A large-scale exome array analysis of venous thromboembolism

Sara Lindström[1,2], Jennifer A. Brody[3], Constance Turman[4], Marine Germain[5], Traci M. Bartz[3,6], Erin N. Smith[7,8], Ming-Huei Chen[9], Marja Puurunen[10], Daniel Chasman[11], Jeffrey Hassler[2], Nathan Pankratz[12], Saonli Basu[13], Weihua Guan[13], Beata Gyorgy[14,15], Manal Ibrahim[16,17,18], Jean-Philippe Empana[19,20], Robert Olaso[21], Rebecca Jackson[22], Sigrid K. Brækkan[8,23], Barbara McKnight[6], Jean-Francois Deleuze[21], Cristopher J. O'Donnell[24], Xavier Jouven[19,25], Kelly A. Frazer[7,8,26], Bruce M. Psaty[1,3,27,28], Kerri L. Wiggins[3], Kent Taylor[29], Alexander P. Reiner[1], Susan R. Heckbert[1], Charles Kooperberg[2], Paul Ridker[11], John-Bjarne Hansen[8,23], Weihong Tang, Andrew D. Johnson[9], Pierre-Emmanuel Morange[16,17,18], David A. Trégouët[5], Peter Kraft[4,31], Nicholas L. Smith[1,28,32], Christopher Kabrhel[33,34,35], on behalf of the INVENT Consortium.

[1] Department of Epidemiology, University of Washington, Seattle, United States.
[2] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, United States.
[3] Department of Medicine, University of Washington, Seattle, United States.
[4] Department of Epidemiology Harvard TH Chan School of Public Health, Boston, United States.
[5] University of Bordeaux, Inserm 1219, Bordeaux Population Health Research Center, Bordeaux, France.
[6] Department of Biostatistics University of Washington, Seattle, United States.
[7] Department of Pediatrics and Rady Children's Hospital University of California, San Diego, La Jolla, United State.
[8] Department of Clinical Medicine, UiT - The Arctic University of Norway, K.G. Jebsen Thrombosis Research and Expertise Center (TREC), Tromsø, Norway.
[9] Population Sciences Branch, National Heart, Lung and Blood Institute's The Framingham Heart Study, Framingham, United States.
[10] School of Medicine, Boston University, Boston, United States.
[11] Division of Preventive Medicine, Brigham and Women's Hospital, Boston, United States.
[12] Department of Laboratory Medicine and Pathology, School of Medicine, University of Minnesota, Minneapolis, USA,
[13] Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, USA.
[14] Team Genomics & Pathophysiology of Cardiovascular Diseases, Sorbonne Universités, UPMC Univ. Paris 06, INSERM, UMR_S 1166, Paris, France.
[15] ICAN Institute for Cardiometabolism and Nutrition, Paris, France.
[16] Laboratory of Haematology, La Timone Hospital, Marseille, France.
[17] Aix-Marseille University, INSERM, INSERM, INRA, C2VN, Marseille, France.
[18] CRB Assistance Publique Hopitaux de Marseille HemoVasc, Marseille, France,
[19] Department of Epidemiology, Université Paris Descartes, Sorbonne Paris Cité, INSERM UMR_S 970, Paris, France.
[20] Faculté de Médecine, Université Paris Descartes, Sorbonne Paris Cité, Paris, France.
[21] Centre National de Recherche en Génomique Humaine (CNRGH), Direction de la Recherche Fondamentale, CEA, Institut de Biologie François Jacob, Evry, France.
[22]Ohio State University, Columbus, United States.
[23] Division of Internal Medicine, University Hospital of North Norway, Tromsø, Norway.
[24] Cardiology, Boston Veteran's Administration Healthcare, Boston, United States.
[25] Department of Cardiology, Georges Pompidou European Hospital, APHP, Paris, France.
[26] Institute for Genomic Medicine, University of California, San Diego, La Jolla, United States.
[27] Department of Health Services, University of Washington, Seattle, United States.
[28] Kaiser Permanente Washington Research Institute, Kaiser Permanente Washington, Seattle, United States.

[29] *LA BioMed, Torrance, United States*
[30] *Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, United States.*
[31] *Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, United States,*
[32]*Department of Veteran Affairs Office of Research and Development, Seattle Epidemiologic Research and Information Center, Seattle, United States.*
[33] *Center for Vascular Emergencies, Department of Emergency Medicine, Massachusetts General Hospital, Boston, United States,*
[34] *Channing Network Medicine, Brigham and Women's Hospital, Boston, United States.*
[35] *Harvard Medical School, Boston, United States*

Running title: Exome array analysis of venous thromboembolism

Corresponding author:
Christopher Kabrhel
55 Fruit Street
Boston, MA 02114-2696
Email: ckabrhel@partners.org
Phone: 617-726-7622

Keywords: Exome, genetic association, venous thromboembolism

Subject Codes: Etiology, Epidemiology, Risk factors, Genetic association studies, Embolism, Thrombosis.

Word count: 2,049

Number of Figures: 1

Number of Tables: 3

TOC category: Clinical and population studies

TOC subcategory: Thrombosis

ABSTRACT

Much of the heritability of venous thromboembolism (VTE) remains unexplained. Although recent genome-wide association studies (GWAS) have identified novel associations for common variants, there has been no comprehensive exome-wide search for low-frequency variants that may affect the risk of VTE. We conducted a meta-analysis of 11 studies comprising a total of 8,332 cases and 16,087 controls of European ancestry and 382 cases and 1,476 controls of African-American ancestry genotyped with the Illumina HumanExome BeadChip. We used the seqMeta package in R to conduct single variant and gene-based rare variant tests. In the single variant analysis, we limited our analysis to the 64,794 variants that had at least 40 minor alleles across studies (corresponding to a minor allele frequency of ~ 0.08%). We confirmed associations with previously identified VTE loci, including *ABO, F5, F11,* and *FGA*. After adjusting for multiple testing, we observed no novel significant findings in either single variant or gene-based analysis. Given our sample size, we had >80% power to detect minimum odds ratios >1.5 and 1.8 for a single variant with minor allele frequency of 0.01 and 0.005, respectively. Beyond already known associations, we did not observe evidence for additional rare coding variants with moderate-to-large effects contributing to VTE risk. Larger studies and sequence data may be needed to identify novel low-frequency and rare variants associated with VTE risk.

INTRODUCTION

Candidate gene studies and genome-wide association studies (GWAS) have identified multiple genetic variants that are associated with venous thromboembolism (VTE), a condition spanning both pulmonary embolism (PE) and deep vein thrombosis (DVT). The majority of genetic variants associated with VTE have been located in genes known to be involved in hemostasis, such as *ABO, F2, F5, F11, FGG and PROCR*[1]. Associations have also been observed for variants located in genes outside known hemostasis pathways, such as *TSPAN15* and *SLC44A2* [2], and the exact mechanisms by which these genes influence VTE risk have yet to be determined.

Despite these successes, much of the heritability of VTE remains unexplained. A recent study based on 3,290 VTE cases and 116,868 controls from the UK Biobank estimated the heritability due to genotyped and imputed SNPs to be ~30%[3], and twin studies have estimated VTE heritability to be as high as ~ 50%[4]. However, the UK Biobank study also noted that known variants only explain 5% of VTE heritability. Thus, additional genetic variants that contribute to VTE risk remain to be discovered.

Recently developed exome arrays[5] allow for cost-efficient genotyping of 240,000 coding variants identified through the NHLBI Exome Sequencing Project[6]. Based on exome and whole-genome sequencing data from 9,000 subjects of European ancestry, 2,000 subjects of African ancestry and 500 subjects each of Hispanic and Asian ancestry, 240,000 SNPs were selected for inclusion on the exome array. To be selected, non-synonymous variants had to be seen at least three times in at least two datasets whereas splice and stop variants had to be seen at least two times and in at least two datasets. The exome array has proven to be an efficient tool for identifying low-frequency coding variants associated with blood and cardiovascular traits including: hypertension[7,8], hematological traits[9,10,11,12], lipid levels[13], coronary artery disease[14], and atrial fibrillation[15]. However, no study has published a comprehensive investigation of the association between low-frequency exonic variants and VTE.

We hypothesized that exonic, low-frequency genetic variation would be associated with VTE. We meta-analyzed exome array genotype data from 11 European and US studies, totaling 8,723 VTE cases and 17,563 controls. We conducted both single-variant and gene-based tests to identify novel genetic variants associated with VTE risk.

MATERIALS AND METHODS

Study participants

All study participants were either of European or African-American ancestry and came from eight US-based cohorts (ARIC, CHS, FHS, HPFS, NHS, NHSII, WGHS and WHI), one US-based case-control study (HVH), one Norwegian case-control study (Tromsø) and one French case-control study (MARTHA)[16-28] **(Table 1)**. Details of each study have been previously published[16-27]. Physician-diagnosed VTE was identified either through hospital records or validated self-reports, supplemented by review of medical records**.** A detailed description of the study-specific design and characteristics is presented in **Supplementary Table 1**. All participating studies were approved by their respective institutional review board and informed consent for genetic analyses was obtained from each study participant.

Genotyping and quality control

Genotyping was conducted using either the Illumina HumanExome BeadChip v1.0, Illumina HumanExome BeadChip v1.1 or the Illumina HumanCore Exome BeadChip v1.1, depending on study. Genotypes from 765 samples from the Tromsø study were obtained from exome sequencing rather than genotyping. Genotypes were called using either GenomeStudio or Zcall. Each study conducted data cleaning and quality assurance checks following a common protocol. Details of study-specific genotype calling and quality control can be found in **Supplementary Table 1.**

One of the included studies (MARTHA) genotyped cases and controls on separate platforms (cases were genotyped on Illumina HumanExome 12v.1-2_A and controls were genotyped with the Illumina HumanExome 12v.1_A array). We identified a moderate inflation in test statistics ($\lambda_{1000} = 1.09$) in MARTHA for the single variant analysis with MAF>0.005. Therefore, we re-meta-analyzed the data while excluding MARTHA, leaving a total of 6,095 cases and 14,149 controls in a sensitivity analysis.

Statistical Methods

Each study conducted individual analysis following a common protocol. To avoid type-1 error inflation[29], studies with more than four controls per case randomly selected a maximum of four controls for each case (i.e. 1:4 case:control ratio). Those studies that performed control selection (ARIC, CHS, FHS, WGHS and WHI) reviewed the distribution of age and sex following control selection to ensure roughly equal distributions among cases and controls. Each study conducted both single variant and gene-based analysis. Association analysis were based on logistic regression adjusting for age, sex, principal components and other study-specific variables (as needed). Analyses were conducted using the seqMeta[30] package in R which produces study-specific results. Each study sent their study-specific results to the coordinating center at Harvard T.H. Chan School of Public Health where the meta-analyses took place.

For single variant analysis, we conducted a meta-analysis of the study-specific score statistics based on an additive coding. We limited our analysis to the 64,794 variants that had at least 40 minor alleles across studies (corresponding to a minor allele frequency of ~ 0.08%). Bonferroni correction for the number of variants tested was used to set the significance threshold for the analysis corresponding to $P<7.7\text{x}10^{-7}$ (0.05/64,794 variants).

For the gene-based rare variant analyses, we conducted two tests: Weighted-Sum Burden[31] (WSB) test as implemented in seqMeta[30] and SKAT[32]. We applied two different sets of criteria to select variants, based on coding variant annotation from five prediction algorithms (PolyPhen2, HumDiv and HumVar, LRT, MutationTaster and SIFT)[33]. The 'broad' definition included variants with a minor allele frequency (MAF) < 0.01 that were nonsense, stop-loss, splice site, as well as missense variants that are annotated as damaging by at least one prediction algorithm. The 'strict' definition included only variants with a MAF < 0.01 that were nonsense, stop-loss, splice-site, as well as missense variants annotated as damaging by all five algorithms. For the SKAT analysis, variants were weighted according to the beta density function as previously described[32]. We excluded all genes that had fewer than two variants included in the analysis. In total, we tested 15,041 genes using the broad definition and 5,749 genes using the strict definition. Bonferroni correction for the number of genes and tests performed was used to set the significance threshold for the gene-based analysis corresponding to $P<1.2\text{x}10^{-6}$ (0.05/41,580 tests [(5,749+15,041) genes × 2 tests]).

We conducted gene set enrichment analysis using the GSEAPreranked algorithm as implemented in the GenePattern software and the KEGG, Gene Ontology, and Hallmarks pathway sets[34,35]. We applied GSEAPreranked to four sets of results (1) Burden test of variants using a broad definition, (2) Burden test of variants using a strict definition, (3) SKAT test of variants using a broad definition, (4) SKAT test of variants using a strict definition.

One of the included studies (MARTHA) genotyped cases and controls on separate platforms (cases were genotyped on Illumina HumanExome 12v.1-2_A and controls were genotyped with the Illumina HumanExome 12v.1_A array. We identified a moderate inflation in test statistics ($\lambda_{1000} = 1.09$) in MARTHA for the single variant analysis with MAF>0.005. Therefore, we reran

the analysis excluding MARTHA, leaving a total of 6,095 cases and 14,149 controls in a sensitivity analysis.

RESULTS

Single Variant Analysis

After excluding all variants with MAC<40 (corresponding to a minor allele frequency of ~ 0.08%) in the combined study population, single variant meta-analysis showed no sign of genomic inflation (λ_{1000}=1.03, Supplementary Figure 1). The strongest association was observed for rs635634 at the *ABO* locus (OR=1.60, 95% CI: 1.52-1.68, *P*=1.51x10^{-73}). In addition, we observed significant associations for previously known genes including the *F5, FGG, and F11* **(Supplementary Table 2)**. The most strongly associated rare (MAF<0.01) variant we observed was rs121918472 (OR=1.93, 95% CI: 1.46-2.56, *P*=3.55x10^{-6}), a non-synonymous variant located in the Protein S (*PROS1*) gene, also known as the p.Ser501Pro or PS Herleen mutation, MAF=0.005. This variant is also known to be associated with VTE[36]. After excluding known loci, only one single variant remained significant after adjusting for multiple testing but this signal was driven exclusively by the MARTHA study (p=1.96x10^{-15} when including MARTHA, p=0.37 after excluding MARTHA). As the signal at this locus is most likely due to technical issues, we removed it from further analysis. No other variant reached genome-wide significance (*P*<7.7x10^{-7}). The strongest sub-threshold association was observed for rs755109, a common (MAF=0.37) variant previously associated with thyroid stimulating hormones (OR=1.10, 95% CI: 1.06-1.16, *P*=3.31x10^{-6}).

Weighted-Sum Burden (WSB) Rare Variant Analysis

No gene reached the pre-determined significance threshold of *P*<1.2x10^{-6} **(Figure 1).** The top three associated genes using the 'broad' and 'strict' definitions are shown in **Table 2** and all associations with p<0.01 are shown in **Supplementary Tables 3 & 4.** The *SERPINA10* gene

on chromosome 14 was the third strongest associated gene using the broad (p=0.0002) and the strict (p=0.0007) definition. *SERPINA10* is expressed primarily in the liver and mutations in this gene have previously been linked to VTE[37,38].

SKAT Rare Variant Analysis

No gene reached the pre-determined significance threshold of $P<1.2x10^{-6}$ **(Figure 1).** The top three associated genes using the using the 'broad' and 'strict' definitions are shown in **Table 2** and all associations with p<0.01 are shown in **Supplementary Tables 5 & 6**.

Gene Set Enrichment Analysis

Gene set enrichment analysis based on the results obtained from WSB and SKAT did not yield any significant pathway after adjusting for number of pathways tested (all FDR q>0.08, data not shown).

DISCUSSION

To assess the contribution of rare coding variation to VTE risk, we combined data from 11 studies spanning four countries, resulting in exome array data on 8,723 VTE cases and 17,563 controls. By comparison, the number of cases included in this study is larger than the largest GWAS of VTE published to date[2]. Beyond known associations, we did not observe evidence that additional low-frequency and rare coding variants with moderate-to-large effects contribute to VTE risk.

Although our study is the largest genomic study of VTE to date, our ability to identify rare variants associated with VTE was limited by low statistical power. Given our sample size, we had >80% power to detect minimum odds ratios of 1.56 and 1.81 for a single variant with MAF = 0.01 and 0.005, respectively. It is estimated that the exome array includes 97-98% of non-

synonymous variants and 94-95% of stop variants that would have been detected in an average genome through exome sequencing (https://genome.sph.umich.edu/wiki/Exome_Chip_Design). Thus, it is possible that we missed coding variants associated with VTE, especially if such variants are particularly rare in individuals who are not affected with VTE. Another limitation with this study is the limited contribution of non-European populations to our analyses, with 93% of our study population being of European ancestry.

Identifying risk factors for VTE, including genetic risk factors, is of great public health importance. VTE affects 1-2/1000 Americans yearly. The incidence has been increasing and mortality from PE remains high[39-42]. The mortality of VTE is greatest in the first 24 hours, and for one-fourth of PE patients, the initial clinical presentation is sudden death[43-45]. Therefore, our ability to improve mortality hinges on primary prevention, identifying patients at risk for VTE, and understanding the underlying pathophysiology of the disease. Despite the accumulated evidence that genetic factors play a major role in the pathophysiology of VTE, only 35% of VTE patients undergoing testing for thrombophilia carry a polymorphism known to increase VTE risk[46]. Additional efforts to identify genetic variants associated with VTE risk are still needed.

The INVENT Consortium is a well-established collaboration of genetic studies of VTE and, to our knowledge, our meta-analysis includes data from the vast majority of exome array studies of VTE. Additional large studies, potentially including comprehensive sequencing data, may be needed to identify novel low-frequency rare coding variants associated with VTE. Further research into the genetic basis of VTE is needed to aid in the primary prevention of this potentially fatal disease.

ACKNOWLEDGEMENTS

ARIC: The Atherosclerosis Risk in Communities (ARIC) 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 (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. Funding for LITE was supported by R01HL59367. Funding support for "Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium" was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419).

CHS: Cardiovascular Health Study: This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants U01HL080295, R01HL087652, R01HL105756, R01HL103612, R01HL120393, R01HL130114, and R01HL068986 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 UL1TR000124, 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.

FHS: The Framingham Heart Study is conducted and supported by the NHLBI in collaboration with Boston University (Contract No. N01-HC-25195). Genotyping, quality control, and calling of the Illumina HumanExome BeadChip in the Framingham Heart Study were supported by funding from the National Heart, Lung and Blood Institute Division of Intramural Research (Daniel Levy and Christopher J. O'Donnell, Principal Investigators). Support for the centralized genotype calling was provided by Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium through the National Institutes of Health (NIH) American Recovery and Reinvestment Act of 2009 (5RC2HL102419). M.H.C. and A.D.J. were supported by National Heart, Lung and Blood Institute Division of Intramural Research funds. 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.

HVH: The research of the Heart and Vascular Health Studies has been funded in part by National Institute of Health grants HL040628, HL043201, HL053375, HL060739, HL68639, HL068986, HL073410, HL74745, HL085251, HL095080.

MARTHA genetics project was supported by the GENMED Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013), the French Clinical Research Infrastructure Network on Venous Thrombo-Embolism (F-CRIN INNOVTE) and the ICAN Institute for Cardiometabolism and Nutrition (ANR-10-IAHU), three research programs managed by the National Research Agency (ANR) part of the French Investment for the Future initiative.

NHS/NHSII/HPFS - These studies received grant supports P01CA87969, R01CA49449, R01HL034594, R01HL088521, R01CA50385, R01CA67262, P01CA055075, R01HL35464, R01HL116854. We thank David Hunter, Frank B. Hu, Eric B. Rimm, RullaTamimi, Hyon Choi,

Charlie Fuchs, Louis Pasquale, Immaculata DeVivo, Andrew Chan, Daniel Cramer and Gary Curhan for their work on the component GWAS sets. We thank Hongyan Huang, Jihye Kim, Laura Harrington and Kaitlin Hagan for their support.

Tromsø: This work was supported by an independent grant from Stiftelsen Kristian Gerhard Jebsen in Norway.

WGHS: The Women's Genome Health Study is funded by the National Heart, Lung, and Blood Institute (HL043851, HL080467, HL099355) and the National Cancer Institute (CA047988 and UM1CA182913),, with support for genotyping provided by Amgen.

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 HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C.

REFERENCES

1. Crous-Bou, M., Harrington, L.B. & Kabrhel, C. Environmental and Genetic Risk Factors Associated with Venous Thromboembolism. *Semin Thromb Hemost* **42**, 808-820 (2016).
2. 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**, 532-42 (2015).
3. Klarin, D. *et al.* Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor. *Circ Cardiovasc Genet* **10**(2017).
4. Heit, J.A. *et al.* Familial segregation of venous thromboembolism. *J Thromb Haemost* **2**, 731-6 (2004).
5. Huyghe, J.R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* **45**, 197-201 (2013).
6. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
7. Surendran, P. *et al.* Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. *Nat Genet* **48**, 1151-1161 (2016).
8. Liu, C. *et al.* Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat Genet* **48**, 1162-70 (2016).
9. Eicher, J.D. *et al.* Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293 Individuals. *Am J Hum Genet* **99**, 40-55 (2016).
10. Chen, M.H. *et al.* Exome-chip meta-analysis identifies association between variation in ANKRD26 and platelet aggregation. *Platelets*, 1-10 (2017).
11. Chami, N. *et al.* Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. *Am J Hum Genet* **99**, 8-21 (2016).
12. Mousas, A. *et al.* Rare coding variants pinpoint genes that control human hematological traits. *PLoS Genet* **13**, e1006925 (2017).
13. Peloso, G.M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet* **94**, 223-32 (2014).
14. Myocardial Infarction, G. *et al.* Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. *N Engl J Med* **374**, 1134-44 (2016).
15. Christophersen, I.E. *et al.* Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat Genet* **49**, 946-952 (2017).
16. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol* **129**, 687-702 (1989).
17. Fried, L.P. *et al.* The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* **1**, 263-76 (1991).
18. Tell, G.S. *et al.* Recruitment of adults 65 years and older as participants in the Cardiovascular Health Study. *Ann Epidemiol* **3**, 358-66 (1993).
19. Germain, M. *et al.* Genetics of venous thrombosis: insights from a new genome wide association study. *PLoS One* **6**, e25581 (2011).
20. Hankinson, S.E. *et al.* Reproductive factors and family history of breast cancer in relation to plasma estrogen and prolactin levels in postmenopausal women in the Nurses' Health Study (United States). *Cancer Causes Control* **6**, 217-24 (1995).
21. Tworoger, S.S., Sluss, P. & Hankinson, S.E. Association between plasma prolactin concentrations and risk of breast cancer among predominately premenopausal women. *Cancer Res* **66**, 2476-82 (2006).
22. Ridker, P.M. *et al.* Rationale, design, and methodology of the Women's Genome Health Study: a genome-wide association study of more than 25,000 initially healthy american women. *Clin Chem* **54**, 249-55 (2008).

23. Solomon, T. *et al.* Associations Between Common and Rare Exonic Genetic Variants and Serum Levels of 20 Cardiovascular-Related Proteins: The Tromso Study. *Circ Cardiovasc Genet* **9**, 375-83 (2016).
24. Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N. & Stokes, J., 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. *Ann Intern Med* **55**, 33-50 (1961).
25. Giovannucci, E. *et al.* Intake of carotenoids and retinol in relation to risk of prostate cancer. *J Natl Cancer Inst* **87**, 1767-76 (1995).
26. Heckbert, S.R., Li, G., Cummings, S.R., Smith, N.L. & Psaty, B.M. Use of alendronate and risk of incident atrial fibrillation in women. *Arch Intern Med* **168**, 826-31 (2008).
27. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials* **19**, 61-109 (1998).
28. Psaty, B.M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* **2**, 73-80 (2009).
29. Ma, C., Blackwell, T., Boehnke, M., Scott, L.J. & Go, T.D.i. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol* **37**, 539-50 (2013).
30. Voorman A, B.J., Chen H, Lumley T, David B. seqMeta: An R package for meta-analyzing region-based tests of rare DNA variants. *R package version 1.6.7* (2017).
31. Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).
32. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
33. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-90 (2014).
34. Mootha, V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-73 (2003).
35. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
36. Suchon, P. *et al.* Protein S Heerlen mutation heterozygosity is associated with venous thrombosis risk. *Sci Rep* **7**, 45507 (2017).
37. Van de Water, N. *et al.* Mutations within the protein Z-dependent protease inhibitor gene are associated with venous thromboembolic disease: a new form of thrombophilia. *Br J Haematol* **127**, 190-4 (2004).
38. Corral, J. *et al.* A nonsense polymorphism in the protein Z-dependent protease inhibitor increases the risk for venous thrombosis. *Blood* **108**, 177-83 (2006).
39. Arshad, N., Isaksen, T., Hansen, J.B. & Braekkan, S.K. Time trends in incidence rates of venous thromboembolism in a large cohort recruited from the general population. *Eur J Epidemiol* **32**, 299-305 (2017).
40. Heit, J.A. *et al.* Reasons for the persistent incidence of venous thromboembolism. *Thromb Haemost* **117**, 390-400 (2017).
41. 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**, 585-93 (1998).
42. Heit, J.A. Venous thromboembolism: disease burden, outcomes and risk factors. *J Thromb Haemost* **3**, 1611-7 (2005).
43. Goldhaber, S.Z., Visani, L. & De Rosa, M. Acute pulmonary embolism: clinical outcomes in the International Cooperative Pulmonary Embolism Registry (ICOPER). *Lancet* **353**, 1386-9 (1999).

44. Courtney, D.M. & Kline, J.A. Identification of prearrest clinical factors associated with outpatient fatal pulmonary embolism. *Acad Emerg Med* **8**, 1136-42 (2001).
45. Cohen, A.T. *et al.* Venous thromboembolism (VTE) in Europe. The number of VTE events and associated morbidity and mortality. *Thromb Haemost* **98**, 756-64 (2007).
46. Cushman, M. Inherited risk factors for venous thrombosis. *Hematology Am Soc Hematol Educ Program*, 452-7 (2005).

Table 1: Studies included in the VTE exome array analysis.

Ethnicity	Study	Country	Cases	Controls
African-American	ARIC	US	202	807
African-American	CHS	US	30	120
African-American	HVH	US	58	181
African-American	WHI	US	92	368
European	ARIC	US	433	1,734
European	CHS	US	112	448
European	FHS	US	212	848
European	HPFS/NHS/NHSII	US	2,321	2,301
European	HVH	US	841	1,788
European	MARTHA	France	2,628	3,414
European	Tromsø	Norway	528	526
European	WGHS	US	610	2,404
European	WHI	US	656	2,624
	Total		**8,341**	**16,087**

Table 2: Association results for the three strongest associations from the rare variant WSB test.

Variant Inclusion	**Gene**	**Beta**	**SE**	***P***	**CMAF***	**# Variants****
Broad	*FAM71C*	3.86	0.871	9.30E-06	0.0002	2
	FOXB2	1.56	0.372	2.71E-05	0.0009	4
	SERPINA10	0.28	0.077	0.0002	0.017	8
Strict	*DGAT2*	0.64	0.179	0.0004	0.003	4
	NUDT12	3.11	0.877	0.0004	0.0002	2
	SERPINA10	0.28	0.082	0.0007	0.015	6

* CMAF = Cumulative MAF for SNPs included in the analysis

** Number of Variants included in the analysis

Table 3: Association results for the three strongest associations from the rare variant SKAT test.

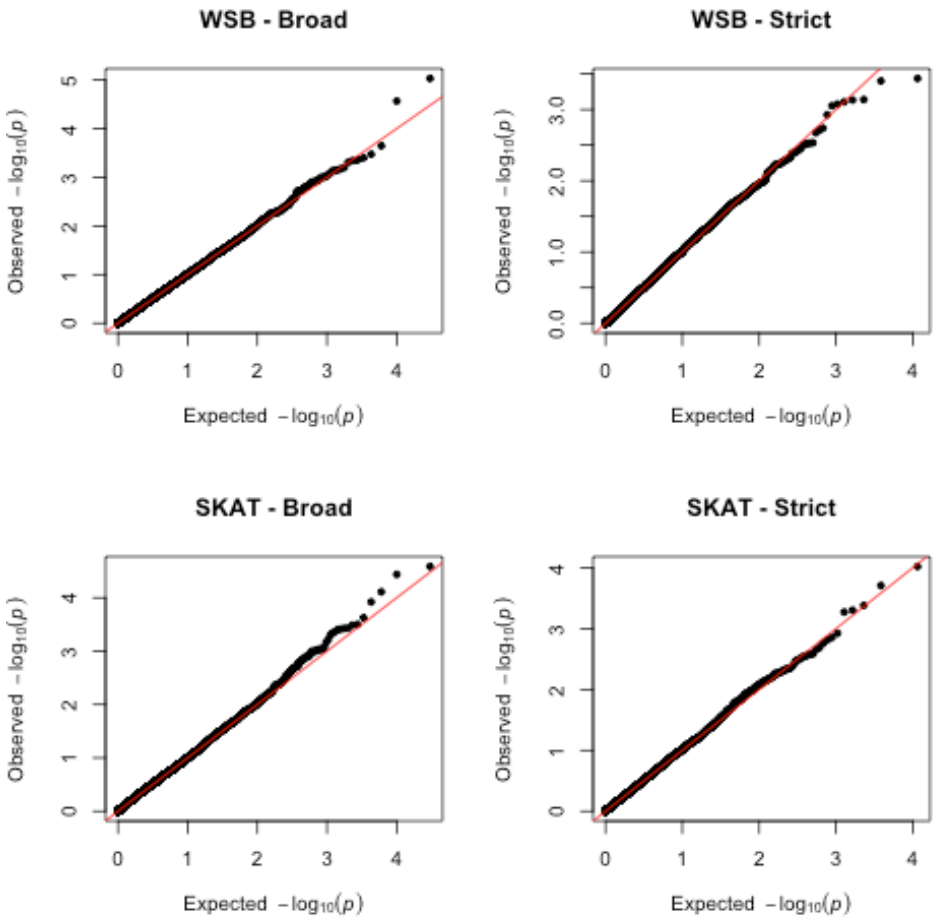
Variant Inclusion	**Gene**	**Qmeta***	***P***	**CMAF****	**# Variants*****
Broad	*CREB3L1*	596343.13	2.59E-05	0.016	7
	FAM71C	9844.60	3.65E-05	0.0002	2
	PHC3	514613.94	7.76E-05	0.017	10
Strict	*SRR*	37032.58	9.52E-05	0.0007	4
	ABCF3	15590.02	0.0002	0.0007	2
	DSC1	50186.59	0.0004	0.0014	5

***** The SKAT Q statistic, defined as $\sum_j w_j U_j^2$, where w_j is the weight given to SNP j, and U_j^2 is the associated score statistic

** CMAF = Cumulative MAF for SNPs included in the analysis

*** Number of Variants included in the analysis

Figure 1: QQ plots for gene burden tests including non-synonymous variants with MAF≤0.01. The WSB test using a broad definition of variant inclusion (upper left panel), the WSB test using a strict definition of variant inclusion (upper right panel), the SKAT test using a broad definition of variant inclusion (lower left panel), the SKAT test using a strict definition of variant inclusion (lower right panel).



SUPPLEMENTARY INFORMATION

Supplementary Figure 1: QQ-plot for single variant analysis based on meta-analysis.
Supplementary Figure 2: QQ-plot for single variant analysis based on meta-analysis with known loci excluded.

Supplementary Table 1: Characteristics and details about genotyping for included studies
Supplementary Table 2: Single variant results that reached statistical significance after adjusting for multiple testing.
Supplementary Table 3: Association results ($p<0.01$) from the rare variant WSB test. SNPs were included based on the broad definition (see text).
Supplementary Table 4: Association results ($p<0.01$) from the rare variant WSB test. SNPs were included based on the strict definition (see text).
Supplementary Table 5: Association results ($p<0.01$) from the rare variant SKAT test. SNPs were included based on the broad definition (see text).
Supplementary Table 6: Association results ($p<0.01$) from the rare variant SKAT test. SNPs were included based on the strict definition (see text).