Title: Cross-Ancestry Investigation of Venous Thromboembolism Genomic Predictors

Florian Thibord PhD[1,2]*, Derek Klarin MD[3]*, Jennifer A. Brody BA[4], Ming-Huei Chen PhD[1,2], Michael G. Levin MD[5,6], Daniel I. Chasman PhD[7,8], Ellen L. Goode PhD MPH[9], Kristian Hveem PhD[10,11], Maris Teder-Laving MS[12], Angel Martinez-Perez MS[13], Dylan Aïssi PhD[14,15], Delphine Daian-Bacq PhD[16,17], Kaoru Ito PhD MD[18], Pradeep Natarajan MD MMSc[19,20,21], Pamela L. Lutsey PhD[22], Girish N. Nadkarni MD MPH[23,24,25], Paul S. de Vries PhD[26], Gabriel Cuellar-Partida PhD[27], Brooke N. Wolford PhD[28], Jack W. Pattee PhD[29,30], Charles Kooperberg PhD[31], Sigrid K. Braekkan PhD[32,33], Ruifang Li-Gao PhD[34], Noemie Saut PhD[35], Corriene Sept PhD[36], Marine Germain MS[14,15,17], Renae L. Judy MS[37], Kerri L. Wiggins MS RD[4], Darae Ko MD[38,2], Christopher O'Donnell PhD MD MPH[39], Kent D. Taylor PhD[40], Franco Giulianini PhD[7], Mariza De Andrade PhD[9], Therese H. Nøst PhD[11], Anne Boland PhD[16,17], Jean-Philippe Empana PhD[41,42], Satoshi Koyama PhD MD[18,43,19], Thomas Gilliland MD[19,43,21], Ron Do PhD[23,24,44], Jennifer E. Huffman PhD[45], Xin Wang PhD[27], Wei Zhou PhD[46], Jose Manuel Soria PhD[13], Juan Carlos Souto MD PhD[47,13], Nathan Pankratz PhD[48], Jeffery Haessler MS[31], Kristian Hindberg PhD[32], Frits R. Rosendaal MD PhD[49], Constance Turman MS[36], Robert Olaso PhD[16,17], Rachel L. Kember PhD[50], Traci M. Bartz MS[51], Julie A. Lynch PhD RN MBA[52,53], Susan R. Heckbert MD MPH[54], Sebastian M. Armasu MS[9], Ben Brumpton PhD[11], David M. Smadja MD PhD[55,56], Xavier Jouven MD PhD[41,42], Issei Komuro PhD MD[57], Katharine Clapham MD[58,20,21], Ruth J.F. Loos PhD[23], Cristen J. Willer PhD[28], Maria Sabater-Lleal PhD[13,59], James S. Pankow PhD[22], Alexander P. Reiner MD MSc[60,31], Vania M. Morelli MD PhD[32,33], Paul M. Ridker MD MPH[7,8], Astrid van Hylckama Vlieg PhD[49], Jean-François Deleuze PhD[16,61,62], Peter Kraft PhD[36], Daniel J. Rader MD[63], Barbara McKnight PhD[51], Global Biobank Meta-Analysis Initiative, Estonian Biobank Research Team, 23andMe Research Team, Biobank Japan, CHARGE Hemostasis Working Group, Kyung Min Lee PhD[52], Bruce M. Psaty MD PhD[4,64,65], Anne Heidi Skogholt PhD[11], Joseph Emmerich MD PhD[66,67], Pierre Suchon MD PhD[35,68], Stephen S. Rich PhD[69], Ha My T. Vy PhD[23,24], Weihong Tang MD PhD[70], Rebecca D. Jackson MD[71], John-Bjarne Hansen MD PhD[32,33], Pierre-Emmanuel Morange MD PhD[35,68], Christopher Kabrhel MD MPH[72,73], David-Alexandre Trégouët PhD[14,15,17]*, Scott Damrauer MD[6]*, Andrew D. Johnson PhD[1,2]*, Nicholas L. Smith PhD[51,74]*

*Denotes equal contribution

[1]Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, 73 Mt. Wayte, Suite #2, Framingham, MA, 01702, USA,
[2]The Framingham Heart Study, Boston University and NHLBI, 73 Mt. Wayte Ave, Suite #2, Framingham, MA, 01702, USA,
[3]Division of Vascular Surgery, Stanford University School of Medicine, Palo Alto, CA, 94305, USA,
[4]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, 1730 Minor Ave, Suite 1360, Seattle, WA, 98101, USA,
[5]Division of Cardiovascular Medicine, Department of Medicine, University of Pennsylvania, 3400 Spruce Street, PA, 19104, USA,
[6]Medicine, Corporal Michael J. Crescenz Philadelphia VA Medical Center, 3900 Woodland Ave, PA, 19104, USA,

[7]Division of Preventive Medicine, Brigham and Women's Hospital, 900 Commonwealth Ave, Boston, MA, 02215, USA,
[8]Harvard Medical School, Boston, MA, 02115, USA,
[9]Department of Quantitative Health Sciences, Mayo Clinic, 200 First Street SW, Rochester, MN, 55905, USA,
[10]HUNT Research Center, Department of Public Health and Nursing, Norwegian University of Science and Technology, Forskningsvegen 2, Levanger, 7600, Norway,
[11]K.G. Jebsen Centre for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Håkon Jarls gate 11, Trondheim, 7030, Norway,
[12]Institute of Genomics, University of Tartu, Riia 23b, Tartu, Tartu, 51010, Estonia,
[13]Genomics of Complex Disease Unit, Sant Pau Biomedical Research Institute (IIB Sant Pau), Barcelona, Spain, St Quinti 77-79, Barcelona, 8041, Spain,
[14]Bordeaux Population Health Research Center, University of Bordeaux, 146 rue Léo Saignat, Bordeaux, 33076, France,
[15]UMR1219, INSERM, 146 rue Léo Saignat, Bordeaux, 33076, France,
[16]Centre National de Recherche en Génomique Humaine, CEA, Université Paris-Saclay, 2 Rue Gaston Crémieux, Evry, 91057, France,
[17]Laboratory of Excellence on Medical Genomics, France,
[18]Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehirocho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan,
[19]Cardiovascular Research Center, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA, 02446, USA,
[20]Program in Medical and Population Genetics and the Cardiovascular Disease Initiative, Broad Institute of Harvard & MIT, 75 Ames St, Cambridge, MA, USA,
[21]Department of Medicine, Harvard Medical School, Shattuck St, Boston, MA, USA,
[22]Division of Epidemiology and Community Health, University of Minnesota, 1300 South Second Street, MN, 55454, USA,
[23]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY, 10029, USA,
[24]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY, 10029, USA,
[25]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY, 10029, USA,
[26]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, 1200 Pressler St, Houston, TX, 77030, USA,
[27]23andMe, Inc., 223 N Mathilda Ave, CA, 94086, USA,
[28]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, 48109, USA,
[29]Division of Biostatistics, University of Minnesota, 420 Delaware St. SE, MN, 55455, USA,
[30]Center for Innovative Design & Analysis and Department of Biostatistics & Informatics, Colorado School of Public Health, 13001 East 17th Place, CO, 80045, USA,

[31]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA, 98109, USA,
[32]Thrombosis Research Center (TREC), UiT - The Arctic University of Norway, Universitetsvegen 57, Tromsø, 9037, Norway,
[33]Division of internal medicine, University Hospital of North Norway, Tromsø, 9038, Norway,
[34]Clinical Epidemiology, Leiden University Medical Center, PO Box 9600, 2300 RC, The Netherlands,
[35]Hematology Laboratory, La Timone University Hospital of Marseille, 264 Rue Saint-Pierre, Marseille, 13385, France,
[36]Department of Epidemiology, Harvard TH Chan Harvard School of Public Health, 655 Huntington Ave., Building II, Boston, MA, 02115, USA,
[37]Surgery, University of Pennsylvania, 3401 Walnut Street, PA, 19104, USA,
[38]Section of Cardiovascular Medicine, Boston University School of Medicine, 85 East Newton Street, Boston, MA, 02118, USA,
[39]Cardiology, VA Boston Healthcare System, Boston, MA, 02130, USA,
[40]Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation, 1124 W Carson St., Torrance, CA 90502, CA, USA,
[41]Integrative Epidemiology of cardiovascular diseases, Université Paris Descartes, Sorbonne Paris Cité, 56 rue Leblanc, Paris, 75015, France,
[42]Paris Cardiovascular Research Center, Inserm U970, Université Paris Descartes, Sorbonne Paris Cité, 20 rue Leblanc, Paris, 75015, France,
[43]Program in Medical and Population Genetics and the Cardiovascular Disease Initiative, Broad Institute of Harvard & MIT, 75 Ames St, Cambridge, MA, 02142, USA,
[44]BioMe Phenomics Center, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY, 10029, USA,
[45]MAVERIC, VA Boston Heathcare System, 2 Avenue de Lafayette, Boston, MA, 02111, USA,
[46]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA, 02114, USA,
[47]Unit of Thrombosis and Hemostasis, Hospital de la Santa Creu i Sant Pau, St Quinti 89, Barcelona, 8041, Spain,
[48]Department of Laboratory Medicine and Pathology, University of Minnesota, 420 Delaware St. SE, MN, 55455, USA,
[49]Clinical Epidemiology, Leiden University Medical Center, PO Box 9600, Leiden, The Netherlands,
[50]Psychiatry, University of Pennsylvania, 3401 Walnut Street, PA, 19104, USA,
[51]Cardiovascular Health Research Unit, Departments of Biostatistics and Medicine, University of Washington, 1730 Minor Ave, Suite 1360, Seattle, WA, 98101, USA,
[52]VA Informatics & Computing Infrastructure, VA Salt Lake City Healthcare System, 500 Foothills Drive, UT, 01730, USA,
[53]Epidemiology, University of Utah, 500 Foothills Drive, UT, 01730, USA,
[54]Department of Epidemiology, University of Washington, 1730 Minor Ave, Suite 1360, Seattle, WA, 98101, USA,

[55]Hematology Department and Biosurgical Research Lab (Carpentier Foundation), European Georges Pompidou Hospital, Assistance Publique Hôpitaux de Paris, 20 rue Leblanc, Paris, 75015, France,
[56]Innovative Therapies in Haemostasis, INSERM, Université de Paris, 4 avenue de l'Observatoire, Paris, 75270, France,
[57]Department of Cardiovascular Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Tokyo, 113-8655, Japan,
[58]Division of Cardiovascular Medicine, Brigham & Women's Hospital, MA, USA,
[59]Cardiovascular Medicine Unit, Department of Medicine, Karolinska Institutet, Center for Molecular Medicine, Stockholm, 17176, Sweden,
[60]Department of Epidemiology, University of Washington, 3980 15th Ave NE, WA, 98195, USA,
[61]Centre D'Etude du Polymorphisme Humain, Fondation Jean Dausset, 27 rue Juliette Dodu, Paris, 75010, France,
[62]Laboratory of Excellence on Medical Genomics,
[63]Cardiology, Medical Genetics, University of Pennsylvania, 3401 Walnut Street, PA, 19104, USA,
[64]Department of Epidemiology, University of Washington, 1730 Minor Ave, Suite #1360, Seattle, WA, 98101, USA,
[65]Department of Health Systems and Population Heath, University of Washington, 1730 Minor Ave, Suite #1360, Seattle, WA, USA,
[66]Department of vascular medicine, Paris Saint-Joseph Hospital Group, University of Paris, 185 rue Raymond Losserand, Paris, 75674, France,
[67]UMR1153, INSERM CRESS, 185 rue Raymond Losserand, Paris, 75674, France,
[68]C2VN, INSERM, INRAE, Aix-Marseille University, 27, bd Jean Moulin, Marseille, 13385, France,
[69]Center for Public Health Genomics, University of Virginia, 3242 West Complex, Charlottesville, VA, 22908-0717, USA,
[70]Division of Epidemiology and Community Health, University of Minnesota, 1300 South Second Street, MN, 55454, USS,
[71]College of Medicine, Ohio State University, 376 W. 10th Ave, Columbus, OH, 43210, USA,
[72]Emergency Medicine, Massachusetts General Hospital, Zero Emerson Place, Suite 3B, Boston, MA, 02114, USA,
[73]Emergency Medicine, Harvard Medical School, Zero Emerson Place, Suite 3B, Boston, 02114, MA,
[74]Department of Surgery, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, 19104, USA

Short Title: VTE GWAS

Word counts: total = 8993; manuscript body = 4,9496

Contact: Nicholas L. Smith, Department of Epidemiology, University of Washington, Seattle WA, 98195, USA. (nlsmith@u.washington.edu)

Keywords: venous thrombosis, venous thromboembolism, genetic, genome-wide association studies, transcriptome-wide association study

Version: June 12, 2022

ABSTRACT (332/350 words)

Background: Venous thromboembolism (VTE) is a life-threatening vascular event with environmental and genetic determinants. Recent VTE genome-wide association studies (GWAS) meta-analyses involved nearly 30,000 VTE cases and identified up to 40 genetic loci associated with VTE risk, including loci not previously suspected to play a role in hemostasis. The aims of our research was to expand discovery of new genetic loci associated with VTE by using cross-ancestry genomic resources.

Methods: We present new cross-ancestry meta-analyzed GWAS results involving up to 81,669 VTE cases from 30 studies, with replication of novel loci in independent populations and loci characterization through *in silico* genomic interrogations.

Results: In our genetic discovery effort that included 55,330 participants with VTE (47,822 European, 6,320 African, and 1,188 Hispanic ancestry), we identified 48 novel associations of which 34 replicated after correction for multiple testing. In our combined discovery-replication analysis (81,669 VTE participants) and ancestry-stratified meta-analyses (European, African and Hispanic), we identified another 44 novel associations, which are new candidate VTE-associated loci requiring replication. In total, across all GWAS meta-analyses, we identified 135 independent genomic loci significantly associated with VTE risk. A genetic risk score of the 135 loci identified a 6-fold increase in risk for those in the top 1% of scores compared with those with average scores. We also identified 31 novel transcript associations in transcriptome-wide association studies and 8 novel candidate genes with protein quantitative-trait locus Mendelian randomization analyses. *In silico* interrogations of hemostasis and hematology traits and a large phenome-wide association analysis of the 135 GWAS loci provided insights to biological pathways contributing to VTE, with some loci contributing to VTE through well-characterized coagulation pathways while others providing new data on the role of hematology traits, particularly platelet function. Many of the replicated loci are outside of known or currently hypothesized pathways to thrombosis.

Conclusions: Our cross-ancestry GWAS meta-analyses identified new loci associated with VTE. These findings highlight new pathways to thrombosis and provide novel molecules that may be useful in the development of improved antithrombosis treatments.

Clinical Perspective

What is new: (100 words max)

- Our VTE genetic analyses revealed 135 loci associated with VTE, of which 92 were novel. While novel VTE associated variants were typically non-coding and displayed small odds ratios, they point at novel biological pathways involved in VTE.
- In particular, a large number of VTE variants are shared with platelets traits and located in loci with known roles in hematopoiesis or megakaryocyte development, which suggests that platelet generation, turnover or reactivity may be a feature of VTE pathogenesis.

What are the clinical implications: (100 words max)

- These results constitute a valuable resource for thrombosis researchers and for the discovery of new VTE therapeutic targets.
- A genetic risk score constructed from the European specific results and applied to the UK Biobank participants of European ancestry explained ~5% of the phenotypic variance, and displayed a significant predictive ability.

INTRODUCTION

Venous thrombosis is a vascular event resulting from an imbalance in the regulation of hemostasis, with subsequent pathologic coagulation and vascular thrombosis formation. Clinically, venous thrombosis can manifest as deep vein thrombosis (DVT), when occurring in the deep veins primarily of the legs and trunk, or as a pulmonary embolism (PE), when the thrombus embolizes and obstructs the pulmonary arteries. Collectively, these events are known as venous thromboembolism (VTE), a life-threatening condition with an incidence of 1-2 events per 1,000 person-years.[1–3] VTE is a complex disease with both environmental and genetic determinants. Family studies, candidate-gene approaches, and early genome-wide association studies (GWAS) primarily identified genetic risk factors in loci with well characterized effects on coagulation (*F2*, *F5*, *F11*, *FGG*, *ABO*, *SERPINC1*, *PROCR*, *PROC*, *PROS1*), supporting current therapeutic strategies that mainly target the coagulation cascade.[4–8] In recent years, larger GWAS meta-analyses revealed unanticipated loci, such as *SLC44A2*,[9] which was later characterized as a choline transporter involved in platelet activation,[10] and in the adhesion and activation of neutrophils.[11,12] Thus, genetic associations with VTE in larger and more diverse populations may uncover new biological pathways and molecular events contributing to the disease and potentially help identify novel targets for treatment. Most recently, 2 large efforts involving up to 30,000 VTE cases, led by the International Network Against Venous Thrombosis (INVENT) consortium[13] and the Million Veteran Program[14] (MVP), identified up to 43 genetic loci associated with VTE. To expand discovery of novel VTE risk loci, we conducted a large, cross-ancestry GWAS meta-analysis involving more than 80,000 VTE cases, along with a replication of novel loci and their characterization through downstream analyses.

METHODS

The data that support the findings of this study will be available through dbGaP.

Design and Study Participants

The study design (see **Figure 1**) included a cross-ancestry discovery meta-analysis of GWAS summary data from 4 consortium/studies (INVENT-2019, MVP, FinnGen, EGP) followed by a replication of discovery loci that exceeded the genome-wide significance threshold ($P<5.0\times10^{-8}$). The replication population involved 12 studies, limiting data to non-overlapping studies with our discovery.[15] The combined discovery and replication data (when available) were then meta-analyzed, and ancestry-stratified meta-analyses were performed for African-ancestry (AA), European-ancestry (EA), and Hispanic-ancestry (HIS) participants to enable further downstream ancestry-specific analyses, such as fine mapping. Participants from studies provided written informed consent for use of their genetic and health information for analysis, and the studies were individually approved by the appropriate Institutional Review Boards (see **Supplemental Materials**).

Study-specific GWAS

Each study performed association analyses and provided summary data for meta-analysis. Genotyping arrays, imputation panels, and analyses performed by each participating study are detailed in **Table S1**. Additional study specific are available as **Supplemental Materials**.

Discovery, Replication, and Combined GWAS Meta-analyses

All GWAS meta-analyses were conducted with METAL,[16] using a fixed-effects inverse-variance weighted model. All variants were included and there was no lower minor allele frequency (MAF) limit beyond study-specific minor allele count. Genome-wide significant variants ($P<5.00\times10^{-8}$) were kept if a concordant effect direction was observed in 2 or more studies and grouped into the same locus if they were within 1Mb. We used the closest gene to the lead variant to refer to each locus, except at known loci where the causal gene has been previously identified and is different from the closest gene (such as *PROCR* or *PROS1*). We defined a locus as novel if a genetic association with VTE has not been previously observed in the region according to our review of peer-reviewed published reports.

Discovery Meta-Analysis: For the discovery cross-ancestry GWAS meta-analysis, we meta-analyzed data from 4 consortium/studies: INVENT-2019, MVP, FinnGen and EGCUT. Participants were EA, AA, and HIS adult men and women VTE cases (either DVT and/or PE cases) and controls. At each locus with a genome-wide significant signal, the lead variant was extracted and tested in an independent replication meta-analysis.

Replication: The replication GWAS meta-analysis consisted of the remaining 10 participating studies, as well as 2 external collaborators (GBMI[15] and 23andMe[17]). Replicating variants from the discovery were defined as those that had concordant effect direction in the discovery and the replication, and reached a Bonferroni-corrected p-value threshold in the replication population corresponding to the number of variants tested for replication with a 1-sided hypothesis: p-value threshold = [(0.05*2)/number of variants tested for replication] in the replication analysis.

Combined GWAS Meta-Analysis and Stratification by Ancestry: We performed a combined, cross-ancestry GWAS meta-analysis of discovery and replication data (when available) using participating studies with genome-wide summary data. We included variants with MAF≥0.01 to maintain adequate statistical power by reducing the number of low-powered tests since replication was not available. We estimated the heterogeneity associated with each variant using Cochran's Q test and the corresponding I^2 statistic. We assessed the genomic inflation with the lambda genomic control.[18] We report on variants exceeding the genome-wide threshold ($P<5.00\times10^{-8}$) and view these as candidate novel loci associated with VTE and needing future replication.

We then stratified the analyses by ancestry and limited strata to EU, AA, and HIS as the remaining ancestries had too few VTE events to be informative. As above, we estimated heterogeneity and genomic inflation; the LD-score intercept was computed for EA analysis, using the recommended Hapmap3 variants.[19] We report all additional ancestry-specific variants

exceeding the genome-wide threshold ($P<5.00\times10^{-8}$) and view these as ancestry-specific candidate loci associated with VTE and needing future replication.

Ancestry-Stratified Analyses: Conditional Analyses and Fine-mapping

To estimate the presence of independent signals, we performed conditional analyses with GCTA-COJO[20] at each locus with significant signals in EA, AA, and HIS GWAS meta-analyses. The Trans-Omics for Precision Medicine (TOPMed) ancestry-specific sequence data were used as reference panels.[21] Conditional analyses were performed at each locus, using a window that encompassed at least the genome-wide significant variants present in the locus with an additional buffer of ±100 kb. A stepwise joint regression model was used to identify secondary signals with joint p-values $<5.00\times10^{-8}$ and a linkage disequilibrium (LD) $r^2<0.2$ with selected variants. In addition, for each locus and for each ancestry-specific GWAS meta-analysis, we produced forest plots, Manhattan plots, and regional association plots to visually inspect the local genetic architecture (available as **Figures S1-S9**).[22,23] Additional information is found in the **Supplemental Materials**.

Genetic Risk Score (GRS)

We constructed an ancestry-specific GRS derived from the genome-wide significant lead variants observed in the EA meta-analysis and evaluated it for UKB EA participants. The GRS for AA and HIS were not constructed due to a lack of availability of a large-scale dataset with accessible genotype data for other ancestries. The EA GRS was calculated for each individual as a summation of log(OR)-weighted genotypes. We then performed logistic regression to measure the association of the GRS with VTE status, while correcting for age, sex, and the top 10 genetic principal components (PCs). The predictive ability of the score was estimated by calculating the area under the curve (AUC), using the *pROC* R library.[24] Additional information is available in the **Supplemental Materials**.

Transcriptome-Wide Association Studies (TWAS)

We performed TWAS with the FUSION pipeline[25] using the EA meta-analysis results. We first performed a series of single-tissue TWAS using gene expression from eQTL datasets relevant to blood and thrombosis disorders: whole blood, peripheral blood, liver, lung, and spleen.[26–28] All associations reaching a Bonferroni corrected significance threshold corresponding to the number of gene tested (N=14,219, $P<3.52\times10^{-6}$) were deemed statistically significant. Additional details are available in the **Supplemental Materials**.

Protein QTL Mendelian Randomization

Using the combined, cross-ancestry VTE GWAS meta-analysis results, we performed a proteome Mendelian randomization (MR) analysis with high-confidence genomic instruments corresponding to protein QTL (pQTL) for 1,216 circulating plasma proteins that passed consistency and pleiotropy filters, as previously described.[29] Additional information is available in the **Supplemental Materials**. To account for multiple testing, associations passing the Bonferroni corrected threshold (N=1,256, $P<3.98\times10^{-5}$) were considered statistically significant.

Association of VTE Loci with Hemostasis and Hematology Traits

We conducted a series of *in silico* investigations involving hemostasis and hematology traits to better characterize the VTE-associated variants from the GWAS meta-analyses. To better understand if novel VTE-associated variants operate through hemostasis pathways, we extracted associations from published GWAS of 10 hemostatic traits: fibrinogen;[30] fibrin D-dimer;[31] coagulation factors VII (FVII),[32] VIII (FVIII),[33] and XI (FXI);[34] von Willebrand factor (vWF);[33] tissue plasminogen activator (tPA);[35] plasminogen-activator inhibitor 1 (PAI-1);[36] activated partial thromboplastin time (aPTT); and prothrombin time (PT).[37] Since each variant-association was investigated in 10 hemostasis traits, we set a p-value threshold of 0.005 (0.05/10 traits tested for each lead variant of a locus) to separate associations of interest from other associations.

Similarly, we extracted associations with complete blood count (CBC) measures using summary data from nearly 750,000 individuals on 15 leukocyte, erythrocyte, and platelet traits.[38] Given

the large sample size and high statistical power of these analyses, we used a more stringent threshold of interest that was a Bonferroni correction corresponding to the number of look-ups performed ($P<1.92\times10^{-5}$). We further performed colocalization analyses with the *coloc*[39] R library for significant associations, using the discovery, combined, EA and AA VTE meta-analyses.

Phenome-wide Association Testing

To explore associations between VTE-associated variants and other traits agnostically, we performed a phenome-wide association study (PheWAS) using the MRC IEU infrastructure and the associated *ieugwasr* R library.[40] Lead variants identified in our VTE meta-analyses were queried in 2 sources of GWAS (using the PheWAS codes 'ukb-a' and 'ukb-d') which correspond to 1,500 UKB analyses performed by the Neale lab on 337,000 individuals of British. We then retrieved associations reaching genome-wide significance ($P<5.00x10^{-8}$) for each of the 1,500 traits investigated.

RESULTS

Discovery Cross-Ancestry Meta-analysis and Replication

The primary cross-ancestry discovery analysis included 55,330 participants among 3 ancestry groups with VTE (47,822 EU, 6,320 AA, and 1,188 HIS) and 1,081,973 participants without VTE (918,195 EU, 118,144 AA, and 45,634 HIS). Over the 22 autosomal and X chromosomes, 35.5 million variants were analyzed, and the observed lambda was 1.06. We identified 10,493 variants reaching genome-wide significance, corresponding to 85 loci, of which 48 have not been identified in previous genetic studies of VTE (see **Table S2**).

We tested lead variants from these 85 loci for replication in 91,230 cases and 3,322,939 controls from the independent replication data. After meta-analyzing the results of these 85 tests in the replication population, we identified 83 variants with a concordant effect direction between the discovery and the replication, of which 68 replicated at the 1-sided Bonferroni corrected significance threshold ($p<0.1/83 = 0.0012$) (**Table 1**, **Figure 2**, **Table S2**). The

successfully replicated signals corresponded to 34 known and 34 novel loci. Among the 34 novel loci that replicated, heterogeneity was minimal (heterogeneity P>0.05), odds ratios (ORs) ranged between 0.84-0.98 and 1.03-1.18, and MAFs were all ≥0.021. The majority of variants were gene-centric (4 exonic, 16 intronic, and 3 in 3' or 5' UTR regions or immediately downstream), 3 were linked to intronic non-coding RNA, and 8 were considered intergenic. Among the 17 variants and their associated loci that failed replication, 14 were novel and remain candidate loci that merit additional replication while 3 were known loci.

Combined Cross-Ancestry GWAS Meta-analysis and Ancestry-Stratified Results

Combined: The combined, cross-ancestry meta-analysis of the studies with genome-wide markers included 81,669 individuals with VTE and 1,426,717 individuals without VTE. We analyzed 19.1 million common variants (MAF≥0.01) and observed lambda of 1.16 which is slightly elevated but expected for large scale meta-analyses of polygenic traits.[41] We identified 16,550 variants reaching genome-wide significance located in 111 loci, of which 41 were not observed in the discovery analysis (**Table S3, Figure 2**). Of these 41 additional loci, 1 corresponded to a common variant at the known *SERPINC1* locus (rs6695940) which encodes antithrombin, 4 were previously identified in the INVENT-2019[13] or MVP[14] meta-analyses at the *PEPD*, *ABCA5*, *MPHOSPH9*, and *ARID4A* loci, and 1 was a known pathogenic missense variant located in *SERPINA1* (rs28929474, p.Glu366Lys).[42] The remaining 35 loci were novel associations and are presented in **Table 2**. Among these 35 candidate loci, all had ORs with ranges of 0.93-0.97 and 1.03-1.15 and had a minimum MAF of 0.021. The majority of the variants were gene-centric (18 intronic and 3 in 3' UTR regions), 3 were intronic in non-coding RNA, and 11 were considered intergenic.

European Ancestry: The EA meta-analysis, which included 71,771 participants with VTE and 1,059,740 participants without VTE, had a lambda of 1.22. As population stratification might be introduced by founder effects in Finnish participants from FinnGen, we did a sensitivity analysis by removing this cohort, and observed a similar genomic factor of 1.19. We also observed an LD-score intercept of 1.07, indicating an inflation mainly due to polygenic architecture, and

possibly slight residual stratification. Of the 11.1 million variants analyzed, 16,867 were genome-wide significant and clustered into 100 regions, of which 7 did not overlap with loci identified in the discovery or combined meta-analysis (**Table 2**, **Figure 2**, **Table S4**). For these 7 additional candidate loci, the ORs ranged from 0.94-0.97 to 1.04-1.07 and the minimum MAF was 0.058. Conditional analyses were performed at each of the 100 significant loci and revealed a subset of 21 loci with multiple independent signals (**Table S5**) and included 3 of the novel loci.

African Ancestry: The AA meta-analysis included 7,482 participants with VTE and 129,975 participants without VTE from 7 cohorts and had a lambda of 1.05. Here, 17.1 million variants were analyzed, of which 752 were genome-wide significant and located within 13 loci, of which 2 corresponded to novel ancestry-specific signals at *RBFOX1* (OR=0.56; MAF=0.04) and *COL6A2* (OR=2.16; MAF=0.011) (**Table 2**, **Figure 2**, **Table S6**). Conditional analyses were performed at each of the 13 significant loci and revealed 3 loci with additional independent signals (**Table S7**).

Hispanic Ancestry: The HIS meta-analysis included 1,720 participants with VTE and 57,367 participants without VTE from 4 cohorts and had a lamba of 1.02. We analyzed 11.1 million variants, of which 58 were genome-wide significant, all located at the *ABO* locus with rs2519093 as lead variant (OR=1.49, MAF=0.15, $P=3.08\times10^{-15}$). The conditional analysis revealed a secondary signal at this locus (**Table S7**).

Comparison of Ancestry-Specific and Cross-Ancestry Meta-Analysis Results: We then investigated the lead variants from the AA and EA meta-analyses at the 11 loci (all known) identified in both analyses. At 5 loci, none of the AA lead variants were available in the EA analyses, due to their low frequency in EAs (MAF<0.0006 for all 5 lead variants in non-Finnish Europeans according to gnomAD[43]). At the remaining 6 loci, the lead variants from the AA analysis were also genome wide significant in the EA analysis, and shared similar effect sizes.

Across the discovery, combined, EA, AA and HIS meta-analyses, we identified 135 distinct loci (**Figure 2**). A summary of each locus, including LD patterns between lead variants from each

meta-analysis as well as independent signals and association test results across all meta-analyses, is available in **Table S8**.

Genetic Risk Score

Using the 100 lead variants identified in the EA meta-analysis, we developed a GRS that was applied to independent UKB EA participants, which included 18,516 cases and 92,929 controls (**Figure 3.A** and **3.B**). The GRS was significantly associated with VTE status (OR=1.55, 95% confidence interval [CI]=[1.53-1.58]) and the phenotypic variance explained by the score was estimated at 0.051. To assess the predictive ability of the score, we first calculated the AUC of the base model, which included the age, sex and 10 genetic PCs, and obtained AUC_{base}=0.516 (CI=[0.511-0.520]). After adding the GRS to the model, the AUC reached AUC_{GRS}=0.620 (CI=[0.616-0.625]), an increase of Δ-AUC=0.104 over the base model. Compared to individuals with a score in the middle stratum (44 to 55%), participants with a GRS in the top 1% had a significantly higher risk (OR=6.07, CI=[5.33-6.91]), while individuals in the bottom 1% had a significant risk reduction (OR=0.52, CI=[0.42-0.65]) (**Figure 3**, **Table S9**).

Gene Prioritization with TWAS and Protein QTL MR

Transcriptome Wide Association Study: Across the 6 single-tissue and 3 cross-tissues datasets analyzed, we identified 166 significant ($P<3.52\times10^{-6}$) and conditionally independent associations with a high posterior probability of colocalization (>0.75) between gene expression and VTE risk (see **Table S10**). These associations involved 108 genes, of which 77 were mapped to 46 genome-wide significant GWAS loci, leaving an additional 31 novel candidate genes that mapped outside of genome-wide significant GWAS loci (**Table S11**). At 33 GWAS loci, an associated gene matched the gene closest to the lead variant, supporting a role as a causal gene, while associated genes at the remaining 13 GWAS loci indicate genes for further investigation.

Protein QTL Mendelian Randomization: We performed agnostic MR of 1,216 plasma circulating pQTL using the combined VTE meta-analysis results and identified 23 proteins with a significant

association ($P<3.98\times10^{-5}$, **Figure 4**, **Table S12**). For 13 proteins, the gene coordinates matched a genome-wide significant GWAS locus and included 5 of the novel GWAS loci.

Association of VTE-associated Variants with Hemostasis and Hematology Traits

The association of any lead or conditionally independent variant at the 135 GWAS loci with hemostasis traits is presented in **Figure 5.A** and **Table S13**. Among the 92 novel (replicated and candidate) loci reported above, 18 (19%) had a variant associated with 1 or more of the 10 hemostasis traits (**Figure S10.A**).

Next, we investigated associations of the 135 GWAS loci with hematology traits, presented in **Figure 5.B** and **Table S14**. Across all 15 CBC measures and among the 92 novel loci, we observed at least 1 association at 55 (59%) novel (replicated and candidate) loci (**Figure S10.B**). Loci shared between hemostatic factors and VTE mostly displayed biologically consistent effect directions, with the exception of FVII, which shared 4 loci with the same effect direction than VTE, 4 with an opposite direction, and 1 with 2 independent variants that displayed the same direction for the first and an opposite direction for the second. Hematology traits displayed less consistent directions of effect with VTE across shared loci.

Phenome-wide Association Studies

We performed a pheWAS of lead and conditionally independent variants at the 135 significantly associated loci across 1,500 publicly available phenotypes involving EA UKB participants (**Table S15**). For each trait, only genome-wide significant variants were retrieved, and we restricted our analyses on traits sharing at least 10 loci with VTE (**Figure 6**, **Table S16**), which might indicate common biological pathways. Hematology traits, in particular platelet traits, shared the most loci with VTE (for example 33 for platelet count), consistent with our observations from the larger CBC GWAS (n~750,000) sample (**Figure 5.B**). Several traits correspond to height and weight measurements, as well as enzymes mainly produced by the liver (such as albumin, sex-hormone binding globulin, or insulin growth factor-1), and plasma lipid-related traits (Apolipoprotein-A and B, HDL cholesterol, or triglycerides). Blood pressure (systolic and

diastolic), glycated hemoglobin, calcium, cystatin C, and C-reactive protein levels were among additional traits sharing at least 10 loci with VTE. Few traits had a consistent direction of effect with respect to VTE risk across shared loci (**Figure 6**). For example, out of 10 loci shared between bilirubin levels and VTE, 9 (90%) were associated with an increase of both bilirubin levels and VTE risk. For albumin levels, glycated hemoglobin, and systolic blood pressure, an opposite direction of effect between these traits and VTE risk was observed at more than 75% of shared loci.

DISCUSSION

We identified 135 independent genomic loci and 39 additional genes from TWAS and pQTL associated with an increased or decreased risk of VTE. This reflects a substantial increase in the number of validated and candidate loci for VTE risk beyond past genetic mapping efforts.[13,14] While the novel VTE associated variants were typically non-coding and displayed small effect sizes, they may provide valuable insights into genetic loci not previously suspected to play a role in VTE. Our results highlight genetic variation across the rare-to-common allele frequency spectrum in multiple ancestry groups and add new evidence of biologic predictors of VTE pathogenesis for further investigation. The *in silico* interrogations provide valuable clues regarding the putative causal gene at each locus and additional insights to biological pathways shared with VTE.

Biological Insights

Novel Replicated Loci: Our strongest evidence supports 34 loci with novel VTE associations. Except for *TFPI* and *SERPINE2*, the novel genetic loci were not in established VTE pathophysiology pathways. A subset of these 34 loci (12 loci, 35%) was associated with plasma levels of the hemostasis traits interrogated and most (26 loci, 76%) were associated with a hematology trait. This contrast should be interpreted with caution as statistical power for the hemostasis traits was much smaller than for the hematology traits.

While most of the novel associations reported had an OR in the range of [0.90-0.98; 1.03-1.10], we were able to identify and replicate 3 uncommon variants with larger estimated effects: an intronic variant (MAF=0.021) in the glycosyltransferase *ST3GAL4* ($OR_{discovery}$ OR=1.21, $OR_{replication}$=1.18), which was also associated with increased vWF and FVIII levels, an intronic variant (MAF=0.029) in the transcriptional co-activator *ZMIZ1* ($OR_{discovery}$=1.15, $OR_{replication}$=1.11), and an exonic variant (MAF=0.027) in *MAP1A* (p.Pro2349Leu, $OR_{discovery}$=0.87, $OR_{replication}$=0.84), which was also associated with decreased levels of vWF and fibrinogen, and had a protective effect against VTE.

Variants associated with hemostasis traits provide clues that the causal gene at these loci might directly or indirectly perturb the coagulation cascade. For instance, *XXYLT1* encodes a xylosyltransferase known to interact with coagulation factors[44] and had a nearby variant ($OR_{discovery}$ =1.06, $OR_{replication}$ =1.06) also associated with decreased FVII levels. Another example is *FUT2*, a fucosyltransferase gene involved in the synthesis of the H antigen, a building block for the production of antigens within the ABO blood group. *FUT2* had a downstream variant ($OR_{discovery}$ =0.96, $OR_{replication}$ =0.96) that was also associated with decreased vWF levels, mirroring results observed with vWF at the *ABO* locus. In addition, some variants were associated with several hematology traits, suggesting common genetic regulatory pathways affecting hematopoiesis, such as the replicated *RCOR1* signal on chromosome 14, and the candidate gene *REST* on chromosome 4 identified in the combined meta-analysis, 2 genes that form the transcriptional repressor CoREST, known to mediate hematopoiesis.[45]

Among the 34 loci, 17 (50%) had TWAS evidence linking transcript expression with a gene in the locus and 3 were linked to protein measures. These results may help to prioritize biologically relevant genes for further investigations. Notably, at the *COPZ1* locus, the lead variant was associated with several CBC measures, including platelet count and red blood cell count, and the TWAS revealed an association with *NFE2*, known to regulate erythroid and megakaryocyte maturation.[46]

Other Replicated and Non-Replicated Loci: Replicated variants included 2 rare variants at the known *EPHA3* (intergenic, MAF=0.0024, OR=2.40) and *FADS2B* (intronic, MAF=0.0047, OR=0.64) loci. Among the 17 failed replications, 7 reached nominal significance (P<0.05), suggesting that these variants might need a larger replication sample to be validated. See Supplemental Materials for more details.

Novel Candidate Loci: Across the multiple interrogation approaches, we identified several scores of candidate loci with evidence to support their association with VTE, though not yet replicated. This included 35 candidates from the combined GWAS, 7 candidates from the EA GWAS, and 2 candidates from the AA GWAS. Interestingly, the 2 variants (MAF 0.04 and 0.011) in the AA population were not present in EU participants and were associated with nearly 2-fold changes in risk of VTE. However, these 2 variants were only detected in a subset of studies, which included only 882 AA VTE cases out of 7,482, warranting additional investigations to confirm these 2 signals in *RBFOX1* (an RNA-binding protein) and *COL6A2* (a collagen-generating gene that contains several domains similar to *VWF* type A domains). For the remaining candidate GWAS loci, we saw similar attributes and associations as we did with the replicated loci. With additional replication resources in the future, these candidates may become fully replicated genetic associations.

In addition, the conditional analyses revealed independently associated variants mapping to distinct genes that may be of interest for further investigations, such as *BRD3* at the *ABO* locus, a chromatin reader known to associate with the hematopoietic transcription factor *GATA1*.[47] At the *EPHA3* locus, we also noted that the lead GWAS variant and the conditionally independent variant mapped upstream and downstream of *PROS2P*, a protein S pseudogene that might be of interest.

At these candidate loci, gene prioritized by the TWAS may also provide putative genes at these loci. For example *ZBTB7B*, a zinc-finger protein that represses the expression of extracellular matrix genes such as fibronectin and collagen[48] was identified by TWAS at the GWAS candidate

locus *DCST2*. The 31 candidate genes identified in the TWAS as well as the additional 8 from the pQTL MR analyses, although lacking a significant genetic association at these loci, might indicate relevant genes for future investigations. For instance, SYK is a critical platelet-activation protein and tyrosine kinase inhibitors of SYK have been explored for platelet inhibition.[49,50]

Clinical Implications

The GRS provided VTE risk discrimination in our EA population and those at the extremes of the score distribution experienced multi-fold risk differences. We were not able to integrate or to compare non-genetic risk factors with the GRS.

Current anticoagulation therapy to prevent or treat VTE operate through the modulation of proteins produced in the liver (coumarin-based therapies) or through direct inhibition of coagulation factors IIa (thrombin) and Xa. Although the safety profile of anticoagulation treatments has evolved, bleeding remains a life-threatening off-target outcome. New approaches to preventing thrombosis while minimizing bleeds are in development, including a focus on contact (intrinsic) pathway proteins factor XI, factor XII, prekallikrein, and high-molecular-weight kininogen.[51] Agnostic interrogations such as these may lead to discovery of novel proteins that "break the inexorable link between antithrombotic therapy and bleeding risk."[52]

The hematology traits investigations and the pheWAS established that CBC measures share a large number of loci with VTE, and platelet phenotypes in particular are the most frequent traits shared with VTE variants: 51 loci were associated with either platelet count, mean platelet volume, plateletcrit or platelet distribution width in the pheWAS, and 35 of these loci are novel, which represents more than a third of all novel genetic associations. Several loci associated with VTE harbor genes with known roles in hematopoiesis and megakaryocyte development, or platelet turnover,[45,46,53–60] or platelet aggregation (see Supplemental Materials).[10,61–71] Altered platelet generation, turnover or reactivity may be a feature of VTE pathogenesis. For one, past prospective studies[72] and case-control studies[73,74] suggest that

enlarged platelets, as measured by MPV, are associated with VTE and VTE outcomes. Studies of platelet function measures with VTE have been less conclusive which may relate to the limitations of these studies in assessing comprehensive and standardized platelet reactivity mechanisms.[75–77] Collectively, these results suggest that treatments inhibiting platelet activation such as aspirin might be beneficial in the prevention of VTE, although previous studies and trials on aspirin and combinations with anticoagulants offered mixed results.[78] Different antiplatelets, such as more targeted thrombin, PAR1 or PAR4 inhibitors, or intracellular PDE platelet signaling inhibitors like cilostazol, could be worthwhile for further study in VTE prevention.

Strengths and Limitations

The major strength of this genetic discovery effort is the large sample size of the populations contributing to the genetic variation interrogations. We increased statistical power compared with previous VTE GWAS meta-analysis efforts and increased our ability to detect new associations, many of which were replicated, and less common genetic variation. The cross-ancestry meta-analyses also increased discovery potential where allele frequencies were more common in some populations compared with others.

Several limitations deserve mention. Case ascertainment varied by study and some studies provided validated VTE events while others relied on information from electronic health records. Further, some studies only included hospitalized VTE events and did not capture events in the outpatient setting. These differences may have introduced some bias if case ascertainment and hospitalization status have genetic determinants. We included all VTE cases and did not stratify by provoked status in order to increase statistical power. Furthermore, many of the studies had not classified the VTE events as provoked and unprovoked. In addition, although the cross-ancestry approach provided benefits as described above, the numbers of VTE cases were not evenly distributed by ancestry, thus reducing our ability to detect ancestry-specific VTE variants in the under-represented ancestry groups with more modest case counts. Due to the diversity of imputation panels used by the participating studies, genetic variants had

variable coverage across studies which weakened our power to detect associations. Another limitation of our approach that used summary GWAS statistics from meta-analyses is the absence of participant-specific genotype-level information. This required us to rely on LD information extracted from external datasets, which can result in variants being missed and LD patterns not accurately captured. This may have introduced some bias in analyses that relied on LD, such as the conditional analyses and the TWAS. Further, *in silico* work was performed using external data sets such as the hemostatic factors and hematology traits summary statistics, where the size (and statistical power) of the datasets varied greatly. Although different significance thresholds were employed for significance, this may have biased the detection of significant associations to those traits that had large sample sizes. In addition, the pQTL MR analyses relied in some cases on a single genetic instrument, such as the *KLKB1* analysis, and these results should be considered hypothesis generating.

Summary

These cross-ancestry GWAS meta-analyzes identified 34 loci that replicated discovery findings. Some of the novel loci may contribute to VTE through well-characterized coagulation pathways while others provide new data on the role of hematology traits, particularly platelet function. Many of the replicated loci are outside of known or currently hypothesized pathways to thrombosis. We also provided a list of 44 new candidate loci including candidates from the combined cross-ancestry GWAS, from the EA GWAS, from the AA GWAS, and also 39 candidate genes from the TWAS and pQTL MR. These findings highlight new pathways to thrombosis and provide novel molecules that may be useful in the development of antithrombosis treatment that reduce bleeding adverse occurrences.

Acknowledgments

The INVENT Consortium would like to acknowledge all the participants across studies that provided their health information to support these analyses.

Sources of Funding

The INVENT Consortium is supported in part by HL134894 and HL154385. The Analysis Commons was funded by R01HL131136. Infrastructure for the CHARGE Consortium is supported in part by the National Heart, Lung, and Blood Institute grant R01HL105756. Study-specific acknowledgements and funding can be found in the **Supplemental Materials**. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung and Blood Institute, the National Institute of Health, Department of Veterans Affairs, or the U.S. Department of Health and Human Services.

Disclosures

B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. P.M.R. has received investigator initiated research grant support for unrelated projects from NHLBI, Operation Warp Speed, Novartis, Kowa, Amarin, and Pfizer; and has served as a consultant on unrelated issues to Novo Nordisk, Flame, Agepha, Uppton, Novartis, Jansen, Health Outlook, Civi Biopharm, Alnylam, and SOCAR. P.N. reports grants from Amgen, Apple, AstraZeneca, Boston Scientific, and Novartis, personal fees from Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Genentech / Roche, Novartis, and TenSixteen Bio, equity in geneXwell, TenSixteen Bio, and Zizi, co-founder of TenSixteen Bio, and spousal employment at Vertex, all unrelated to the present work. G.C-P and X.W are employed by and hold stock or stock options in 23andMe, Inc. The spouse of C.J.W. works at Regeneron Pharmaceuticals.

References

1. Silverstein, M. D. *et al.* Trends in the incidence of deep vein thrombosis and pulmonary embolism: a 25-year population-based study. *Arch Intern Med* **158**, 585–593 (1998).
2. Ghanima, W. *et al.* Incidence and prevalence of venous thromboembolism in Norway 2010-2017. *Thromb Res* **195**, 165–168 (2020).
3. Delluc, A. *et al.* Current incidence of venous thromboembolism and comparison with 1998: a community-based study in Western France. *Thromb Haemost* **116**, 967–974 (2016).
4. Smith, N. L. *et al.* Association of genetic variations with nonfatal venous thrombosis in postmenopausal women. *JAMA* **297**, 489–498 (2007).
5. Bezemer, I. D. *et al.* Gene variants associated with deep vein thrombosis. *JAMA* **299**, 1306–1314 (2008).
6. Heit, J. A. *et al.* A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *J. Thromb. Haemost.* **10**, 1521–1531 (2012).
7. Buil, A. *et al.* C4BPB/C4BPA is a new susceptibility locus for venous thrombosis with unknown protein S-independent mechanism: results from genome-wide association and gene expression analyses followed by case-control studies. *Blood* **115**, 4644–4650 (2010).
8. Tang, W. *et al.* A genome-wide association study for venous thromboembolism: the extended cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium. *Genet Epidemiol* **37**, 512–521 (2013).
9. 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.* **96**, 532–542 (2015).

10. Bennett, J. A. *et al.* The choline transporter Slc44a2 controls platelet activation and thrombosis by regulating mitochondrial function. *Nat Commun* **11**, 3479 (2020).
11. Constantinescu-Bercu, A. *et al.* Activated αIIbβ3 on platelets mediates flow-dependent NETosis via SLC44A2. *Elife* **9**, e53353 (2020).
12. Zirka, G. *et al.* Impaired adhesion of neutrophils expressing Slc44a2/HNA-3b to VWF protects against NETosis under venous shear rates. *Blood* **137**, 2256–2266 (2021).
13. Lindström, S. *et al.* Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood* **134**, 1645–1657 (2019).
14. Klarin, D. *et al.* Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nat Genet* **51**, 1574–1579 (2019).
15. Zhou, W. *et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human diseases*. 2021.11.19.21266436 https://www.medrxiv.org/content/10.1101/2021.11.19.21266436v1 (2021) doi:10.1101/2021.11.19.21266436.
16. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
17. Hinds, D. A. *et al.* Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. *Hum Mol Genet* **25**, 1867–1874 (2016).
18. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

19. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
20. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369–375, S1-3 (2012).
21. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* **15**, e1008500 (2019).
22. Sabik, O. L. & Farber, C. R. *RACER: A data visualization strategy for exploring multiple genetic associations*. 495366 https://www.biorxiv.org/content/10.1101/495366v3 (2018) doi:10.1101/495366.
23. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Front Genet* **11**, 157 (2020).
24. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
25. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245–252 (2016).
26. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
27. Nuotio, J. *et al.* Cardiovascular risk factors in 2011 and secular trends since 2007: the Cardiovascular Risk in Young Finns Study. *Scand J Public Health* **42**, 563–571 (2014).

28. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat Genet* **46**, 430–437 (2014).

29. Zhao, H. *et al.* Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. 2022.01.09.21268473 (2022) doi:10.1101/2022.01.09.21268473.

30. de Vries, P. S. *et al.* A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. *Hum. Mol. Genet.* **25**, 358–370 (2016).

31. Smith, N. L. *et al.* Genetic predictors of fibrin D-dimer levels in healthy adults. *Circulation* **123**, 1864–1872 (2011).

32. de Vries, P. S. *et al.* A genome-wide association study identifies new loci for factor VII and implicates factor VII in ischemic stroke etiology. *Blood* **133**, 967–977 (2019).

33. Sabater-Lleal, M. *et al.* Genome-Wide Association Transethnic Meta-Analyses Identifies Novel Associations Regulating Coagulation Factor VIII and von Willebrand Factor Plasma Levels. *Circulation* **139**, 620–635 (2019).

34. 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.* **26**, 637–649 (2017).

35. Huang, J. *et al.* Genome-wide association study for circulating tissue plasminogen activator levels and functional follow-up implicates endothelial STXBP5 and STX2. *Arterioscler. Thromb. Vasc. Biol.* **34**, 1093–1101 (2014).

36. Huang, J. *et al.* Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation. *Blood* **120**, 4873–4881 (2012).

798 37. Tang, W. *et al.* Genetic associations for activated partial thromboplastin time and
799 prothrombin time, their gene expression profiles, and risk of coronary artery disease. *Am J*
800 *Hum Genet* **91**, 152–162 (2012).
801 38. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667
802 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).
803 39. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic
804 Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).
805 40. Elsworth, B. *et al. The MRC IEU OpenGWAS data infrastructure*. 2020.08.10.244293
806 https://www.biorxiv.org/content/10.1101/2020.08.10.244293v1 (2020)
807 doi:10.1101/2020.08.10.244293.
808 41. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *European Journal of*
809 *Human Genetics* **19**, 807 (2011).
810 42. Riis, J., Nordestgaard, B. G. & Afzal, S. α1 -Antitrypsin Z allele and risk of venous
811 thromboembolism in the general population. *J Thromb Haemost* **20**, 115–125 (2022).
812 43. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
813 141,456 humans. *Nature* **581**, 434–443 (2020).
814 44. Minamida, S. *et al.* Detection of UDP-D-xylose: alpha-D-xyloside alpha 1--
815 >3xylosyltransferase activity in human hepatoma cell line HepG2. *J Biochem* **120**, 1002–
816 1006 (1996).
817 45. Saleque, S., Kim, J., Rooke, H. M. & Orkin, S. H. Epigenetic regulation of hematopoietic
818 differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. *Mol Cell*
819 **27**, 562–572 (2007).

820 46. Schulze, H. & Shivdasani, R. A. Mechanisms of thrombopoiesis. *J Thromb Haemost* **3**, 1717–
821 1724 (2005).
822 47. Lamonica, J. M. *et al.* Bromodomain protein Brd3 associates with acetylated GATA1 to
823 promote its chromatin occupancy at erythroid target genes. *Proc Natl Acad Sci U S A* **108**,
824 E159-168 (2011).
825 48. Widom, R. L., Lee, J. Y., Joseph, C., Gordon-Froome, I. & Korn, J. H. The hcKrox gene family
826 regulates multiple extracellular matrix genes. *Matrix Biol* **20**, 451–462 (2001).
827 49. Perrella, G. *et al.* Role of Tyrosine Kinase Syk in Thrombus Stabilisation at High Shear. *Int J*
828 *Mol Sci* **23**, 493 (2022).
829 50. Zheng, T. J. *et al.* Assessment of the effects of Syk and BTK inhibitors on GPVI-mediated
830 platelet signaling and function. *Am J Physiol Cell Physiol* **320**, C902–C915 (2021).
831 51. Fredenburgh, J. C. & Weitz, J. I. New anticoagulants: Moving beyond the direct oral
832 anticoagulants. *J Thromb Haemost* **19**, 20–29 (2021).
833 52. Mackman, N., Bergmeier, W., Stouffer, G. A. & Weitz, J. I. Therapeutic strategies for
834 thrombosis: new targets and approaches. *Nat Rev Drug Discov* **19**, 333–352 (2020).
835 53. Han, L. *et al.* Chromatin remodeling mediated by ARID1A is indispensable for normal
836 hematopoiesis in mice. *Leukemia* **33**, 2291–2305 (2019).
837 54. Scheicher, R. *et al.* CDK6 as a key regulator of hematopoietic and leukemic stem cell
838 activation. *Blood* **125**, 90–101 (2015).
839 55. Maslah, N., Cassinat, B., Verger, E., Kiladjian, J.-J. & Velazquez, L. The role of LNK/SH2B3
840 genetic alterations in myeloproliferative neoplasms and other hematological disorders.
841 *Leukemia* **31**, 1661–1670 (2017).

842 56. Mancini, E. *et al.* FOG-1 and GATA-1 act sequentially to specify definitive megakaryocytic
843 and erythroid progenitors. *EMBO J* **31**, 351–365 (2012).
844 57. Krosl, J. *et al.* A mutant allele of the Swi/Snf member BAF250a determines the pool size of
845 fetal liver hemopoietic stem cell populations. *Blood* **116**, 1678–1684 (2010).
846 58. Ayoub, E. *et al.* EVI1 overexpression reprograms hematopoiesis via upregulation of Spi1
847 transcription. *Nat Commun* **9**, 4239 (2018).
848 59. Fonseca-Pereira, D. *et al.* The neurotrophic factor receptor RET drives haematopoietic stem
849 cell survival and function. *Nature* **514**, 98–101 (2014).
850 60. Gregory, G. D. *et al.* FOG1 requires NuRD to promote hematopoiesis and maintain lineage
851 fidelity within the megakaryocytic-erythroid compartment. *Blood* **115**, 2156–2166 (2010).
852 61. Keramati, A. R. *et al.* Genome sequencing unveils a regulatory landscape of platelet
853 reactivity. *Nat Commun* **12**, 3626 (2021).
854 62. Mitsui, T. *et al.* ALOX12 mutation in a family with dominantly inherited bleeding diathesis. *J*
855 *Hum Genet* **66**, 753–759 (2021).
856 63. Fukami, K. Structure, regulation, and function of phospholipase C isozymes. *J Biochem* **131**,
857 293–299 (2002).
858 64. Johnson, A. D. *et al.* Genome-wide meta-analyses identifies seven loci associated with
859 platelet aggregation in response to agonists. *Nat Genet* **42**, 608–613 (2010).
860 65. Radomski, A. *et al.* Identification, regulation and role of tissue inhibitor of
861 metalloproteinases-4 (TIMP-4) in human platelets. *Br J Pharmacol* **137**, 1330–1338 (2002).

66. Moore, S. F., Smith, N. R., Blair, T. A., Durrant, T. N. & Hers, I. Critical roles for the phosphatidylinositide 3-kinase isoforms p110β and p110γ in thrombopoietin-mediated priming of platelet function. *Sci Rep* **9**, 1468 (2019).
67. Kuijpers, M. J. E. *et al.* Platelet CD40L Modulates Thrombus Growth Via Phosphatidylinositol 3-Kinase β, and Not Via CD40 and IκB Kinase α. *Arterioscler Thromb Vasc Biol* **35**, 1374–1381 (2015).
68. Rodriguez, B. A. T. *et al.* A Platelet Function Modulator of Thrombin Activation Is Causally Linked to Cardiovascular Disease and Affects PAR4 Receptor Signaling. *Am J Hum Genet* **107**, 211–221 (2020).
69. Antl, M. *et al.* IRAG mediates NO/cGMP-dependent inhibition of platelet aggregation and thrombus formation. *Blood* **109**, 552–559 (2007).
70. Schinner, E., Salb, K. & Schlossmann, J. Signaling via IRAG is essential for NO/cGMP-dependent inhibition of platelet activation. *Platelets* **22**, 217–227 (2011).
71. van Geffen, J. P. *et al.* High-throughput elucidation of thrombus formation reveals sources of platelet function variability. *Haematologica* **104**, 1256–1267 (2019).
72. Braekkan, S. K. *et al.* Mean platelet volume is a risk factor for venous thromboembolism: the Tromsø Study, Tromsø, Norway. *J Thromb Haemost* **8**, 157–162 (2010).
73. Ghaffari, S. *et al.* Prognostic value of platelet indices in patients with acute pulmonary thromboembolism. *J Cardiovasc Thorac Res* **12**, 56–62 (2020).
74. Farah, R., Nseir, W., Kagansky, D. & Khamisy-Farah, R. The role of neutrophil-lymphocyte ratio, and mean platelet volume in detecting patients with acute venous thromboembolism. *J Clin Lab Anal* **34**, e23010 (2020).

884 75. Puurunen, M. K., Hwang, S.-J., O'Donnell, C. J., Tofler, G. & Johnson, A. D. Platelet function
885 as a risk factor for venous thromboembolism in the Framingham Heart Study. *Thromb Res*
886 **151**, 57–62 (2017).
887 76. Sokol, J., Skerenova, M., Ivankova, J., Simurda, T. & Stasko, J. Association of Genetic
888 Variability in Selected Genes in Patients With Deep Vein Thrombosis and Platelet
889 Hyperaggregability. *Clin Appl Thromb Hemost* **24**, 1027–1032 (2018).
890 77. Panova-Noeva, M. *et al.* Comprehensive platelet phenotyping supports the role of platelets
891 in the pathogenesis of acute venous thromboembolism - results from clinical observation
892 studies. *EBioMedicine* **60**, 102978 (2020).
893 78. Diep, R. & Garcia, D. Does aspirin prevent venous thromboembolism? *Hematology Am Soc*
894 *Hematol Educ Program* **2020**, 634–641 (2020).
895 79. Klarin, D., Emdin, C. A., Natarajan, P., Conrad, M. F. & Kathiresan, S. Genetic Analysis of
896 Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates
897 Obesity as a Causal Risk Factor. *Circ Cardiovasc Genet* **10**, (2017).
898 80. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide
899 Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).
900 81. Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using
901 population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet*
902 **25**, 869–876 (2017).
903 82. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol*
904 **27**, S2–S8 (2017).

83. Smoller, J. W. *et al.* An eMERGE Clinical Center at Partners Personalized Medicine. *J Pers Med* **6**, E5 (2016).
84. Antoni, G. *et al.* A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism. *J Thromb Haemost* **8**, 2671–2679 (2010).
85. Ibrahim-Kosta, M. *et al.* Minor allele of the factor V K858R variant protects from venous thrombosis only in non-carriers of factor V Leiden mutation. *Sci Rep* **9**, 3750 (2019).
86. Vázquez-Santiago, M. *et al.* Short closure time values in PFA-100® are related to venous thrombotic risk. Results from the RETROVE Study. *Thromb Res* **169**, 57–63 (2018).
87. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol* **156**, 871–881 (2002).
88. Souto, J. C. *et al.* Genetic determinants of hemostasis phenotypes in Spanish families. *Circulation* **101**, 1546–1551 (2000).
89. Souto, J. C. *et al.* Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. *Am J Hum Genet* **67**, 1452–1459 (2000).
90. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol* **129**, 687–702 (1989).
91. Fried, L. P. *et al.* The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* **1**, 263–276 (1991).
92. Tell, G. S. *et al.* Recruitment of adults 65 years and older as participants in the Cardiovascular Health Study. *Ann Epidemiol* **3**, 358–366 (1993).

927 93. Trégouët, D.-A. *et al.* Common susceptibility alleles are unlikely to contribute as strongly as
928 the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood* **113**, 5298–5303
929 (2009).
930 94. McCarty, C. A. *et al.* The eMERGE Network: a consortium of biorepositories linked to
931 electronic medical records data for conducting genomic studies. *BMC Med Genomics* **4**, 13
932 (2011).
933 95. Milani, L., Leitsalu, L. & Metspalu, A. An epidemiological perspective of personalized
934 medicine: the Estonian experience. *J Intern Med* **277**, 188–200 (2015).
935 96. Zhu, T. *et al.* Association of influenza vaccination with reduced risk of venous
936 thromboembolism. *Thromb Haemost* **102**, 1259–1264 (2009).
937 97. Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. & Castelli, W. P. An
938 investigation of coronary heart disease in families. The Framingham offspring study. *Am J*
939 *Epidemiol* **110**, 281–290 (1979).
940 98. Feinleib, M., Kannel, W. B., Garrison, R. J., McNamara, P. M. & Castelli, W. P. The
941 Framingham Offspring Study. Design and preliminary data. *Prev Med* **4**, 518–525 (1975).
942 99. Smith, N. L. *et al.* Esterified estrogens and conjugated equine estrogens and the risk of
943 venous thrombosis. *JAMA* **292**, 1581–1587 (2004).
944 100. Holmen m.fl, J. The Nord-Trøndelag Health Study 1995-97 (HUNT 2). *Nor J Epidemiol* **13**,
945 (2011).
946 101. Glynn, R. J. *et al.* A randomized trial of rosuvastatin in the prevention of venous
947 thromboembolism. *N Engl J Med* **360**, 1851–1861 (2009).

948 102. Ridker, P. M. *et al.* Rosuvastatin to prevent vascular events in men and women with
949 elevated C-reactive protein. *N Engl J Med* **359**, 2195–2207 (2008).
950 103. Chasman, D. I. *et al.* Genetic Determinants of Statin-Induced Low-Density Lipoprotein
951 Cholesterol Reduction: The Justification for the Use of Statins in Prevention: An
952 Intervention Trial Evaluating Rosuvastatin (JUPITER) Trial. *Circ Cardiovasc Genet* **5**, 257–264
953 (2012).
954 104. Oudot-Mellakh, T. *et al.* Genome wide association study for plasma levels of natural
955 anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. *Br J*
956 *Haematol* **157**, 230–239 (2012).
957 105. 3C Study Group. Vascular factors and risk of dementia: design of the Three-City Study
958 and baseline characteristics of the study population. *Neuroepidemiology* **22**, 316–325
959 (2003).
960 106. Blom, J. W., Doggen, C. J. M., Osanto, S. & Rosendaal, F. R. Malignancies, prothrombotic
961 mutations, and the risk of venous thrombosis. *JAMA* **293**, 715–722 (2005).
962 107. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic
963 influences on health and disease. *J Clin Epidemiol* **70**, 214–223 (2016).
964 108. Hankinson, S. E. *et al.* Reproductive factors and family history of breast cancer in
965 relation to plasma estrogen and prolactin levels in postmenopausal women in the Nurses'
966 Health Study (United States). *Cancer Causes Control* **6**, 217–224 (1995).
967 109. Tworoger, S. S., Sluss, P. & Hankinson, S. E. Association between plasma prolactin
968 concentrations and risk of breast cancer among predominately premenopausal women.
969 *Cancer Res* **66**, 2476–2482 (2006).

970 110. Jacobsen, B. K., Eggen, A. E., Mathiesen, E. B., Wilsgaard, T. & Njølstad, I. Cohort profile:
971 the Tromso Study. *Int J Epidemiol* **41**, 961–967 (2012).
972 111. Braekkan, S. K. *et al.* Family history of myocardial infarction is an independent risk factor
973 for venous thromboembolism: the Tromsø study. *J Thromb Haemost* **6**, 1851–1857 (2008).
974 112. Design of the Women's Health Initiative clinical trial and observational study. The
975 Women's Health Initiative Study Group. *Control Clin Trials* **19**, 61–109 (1998).
976 113. Anderson, G. L. *et al.* Implementation of the Women's Health Initiative study design.
977 *Ann Epidemiol* **13**, S5-17 (2003).
978 114. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-
979 analyses. *Nat Protoc* **9**, 1192–1212 (2014).
980 115. Wolfe, D., Dudek, S., Ritchie, M. D. & Pendergrass, S. A. Visualizing genomic information
981 across chromosomes with PhenoGram. *BioData Min* **6**, 18 (2013).
982 116. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A better coefficient of
983 determination for genetic profile analysis. *Genet Epidemiol* **36**, 214–224 (2012).
984 117. The 1000 Genomes Project Consortium. A global reference for human genetic variation.
985 *Nature* **526**, 68–74 (2015).
986 118. Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African
987 ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat Genet*
988 **54**, 593–602 (2022).
989 119. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79
990 (2018).

120. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet* **13**, e1006706 (2017).

121. Desch, K. C. *et al.* Whole-exome sequencing identifies rare variants in STAB2 associated with venous thromboembolic disease. *Blood* **136**, 533–541 (2020).

122. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

123. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

Figures and legends

Figure 1: Analyses Workflow

Workflow of genetic analyses conducted for this study.

Figure 2: Genetic loci associated with VTE

This figure presents the 135 loci significantly associated with VTE identified across all 4 meta-analyses: the Discovery (in red), the overall meta-analysis (in green), the analysis restricted to individuals of European ancestry (in purple), African ancestry (in orange) and Hispanic ancestry (in blue) . Novel loci are represented with circles and known loci with diamonds. Loci with replication evidence are indicated with a red '*'.

Figure 3: Genetic risk score analysis

Distribution of the GRS in VTE cases (in green) and controls (in purple) as a density plot (A) and a boxplot (B). (C) Presentation of the VTE risk as odds ratios and associated 95% confidence intervals (y-axis) for different percentiles ranges of the GRS score (x-axis) relative to the middle range (45-55%).

Figure 4: Significant associations of protein QTL Mendelian Randomization

23 proteins significantly associated with VTE, out of 1,216 plasma protein analyzed, using the combined VTE summary statistics.

Figure 5: VTE genetic loci shared with hemostatic factors and blood traits

(A) Number of known and novel VTE loci shared with each of the 10 hemostatic factors investigated. Loci with shared variants that had an opposite effect direction between the trait and VTE are indicated in orange, while those that had the same effect direction are presented in blue. Loci with several independent shared variants and no consistent effect direction between the trait and VTE are indicated in gray. (B) Same analysis with complete blood count traits: PLT (platelet count), MPV (mean platelet volume), RBC (red blood cell count), MCV

(mean corpuscular volume), HCT (hematocrit), MCH (mean corpuscular hemoglobin), MCHC (MCH concentration), HGB (hemoglobin concentration), RDW (red cell distribution width), WBC (white blood cell count), MONO (monocyte count), NEU (neutrophil count), EOS (eosinophil count), BASO (basophil count), LYM (lymphocyte count).

Figure 6: PheWAS traits sharing at least 10 loci with VTE

This figure presents the pheWAS traits sharing at least 10 loci with VTE. Shape and color represent one of 5 categories: Complete Blood Count (CBC) traits, lipid traits, liver enzyme, height and weight traits, or other (if the trait did not fit in one of the aforementioned categories). The x-axis indicates the number of loci shared between VTE and the pheWAS trait, while the y-axis indicates the proportion of loci where the direction of effect was the same between the pheWAS trait and VTE. As a result, traits close to 100% have the same direction of effect than VTE at most shared loci, while traits close to 0% have an opposite direction than VTE at most shared loci.

Tables

Table 1: 68 Lead variants from the Discovery meta-analysis successfully replicated

rsID	CHR:POS:EA:NEA	EAF.Disc	OR.Disc	P.Disc	OR.Repl	P.Repl	Locus.Context	Locus.Gene
rs9442580	1:9339467:T:C	0.1551	1.06	1.83E-08	1.03	9.70E-05	intergenic	***H6PD;SPSB1***
rs3767812	1:118155620:A:G	0.2437	1.05	9.64E-11	1.06	1.03E-20	intronic	***TENT5C***
rs6025	1:169519049:T:C	0.0259	3.02	8.40E-811	3.59	9.29E-3103	exonic	*F5* (p.Q534Q)
rs2842700	1:207282149:A:C	0.1092	1.11	5.95E-17	1.12	1.19E-25	intronic	*C4BPA*
rs3811444	1:248039451:T:C	0.3324	0.96	5.70E-09	0.95	1.53E-20	exonic	***TRIM58*** (p.T374M)
rs7600986	2:68636923:A:T	0.2819	1.06	3.54E-12	1.05	9.18E-19	intergenic	*PLEK;FBXO48*
rs182293241	2:128029746:A:G	0.0195	1.89	1.83E-27	1.55	0.0001063	intronic	*ERCC3*
rs6719550	2:188272460:T:C	0.6639	1.04	7.56E-09	1.05	1.93E-17	intronic	***CALCRL***
rs715	2:211543055:T:C	0.7022	0.95	3.51E-09	0.95	1.43E-17	UTR3	***CPS1***
rs13412535	2:224874874:A:G	0.2047	1.06	3.05E-10	1.08	1.10E-36	intronic	***SERPINE2***
rs13084580	3:39188182:T:C	0.1076	1.09	2.89E-15	1.08	9.10E-22	exonic	*CSRNP1* (p.G18S)
rs562281690	3:90177913:T:G	0.0024	2.01	6.45E-15	2.40	8.68E-31	intergenic	*EPHA3;NONE*
rs62282204	3:138584405:T:C	0.5784	0.96	1.87E-08	0.98	6.73E-05	intergenic	***PIK3CB;LINC01391***
rs7613621	3:169191186:A:G	0.4467	1.04	3.21E-09	1.03	5.33E-09	intronic	***MECOM***
rs710446	3:186459927:T:C	0.5799	0.96	5.92E-11	0.96	1.41E-16	exonic	*KNG1* (p.I581I)
rs6797948	3:194784705:T:C	0.7983	1.06	2.99E-11	1.05	7.59E-16	intergenic	***LINC01968;XXYLT1***
rs6826579	4:83785031:T:C	0.7914	1.05	2.38E-08	1.03	2.44E-07	intronic	***SEC31A***
rs17010957	4:86719165:T:C	0.8581	1.06	3.99E-09	1.05	1.00E-11	intronic	***ARHGAP24***
rs2066864	4:155525695:A:G	0.2585	1.23	1.98E-172	1.23	1.94E-284	UTR3	*FGG*
rs3756011	4:187206249:A:C	0.3903	1.23	7.48E-198	1.24	9.26e-398	intronic	*F11*
rs16867574	5:38708554:T:C	0.6673	0.95	2.78E-11	0.95	5.67E-16	ncRNA_intronic	*OSMR-AS1*
rs38032	5:96321887:T:C	0.6049	1.04	8.74E-09	1.03	1.49E-09	intronic	***LNPEP***
rs9268881	6:32431606:A:T	0.5727	0.96	4.17E-10	0.97	6.73E-09	intergenic	***HLA-DRA;HLA-DRB5***
rs145294670	6:34622561:A:AG	0.1385	1.06	6.11E-10	1.04	6.89E-06	intronic	***ILRUN***
rs9390460	6:147694334:T:C	0.4957	0.95	2.49E-13	0.95	1.01E-20	intronic	*STXBP5*
rs67694436	8:6654220:T:C	0.3486	0.96	3.94E-08	0.98	0.0001105	intergenic	***AGPAT5;XKR5***
rs2685417	8:27807434:C:G	0.2562	1.06	1.57E-14	1.06	2.84E-25	intronic	*SCARA5*
rs6993770	8:106581528:A:T	0.7142	1.08	4.48E-25	1.09	3.55E-48	intronic	*ZFPM2*
rs35208412	9:99194509:A:AT	0.8298	1.09	1.56E-08	1.04	5.54E-06	intergenic	***ZNF367;HABP4***
rs505922	9:136149229:T:C	0.6334	0.74	1.11E-425	0.69	1.55E-1043	intronic	*ABO*
rs1887091	10:14535113:T:C	0.4936	0.96	4.77E-08	0.98	0.001107	intergenic	***MIR1265;FAM107B***
rs17490626	10:71218646:C:G	0.1136	0.80	1.02E-79	0.80	3.23E-160	intronic	*TSPAN15*
rs16937003	10:80938499:A:G	0.0287	1.15	1.07E-08	1.11	2.11E-11	intronic	***ZMIZ1***
rs2274224	10:96039597:C:G	0.4414	1.04	2.55E-09	1.03	1.29E-10	exonic	***PLCE1*** (p.R1267P)
rs10886430	10:121010256:A:G	0.8897	0.89	7.34E-25	0.88	2.76E-64	intronic	*GRK5*
rs11032074	11:32993887:A:G	0.7792	1.05	5.37E-09	1.03	3.24E-06	intronic	*QSER1*
rs1799963	11:46761055:A:G	0.0136	2.05	2.19E-135	2.09	6.86E-420	UTR3	*F2*
rs141687379	11:56666822:A:G	0.9953	0.52	3.56E-31	0.64	1.06E-42	intronic	*FADS2B*
rs174551	11:61573684:T:C	0.6583	1.07	1.65E-19	1.07	4.90E-35	intronic	*FADS1*
rs35257264	11:126296816:T:C	0.0212	1.21	2.88E-14	1.18	2.28E-24	intronic	***ST3GAL4***
rs1558519	12:6153738:A:G	0.6175	0.93	7.73E-24	0.92	1.42E-55	intronic	*VWF*
rs7311483	12:9053661:T:C	0.3589	0.96	2.74E-09	0.97	2.73E-07	intergenic	***A2ML1;PHC1***
rs6580981	12:54723028:A:G	0.5081	0.96	3.71E-09	0.95	2.26E-23	intronic	***COPZ1***
rs3184504	12:111884608:T:C	0.4520	1.05	1.18E-11	1.04	3.30E-12	exonic	***SH2B3*** (p.T178T)
rs3211752	13:113787459:A:G	0.5527	0.95	1.69E-12	0.94	3.49E-25	intronic	*F10*
rs57035593	14:92268096:T:C	0.3202	1.07	1.08E-20	1.07	2.64E-38	intronic	*TC2N*
rs8013957	14:103140254:T:C	0.3699	1.04	5.33E-09	1.03	2.23E-07	intronic	***RCOR1***
rs55707100	15:43820717:T:C	0.0270	0.87	2.90E-08	0.84	2.49E-27	exonic	***MAP1A*** (p.P2349L)

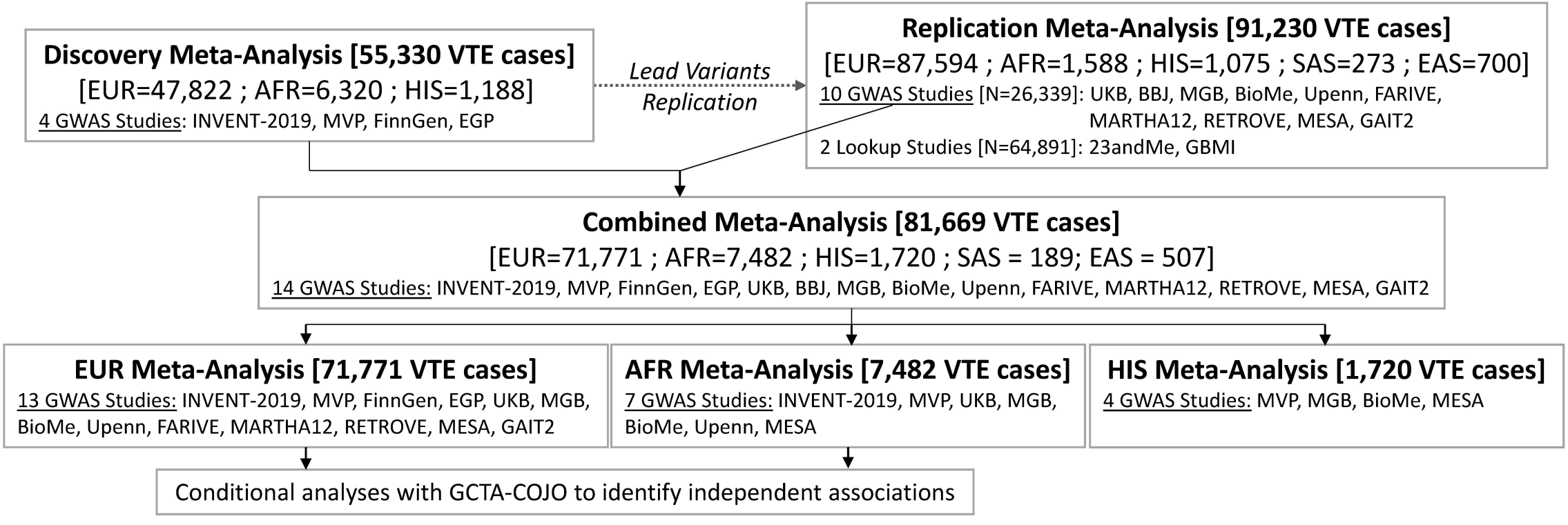
rs59442804	15:60899031:G:GAAAT	0.6438	0.96	4.67E-08	0.97	5.42E-10	ncRNA_intronic	***RORA-AS1***
rs12443808	16:30996871:C:G	0.4668	1.06	3.89E-14	1.03	1.85E-07	UTR5	***HSD3B7***
rs56943275	16:81898152:T:G	0.2446	1.08	4.15E-13	1.07	1.20E-26	intronic	*PLCG2*
rs28634651	16:88553198:T:C	0.6191	1.06	9.20E-13	1.04	7.62E-14	intronic	***ZFPM1***
rs6503222	17:1977862:A:G	0.6188	1.05	1.59E-12	1.04	5.21E-06	intronic	*SMG6*
rs7225756	17:6893691:A:G	0.4877	0.96	3.57E-08	0.98	1.20E-06	ncRNA_intronic	***ALOX12-AS1***
rs62054822	17:43927708:A:G	0.8028	0.95	6.39E-09	0.95	7.11E-19	ncRNA_intronic	***MAPT-AS1***
rs142140545	17:64191540:CTATT:C	0.1169	0.93	2.27E-08	0.95	7.83E-07	intergenic	***CEP112;APOH***
rs59277920	19:6077231:A:G	0.8210	0.94	1.47E-09	0.96	8.52E-06	intronic	***RFX2***
rs8110055	19:10739143:A:C	0.2000	0.89	5.36E-44	0.89	6.50E-70	intronic	*SLC44A2*
rs34783010	19:46180414:T:G	0.2132	0.95	3.25E-09	0.96	4.87E-10	intronic	***GIPR***
rs1688264	19:49209560:T:G	0.5341	0.96	2.07E-10	0.96	3.02E-15	downstream	***FUT2***
rs1654425	19:55538980:T:C	0.1468	0.91	2.65E-18	0.94	4.21E-14	exonic	*GP6* (p.S192S)
rs79388863	20:23168500:A:G	0.1521	0.92	1.74E-18	0.92	4.48E-27	intergenic	*LINC00656;NXT1*
rs6060288	20:33772243:A:G	0.3417	1.12	8.19E-54	1.13	1.52E-102	intronic	*MMP24-AS1-EDEM2*
rs4820093	22:33160208:T:C	0.2693	1.05	1.04E-08	1.04	5.39E-14	intronic	***SYN3***
rs9611844	22:43115776:C:G	0.1286	1.10	2.09E-21	1.07	7.54E-20	intronic	*A4GALT*
rs3002416	23:39710195:T:C	0.3638	0.95	2.20E-18	0.93	2.23E-23	intergenic	*MIR1587;BCOR*
rs6048	23:138633280:A:G	0.7215	1.07	1.09E-25	1.08	1.59E-46	exonic	*F9* (p.T156T)
rs2084408	23:154346709:T:G	0.3764	0.94	5.36E-19	0.94	6.27E-09	intronic	*BRCC3*

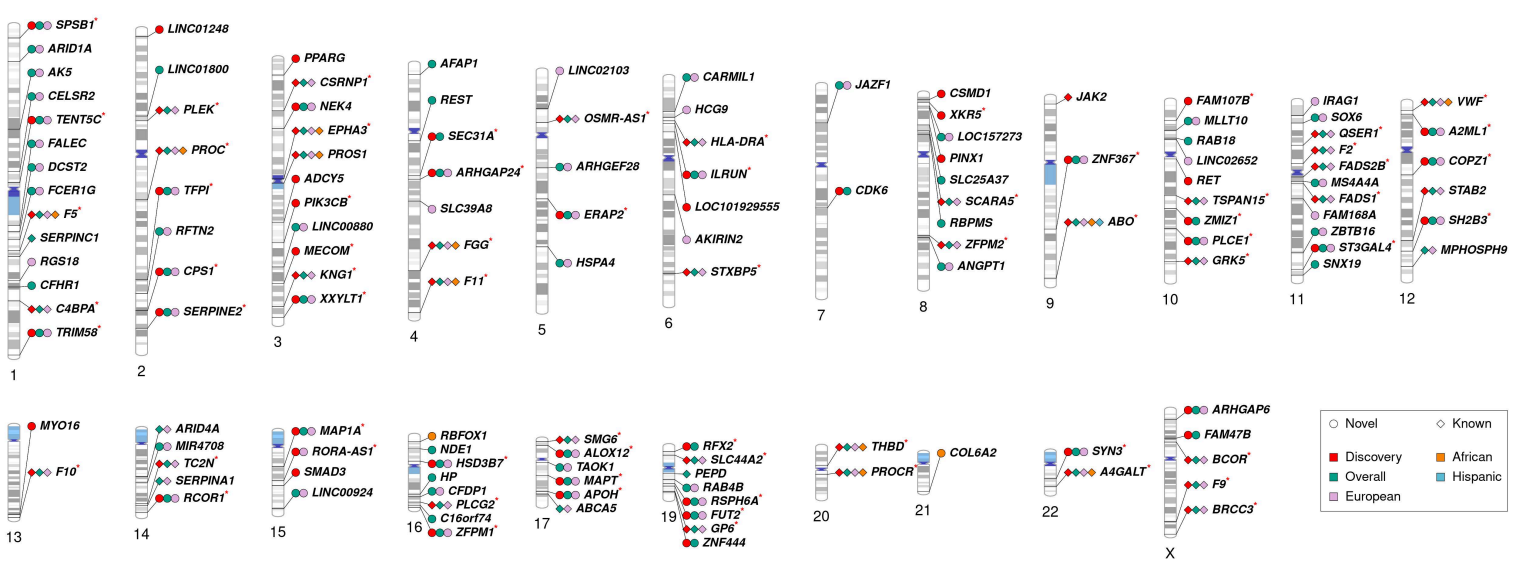
CHR: chromosome; POS: position (hg19 build); EA: effect allele; NEA: non effect allele; EAF: effect allele frequency; OR: odds ratio; P: P-value. Results from the discovery are in presented in columns suffixed with “Disc”, while results from the replication are in columns suffixed with “Repl”. Novel genetic associations are indicated as bold gene names.

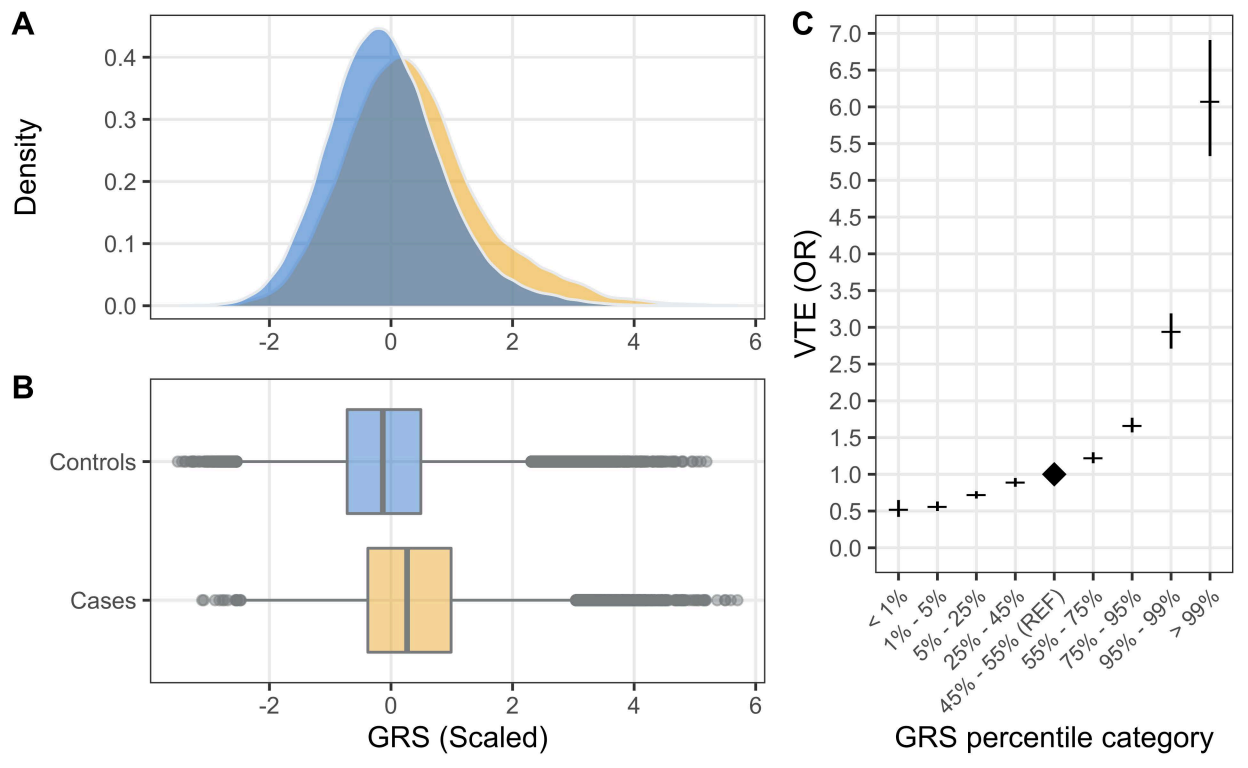
Table 2: Additional 44 candidate novel loci identified in the Overall, European and African meta-analyses

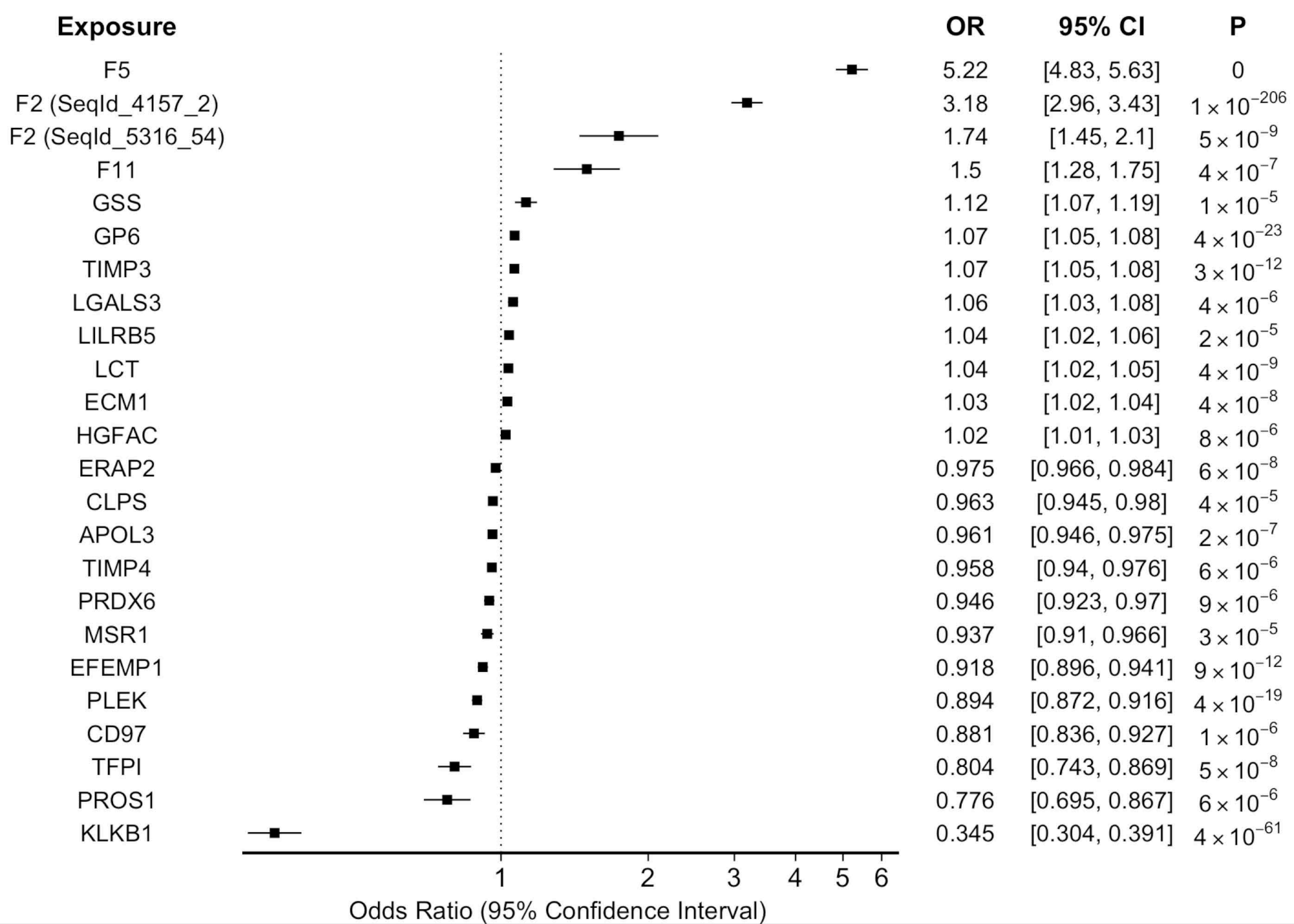
rsID	CHR:POS:EA:NEA	EAF	EFFECT	SE	OR	P	Locus.context	Locus.Gene
Novel loci identified in the overall meta-analysis								
rs551176418	1:27107263:T:TC	0.9248	0.0759	0.0132	1.08	9.61E-09	UTR3	*ARID1A*
rs6695572	1:77945635:A:G	0.1938	0.0424	0.0072	1.04	4.28E-09	intronic	*AK5*
rs3832016	1:109818158:CT:C	0.7627	-0.0449	0.0066	0.96	8.95E-12	UTR3	*CELSR2*
rs1267881263	1:150496127:CA:C	0.5468	0.0426	0.0076	1.04	2.36E-08	intergenic	*FALEC;ADAMTSL4*
rs905938	1:154991389:T:C	0.7448	-0.0346	0.0063	0.97	3.70E-08	intronic	*DCST2*
rs3557	1:161188893:T:G	0.9182	0.0654	0.0106	1.07	7.70E-10	UTR3	*FCER1G*
rs143410348	1:196809316:T:TAA	0.5434	0.0415	0.0074	1.04	2.44E-08	intergenic	*CFHR1;CFHR4*
rs78475244	2:65086804:T:C	0.0542	-0.0713	0.0128	0.93	2.52E-08	ncRNA_intronic	*LINC01800*
rs78872368	2:198545250:C:G	0.1919	-0.0412	0.0071	0.96	7.27E-09	intergenic	*RFTN2;MARS2*
rs900399	3:156798732:A:G	0.6205	0.0382	0.0060	1.04	1.46E-10	intergenic	*LINC02029;LINC00880*
rs9654093	4:7903763:C:G	0.1504	0.0492	0.0081	1.05	1.03E-09	intronic	*AFAP1*
rs781656	4:57778645:A:G	0.1963	0.0389	0.0070	1.04	2.26E-08	intronic	*REST*
rs7730244	5:72957088:T:C	0.5245	-0.0328	0.0057	0.97	1.04E-08	intronic	*ARHGEF28*
rs147133967	5:132426851:G:GTT	0.0810	-0.0659	0.0110	0.94	2.43E-09	intronic	*HSPA4*
rs214059	6:25536937:T:C	0.4331	0.0357	0.0055	1.04	1.01E-10	intronic	*CARMIL1*
rs2394251	6:29943688:G:C	0.7331	-0.0405	0.0063	0.96	1.43E-10	ncRNA_intronic	*HCG9*
rs1513275	7:28259233:T:C	0.7453	0.0449	0.0070	1.05	1.40E-10	ncRNA_intronic	*JAZF1-AS1*
rs10099512	8:9178821:C:G	0.1105	0.0608	0.0105	1.06	6.98E-09	intergenic	*LOC101929128;LOC157273*
rs2048528	8:23373680:A:G	0.3089	-0.0347	0.0060	0.97	5.77E-09	intergenic	*ENTPD4;SLC25A37*
rs2915595	8:30402817:A:G	0.2391	0.0365	0.0065	1.04	2.52E-08	intronic	*RBPMS*
rs4236786	8:108291878:C:G	0.2492	0.0353	0.0064	1.04	3.93E-08	intronic	*ANGPT1*
rs1243187	10:21907016:T:C	0.6920	-0.0341	0.0061	0.97	2.53E-08	intronic	*MLLT10*
rs4272700	10:27881308:A:T	0.2726	0.0395	0.0064	1.04	7.75E-10	intergenic	*RAB18;MKX*
rs2030291	11:16251251:A:T	0.6077	-0.0325	0.0056	0.97	8.19E-09	intronic	*SOX6*
rs4354705	11:60088159:C:G	0.3635	0.0315	0.0058	1.03	4.83E-08	intergenic	*MS4A4A;MS4A6E*
rs2846027	11:114003415:T:C	0.3112	-0.0344	0.0061	0.97	1.42E-08	intronic	*ZBTB16*
rs7107568	11:130779668:T:C	0.5610	-0.0303	0.0056	0.97	4.71E-08	intronic	*SNX19*
rs2127869	14:65794352:T:C	0.3350	-0.0340	0.0062	0.97	4.68E-08	intergenic	*LINC02324;MIR4708*
rs7183672	15:96101018:A:G	0.6432	-0.0358	0.0062	0.96	7.34E-09	intergenic	*LINC00924;LOC105369212*
rs71376077	16:15738114:C:G	0.9728	0.1408	0.0249	1.15	1.57E-08	intronic	*NDE1*
rs7197453	16:72079127:C:G	0.3572	0.0315	0.0057	1.03	3.19E-08	intergenic	*DHODH;HP*
rs77246010	16:75429853:T:C	0.4489	0.0408	0.0069	1.04	4.12E-09	intronic	*CFDP1*
rs8049403	16:85778651:A:G	0.0214	0.1365	0.0248	1.15	3.91E-08	intronic	*C16orf74*
rs71138827	17:27833678:A:AGATT	0.4288	0.0336	0.0058	1.03	5.89E-09	intronic	*TAOK1*
rs2545774	19:41287674:T:C	0.2528	-0.0378	0.0065	0.96	6.80E-09	intronic	*RAB4B*
Additional novel loci identified in the European meta-analysis								
rs4540639	1:192104320:C:G	0.4675	0.0346	0.0060	1.04	6.88E-09	intergenic	*LINC01680;RGS18*
rs35225200	4:103146888:A:C	0.9190	-0.0645	0.0115	0.94	1.89E-08	intergenic	*BANK1;SLC39A8*
rs112367053	5:28379046:T:G	0.6662	0.0586	0.0107	1.06	4.07E-08	intergenic	*LINC02103;LSP1P3*
rs2754251	6:88385949:A:G	0.0584	0.0715	0.0129	1.07	2.65E-08	intronic	*AKIRIN2*
rs10763665	10:28771491:C:G	0.5783	-0.0342	0.0062	0.97	3.13E-08	ncRNA_intronic	*LINC02652*
rs7122100	11:10732560:A:C	0.2411	0.0410	0.0075	1.04	4.93E-08	intergenic	*IRAG1;CTR9*
rs1145656	11:73305859:A:C	0.8171	-0.0442	0.0079	1.05	2.00E-08	upstream	*FAM168A*
Additional novel loci identified in the African meta-analysis								
rs76668186	16:6686083:A:T	0.9597	-0.5776	0.1056	0.56	4.52E-08	intronic	*RBFOX1*
rs114102448	21:47523605:A:G	0.0114	0.9527	0.1725	2.60	4.11E-08	intronic	*COL6A2*

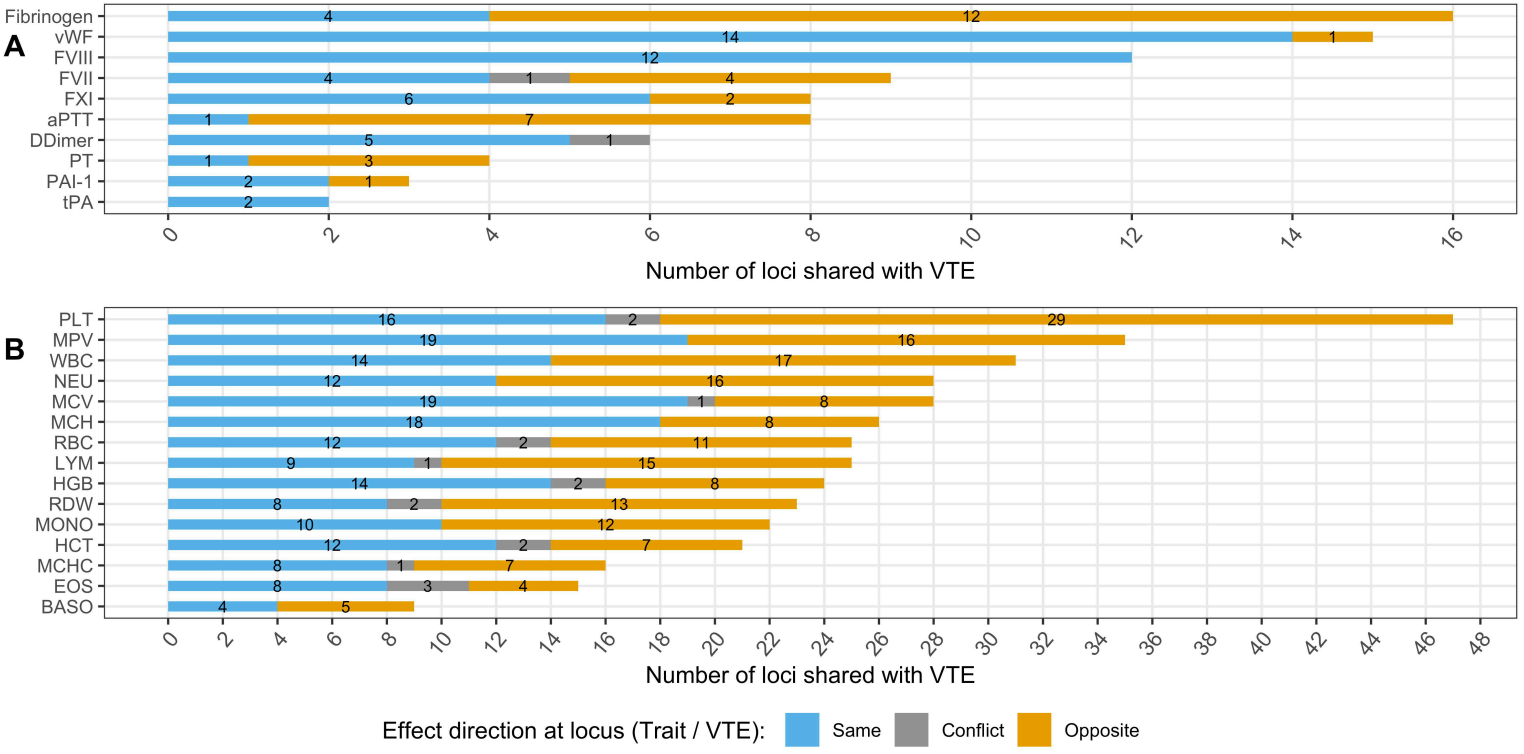
CHR: chromosome; POS: position (hg19 build); EA: effect allele; NEA: non effect allele; EAF: effect allele frequency; SE: Standard Error of Effect; OR: odds ratio; P: P-value.

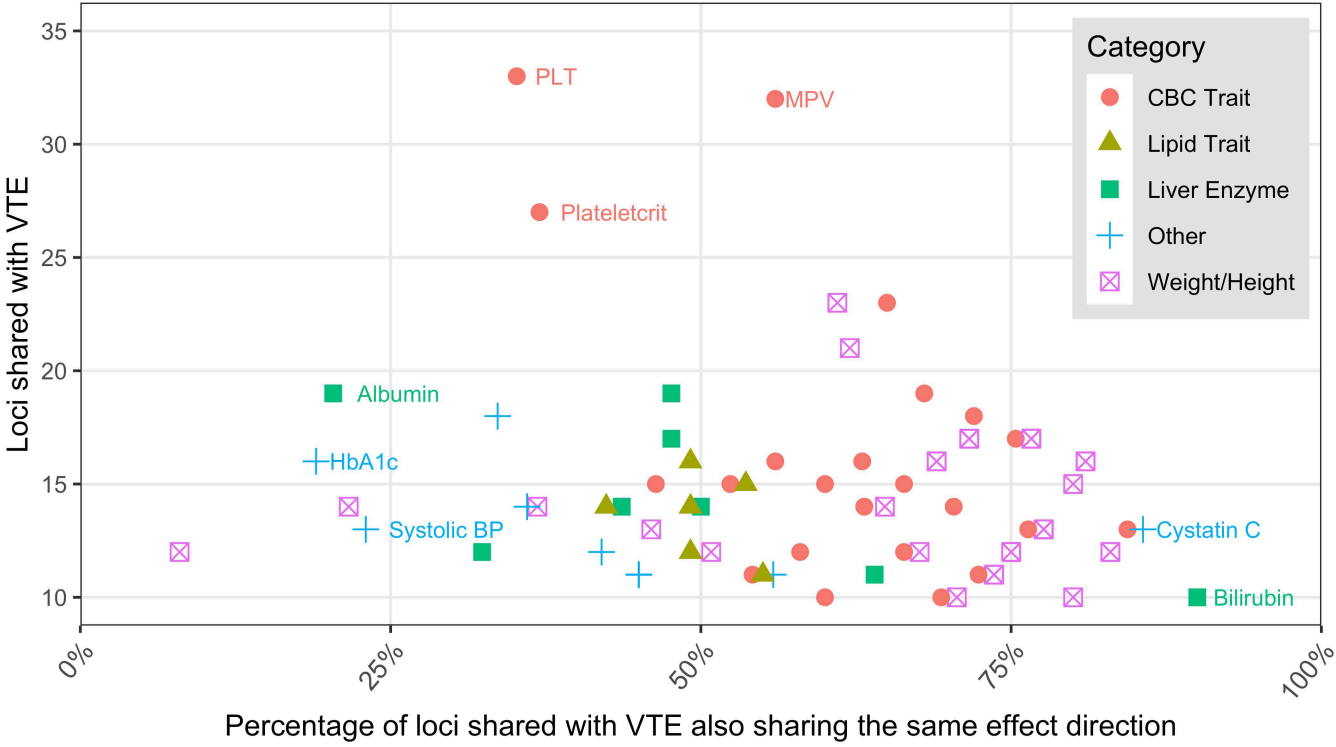

Discovery Meta-Analysis [55,330 VTE cases]
[EUR=47,822 ; AFR=6,320 ; HIS=1,188]
4 GWAS Studies: INVENT-2019, MVP, FinnGen, EGP
Lead Variants
Replication
Replication Meta-Analysis [91,230 VTE cases]
[EUR=87,594 ; AFR=1,588 ; HIS=1,075 ; SAS=273 ; EAS=700]
10 GWAS Studies [N=26,339]: UKB, BBJ, MGB, BioMe, Upenn, FARIVE, MARTHA12, RETROVE, MESA, GAIT2
2 Lookup Studies [N=64,891]: 23andMe, GBMI
Combined Meta-Analysis [81,669 VTE cases]
[EUR=71,771 ; AFR=7,482 ; HIS=1,720 ; SAS = 189; EAS = 507]
14 GWAS Studies: INVENT-2019, MVP, FinnGen, EGP, UKB, BBJ, MGB, BioMe, Upenn, FARIVE, MARTHA12, RETROVE, MESA, GAIT2
EUR Meta-Analysis [71,771 VTE cases]
13 GWAS Studies: INVENT-2019, MVP, FinnGen, EGP, UKB, MGB, BioMe, Upenn, FARIVE, MARTHA12, RETROVE, MESA, GAIT2
AFR Meta-Analysis [7,482 VTE cases]
7 GWAS Studies: INVENT-2019, MVP, UKB, MGB, BioMe, Upenn, MESA
HIS Meta-Analysis [1,720 VTE cases]
4 GWAS Studies: MVP, MGB, BioMe, MESA
Conditional analyses with GCTA-COJO to identify independent associations


SPSB1*
ARID1A
AK5
CELSR2
TENT5C*
FALEC
DCST2
FCER1G
F5*
SERPINC1
RGS18
CFHR1
C4BPA*
TRIM58*
1
LINC01248
LINC01800
PLEK*
PROC*
TFPI*
RFTN2
CPS1*
SERPINE2*
2
PPARG
CSRNP1*
NEK4
EPHA3*
PROS1
ADCY5
PIK3CB*
LINC00880
MECOM*
KNG1*
XXYLT1*
3
AFAP1
REST
SEC31A*
ARHGAP24*
SLC39A8
FGG*
F11*
4
LINC02103
OSMR-AS1*
ARHGEF28
ERAP2*
HSPA4
5
CARMIL1
HCG9
HLA-DRA*
ILRUN*
LOC101929555
AKIRIN2
STXBP5*
6
JAZF1
CDK6
7
CSMD1
XKR5*
LOC157273
PINX1
SLC25A37
SCARA5*
RBPMS
ZFPM2*
ANGPT1
8
JAK2
ZNF367*
ABO*
9
FAM107B*
MLLT10
RAB18
LINC02652
RET
TSPAN15*
ZMIZ1*
PLCE1*
GRK5*
10
IRAG1
SOX6
QSER1*
F2*
FADS2B*
MS4A4A
FADS1*
FAM168A
ZBTB16
ST3GAL4*
SNX19
11
VWF*
A2ML1*
COPZ1*
STAB2
SH2B3*
MPHOSPH9
12
MYO16
F10*
13
ARID4A
MIR4708
TC2N*
SERPINA1
RCOR1*
14
MAP1A*
RORA-AS1*
SMAD3
LINC00924
15
RBFOX1
NDE1
HSD3B7*
HP
CFDP1
PLCG2*
C16orf74
ZFPM1*
16
SMG6*
ALOX12*
TAOK1
MAPT*
APOH*
ABCA5
17
RFX2*
SLC44A2*
PEPD
RAB4B
RSPH6A*
FUT2*
GP6*
ZNF444
19
THBD*
PROCR*
20
COL6A2
21
SYN3*
A4GALT*
22
ARHGAP6
FAM47B
BCOR*
F9*
BRCC3*
X
Novel
Known
Discovery
African
Overall
Hispanic
European


A
Density
0.4
0.3
0.2
0.1
0.0
-2
0
2
4
6
B
Controls
Cases
-2
0
2
4
6
GRS (Scaled)
C
VTE (OR)
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
< 1%
1% - 5%
5% - 25%
25% - 45%
45% - 55% (REF)
55% - 75%
75% - 95%
95% - 99%
> 99%
GRS percentile category


Exposure
OR
95% CI
P
F5
5.22
[4.83, 5.63]
0
F2 (SeqId_4157_2)
3.18
[2.96, 3.43]
1×10^{-206}
F2 (SeqId_5316_54)
1.74
[1.45, 2.1]
5×10^{-9}
F11
1.5
[1.28, 1.75]
4×10^{-7}
GSS
1.12
[1.07, 1.19]
1×10^{-5}
GP6
1.07
[1.05, 1.08]
4×10^{-23}
TIMP3
1.07
[1.05, 1.08]
3×10^{-12}
LGALS3
1.06
[1.03, 1.08]
4×10^{-6}
LILRB5
1.04
[1.02, 1.06]
2×10^{-5}
LCT
1.04
[1.02, 1.05]
4×10^{-9}
ECM1
1.03
[1.02, 1.04]
4×10^{-8}
HGFAC
1.02
[1.01, 1.03]
8×10^{-6}
ERAP2
0.975
[0.966, 0.984]
6×10^{-8}
CLPS
0.963
[0.945, 0.98]
4×10^{-5}
APOL3
0.961
[0.946, 0.975]
2×10^{-7}
TIMP4
0.958
[0.94, 0.976]
6×10^{-6}
PRDX6
0.946
[0.923, 0.97]
9×10^{-6}
MSR1
0.937
[0.91, 0.966]
3×10^{-5}
EFEMP1
0.918
[0.896, 0.941]
9×10^{-12}
PLEK
0.894
[0.872, 0.916]
4×10^{-19}
CD97
0.881
[0.836, 0.927]
1×10^{-6}
TFPI
0.804
[0.743, 0.869]
5×10^{-8}
PROS1
0.776
[0.695, 0.867]
6×10^{-6}
KLKB1
0.345
[0.304, 0.391]
4×10^{-61}
1
2
3
4
5
6
Odds Ratio (95% Confidence Interval)


A
Fibrinogen
vWF
FVIII
FVII
FXI
aPTT
DDimer
PT
PAI-1
tPA
4
12
14
1
12
4
1
4
6
2
1
7
5
1
1
3
2
1
2
0
2
4
6
8
10
12
14
16
Number of loci shared with VTE
B
PLT
MPV
WBC
NEU
MCV
MCH
RBC
LYM
HGB
RDW
MONO
HCT
MCHC
EOS
BASO
16
2
29
19
16
14
17
12
16
19
1
8
18
8
12
2
11
9
1
15
14
2
8
8
2
13
10
12
12
2
7
8
1
7
8
3
4
4
5
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
Number of loci shared with VTE
Effect direction at locus (Trait / VTE):
Same
Conflict
Opposite


Loci shared with VTE
35
30
25
20
15
10
PLT
MPV
Plateletcrit
Albumin
HbA1c
Systolic BP
Cystatin C
Bilirubin
Category
CBC Trait
Lipid Trait
Liver Enzyme
Other
Weight/Height
0%
25%
50%
75%
100%
Percentage of loci shared with VTE also sharing the same effect direction

Supplemental Materials

SUPPLEMENTAL METHODS

Design and Study Participants

The current cross-ancestry GWAS meta-analysis is comprised of new analyses of data from 13 studies, including the Department of Veterans Affairs Million Veteran Program (MVP),[14] UK Biobank (UKB),[79,80] FinnGen, Estonian Biobank (EGP),[81] Biobank Japan (BBJ),[82] Mass General Brigham biobank (MGB),[83] BioMe, Penn Medicine BioBank (UPenn), FARIVE,[84] MARTHA12,[85] RETROVE,[86] Multi-Ethnic Study of Atherosclerosis (MESA),[87] and GAIT2,[88,89] as well as previously published data from the INVENT consortium, a 17 study analysis of prospective cohorts and case-control data (designated INVENT-2019).[13] A detailed description of participating studies is provided in **Table S1**.

Study Descriptions

The **23andMe, Inc. cohort** is a population-based cohort. Participants provided informed consent to participate in research, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (E&I Review). Cases were defined as those individuals that reported having had DVT through the survey questions "*What types of blood clot or stroke were you diagnosed with? Answer: A blood clot in your arms or legs (deep vein thrombosis or DVT)*" and "*Have you ever been diagnosed with deep vein thrombosis? Answer: Yes*". Controls were those individuals who reported not having had DVT. For the analysis, a set of unrelated individuals was chosen using a segmental identity-by-descent (IBD) estimation algorithm (https://www.23andme.com/ancestry-composition-guide/). Individuals were defined as related if they shared more than 700 cM IBD. The selection process was done by preferentially retaining cases over controls to maximize statistical power. A total of 59,143 DVT cases and 2,835,159 controls were included in this study. The variant-level data for the 23andMe replication dataset are fully disclosed in the manuscript. Individual-level data are not publicly available due to participant confidentiality, and in accordance with the IRB-approved protocol under which the study was conducted.
Participants provided informed consent and participated in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (E&I Review). Participants were included in the analysis on the basis of consent status as checked at the time data analyses were initiated.

The **Atherosclerosis Risk in Communities (ARIC)** study has been described in detail previously.[90] Men and women aged 45-64 years at baseline were recruited from four communities: Forsyth County, North Carolina; Jackson, Mississippi; Minneapolis, Minnesota; and Washington County, Maryland. A total of 15,792 individuals, predominantly White and African American, participated in the baseline examination in 1987-1989, with 6 reexamination visits conducted from 1990-2019.

The **BioBank Japan (BBJ)** is a hospital-based Japanese national biobank project including data from approximately 200,000 patients enrolled between 2003 and 2007.[82,82] Participants were recruited at 12 medical institutes throughout Japan (Osaka Medical Center for Cancer and Cardiovascular Diseases, Cancer Institute Hospital of Japanese Foundation for Cancer Research, Juntendo University, Tokyo Metropolitan Geriatric Hospital, Nippon Medical School, Nihon University School of Medicine, Iwate Medical University, Tokushukai Hospitals, Shiga University of Medical Science, Fukujuji Hospital, National Hospital Organization Osaka National Hospital and Iizuka Hospital).

BioME: Mount Sinai's BioMe Biobank is an electronic health record (EHR)-linked clinical care cohort ('Biobank'), consisting of ~60,000 participants from diverse ancestries characterized by a broad spectrum of (longitudinal) biomedical traits. Enrolled participants consent to be followed throughout their clinical care (past, present, and future) at Mount Sinai in real-time, integrating their genomic information with their electronic health records for discovery research. Recontacting of participants is permitted by consent, enabling a broad range of additional studies including in-depth clinical and OMICS phenotyping, mobile Health applications,

recruitment for prospective studies/trials, and return-of-results clinical care implementation projects. The BioMe Cohort, launched in September 2007, is an ongoing, consented EHR-linked bio- and data repository that enrolls participants non-selectively from the Mount Sinai Health System patient population. Starting December of 2014, we began in addition enrollment of participants at the Institute of Family Health, a network of 19 full-time Federally-Qualified Community Health Centers (FQHC) throughout New York City with an unified implementation of electronic health records (Epic systems) under a single administration. The Institute of Family Health facilities has over 20 years of successful involvement in community-based research to address health disparities. With IFH participation, BioMe further affords unique opportunities to extend comprehensive studies of genome sequence variation underlying rare and common diseases in underserved, urban communities. As of Oct 2018, 45,479 adult participants were enrolled. On average 1000 new participants are consented each month. BioMe participants represent a broad ancestral, ethnic and socioeconomic diversity with a distinct and population-specific disease burden, characteristic of Northern Manhattan communities served by Mount Sinai. BioMe participants are predominantly of African (35%) or Hispanic/Latino (36%) ancestry. Participants who self-identify as Hispanic/Latino further report to be of Puerto Rican (39%), Dominican (23%), Central/South American (17%), Mexican (5%) or other Hispanic (16%) ancestry.

The **Cardiovascular Health Study (CHS)** is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers.[91,92] The original predominantly European -ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists; subsequently, an additional predominantly African-American cohort of 687 persons were enrolled for a total sample of 5,888. Blood samples were drawn from all participants at their baseline examination and DNA was subsequently extracted from available samples. CHS was approved by institutional review committees at each field center. Participants included in the present analyses had available DNA and gave informed consent including consent to use of genetic information for the study of cardiovascular disease.

The **Early Onset Venous Thrombosis Study (EOVT):** In the EOVT study, 453 patients of European origin with early age of onset of VT (<50 years) recruited in 4 different French centers (Grenoble, Marseille, Montpellier and Paris) were compared to 1,327 healthy French subjects from the Suvimax study.[93] VT patients had a first DVT and/or PE event documented by venography, Doppler ultrasound, angiography, and/or ventilation/perfusion lung scan. They were free of any acquired risk factors at the time of VTE (including surgery, hospitalization, pregnancy, puerperium, oral contraception, cancer, and autoimmune disease); and strong known genetic risk factors, including anti-thrombin, protein C or protein S deficiency, homozygosity for FV Leiden or FII 20210A, and lupus anticoagulant. Criteria for inclusion of the healthy controls were European origin, no chronic conditions, and no regular medicines.

The **eMERGE (electronic MEdical Records GEnomics) Network** is a national initiative to combine biobanks with electronic medical records.[7] As part of eMERGE, more than 83,000 individuals across 13 sites have been genotyped. Recruitment criteria varied between sites.[94] For this study, four sites contributed data to the European ancestry analysis (Geisinger Health System, Group Health, Marshfield Clinic and Vanderbilt University). In total, 1,558 cases and 10,027 controls of European ancestry were included. For the African-American studies, two eMERGE sites contributed to the analyses (Vanderbilt and Mount Sinai). In total, 436 cases and 14,353 controls of African-American ancestry were included. VTE was ascertained through a NPL-based Electronic Health Records-driven algorithm that leveraged structured data (ICD codes) and unstructured data (clinical notes).[83]

The **Estonian Biobank** is a population-based cohort of the Estonian Genome Center at the University of Tartu (EGCUT), Estonia.[95] The current cohort size is ca 200000, from 18 years of age and up and reflects closely the age distribution in the adult Estonian population. This project is conducted according to the Estonian Gene Research Act and all participants have signed the broad informed consent. Upon recruitment, the biobank participants filled out a thorough questionnaire, covering lifestyle, diet and clinical diagnoses defined according

to the ICD10 coding. Illumina GSA arrays were used for genotyping and imputation was performed by using the Estonian-specific reference panel.[81]

The **FARIVE Study** is a multicenter case-control study of 607 patients with a first episode of proximal deep VT and/or pulmonary embolism. Patients younger than 18 years, with previous VT event, that had a diagnosis of active cancer or a history of malignancy less than 5 years previously, or have a short life expectancy because of other causes, were excluded. The control group consists of age- and sex-matched individuals free of venous and arterial thrombotic disease. Potential control subjects with cancer, liver or kidney failure, or a history of venous and/or arterial thrombotic disease are ineligible.[96]

The **FinnGen** study is a public–private partnership project. Six regional and three country-wide Finnish biobanks participate in FinnGen. Additionally, data from previously established population and disease-based cohorts are utilized. Participants' health outcomes are followed up by linking to the national health registries (since 1969), which collect information from birth to death. Summary statistics for VTE (from release 5, code I9_VTE) were publicly available and retrieved from https://www.finngen.fi/en/access_results.

The **Framingham Heart Study (FHS)** was started in 1948 with 5,209 randomly ascertained participants from Framingham, Massachusetts, US, who had undergone biannual examinations to investigate cardiovascular disease and its risk factors.[97] In 1971, the Offspring cohort (comprising 5,124 children of the original cohort and the children's spouses) and in 2002, the Third Generation (consisting of 4,095 children of the Offspring cohort) were recruited. FHS participants in this study are of European ancestry. The methods of recruitment and data collection for the Offspring and Third Generation cohorts have been described elsewhere.[98]

The **GAIT (Genetic Analysis of Idiopathic Thrombophilia**) project is a family based study where 935 subjects in 35 extended pedigrees were collected.[88,89] To be included in the study, a family was required to have at least 10 living individuals in 3 or more generations. Families were selected through a proband with idiopathic thrombophilia, which was defined as recurrent thrombotic events (at least one of which was spontaneous), a single spontaneous thrombotic episode plus a first-degree relative also affected, or onset of thrombosis before age 45. Thrombosis in these probands was considered idiopathic when biological causes as antithrombin deficiency, protein S and C deficiencies, activated protein C resistance, plasminogen deficiency, heparin cofactor II deficiency, Factor V Leiden, dysfibrogenemia, lupus anticoagulant and antiphospholipid antibodies, were excluded. Subjects were interviewed by a physician to determine their health and reproductive history, current medications, alcohol consumption, use of sex hormones (oral contraceptives or hormonal replacement therapy) and their smoking history. The study was performed according to the Declaration of Helsinki. All procedures of the study were reviewed by the Institutional Review Board of the Hospital de la Santa Creu i Sant Pau, Barcelona, Spain. Adult subjects gave informed consent for themselves and for their minor children.

The **Global Biobank Meta-analysis Initiative** (GBMI) is a collaborative network of 19 biobanks from 4 continents representing more than 2.1 million consented individuals with genetic data linked to electronic health records. Standard ICD based definitions of VTE were used for identifying VTE cases and excluding related diseases from population based controls. Contributing cohorts for a meta-analysis of VTE for a non-overlapping cohort replication of genome-wide significant findings included the China Kadoorie Biobank, East London Genes and Health, UCLA ATLAS Community Health Initiative, and the Michigan Genomics Initiative.

The **Heart and Vascular Health (HVH) VTE Study** is a case-control study of risk factors for cardiovascular outcomes set at Group Health (GH), an integrated health care delivery system in western Washington State. Cases include venous thromboembolism (VTE), myocardial infarction (MI), stroke, and atrial fibrillation, with a shared common control group frequency matched to MI cases on age (within decade) sex, treated hypertension, and calendar year of identification. Study approval was granted by the human subjects committee at GH, and written informed consent was provided by all study participants. Methods for the study

have been described previously.[4,99] VTE cases and controls with no prior VTE were utilized for these analyses. All study participants were GH members, with women aged 18-89 years old and men aged 30-89 years old. Deep venous thrombosis (DVT) and pulmonary embolism (PE) events were identified using hospital discharge diagnosis codes and urgent care diagnosis codes. Additionally, subjects who received a prescription for low molecular weight heparin were screened as potential cases of DVT. Eligibility and risk factor information were collected by trained medical record abstractors from a review of the GH medical record using only data available prior to the index date. A venous blood sample was collected from all consenting subjects, and DNA was extracted from white blood cells using standard procedures.

The **HUNT (Nord-Trøndelag Health) Study** is a large population-based health study started in 1984 for the inhabitants of Nord-Trøndelag county in central Norway. A comprehensive description of the study population has been previously reported.[100] In brief, approximately every 10 years the entire adult population of Nord-Trøndelag (~90,000 adults in 1995) is invited to attend a health survey which includes comprehensive questionnaires, an interview, clinical examination, and detailed phenotypic measurements (HUNT1 [1984 to 1986]; HUNT2 [1995 to 1997] and HUNT3 [2006 to 2008]). Approximately 90% of participants from HUNT2 and HUNT3 were genotyped in 2015 (n = 69,422). All VTE cases are validated based on thorough review of the electronic health record.

The **JUPITER study** (Justification for the Use of statins in Prevention: an Intervention Trial Evaluating Rosuvastatin) was an international, randomized, placebo-controlled trial of rosuvastatin (20mg/day) in the primary prevention of cardiovascular disease conducted among 17,802 apparently healthy men and women with LDL-C < 130 mg/dL and hsCRP > 2 mg/L conducted between 2003 and 2008, with median and maximum follow-up times of 1.9 and 5.0 years, respectively.[101–103] Approximately 71.2% of JUPITER participants had European ancestry among whom 71.4% provided DNA and consent for genetic analysis. Ascertainment of incident venous thromboembolism (VTE) was pre-specified in the trial protocol and included all cases of diagnosed pulmonary embolism or deep-vein thrombosis as well as corroborating evidence from confirmatory diagnostic tests, the initiation of anticoagulation therapy, or death that was considered likely to have been due to a pulmonary embolism. The genetic sub-sample for the current analysis included 8,749 unrelated individuals with self-reported European ancestry confirmed by genetic analysis.

The **MARseille THrombosis Association (MARTHA)** project has already been described extensively.[9,84,104] It is composed of unrelated subjects of European origin, with the majority being of French ancestry, consecutively recruited at the Thrombophilia center of La Timone hospital (Marseille, France) between January 1994 and October 2012. All patients had a documented history of VT and were free of well characterized genetic risk factors including AT, PC, or PS deficiency, homozygosity for FV Leiden or FII 20210A, and lupus anticoagulant. They were interviewed by a physician on their medical history, which emphasized manifestations of deep vein thrombosis and pulmonary embolism using a standardized questionnaire. The thrombotic events were confirmed by venography, Doppler ultrasound, spiral computed tomographic scanning angiography, and/or ventilation/perfusion lung scan. Controls were healthy individuals randomly selected from the 3C study, a population-based study carried out in 3 French cities composed of 8707 non institutionalized individuals aged over 65 randomly selected from the electoral rolls and free of any chronic diseases and for which biological (DNA, plasma) samples could have been obtained.[105]

The **MARseille THrombosis Association study of 2010-2012 (MARTHA12)** is composed of an independent sample of 1,245 VT patients. Patients have been recruited between 2010 and 2012 according to the same criteria as the MARTHA patients.[9]

The **Mass General Brigham Biobank (MGBB)** is a hospital-based research cohort containing genotypic and clinical data from >105,000 individuals enrolled across 7 regional hospitals with median 3 years of follow-up. Genotyping was performed for ~36,000 MGBB participants. Venous thromboembolism case status was ascertained from EHR query of relevant ICD-9 codes (451.11, 451.19, 453.2, 453.4, 415.1) and ICD-10 codes

(I80.1, I80.2, I82.22, I82.4, I82.5, I26.0, I26.9) when a minimum of two hospital (inpatient or outpatient) encounters had occurred. Control status was defined among the genotyped population as lacking the above ICD-9/10 codes.

The **Mayo VTE Study** recruited consecutive Mayo Clinic outpatients who resided in the upper midwest United States and who were referred to the Mayo Clinic Special Coagulation Laboratory or Thrombophilia Center.[6] We prospectively selected clinic-based controls from persons undergoing outpatient general medical examinations in 2004 - 2009 within the Mayo Clinic Divisions of General Internal Medicine and Primary Care Internal Medicine, Department of Internal Medicine, and general internal medicine practices that care for patients (> 10 000 per year) from the upper Midwest United States. Additional controls were recruited from the Department of Family Medicine and the Mayo Clinic Sports Medicine Center. Controls were frequency matched on the age group, sex, state of residence and myocardial infarction(MI)/stroke status distribution of the cases, and had no previous diagnosis of VT or superficial vein thrombosis.

The **Multiple Environmental and Genetic Assessment of risk factors for venous thrombosis (MEGA)** study is a large population-based case-control study.[106] Data collection and ascertainment of venous thrombotic events have been previously described in detail. In short, patients with a first deep vein thrombosis or pulmonary embolism were recruited at six anticoagulation clinics in the Netherlands between 1999 and 2004. The diagnosis of a deep vein thrombosis was based on compression ultrasonography, whereas a pulmonary embolism was confirmed by perfusion and ventilation scintigraphy, helical computed tomography or pulmonary angiography. Blood samples were taken at least 3 months after discontinuation of vitamin K antagonist treatment, unless patients were still receiving anticoagulant therapy one year after their VT event.

The **Multi-Ethnic Study of Atherosclerosis (MESA)** is a study of the characteristics of subclinical cardiovascular disease (disease detected non-invasively before it has produced clinical signs and symptoms) and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease. MESA researchers study a diverse, population-based sample of 6,814 asymptomatic men and women aged 45-84 from six field centers across the United States. Approximately 38 percent of the recruited participants are white, 28 percent African-American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. Five follow-up exams have been completed since 2000.[87] Blood was collected by venipuncture at baseline and at each follow-up visit. Participants provided informed consent for the use of DNA from the blood sample. Participants in the MESA cohort who consented to genetic analyses and data sharing (dbGaP) were genotyped using the Affymetrix Human SNP Array 6.0 (GWAS array) as part of the NHLBI SHARe (SNP Health Association Resource) project.

The **Million Veteran Program (MVP)** is a Department of Veterans Affair cohort study. We conducted a discovery genetic association analysis using DNA samples and phenotypic data from the Million Veteran Program (MVP). In MVP, individuals aged 18 to over 100 years have been recruited from 63 Veterans Affairs (VA) Medical Centers across the United States.[14,107]

The **Nurse's Health Studies (NHS and NHS-II)** These studies have been described previously and additional information is available at http://www.nurseshealthstudy.org.[108,109] NHS and NHS-II are longitudinal cohort studies of female nurses. In 1976, baseline questionnaires were sent to registered nurses from 11 populous US states, establishing a cohort of 121,700 women aged 30-55. There were no exclusions by race, but the majority (96%) were of European ancestry; corresponding to the demographics of nurses in 1976. NHS participants are mailed a questionnaire every two years that assesses risk factor status and interval disease events. Physician-diagnosed PE has been asked on every biennial NHS questionnaire since 1982. In NHS, the question reads: "Since [year], have you had any of these physician-diagnosed illnesses? ... Pulmonary Embolus." NHS questionnaires also ask whether the nurse had "Any other major diagnosis: ____." In NHS, DVT is captured when a nurse answers that she has had phlebitis or thrombophlebitis (ICD-9=453.x).

Questionnaire-reported VTE diagnoses have proven to be highly accurate, with >95% validation of VTE events. A physician reviews medical records for all reported PEs, validating diagnoses when medical records include: a positive pulmonary angiogram, a high-probability ventilation/perfusion scan, or a positive CT pulmonary angiogram.

The **Penn Medicine BioBank (PMBB)** recruits patients from throughout the University of Pennsylvania Health System for genomic and precision medicine research. Participants actively consent to allow the linkage of biospecimens to their longitudinal EHR. Currently, >60 000 participants are enrolled in the PMBB. A subset of ~45000 individuals who have undergone whole exome sequencing and genotyping, performed through a collaboration with the Regeneron Genetics Center. A further subset of ~12000 subjects with imputed genotype data was used in this analysis.

The **Riesgo de Enfermedad TROmboembólica VEnosa study (RETROVE)** is a prospective case–control study that includes 400 consecutive patients with VTE (cancer associated thrombosis was excluded) and 400 healthy control volunteers. All individuals were ≥ 18 years. The diagnosis was confirmed with Doppler ultrasonography, tomography, magnetic resonance, arteriography, phlebography or pulmonary gammagraphy. Blood samples from the patients were taken at least 6 months after thrombosis 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. Controls were selected according to the age and sex distribution of the Spanish population (2001 census). A total of 5 ml of blood was obtained in a Vacutainer tube (BD Vacutainer Becton Dickinson and Company, New Jersey, USA) containing EDTA as anticoagulant. All individuals were genotyped using Infinium Global Screening Array-24 v3.0 kit from Illumina and imputed using the Haplotype Reference Consortium panel. Written informed consent was obtained for all participants and all procedures were approved by the Institutional Review Board of the Hospital de la Santa Creu i Sant Pau (Barcelona).

The **Tromsø Study** is a single-center, population-based cohort study of the inhabitants of Tromsø, Norway. 27,158 individuals participated in the fourth survey of the Tromsø Study between 1994-1995; baseline characteristics were collected using self-reported questionnaires, physical examinations, and blood samples.[110] Non-fasting blood was drawn from an antecubital vein to gather plasma and whole blood. Whole blood was used to prepare archive quality DNA, and was stored at the HUNT Biobank in Levanger, Norway. All 27,158 participants were followed from the date of enrollment through December 31, 2011. All cohort members that experienced an incident venous thromboembolism (VTE) during the study period were identified by searching the hospital discharge diagnosis registry, the autopsy registry, and the radiology procedure registry at the University Hospital of North Norway, the sole hospital in the Tromsø municipality.[111] The VTE events were thoroughly validated by review of medical records as previously described in detail. For each case, a paired control, matched on age and sex, was randomly sampled from the cohort.

The **UK Biobank** (https://www.ukbiobank.ac.uk/) is a large population-based prospective cohort study of ~500,00 participants residing in the United Kingdom (UK).[80] All participants, aged 40-71 at recruitment in 2006-2010, attended a baseline exam at a local study assessment center and gave informed consent. DNA was extracted from the blood samples drawn at the baseline exam and genotyped at the Affymetrix Research Services Laboratory in Santa Clara, California, USA.

The **Women's Genome Health Study (WGHS):** WGHS is a prospective cohort of female North American health care professionals representing participants in the Women's Health Study (WHS) trial who provided a blood sample at baseline and consent for blood-based analyses. Participants in the WHS were 45 years or older at enrollment and free of cardiovascular disease, cancer or other major chronic illness. The current data are derived from 23,294 WGHS participants for whom whole genome genotype information was available at the time of analysis and for whom self-reported European ancestry could be confirmed by multidimensional scaling analysis of 1,443 ancestry informative markers in PLINK v. 1.06. At baseline, BP and lifestyle habits related to smoking, consumption of alcohol, and physical activity as well as other general clinical information

were ascertained by a self-reported questionnaire, an approach which has been validated in the WGHS demographic, namely female health care professionals.

In the WHS (the WGHS parent cohort), venous thromboembolism events are collected by self-report questionnaire and then confirmed by review of medical records. A diagnosis of deep-vein thrombosis is confirmed by a positive venous ultrasonography or venography report, whereas the diagnosis of pulmonary embolism is confirmed by a positive angiogram or computed tomography scan of the chest, or a ventilation-perfusion scan with 2 or more mismatched defects. Deaths due to pulmonary embolism are confirmed when autopsy reports, symptoms, circumstances, and medical history are consistent with this diagnosis.

The Women's Genome Health Study is supported by the National Heart, Lung, and Blood Institute (HL043851 and HL080467) and the National Cancer Institute (CA047988 and UM1CA182913), with funding for genotyping provided by Amgen.

The **Women's Health Initiative (WHI)** is a long-term national health study that focuses on strategies for preventing common diseases such as heart disease, cancer and fracture in postmenopausal women. A total of 161,838 women aged 50–79 years old were recruited from 40 clinical centers in the US between 1993 and 1998. WHI consists of an observational study, two clinical trials of postmenopausal hormone therapy (HT, estrogen alone or estrogen plus progestin), a calcium and vitamin D supplement trial, and a dietary modification trial.[112,113]

Ethical Oversight

Study	IRB / Ethics Committee Providing Approval
FARIVE	Paris Broussais–Hopital Europeen Georges Pompidou ethics committee in Paris (ethical permit: 2002-034)
GAIT2	Institutional Review Board of the Hospital de la Santa Creu i Sant Pau, Barcelona, Spain.
MARTHA12	Health Department of the General Directorate for the French Ministry of Research and Innovation (Projects DC: 2008-880 and 09.576)
Mass General Brigham	Mass General Brigham Ethics Committee
MESA	The University of Washington IRB Committee D
Million Veteran Program	VA Central Institutional Review Board
Penn Medicine	IRB protocol# 813913
RETROVE	Institutional Review Board of the Hospital de la Santa Creu i Sant Pau, Barcelona, Spain.
23andMe	external AAHRPP accredited IRB, Ethical & Independent Review Services
INVENT-2019 (includes ARIC, CHS, EOVT, eMERGE, FHS, HVH, HUNT, JUPITER, MARTHA, MAYO, MEGA, NHS/NHSII/HPFS, Tromso, WGHS, WHI)	*N/A (Meta-analysis published [PMID:31420334] and available through dbGaP)*
Biobank Japan	*N/A (biobank)*
BioME Biobank	*N/A (biobank)*
Estonian Biobank	*N/A (biobank)*
FinnGen	*N/A (biobank)*
GBMI	*N/A (biobank meta-analysis currently detailed in the following medrxiv preprint: doi.org/10.1101/2021.11.19.21266436)*
UK Biobank	*N/A (biobank)*

Study-specific GWAS

Genotyping arrays, imputation panels, and analyses performed by each participating study are detailed in **Table S1**. Briefly, studies performed association analyses (logistic regression analyses or generalized mixed models for case-control studies and Cox regression for cohort studies) using age and sex as covariates and further adjusting for participant relatedness, genetic principal components, and study site or other study-specific factors when applicable.
As a sensitivity analysis, we also produced a version of the combined meta-analysis where datasets that relied on a Cox regression were excluded. This analysis produced signals with nearly identical effects (**Figure S11**).

For each dataset, quality control was performed using EasyQC[114] to remove variants with missing information (effect and/or standard error), low imputation quality (< 0.3), and rare variants (minor allele count < 5). For studies missing either imputation quality or variant frequency (BioMe, UPenn), a filter was added to remove variants with extreme effects (|beta| > 10). Indels and marker names were then harmonized across all studies.

For the X chromosome, all studies performed sex-stratified GWAS, excluding variants from pseudo-autosomal regions. Results from males and females were then meta-analyzed. As chromosome X is only present in one copy in men, our analysis plan indicated to use 0/2 allele coding for male samples in X-chromosome analysis, instead of the traditional 0/1/2 coding used for autosomes and women X-chromosome analyses. A P-value based meta-analysis of chromosome X was performed in addition to the inverse-variance weighted meta-analysis, as a sensitivity analysis, revealing the same genome-wide significant signals (**Figure S12**).

Discovery, Replication, and Combined GWAS Meta-analyses

All GWAS meta-analyses were conducted with METAL,[16] using a fixed-effects inverse-variance weighted model. All variants were included and there was no lower minor allele frequency (MAF) limit beyond study-specific minor allele count. Genome-wide significant variants ($P < 5.00 \times 10^{-8}$) were kept if a concordant effect direction was observed in 2 or more studies and grouped into the same locus if they were within 1Mb. We used the closest gene to the lead variant to refer to each locus, except at known loci where the causal gene has been previously identified and is different from the closest gene (such as *PROCR* or *PROS1*). We defined a locus as novel if a genetic association with VTE has not been previously observed in the region according to our review of peer-reviewed published reports. We used PhenoGram[115] to visually represent the genomic position of the loci, their significance in EUR- or AFR-ancestry analyses, and their discovery status of novel or known.

Discovery Meta-Analysis: For the discovery cross-ancestry GWAS meta-analysis, we meta-analyzed data from 4 consortium/studies: INVENT-2019, MVP, FinnGen and EGCUT. Participants were adult men and women and included 55,330 VTE cases (either DVT and/or PE cases) and 1,081,973 controls of EUR, AFR, or HIS ancestries. At each locus with a genome-wide significant signal, the lead variant was extracted and tested in an independent replication meta-analysis.

Replication: The replication GWAS meta-analysis consisted of the remaining 10 participating studies, as well as 2 external collaborators (GBMI [15] and 23andMe[17]), for a total of 91,230 VTE cases in replication. Replicating variants from the discovery were defined as those that had concordant effect direction in the discovery and the replication, and reached a Bonferroni-corrected p-value threshold in the replication population corresponding to the number of variants tested for replication with a 1-sided hypothesis: p-value threshold = [(0.05*2)/number of variants tested for replication] in the replication analysis.

Combined GWAS Meta-Analysis and Stratification by Ancestry: We performed a combined, cross-ancestry GWAS meta-analysis of discovery and replication data using participating studies with genome-wide summary data. We included variants with MAF ≥ 0.01 to maintain adequate statistical power by reducing the

number of low-powered tests since replication was not available. Genome-wide data from GBMI and 23andMe data were not available and therefore excluded from combined analyses. We estimated the heterogeneity associated with each variant using Cochran's Q test and the corresponding I^2 statistic. We assessed the genomic inflation with the lambda genomic control.[18] We report on variants exceeding the genome-wide threshold ($P < 5.00 \times 10^{-8}$) and view these as candidate novel loci associated with VTE and needing future replication.

We then stratified the analyses by ancestry and limited strata to EUR, AFR, and HIS ancestries as the remaining ancestries had too few VTE events to be informative: East Asian (EAS) in BBJ, n=507 VTE events; South Asian (SAS) in UKB, n=189 VTE events. As above, we estimated heterogeneity and assessed inflation with lambda genomic control; the LD-score intercept was computed for EUR-ancestry analysis, using the recommended Hapmap3 variants.[19] We report all additional ancestry-specific variants exceeding the genome-wide threshold ($P < 5.00 \times 10^{-8}$) and view these as ancestry-specific candidate loci associated with VTE and needing future replication.

Ancestry-Stratified Analyses

To estimate the presence of multiple independent signals, we performed conditional analyses with GCTA-COJO[20] at each locus with significant signals in the EUR-, AFR- and HIS-ancestry GWAS meta-analyses. The Trans-Omics for Precision Medicine (TOPMed) trans-ancestry sequence data (freeze 8) was used as reference panel, selecting only EUR-ancestry participants from TOPMed (N=34,890) for the EUR conditional analyses, AFR-ancestry participants (N=17,322) for the AFR analyses and HIS-ancestry participants (N=1014) for the HIS analyses.[21] Conditional analyses were performed at each locus, using a window that encompassed at least the genome-wide significant variants present in the locus with an additional buffer of ±100 Kb. A stepwise joint regression model was used to identify secondary signals with joint p-values $< 5.00 \times 10^{-8}$ and a linkage disequilibrium (LD) $r^2 < 0.2$ with selected variants.

In addition, for each locus and for each ancestry-specific GWAS meta-analysis, we produced forest plots with the *forestplot* R library, and regional association plots with the *RACER*[22] and *LDlinkR*[23] R libraries, to visually inspect the local genetic architecture (available as **Figures S1-S8**). We used the 1000 Genomes project EUR-ancestry dataset as reference panel to infer LD patterns for the EUR-ancestry participants and overall meta-analyses and the 1000 Genomes project AFR-ancestry reference panel for the AFR-ancestry meta-analysis. Furthermore, at each locus where distinct lead variants were identified in the different meta-analyses, we also extracted the lead variant from each analysis, as well as additional independent variants identified by the conditional analyses, and computed the LD between each variant (using both EUR- and AFR-ancestry reference panels) to verify the independence of the signals. Manhattan plots were generated for each meta-analysis (**Figures S9**).

Genetic Risk Score (GRS)

We constructed an ancestry specific GRS derived from the genome-wide significant lead variants observed in the EUR specific meta-analysis and evaluated it for UKB participants of EA ancestry. GRS for AFR and HIS ancestries were not constructed due to a lack of availability of a large-scale dataset with accessible genotype data for other ancestries. This score can be calculated for each individual as a summation of log(OR)-weighted genotypes. To avoid overfitting bias, we performed an alternative EUR specific meta-analysis where UKB participants were excluded, and we retrieved the log(OR) of this analysis to establish the GRS. Once the score was obtained for UKB participants, we applied z-transformed to obtain a mean of 0 and a unit SD.
The variance explained by the GRS was estimated by calculating the R2, using the method provided by Lee et al,[116] for ascertained case-control studies, assuming a disease prevalence in the population of 0.001. Next, we performed logistic regression to measure the association of the GRS with VTE status, while correcting for age,

sex, and the 10 first genetic PCs. The predictive ability of the score was estimated by calculating the AUC, using the pROC R library,[24] as well as the delta-AUC improvement over the base model with age sex and 10 PCs.

Transcriptome-Wide Association Studies (TWAS)

We performed TWAS with the FUSION pipeline to accomplish 2 tasks: (1) prioritize genes for those genome-significant signals with ambiguous gene associations; and (2) identify new candidate loci by linking gene expression with VTE risk using GWAS results not reaching genome-wide significance. This analysis was performed using the EUR-ancestry autosomal GWAS meta-analysis results, since FUSION depends on a EUR-ancestry LD reference panel (from 1000 genomes[117]) and does not include data for chromosome X. As several genes can be associated at the same locus, the TWAS results were subjected to a conditional analysis implemented in FUSION to select genes that remained conditionally independent. For each tissue, we further performed a colocalization test with coloc[39] for all significant associations, to identify and select genetic signals shared by both VTE risk and gene expression with high posterior probability (PP4>0.75). Selected genes located farther than 200kb from genetic loci identified in the meta-analyses were considered novel candidate VTE genes.

Protein QTL Mendelian Randomization

The protein QTL (pQTL) MR analyses relied on a plasma proteome analysis performed in 7,213 individuals of European ancestry from the ARIC study,[118] which identified cis-acting pQTL (Quantitative Trait Loci) located at most 500 kb from a protein coding gene. Proteins levels were measured with SOMAmers (for “slow off rate modified aptamer”), short single stranded molecules able to bind specific proteins with high affinity. Using this pQTL dataset, Zhao et al[29] selected 6,144 conditionally independent pQTL for 1,310 proteins to be used as genetic instruments for MR analyses. A 3-step instrument validation process was performed to identify pQTL that fit MR assumptions: (i) to avoid collinearity, LD clumping was used to remove pQTLs displaying high LD (r^2 > 0.6); (ii) the instrument strength was estimated using F-statistics, and instruments with a statistics lower than 10 were removed to avoid weak instrument bias; (iii) the MR Steiger approach was applied to remove pQTLs with potential reverse causality. As a result 5,418 pQTLs for 1,310 proteins were kept (available as supplemental table ST1 here: https://www.medrxiv.org/content/medrxiv/early/2022/01/11/2022.01.09.21268473/DC1/embed/media-1.xlsx?download=true). Further, pQTL were classified into 3 tiers, with tier 1 pQTLs (most reliable) corresponding to instruments showing similar effects in 2 additional independent pQTL studies;[119,120] tier 2 instruments showed potential heterogeneous effect compared to the 2 aforementioned independent pQTL studies; tier 3 instruments showed potential pleiotropic effects (associated with more than 5 proteins)

For our VTE analyses, we relied on tier 1 instruments only. When a single genetic instrument was available for a protein, we performed Wald-ratio MR. When 2 or more genetic instruments were available, we performed inverse variance weighted MR instead, and assessed pleiotropy with MR Egger. The MR analyses were performed with the *TwoSampleMR* R library.

SUPPLEMENTAL DISCUSSION

Biological insights: Other Replicated and Non-Replicated Loci

Replicated variants included 2 rare variants at the known *EPHA3* (intergenic, MAF=0.0024, OR=2.40) and *FADS2B* (intronic, MAF=0.0047, OR=0.64) loci. Among variants that failed replication, only 1 rare variant displayed significant heterogeneity (P=0.0001, *MYO16* locus). 3 variants were located in known loci: *STAB2* was previously identified as associated with VTE in an independent gene-based study using exome sequencing,[121] *ARL13B* (near *PROS1*) was identified in the previous VTE GWAS from MVP,[14] and the *JAK2* V617F variant, which is known to increase the risk of myeloproliferative neoplasm, was recently identified as associated with VTE in an exome study of nearly 450,000 UKB participants.[122] According to gnomAD,[43] the *ARL13B* variant identified is mostly observed in AFR-ancestry individuals (rs79324379, AFR MAF=0.026 against MAF<0.0003 in other ancestries) and was not in LD with the lead variant identified in the previous MVP GWAS (rs6795524, LD r^2=0.01 in AFR); nonetheless, we would need additional information to validate this locus as a truly independent signal—and not just a marker—from any strong, uncharacterized signal in *PROS1*. Similarly, the *STAB2* variant identified is mostly observed in Finns (rs142351376, Finns MAF=0.020 against MAF<0.0003 in other ancestries); the lack of Finns in the replication likely impaired our ability to replicate the association. Out of the other 13 failed replications, 6 involved rare variants (MAF<0.01) that did not reach nominal significance (P<0.05), while 6 of the remaining 7 common variants reached nominal significance, suggesting that these common variants might need a larger replication sample to be validated. One of these signals, located between *SYN2* and *PPARG*, was associated with the protein levels of *TIMP4* in a previous study.[123] This protein, known to inhibit matrix metalloproteinases and involved in platelet aggregation and recruitment,[65] was confirmed by the pQTL MR analysis as a gene associated with VTE risk.

Clinical implications: VTE loci with known roles in hematopoieisis or platelet phenotypes

The hematology traits investigations and the pheWAS established that CBC measures share a large number of loci with VTE, and platelet phenotypes in particular are the most frequent traits shared with VTE variants: 51 loci were associated with either platelet count, mean platelet volume, plateletcrit or platelet distribution width in the pheWAS, and 35 of these loci are novel, which represents more than a third of all novel genetic associations. Several loci associated with VTE harbor genes with known roles in hematopoiesis and megakaryocyte development, or platelet turnover: *ARID1A*, *REST* and its co-repressor *RCOR1*, *CDK6*, *MECOM*, *RBPMS*, *ANGPT1*, *RET*, *NFE2*, *ST3GAL4, SH2B3*, *ZFPM2* and *ZFPM1*,[45,46,53–60] or platelet aggregation: *SLC44A2, VWF, FGG, GP6, RGS18, GRK5, PIK3CB, PLCE1, PLCG2, IRAG1, TIMP4, FCER1G,* and *ALOX12*.[10,61–71]

FUNDING ACKNOWLEDGMENT

23andMe: The variant-level data for the replication dataset are fully disclosed in the manuscript. Individual-level data are not publicly available due participant confidentiality, and in accordance with the IRB-approved protocol under which the study was conducted. We would like to thank the research participants and employees of 23andMe for making this work possible. The following members of the 23andMe Research Team contributed to this study: Stella Aslibekyan, Adam Auton, Elizabeth Babalola, Robert K. Bell, Jessica Bielenberg, Katarzyna Bryc, Emily Bullis, Daniella Coker, Devika Dhamija, Sayantan Das, Sarah L. Elson, Teresa Filshtein, Kipper Fletez-Brant, Pierre Fontanillas, Will Freyman, Pooja M. Gandhi, Karl Heilbron, Barry Hicks, David A. Hinds, Ethan M. Jewett, Yunxuan Jiang, Katelyn Kukar, Keng-Han Lin, Maya Lowe, Jey C. McCreight, Matthew H. McIntyre, Steven J. Micheletti, Meghan E. Moreno, Joanna L. Mountain, Priyanka Nandakumar, Elizabeth S. Noblin, Jared O'Connell, Aaron A. Petrakovitz, G. David Poznik, Morgan Schumacher, Anjali J. Shastri, Janie F. Shelton, Jingchunzi Shi, Suyash Shringarpure, Vinh Tran, Joyce Y. Tung, Wei Wang, Catherine H. Weldon, Peter Wilton, Alejandro Hernandez, Corinna Wong, Christophe Toukam Tchakouté, Gabriel Cuellar Partida, Xin Wang and 23andMe Research Team are employed by and hold stock or stock options in 23andMe, Inc.

Atherosclerosis Risk in Communities (ARIC): The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I, HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. Funding was also supported by R01HL087641, R01HL059367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. Jack W. Pattee was supported by NIH T32GM108557.

BioBank Japan (BBJ): The BioBank was supported by the Tailor-Made Medical Treatment Program of the Ministry of Education, Culture, Sports, Science, and Technology and Japan Agency for Medical Research (AMED) under grant numbers JP17km0305002, JP17km0305001, JP20km0405209 and JP20ek0109487. S.K, K.I and I.K were funded by AMED under Grant Numbers JP20km0405209 and JP20ek0109487.

BioME: The BioMe Biobank is supported by the Andrea and Charles Bronfman Philanthropies.

Cardiovascular Health Study: This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, HHSN268200960009C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006; and NHLBI grants U01HL080295, R01HL085251, R01HL087652, R01HL105756, R01HL103612, R01HL120393, and U01HL130114 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Early Onset Venous Thrombosis Study (EOVT): No grants to acknowledge.

eMERGE Consortium: The eMERGE Network was initiated and funded by NHGRI through the following grants: U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University

Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center). *Group Health Cooperative/University of Washington:* Funding support for Alzheimer's Disease Patient Registry (ADPR) and Adult Changes in Thought (ACT) study was provided by a U01 from the National Institute on Aging (Eric B. Larson, PI, U01AG006781). A gift from the 3M Corporation was used to expand the ACT cohort. DNA aliquots sufficient for GWAS from ADPR Probable AD cases, who had been enrolled in Genetic Differences in Alzheimer's Cases and Controls (Walter Kukull, PI, R01 AG007584) and obtained under that grant, were made available to eMERGE without charge. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Genome-wide association analyses were supported through a Cooperative Agreement from the National Human Genome Research Institute, U01HG004610 (Eric B. Larson, PI). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). *Marshfield Clinic Research Foundation:* Funding support for the Personalized Medicine Research Project (PMRP) was provided through a cooperative agreement (U01HG004608) with the National Human Genome Research Institute (NHGRI), with additional funding from the National Institute for General Medical Sciences (NIGMS) The samples used for PMRP analyses were obtained with funding from Marshfield Clinic, Health Resources Service Administration Office of Rural Health Policy grant number D1A RH00025, and Wisconsin Department of Commerce Technology Development Fund contract number TDF FYO10718. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). *Vanderbilt University:* Funding support for the Vanderbilt Genome-Electronic Records (VGER) project was provided through a cooperative agreement (U01HG004603) with the National Human Genome Research Institute (NHGRI) with additional funding from the National Institute of General Medical Sciences (NIGMS). The dataset and samples used for the VGER analyses were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and by the Vanderbilt CTSA grant UL1RR024975 from NCRR/NIH. Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). *Geisinger Health System:* Samples and data in this obesity study were provided by the non-alcoholic steatohepatitis (NASH) project. Funding for the NASH project was provided by a grant from the Clinic Research Fund of Geisinger Clinic. Funding support for the genotyping of the NASH cohort was provided by a Geisinger Clinic operating funds and an award from the Clinic Research Fund. Samples and data in this study were provided by the abdominal aortic aneurysm (AAA) project. Funding for the AAA project was provided by a grant from the Clinic Research Fund of Geisinger Clinic. Funding support for the genotyping of the AAA cohort was provided by a Geisinger Clinic operating funds and an award from the Clinic Research Fund. Samples and data in this study were provided by the Geisinger MyCode Project. Funding for the MyCode project was provided by a grant from Commonwealth of Pennsylvania and the Clinic Research Fund of Geisinger Clinic. Funding support for the genotyping of the MyCode cohort was provided by Geisinger Clinic operating funds and an award from the Clinic Research Fund. *Mount Sinai School of Medicine:* Samples and data used in this study were provided by the Mount Sinai School of Medicine (MSSM) Biobank Project funded by The Charles R. Bronfman Institute for Personalized Medicine (IPM) at Mount Sinai School of Medicine. The Coronary Artery Disease study (IPM BioBank GWAS) is a genome-wide association study funded by the Charles R. Bronfman Institute for Personalized Medicine. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000888.v1.p1.

Estonian Biobank: This study was funded by EU H2020 grant 692145, Estonian Research Council Grant IUT20-60, IUT24-6, and European Union through the European Regional Development Fund Project No. 2014-2020.4.01.15-0012 GENTRANSMED and 2014-2020.4.01.16-0125. Data analyses were carried out in part in the High-Performance Computing Center of University of Tartu.

FARIVE: The FARIVE study was supported by Fondation pour la Recherche Médicale, Programme Hospitalier de Recherche Clinique (PHRC), Fondation de France, and the Leducq Foundation (LINAT Leducq International Network Against Thrombosis). FARIVE genetics research program was supported and funded by the GenMed LABEX (ANR-10-LBX-0013) and the French INvestigation Network on Venous Thrombo-Embolism (INNOVTE).

FinnGen: The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis AG, and Boehringer Ingelheim. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank, THL Biobank, Helsinki Biobank, Biobank Borealis of Northern Finland, Finnish Clinical Biobank Tampere, Biobank of Eastern Finland, Central Finland Biobank, Finnish Red Cross Blood Service Biobank and Terveystalo Biobank. All Finnish Biobanks are members of BBMRI.fi infrastructure. Finnish Biobank Cooperative -FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the Fingenious® services (https://site.fingenious.fi/en/) managed by FINBB.

Framingham Heart Study (FHS): Framingham Heart Study (FHS) was partially supported by the National Heart, Lung, and Blood Institute's (NHLBI's) Framingham Heart Study (Contract No. N01-HC-25195) and its contract with Affymetrix, Inc. for genotyping services (Contract No. N02-HL-6-4278). A portion of this research utilized the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center. This project has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N92019D00031. The analyses reflect intellectual input and resource development from the Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project. National Heart, Lung and Blood Institute Division of Intramural Research funds supported personnel involved in this work. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Genetic Analysis for Idiopathic Thrombophilia (GAIT) project was supported partially by grants PI-11/0184, PI-14/0582 and Red Investigación Cardiovascular RD12/0042/0032 from the Spanish Health Institute (ISCIII), and acknowledges current support as a Consolidated Research Group from the Government of Catalonia and CERCA Program (2017 SGR 1736), and from the non-profit organization "Associació Activa´TT per la Salut". Maria Sabater-Lleal is supported by a Miguel Servet contract from the ISCIII Spanish Health Institute (CP17/00142) and co-financed by the European Social Fund.

GBMI: See supplement of GBMI flagship paper for contributing cohort funding acknowledgements (https://doi.org/10.1101/2021.11.19.21266436). W.Z. was supported by the National Human Genome Research Institute of the National Institutes of Health under award number T32HG010464. BNW and CJW were supported by the National Heart Lung Blood under award number R35HL135824.

Heart and Vascular Health (HVH) VTE Study: The HVH Study was supported by National Heart, Lung, and Blood Institute grants HL43201, HL60739, HL68986, HL73410, HL74745, HL85251, HL95080, and HL121414.

HUNT: The Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between HUNT Research Center (Faculty of Medicine and Health Sciences, NTNU, Norwegian University of Science and Technology), Nord-Trøndelag County Council, Central Norway Regional Health Authority, and the Norwegian Institute of Public Health. Drs. Ben Brumpton and Kristian Hveem work in a research unit funded by Stiftelsen Kristian Gerhard Jebsen; Faculty of Medicine and Health Sciences, NTNU; The Liaison Committee for education, research and innovation in Central Norway; and the Joint Research Committee between St. Olavs Hospital and

the Faculty of Medicine and Health Sciences, NTNU. The genotyping in HUNT was financed by the National Institute of Health (NIH); University of Michigan; The Research Council of Norway; The Liaison Committee for education, research and innovation in Central Norway; and the Joint Research Committee between St. Olavs Hospital and the Faculty of Medicine and Health Sciences, NTNU. The K.G. Jebsen Center for Genetic Epidemiology is financed by Stiftelsen Kristian Gerhard Jebsen, Faculty of Medicine and Health Sciences Norwegian University of Science and Technology (NTNU) and the Liaison Committee for education, research and innovation in Central Norway. Dr. Brumpton is financed by the Medical Research Council Integrative Epidemiology Unit at the University of Bristol which is supported by the Medical Research Council and the University of Bristol [MC_UU_12013/1].

JUPITER: The JUPITER trial and its genetic substudy were funded by AstraZeneca.

MARTHA & MARTHA12: The MARTHA and MARTHA12 projects were supported by grants from the Program Hospitalier de Recherche Clinique. MARTHA related genetics research programs were supported and funded by the GenMed LABEX (ANR-10-LBX-0013) and the French INvestigation Network on Venous Thrombo-Embolism (INNOVTE). David-Alexandre Trégouët was financially supported by the “EPIDEMIOM-VTE” Senior Chair from the Initiative of Excellence of the University of Bordeaux.

Mass General Brigham Biobank (MGBB): Pradeep Natarajan is supported by grants from the National Heart Lung and Blood Institute (R01HL142711, R01HL148050, R01HL151283, R01HL127564, R01HL148565, R01HL135242, R01HL151152), National Institute of Diabetes and Digestive and Kidney Diseases (R01DK125782), Fondation Leducq (TNE-18CVD04), and Massachusetts General Hospital (Paul and Phyllis Fireman Endowed Chair in Vascular Medicine). Thomas Gilliland is supported by T32 grant 5T32HL125232 of the National Heart, Lung, and Blood Institute.

Mayo VTE Study: The research was funded, in part, by grants from the National Institutes of Health, National Heart, Lung, and Blood Institute (HL66216 and HL83141) and the National Human Genome Research Institute (HG04735, HG06379-07, and HG06379-08) and Mayo Foundation.

Million Veteran Program: The MVP is funded by the Department of Veterans Affairs Office of Research and Development, Million Veteran Program Grant #MVP000. This publication does not represent the views of the Department of Veterans Affairs or the United States Government. MVP was also supported by three additional Department of Veterans Affairs awards (I01-01BX03340, I01-BX003362, and I01-CX001025). Scott M. Damrauer is supported by the Veterans Administration (IK2-CX001780). Pradeep Natarajan is supported by the NIH/NHLBI K08HL140203 R01HL142711 and a Hassenfeld Award from the Massachusetts General Hospital.

Multiple Environmental and Genetic Assessment of risk factors for VT study (MEGA): The MEGA study was supported by Netherlands Heart Foundation (NHS 98.113), the Dutch Cancer Foundation (RUL 99/1992), the Netherlands Organisation for Scientific Research (912–03–033| 2003), and partially by the Laboratory of Excellence in Medical Genomics (GenMed LABEX ANR-10-LABX-0013). We would like to thank all colleagues from the French Centre National de Génotypage for the genotyping contribution

Multi-Ethnic Study of Atherosclerosis (MESA): The Multi-Ethnic Study of Atherosclerosis (MESA) and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881, and DK063491. Funding for SHARe genotyping was provided by National Heart, Lung, and Blood Institute contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, CA) and the Broad Institute of Harvard and the Massachusetts Institute of Technology (Boston, MA) using the Affymetrix Genome-Wide Human SNP Array 6.0. Pamela Lutsey is supported by K24 grant HL159246 of the National Heart, Lung, and Blood Institute.

Nurse's Health Study (NHS and NHS-II) and Health Professionals Follow Up Study: The work for the current study was funded by grants from the NIH: P01CA87969, R01CA49449, R01HL034594, R01HL088521, R01CA50385, R01CA67262, P01CA055075, R01HL35464, R01HL116854.

Penn Medicine Biobank (PMBB): We thank the patient-participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878.

RETROVE: The RETROVE study acknowledges support from the Government of Catalonia and CERCA Program (2017 SGR 1736), from the Spanish Health Institute (ISCIII grants PI12/000612, PI15/00269 and PI18/00434), and from the non-profit organization "Associació Activa´TT per la Salut". Maria Sabater-Lleal is supported by a Miguel Servet contract from the ISCIII Spanish Health Institute (CP17/00142) and co-financed by the European Social Fund. The genotyping service was carried out at CEGEN-PRB3-ISCIII; and it is supported by grant PT17/0019, of the PE I+D+i 2013-2016, funded by ISCIII and ERDF.

Tromsø Study: The Tromsø Study was supported by an independent grant from Stiftelsen Kristian Gerhard Jebsen in Norway (J.B.H.).

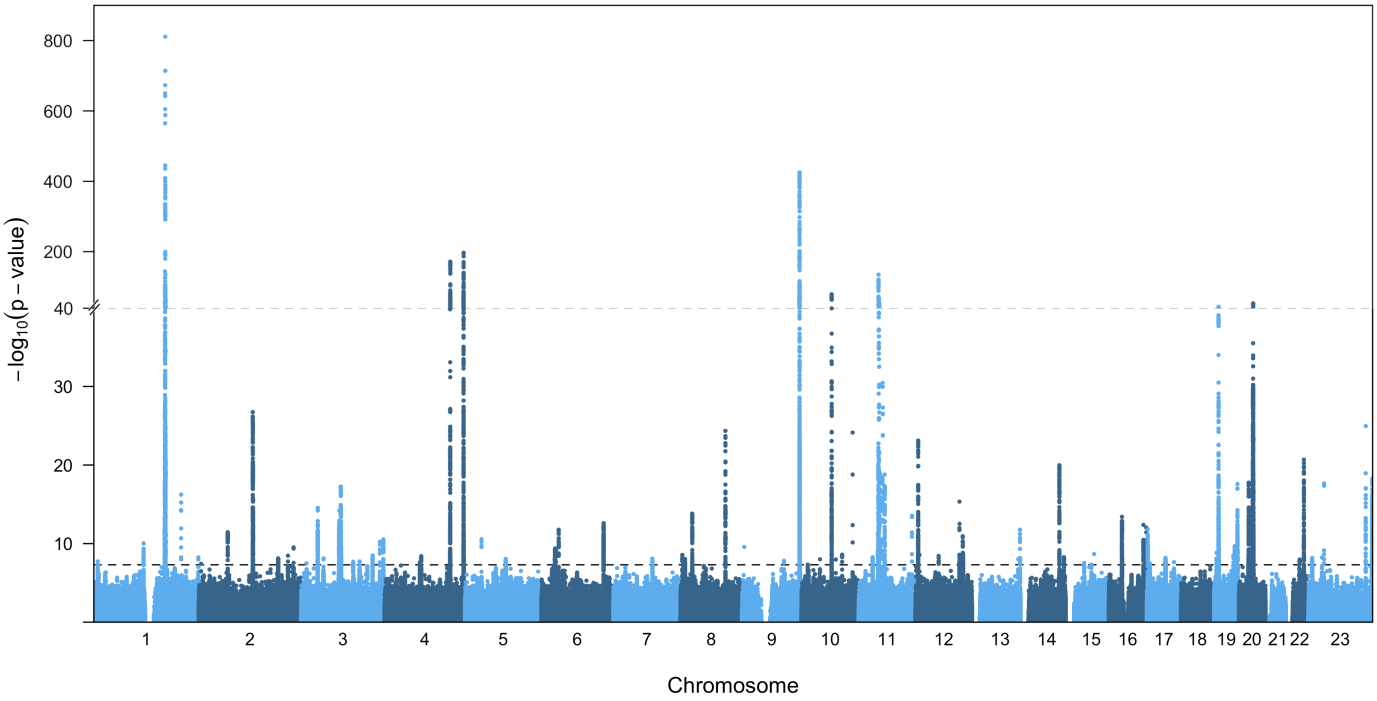
UK Biobank: UK Biobank has received funding from the UK Medical Research Council, Wellcome Trust, UK Department of Health, British Heart Foundation, Cancer Research UK, Diabetes UK, Northwest Regional Development Agency, Scottish Government, and Welsh Assembly Government. This research has been conducted using the UK Biobank Resource under Application number 40713.

Women's Genome Health Study (WGHS): The WGHS is supported by the National Heart, Lung, and Blood Institute (HL043851 and HL080467) and the National Cancer Institute (CA047988 and UM1CA182913), with funding for genotyping provided by Amgen.

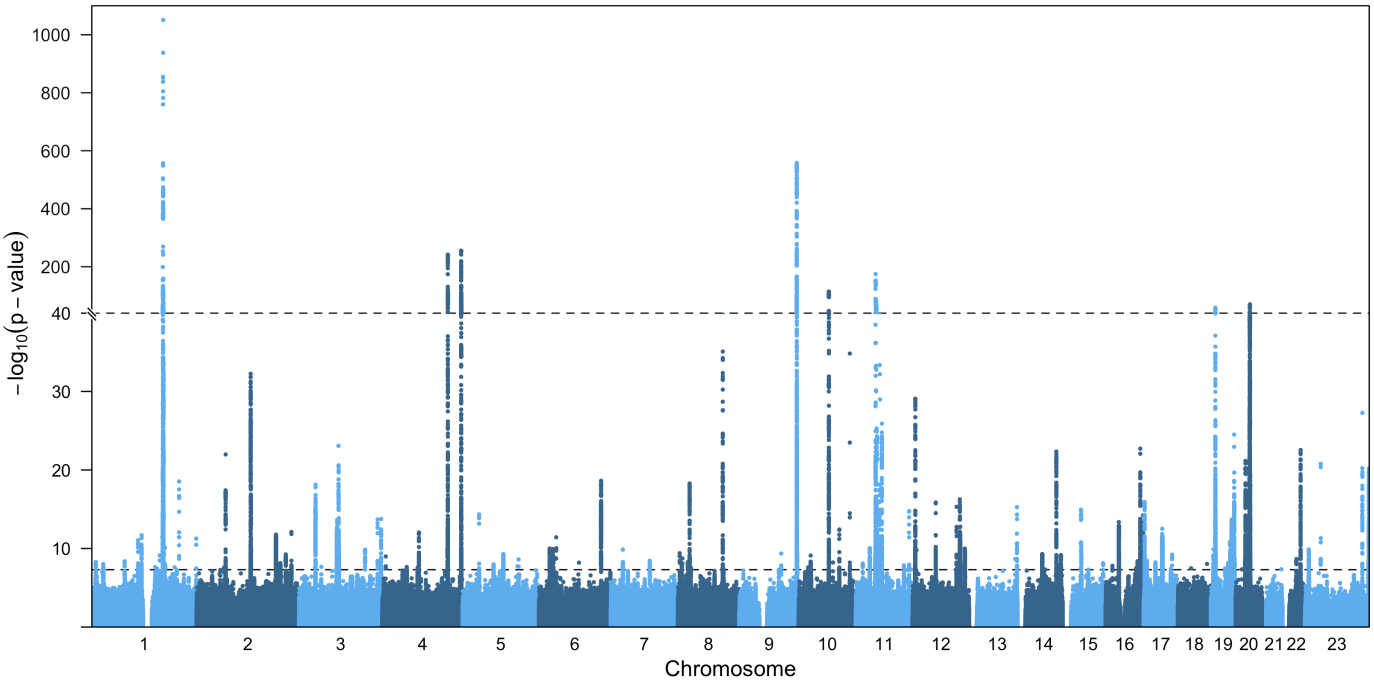
Women's Health Initiative (WHI): The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, and 75N92021D00005.

Supplementary Figure 9.
Manhattan plots.

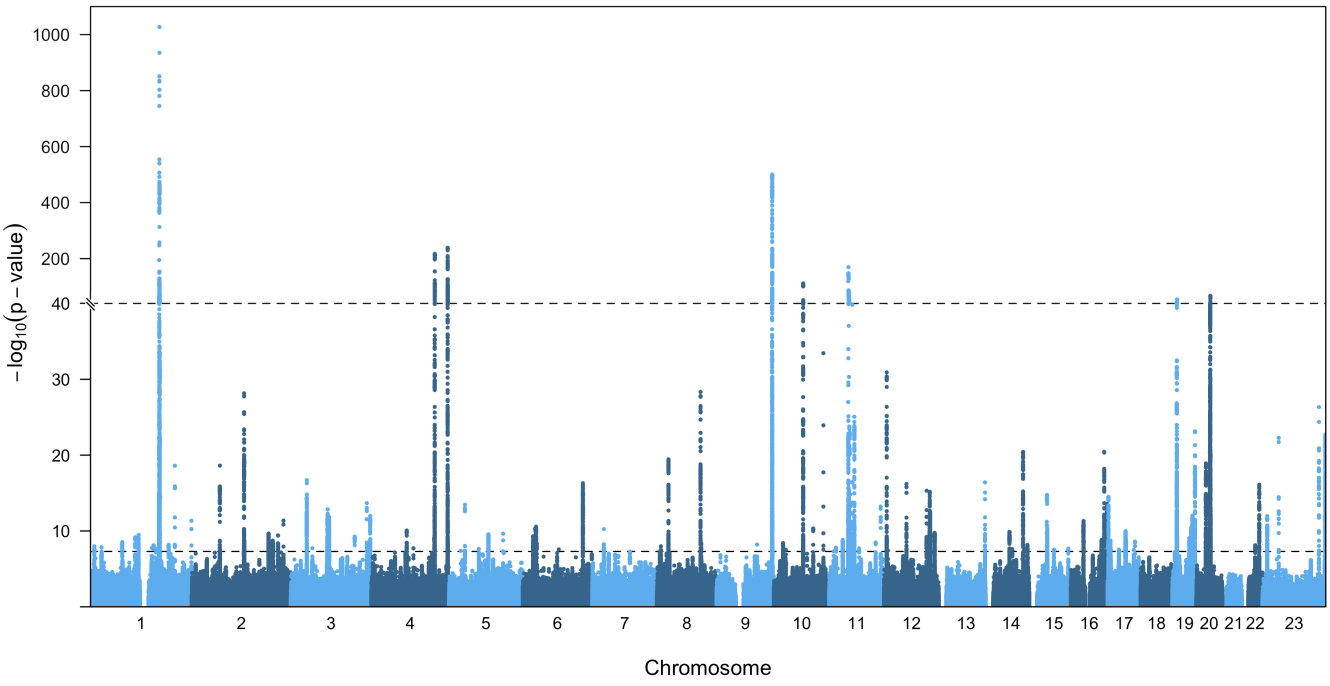
Supplementary Figure 9.1 – Manhattan plot of the Discovery meta-analysis



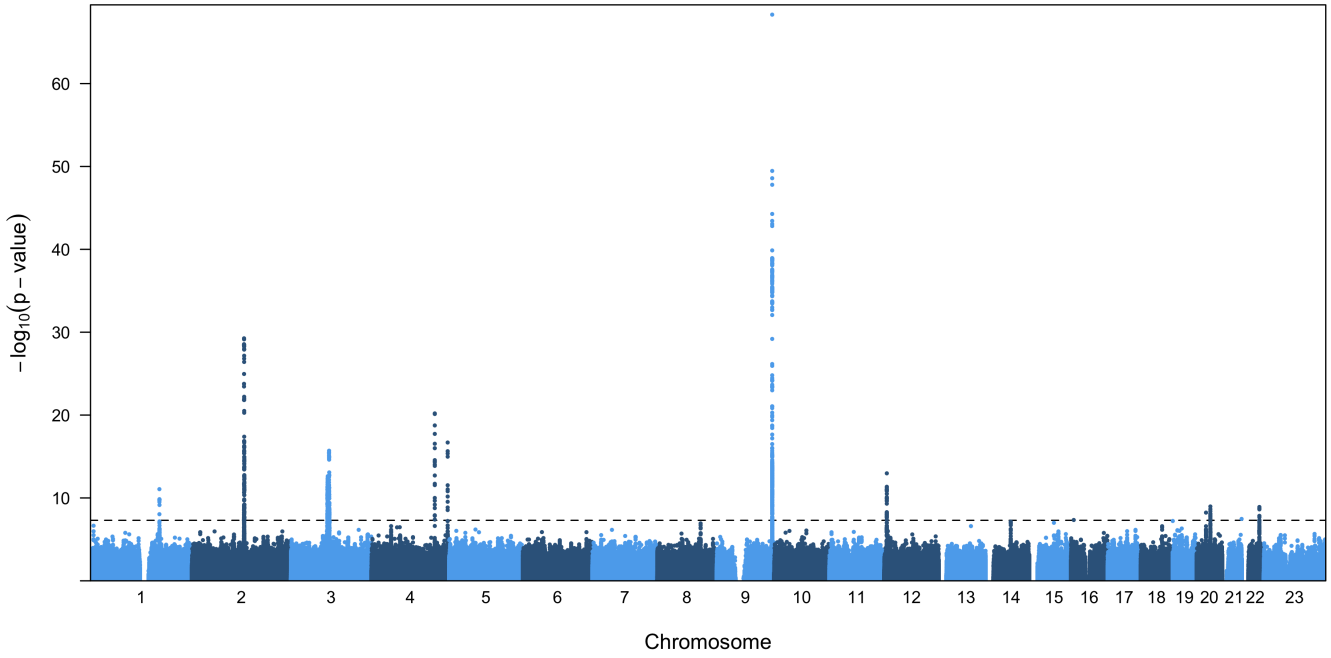
Supplementary Figure 9.2 – Manhattan plot of the Combined meta-analysis



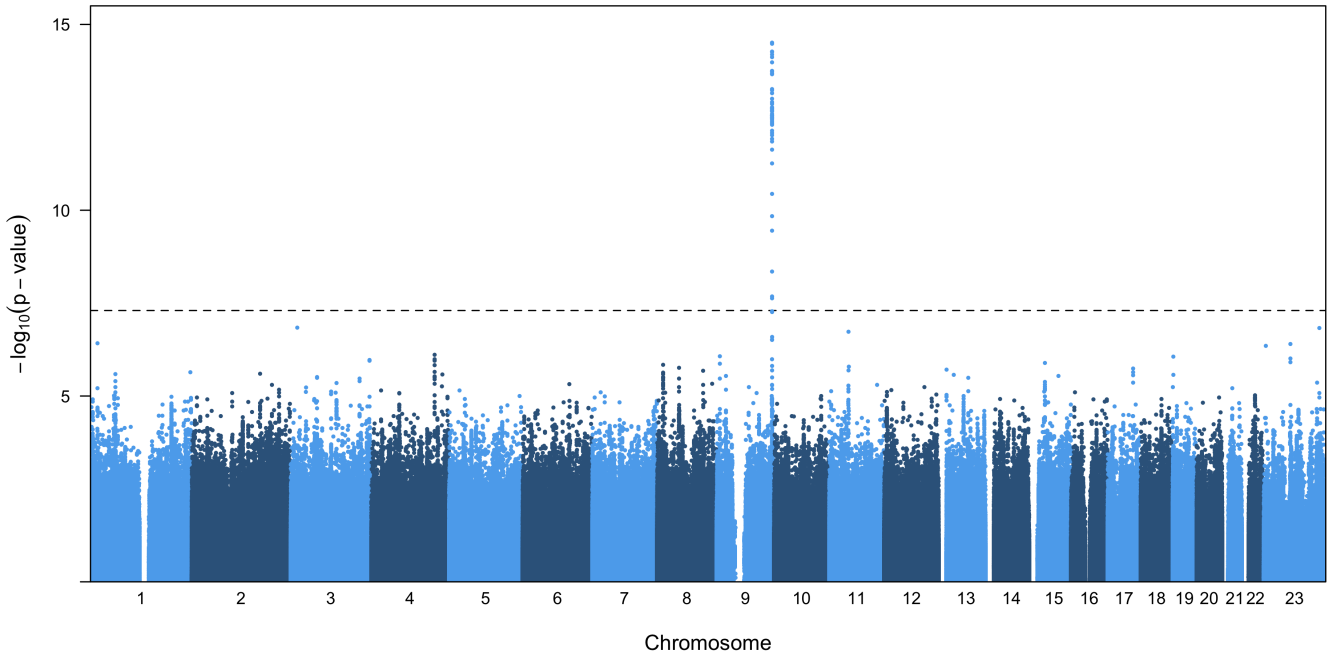
Supplementary Figure 9.3 – Manhattan plot of the European meta-analysis



Supplementary Figure 9.4 – Manhattan plot of the African meta-analysis

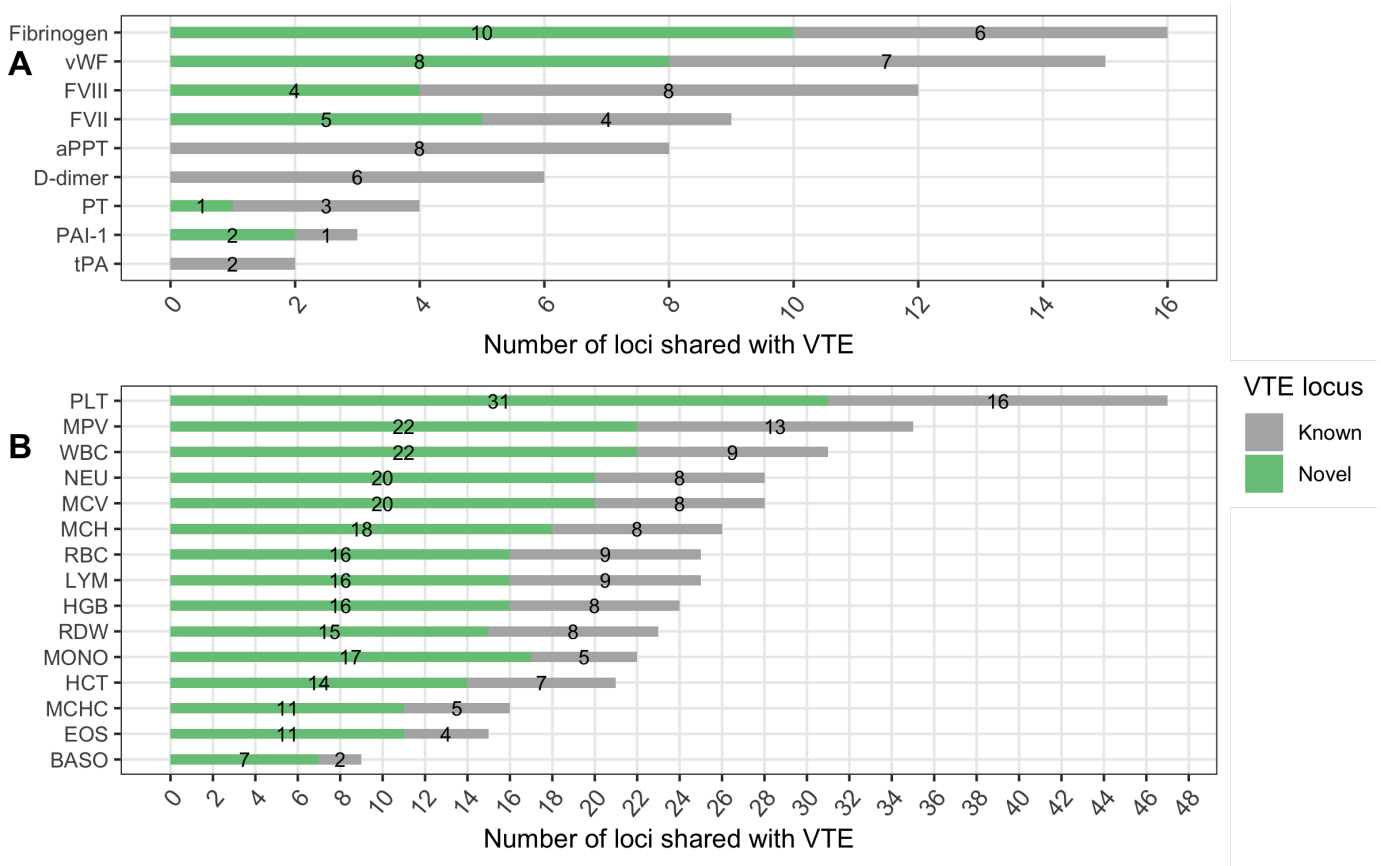


Supplementary Figure 9.5 – Manhattan plot of the Hispanic meta-analysis

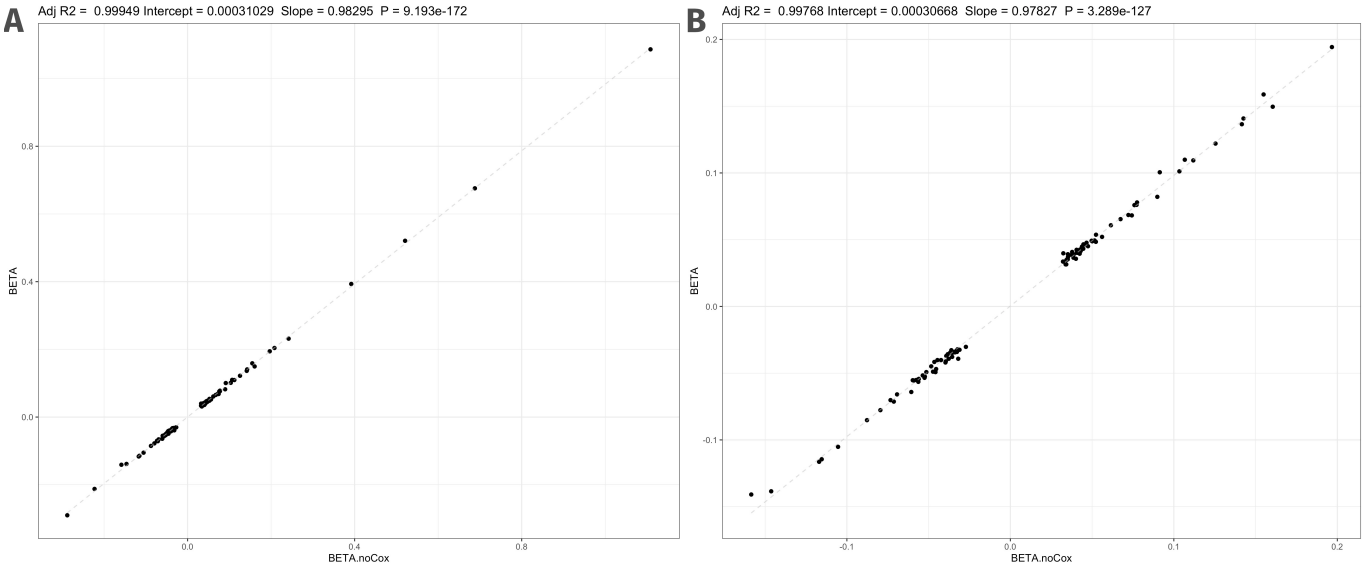


Supplementary Figure 10-12.

Supplementary Figure 10 – VTE genetic loci shared with hemostatic factors and blood traits. (A) Number of known and novel VTE loci shared with each of the 10 hemostatic factors investigated. (B) Same analysis with complete blood count traits: PLT (platelet count), MPV (mean platelet volume), RBC (red blood cell count), MCV (mean corpuscular volume), HCT (hematocrit), MCH (mean corpuscular hemoglobin), MCHC (MCH concentration), HGB (hemoglobin concentration), RDW (red cell distribution width), WBC (white blood cell count), MONO (monocyte count), NEU (neutrophil count), EOS (eosinophil count), BASO (basophil count), LYM (lymphocyte count).



Supplementary Figure 11 – Sensitivity analysis comparing the effects of the Combined meta-analysis lead variants (y-axis) to the effects observed in an alternative Combined meta-analysis that did not include datasets that used Cox regression (x-axis). (A) is showing all 111 lead variants. (B) is the same plot zoomed in between BETA=-0.2 and 0.2



Supplementary Figure 12 – Chromosome X sensitivity analysis of the Combined meta-analysis. The top panel is a Manhattan plot of the P-value based meta-analysis, while the bottom panel is a Manhattan plot of the inverse-variance weighted meta-analysis.