014). Association between venous thrombosis and dyslipidemia. *Medicina Clinica*, *143*(1), 1–5. https://doi.org/10.1016/j.medcli.2013.07.024

Gauthier, M., Agniel, D., Thiébaut, R., & Hejblum, B. P. (2020). dearseq: A variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics*, *2*(4), lqaa093. https://doi.org/10.1093/nargab/lqaa093

Germain, M., Chasman, D. I., de Haan, H., Tang, W., Lindström, S., Weng, L.-C., de Andrade, M., de Visser, M. C. H., Wiggins, K. L., Suchon, P., Saut, N., Smadja, D. M., Le Gal, G., van Hylckama Vlieg, A., Di Narzo, A., Hao, K., Nelson, C. P., Rocanin-Arjo, A., Folkersen, L., … Morange, P.-E. (2015). Meta-analysis of 65,734 Individuals Identifies TSPAN15 and SLC44A2 as Two Susceptibility Loci for Venous Thromboembolism. *American Journal of Human Genetics*, *96*(4), 532–542. https://doi.org/10.1016/j.ajhg.2015.01.019

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., Tam, P. K.-H., Tsui, L.-C., Waye, M. M. Y., Wong, J. T.-F., Zeng, C., Zhang, Q., Chee, M. S., Galver, L. M., Kruglyak, S., … Methods Group. (2003). The International HapMap Project. *Nature*, *426*(6968), Article 6968. https://doi.org/10.1038/nature02168

Glynn, R. J., Ridker, P. M., Goldhaber, S. Z., Zee, R. Y. L., & Buring, J. E. (2007). Effects of Random Allocation to Vitamin E Supplementation on the Occurrence of Venous Thromboembolism. *Circulation*, *116*(13), 1497–1503. https://doi.org/10.1161/CIRCULATIONAHA.107.716407

Goldhaber, S. Z. (2012). Venous thromboembolism: Epidemiology and magnitude of the problem. *Best Practice & Research. Clinical Haematology*, *25*(3), 235–242. https://doi.org/10.1016/j.beha.2012.06.007

Goumidi, L., Thibord, F., Wiggins, K. L., Li-Gao, R., Brown, M. R., van Hylckama Vlieg, A., Souto, J. C., Soria, J. M., Ibrahim-Kosta, M., Saut, N., Daian-Bacq, D., Olaso, R., Amouyel, P., Debette, S., Boland, A., Bailly, P., Morrison, A., Mook-Kanamori, D. O., Deleuze, J.-F., … Morange, P.-E. (2020). Association of ABO haplotypes with the risk of venous thrombosis: Impact on disease risks estimation. *Blood*, *blood.2020008997*. https://doi.org/10.1182/blood.2020008997

Gregson, J., Kaptoge, S., Bolton, T., Pennells, L., Willeit, P., Burgess, S., Bell, S., Sweeting, M., Rimm, E. B., Kabrhel, C., Zöller, B., Assmann, G., Gudnason, V., Folsom, A. R., Arndt, V., Fletcher, A., Norman, P. E., Nordestgaard, B. G., Kitamura, A., … Emerging Risk Factors Collaboration. (2019). Cardiovascular Risk Factors Associated With Venous Thromboembolism. *JAMA Cardiology*, *4*(2), 163–173. https://doi.org/10.1001/jamacardio.2018.4537

Gregson, J., Kaptoge, S., Bolton, T., Pennells, L., Willeit, P., Burgess, S., Bell, S., Sweeting, M., Rimm, E. B., Kabrhel, C., Zöller, B., Assmann, G., Gudnason, V., Folsom, A. R., Arndt, V., Fletcher, A., Norman, P. E., Nordestgaard, B. G., Kitamura, A., … for the Emerging Risk Factors Collaboration. (2019). Cardiovascular Risk Factors Associated With Venous Thromboembolism. *JAMA Cardiology*, *4*(2), 163. https://doi.org/10.1001/jamacardio.2018.4537

Griffin, J. H., Evatt, B., Zimmerman, T. S., Kleiss, A. J., & Wideman, C. (1981). Deficiency of protein C in congenital thrombotic disease. *Journal of Clinical Investigation*, *68*(5), 1370–1373.

Grimnes, G., Isaksen, T., Tichelaar, Y. I. G. V., Brox, J., Brækkan, S. K., & Hansen, J.-B. (2018). C-reactive protein and risk of venous thromboembolism: Results from a population-based case-crossover study. *Haematologica*, *103*(7), 1245–1250. https://doi.org/10.3324/haematol.2017.186957

Harbin, M. M., & Lutsey, P. L. (2020). May-Thurner syndrome: History of understanding and need for defining population prevalence. *Journal of Thrombosis and Haemostasis*, *18*(3), 534–542. https://doi.org/10.1111/jth.14707

Hawez, A., Al-Haidari, A., Madhi, R., Rahman, M., & Thorlacius, H. (2019). MiR-155 Regulates PAD4-Dependent Formation of Neutrophil Extracellular Traps. *Frontiers in Immunology*, *10*, 2462. https://doi.org/10.3389/fimmu.2019.02462

Hawez, A., Taha, D., Algaber, A., Madhi, R., Rahman, M., & Thorlacius, H. (2022). MiR-155 regulates neutrophil extracellular trap formation and lung injury in abdominal sepsis. *Journal of Leukocyte Biology*, *111*(2), 391–400. https://doi.org/10.1002/JLB.3A1220-789RR

Houghton, D., & Moll, S. (2017). HERDOO2 Score: How Long to Treat With Anticoagulation? *The Hematologist*, *14*(4). https://doi.org/10.1182/hem.V14.4.7468

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Inserm, La science pour la santé. (2017). *Thrombose veineuse (Phlébite) ·*. Inserm. https://www.inserm.fr/dossier/thrombose-veineuse-phlebite/

Inserm, La science pour la santé. (2022). *Thérapies à ARN*. Inserm. https://www.inserm.fr/dossier/therapies-a-arn/

Jadaon, M. M. (2011). Epidemiology of Prothrombin G20210A Mutation in the Mediterranean Region. *Mediterranean Journal of Hematology and Infectious Diseases*, *3*(1), e2011054. https://doi.org/10.4084/MJHID.2011.054

Jiao, X., Li, Z., An, S., Huang, J., Feng, M., & Cao, G. (2022). Does diabetes mellitus increase the incidence of early thrombosis in deep vein following unicompartmental knee arthroplasty: A retrospective cohort study. *BMC Geriatrics*, *22*(1), 448. https://doi.org/10.1186/s12877-022-03153-w

Jick, H., Slone, D., Westerholm, B., Inman, W. H., Vessey, M. P., Shapiro, S., Lewis, G. P., & Worcester, J. (1969). Venous thromboembolic disease and ABO blood type. A cooperative study. *Lancet (London, England)*, *1*(7594), 539–542. https://doi.org/10.1016/s0140-6736(69)91955-2

Jørgensen, B., & Paes de Souza, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuar. J.*, *1*.

Kals, M., Kunzmann, K., Parodi, L., Radmanesh, F., Wilson, L., Izzy, S., Anderson, C. D., Puccio, A. M., Okonkwo, D. O., Temkin, N., Steyerberg, E. W., Stein, M. B., Manley, G. T., Maas, A. I. R., Richardson, S., Diaz-Arrastia, R., Palotie, A., Ripatti, S., Rosand, J., … Zafonte, R. (2022).

A genome-wide association study of outcome from traumatic brain injury. *EBioMedicine*, *77*. https://doi.org/10.1016/j.ebiom.2022.103933

Kearon, C., Akl, E. A., Ornelas, J., Blaivas, A., Jimenez, D., Bounameaux, H., Huisman, M., King, C. S., Morris, T. A., Sood, N., Stevens, S. M., Vintch, J. R. E., Wells, P., Woller, S. C., & Moores, L. (2016). Antithrombotic Therapy for VTE Disease: CHEST Guideline and Expert Panel Report. *Chest*, *149*(2), 315–352. https://doi.org/10.1016/j.chest.2015.11.026

Kelly, J., Rudd, A., Lewis, R. R., & Hunt, B. J. (2002). Plasma D-Dimers in the Diagnosis of Venous Thromboembolism. *Archives of Internal Medicine*, *162*(7), 747–756. https://doi.org/10.1001/archinte.162.7.747

Kim, J.-K., Hong, C.-W., Park, M. J., Song, Y. R., Kim, H. J., & Kim, S. G. (2017). Increased Neutrophil Extracellular Trap Formation in Uremia Is Associated with Chronic Inflammation and Prevalent Coronary Artery Disease. *Journal of Immunology Research*, *2017*, 8415179. https://doi.org/10.1155/2017/8415179

Klarin, D., Busenkell, E., Judy, R., Lynch, J., Levin, M., Haessler, J., Aragam, K., Chaffin, M., Haas, M., Lindström, S., Assimes, T. L., Huang, J., Min Lee, K., Shao, Q., Huffman, J. E., Kabrhel, C., Huang, Y., Sun, Y. V., Vujkovic, M., … Veterans Affairs' Million Veteran Program. (2019). Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nature Genetics*, *51*(11), 1574–1579. https://doi.org/10.1038/s41588-019-0519-3

Klok, F. A., & Huisman, M. V. (2020). How I assess and manage the risk of bleeding in patients treated for venous thromboembolism. *Blood*, *135*(10), 724–734. https://doi.org/10.1182/blood.2019001605

Kreidy, R. (2015). Contribution of Recurrent Venous Thrombosis and Inherited Thrombophilia to the Pathogenesis of Postthrombotic Syndrome. *Clinical and Applied Thrombosis/Hemostasis*, *21*(1), 87–90. https://doi.org/10.1177/1076029613497423

Lacut, K., Oger, E., Le Gal, G., Couturaud, F., Louis, S., Leroyer, C., & Mottier, D. (2004). Statins but not fibrates are associated with a reduced risk of venous thromboembolism: A hospital-based case-control study. *Fundamental and Clinical Pharmacology*, *18*(4), 477–482. https://doi.org/10.1111/j.1472-8206.2004.00252.x

Laridan, E., Denorme, F., Desender, L., François, O., Andersson, T., Deckmyn, H., Vanhoorelbeke, K., & De Meyer, S. F. (2017). Neutrophil extracellular traps in ischemic stroke thrombi. *Annals of Neurology*, *82*(2), 223–232. https://doi.org/10.1002/ana.24993

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, *95*(3), Article 3. https://doi.org/10.1038/sj.hdy.6800717

Limperger, V., Kenet, G., Kiesau, B., Köther, M., Schmeiser, M., Langer, F., Juhl, D., Shneyder, M., Franke, A., Klostermeier, U. K., Mesters, R., Rühle, F., Stoll, M., Steppat, D., Kowalski, D., Rocke, A., Kuta, P., Bajorat, T., Torge, A., … Nowak-Göttl, U. (2021). Role of prothrombin 19911 A>G polymorphism, blood group and male gender in patients with venous thromboembolism: Results of a German cohort study. *Journal of Thrombosis and Thrombolysis*, *51*(2), 494–501. https://doi.org/10.1007/s11239-020-02169-6

Lindström, S., Wang, L., Smith, E. N., Gordon, W., van Hylckama Vlieg, A., de Andrade, M., Brody, J. A., Pattee, J. W., Haessler, J., Brumpton, B. M., Chasman, D. I., Suchon, P., Chen, M.-H., Turman, C., Germain, M., Wiggins, K. L., MacDonald, J., Braekkan, S. K., Armasu, S. M., … Smith, N. L. (2019). Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood*, *134*(19), 1645–1657. https://doi.org/10.1182/blood.2019000435

Lutsey, P. L., Cushman, M., Heckbert, S. R., Tang, W., & Folsom, A. R. (2011). Longer legs are associated with greater risk of incident venous thromboembolism independent of total body height: The Longitudinal Study of Thromboembolism Etiology (LITE). *Thrombosis and Haemostasis*, *106*(07), 113–120. https://doi.org/10.1160/TH11-02-0100

Lutsey, P. L., Norby, F. L., Alonso, A., Cushman, M., Chen, L. Y., Michos, E. D., & Folsom, A. R. (2018). Atrial fibrillation and venous thromboembolism: Evidence of bidirectionality in the Atherosclerosis Risk in Communities Study. *Journal of Thrombosis and Haemostasis*, *16*(4), 670–679. https://doi.org/10.1111/jth.13974

Lutsey, P. L., & Zakai, N. A. (2022). Epidemiology and prevention of venous thromboembolism. *Nature Reviews Cardiology*, 1–15. https://doi.org/10.1038/s41569-022-00787-6

Machiela, M. J., & Chanock, S. J. (2015). LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants: Fig. 1. *Bioinformatics*, *31*(21), 3555–3557. https://doi.org/10.1093/bioinformatics/btv402

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*(7), Article 7. https://doi.org/10.1038/s41588-021-00870-7

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., … the Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), Article 10. https://doi.org/10.1038/ng.3643

McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S., & Lin, X. (2020). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, *76*(4), 1262–1272. https://doi.org/10.1111/biom.13214

Mishra, A., Malik, R., Hachiya, T., Jürgenson, T., Namba, S., Posner, D. C., Kamanu, F. K., Koido, M., Le Grand, Q., Shi, M., He, Y., Georgakis, M. K., Caro, I., Krebs, K., Liaw, Y.-C., Vaura, F. C., Lin, K., Winsvold, B. S., Srinivasasainagendra, V., … Debette, S. (2022). Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature*, *611*(7934), Article 7934. https://doi.org/10.1038/s41586-022-05165-3

Mizuno, T., Yoshioka, K., Mizuno, M., Shimizu, M., Nagano, F., Okuda, T., Tsuboi, N., Maruyama, S., Nagamatsu, T., & Imai, M. (2017). Complement component 5 promotes lethal thrombosis. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/srep42714

Morange, P.-E., Bezemer, I., Saut, N., Bare, L., Burgos, G., Brocheton, J., Durand, H., Biron-Andreani, C., Schved, J.-F., Pernod, G., Galan, P., Drouet, L., Zelenika, D., Germain, M., Nicaud, V., Heath, S., Ninio, E., Delluc, A., Münzel, T., … Rosendaal, F. R. (2010). A follow-up study of a genome-wide association scan identifies a susceptibility locus for venous thrombosis on chromosome 6p24.1. *American Journal of Human Genetics*, *86*(4), 592–595. https://doi.org/10.1016/j.ajhg.2010.02.011

Morange, P.-E., Oudot-Mellakh, T., Cohen, W., Germain, M., Saut, N., Antoni, G., Alessi, M.-C., Bertrand, M., Dupuy, A.-M., Letenneur, L., Lathrop, M., Lopez, L. M., Lambert, J.-C., Emmerich, J., Amouyel, P., & Trégouët, D.-A. (2011). KNG1 Ile581Thr and susceptibility to venous thrombosis. *Blood*, *117*(13), 3692–3694. https://doi.org/10.1182/blood-2010-11-319053

Morange, P.-E., Saut, N., Antoni, G., Emmerich, J., & Trégouët, D.-A. (2011). Impact on venous thrombosis risk of newly discovered gene variants associated with FVIII and VWF plasma levels. *Journal of Thrombosis and Haemostasis*, *9*(1), 229–231. https://doi.org/10.1111/j.1538-7836.2010.04082.x

National Human Genome Research Institute. (2018). *Genetics vs. Genomics Fact Sheet*. Genome.Gov. https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics

Obaid, M., El-Menyar, A., Asim, M., & Al-Thani, H. (2020). Prevalence and Outcomes of Thrombophilia in Patients with Acute Pulmonary Embolism. *Vascular Health and Risk Management*, *16*, 75–85. https://doi.org/10.2147/VHRM.S241649

Oger, E., Lacut, K., Le Gal, G., Couturaud, F., Guenet, D., Abalain, J.-H., Roguedas, A.-M., Mottier, D., & THE EDITH COLLABORATIVE STUDY GROUP. (2006). Hyperhomocysteinemia and low B vitamin levels are independently associated with venous thromboembolism: Results from the EDITH study: a hospital-based case-control study. *Journal of Thrombosis and Haemostasis*, *4*(4), 793–799. https://doi.org/10.1111/j.1538-7836.2006.01856.x

Olié, V., Moutengou, E., Barry, Y., Deneux-Tharaux, C., Pessione, F., & Plu-Bureau, G. (2015). Maladie veineuse thromboembolique pendant la grossesse et le post-partum, France, 2009-2014. Numéro thématique. Les femmes au coeur du risque vasculaire. *Bulletin Epidemiologique Hebdomadaire*, *7–8*, 139–147.

Oudot-Mellakh, T., Cohen, W., Germain, M., Saut, N., Kallel, C., Zelenika, D., Lathrop, M., Trégouët, D.-A., & Morange, P.-E. (2012). Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: The MARTHA project. *British Journal of Haematology*, *157*(2), 230–239. https://doi.org/10.1111/j.1365-2141.2011.09025.x

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., & Tanaka, T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, *32*(4), 650–654. https://doi.org/10.1038/ng1047

Perdomo-Sabogal, A., Nowick, K., Piccini, I., Sudbrak, R., Lehrach, H., Yaspo, M.-L., Warnatz, H.-J., & Querfurth, R. (2016). Human Lineage-Specific Transcriptional Regulation through GA-Binding Protein Transcription Factor Alpha (GABPa). *Molecular Biology and Evolution*, *33*(5), 1231–1244. https://doi.org/10.1093/molbev/msw007

Picart, G., Robin, P., Tromeur, C., Orione, C., Raj, L., Ferrière, N., Le Mao, R., Le Roux, P.-Y., Le Floch, P.-Y., Lemarié, C. A., Nonent, M., Leroyer, C., Guegan, M., Lacut, K., Salaün, P.-Y., & Couturaud, F. (2020). Predictors of residual pulmonary vascular obstruction after pulmonary embolism: Results from a prospective cohort study. *Thrombosis Research*, *194*, 1–7. https://doi.org/10.1016/j.thromres.2020.06.004

Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M. A., Oerton, E., Cook, J., Stewart, I. D., Kerrison, N. D., Luan, J., Raffler, J., Arnold, M., Arlt, W., O'Rahilly, S., Kastenmüller, G., Gamazon, E. R., Hingorani, A. D., Scott, R. A., … Langenberg, C. (2021). Mapping the proteo-genomic convergence of human diseases. *Science*, *374*(6569), eabj1541. https://doi.org/10.1126/science.abj1541

Pitte, M. (2019). *Cascade la coagulation*. Soins-Infirmiers.com. https://www.soins-infirmiers.com/ifsi/ue-2.2-cycles-de-la-vie-et-grandes-fonctions/physiologie-hemostase

Poort, S. R., Rosendaal, F. R., Reitsma, P. H., & Bertina, R. M. (1996). 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*(10), 3698–3703.

Prandoni, P., Noventa, F., Ghirarduzzi, A., Pengo, V., Bernardi, E., Pesavento, R., Iotti, M., Tormene, D., Simioni, P., & Pagnan, A. (2007). The risk of recurrent venous thromboembolism after discontinuing anticoagulation in patients with acute proximal deep vein thrombosis or pulmonary embolism. A prospective cohort study in 1,626 patients. *Haematologica*, *92*(2), 199–205. https://doi.org/10.3324/haematol.10516

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rahmani, J., Haghighian Roudsari, A., Bawadi, H., Thompson, J., Khalooei Fard, R., Clark, C., Ryan, P. M., Ajami, M., Rahimi Sakak, F., Salehisahlabadi, A., Abdulazeem, H. M., Jamali, M. R., & Mirzay Razaz, J. (2020). Relationship between body mass index, risk of venous thromboembolism and pulmonary embolism: A systematic review and dose-response meta-analysis of cohort studies among four million participants. *Thrombosis Research*, *192*, 64–72. https://doi.org/10.1016/j.thromres.2020.05.014

Rajasekaran, S., Soundararajan, D. C. R., Nayagam, S. M., Tangavel, C., Raveendran, M., Thippeswamy, P. B., Djuric, N., Anand, S. V., Shetty, A. P., & Kanna, R. M. (2022). Modic changes are associated with activation of intense inflammatory and host defense response pathways—Molecular insights from proteomic analysis of human intervertebral discs. *The Spine Journal: Official Journal of the North American Spine Society*, *22*(1), 19–38. https://doi.org/10.1016/j.spinee.2021.07.003

Robette, N., Génin, E., & Clerget-Darpoux, F. (2022). Heritability: What's the point? What is it not for? A human genetics perspective. *Genetica*, *150*(3–4), 199–208. https://doi.org/10.1007/s10709-022-00149-7

Rodger, M. A., Kahn, S. R., Wells, P. S., Anderson, D. A., Chagnon, I., Le Gal, G., Solymoss, S., Crowther, M., Perrier, A., White, R., Vickars, L., Ramsay, T., Betancourt, M. T., & Kovacs, M. J. (2008). Identifying unprovoked thromboembolism patients at low risk for recurrence who can

discontinue anticoagulant therapy. *Canadian Medical Association Journal*, *179*(5), 417–426. https://doi.org/10.1503/cmaj.080493

Roldan, V., Lecumberri, R., Muñoz-Torrero, J. F. S., Vicente, V., Rocha, E., Brenner, B., & Monreal, M. (2009). Thrombophilia testing in patients with venous thromboembolism. Findings from the RIETE registry. *Thrombosis Research*, *124*(2), 174–177. https://doi.org/10.1016/j.thromres.2008.11.003

Sallah, S. (1997). Inhibitors to clotting factors. *Annals of Hematology*, *75*(1–2), 1–7. https://doi.org/10.1007/s002770050305

Salzberg, S. L. (2018). Open questions: How many genes do we have? *BMC Biology*, *16*, 94. https://doi.org/10.1186/s12915-018-0564-x

Sandset, P. M. (1996). Tissue factor pathway inhibitor (TFPI)—An update. *Haemostasis*, *26 Suppl 4*, 154–165. https://doi.org/10.1159/000217293

Schmechel, D. E., Goldgaber, D., Burkhart, D. S., Gilbert, J. R., Gajdusek, D. C., & Roses, A. D. (1988). Cellular localization of messenger RNA encoding amyloid-beta-protein in normal tissue and in Alzheimer disease. *Alzheimer Disease and Associated Disorders*, *2*(2), 96–111. https://doi.org/10.1097/00002093-198802020-00002

Siegbahn, A., Oldgren, J., Andersson, U., Ezekowitz, M. D., Reilly, P. A., Connolly, S. J., Yusuf, S., Wallentin, L., & Eikelboom, J. W. (2016). D-dimer and factor VIIa in atrial fibrillation—Prognostic values for cardiovascular events and effects of anticoagulation therapy. A RE-LY substudy. *Thrombosis and Haemostasis*, *115*(5), 921–930. https://doi.org/10.1160/TH15-07-0529

Smith, N. L., Rice, K. M., Bovill, E. G., Cushman, M., Bis, J. C., McKnight, B., Lumley, T., Glazer, N. L., van Hylckama Vlieg, A., Tang, W., Dehghan, A., Strachan, D. P., O'Donnell, C. J., Rotter, J. I., Heckbert, S. R., Psaty, B. M., & Rosendaal, F. R. (2011). Genetic variation associated with plasma von Willebrand factor levels and the risk of incident venous thrombosis. *Blood*, *117*(22), 6007–6011. https://doi.org/10.1182/blood-2010-10-315473

Syed, H., Jorgensen, A. L., & Morris, A. P. (2017). SurvivalGWAS_SV: Software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. *BMC Bioinformatics*, *18*(1), 265. https://doi.org/10.1186/s12859-017-1683-z

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), Article 7845. https://doi.org/10.1038/s41586-021-03205-y

Thibord, F., Klarin, D., Brody, J. A., Chen, M.-H., Levin, M. G., Chasman, D. I., Goode, E. L., Hveem, K., Teder-Laving, M., Martinez-Perez, A., Aïssi, D., Daian-Bacq, D., Ito, K., Natarajan, P., Lutsey, P. L., Nadkarni, G. N., de Vries, P. S., Cuellar-Partida, G., Wolford, B. N., … Smith, N. L. (2022). Cross-Ancestry Investigation of Venous Thromboembolism Genomic Predictors. *Circulation*, *146*(16), 1225–1242. https://doi.org/10.1161/CIRCULATIONAHA.122.059675

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, *26*(1), 24–36. https://doi.org/10.2307/1907382

Tosetto, A., Iorio, A., Marcucci, M., Baglin, T., Cushman, M., Eichinger, S., Palareti, G., Poli, D., Tait, R. C., & Douketis, J. (2012). Predicting disease recurrence in patients with previous unprovoked venous thromboembolism: A proposed prediction score (DASH): Clinical prediction of VTE recurrence. *Journal of Thrombosis and Haemostasis*, *10*(6), 1019–1025. https://doi.org/10.1111/j.1538-7836.2012.04735.x

Trégouët, D.-A., Heath, S., Saut, N., Biron-Andreani, C., Schved, J.-F., Pernod, G., Galan, P., Drouet, L., Zelenika, D., Juhan-Vague, I., Alessi, M.-C., Tiret, L., Lathrop, M., Emmerich, J., & Morange, P.-E. (2009). Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: Results from a GWAS approach. *Blood*, *113*(21), 5298–5303. https://doi.org/10.1182/blood-2008-11-190389

Tromeur, C., Sanchez, O., Presles, E., Pernod, G., Bertoletti, L., Jego, P., Duhamel, E., Provost, K., Parent, F., Robin, P., Deloire, L., Leven, F., Mingant, F., Bressollette, L., Le Roux, P.-Y., Salaun, P.-Y., Nonent, M., Pan-Petesch, B., Planquette, B., … PADIS-PE Investigators18. (2018). Risk factors for recurrent venous thromboembolism after unprovoked pulmonary embolism: The PADIS-PE randomised trial. *The European Respiratory Journal*, *51*(1), 1701202. https://doi.org/10.1183/13993003.01202-2017

Tweedie, M. C. K. (1984). *An index which distinguishes between some important exponential families. Statistics: Applications and New Directions*. 579–604.

Uitte de Willige, S., de Visser, M. C. H., Houwing-Duistermaat, J. J., Rosendaal, F. R., Vos, H. L., & Bertina, R. M. (2005). Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen gamma' levels. *Blood*, *106*(13), 4176–4183. https://doi.org/10.1182/blood-2005-05-2180

van der Net, J. B., Janssens, A. C. J. W., Eijkemans, M. J. C., Kastelein, J. J. P., Sijbrands, E. J. G., & Steyerberg, E. W. (2008). Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. *European Journal of Human Genetics*, *16*(9), Article 9. https://doi.org/10.1038/ejhg.2008.59

van Hylckama Vlieg, A., Flinterman, L. E., Bare, L. A., Cannegieter, S. C., Reitsma, P. H., Arellano, A. R., Tong, C. H., Devlin, J. J., & Rosendaal, F. R. (2014). Genetic Variations Associated With Recurrent Venous Thrombosis. *Circulation: Cardiovascular Genetics*, *7*(6), 806–813. https://doi.org/10.1161/CIRCGENETICS.114.000682

Vayne, C., Nguyen, P., & Gruel, Y. (2017). Neutrophil extracellular traps, hemostatsis and thrombosis. *Hématologie*, *23*(2), 108–121. https://doi.org/10.1684/hma.2017.1249

Waheed, S. M., Kudaravalli, P., & Hotwagner, D. T. (2022). Deep Vein Thrombosis. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK507708/

Wang, J., Wang, L., Shang, H., Yang, X., Guo, S., Wang, Y., & Cui, C. (2020). Jugular venous catheter-associated thrombosis and fatal pulmonary embolism. *Medicine*, *99*(26), e20873. https://doi.org/10.1097/MD.0000000000020873

Warwick, C. A., Keyes, A. L., Woodruff, T. M., & Usachev, Y. M. (2021). The complement cascade in the regulation of neuroinflammation, nociceptive sensitization, and pain. *Journal of Biological Chemistry*, *297*(3). https://doi.org/10.1016/j.jbc.2021.101085

Williams, S. R., Hsu, F.-C., Keene, K. L., Chen, W.-M., Nelson, S., Southerland, A. M., Madden, E. B., Coull, B., Gogarten, S. M., Furie, K. L., Dzhivhuho, G., Rowles, J. L., Mehndiratta, P., Malik, R., Dupuis, J., Lin, H., Seshadri, S., Rich, S. S., Sale, M. M., & Worrall, B. B. (2016). Shared genetic susceptibility of vascular-related biomarkers with ischemic and recurrent stroke. *Neurology*, *86*(4), 351–359. https://doi.org/10.1212/WNL.0000000000002319

Winter, M.-P., Schernthaner, G. H., & Lang, I. M. (2017). Chronic complications of venous thromboembolism. *Journal of Thrombosis and Haemostasis: JTH*, *15*(8), 1531–1540. https://doi.org/10.1111/jth.13741

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. *American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, *46*(2), 100–106. https://doi.org/10.1038/ng.2876

Zakai, N. A., & McCLURE, L. A. (2011). Racial differences in venous thromboembolism. *Journal of Thrombosis and Haemostasis*, *9*(10), 1877–1882. https://doi.org/10.1111/j.1538-7836.2011.04443.x

Zhang, S., Cao, Y., Du, J., Liu, H., Chen, X., Li, M., Xiang, M., Wang, C., Wu, X., Liu, L., Wang, C., Wu, Y., Li, Z., Fang, S., Shi, J., & Wang, L. (2021). Neutrophil extracellular traps contribute to tissue plasminogen activator resistance in acute ischemic stroke. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, *35*(9), e21835. https://doi.org/10.1096/fj.202100471RR

Zöller, B., Li, X., Sundquist, J., & Sundquist, K. (2011). Age- and Gender-Specific Familial Risks for Venous Thromboembolism. *Circulation*, *124*(9), 1012–1020. https://doi.org/10.1161/CIRCULATIONAHA.110.965020

Zuo, Y., Yalavarthi, S., Shi, H., Gockman, K., Zuo, M., Madison, J. A., Blair, C., Weber, A., Barnes, B. J., Egeblad, M., Woods, R. J., Kanthi, Y., & Knight, J. S. (2020). Neutrophil extracellular traps in COVID-19. *JCI Insight*, *5*(11), e138999, 138999. https://doi.org/10.1172/jci.insight.138999