LETTERS
https://doi.org/10.1038/s41588-019-0519-3
nature genetics

Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease

Derek Klarin[1,2,3,4], Emma Busenkell[5], Renae Judy[6,7], Julie Lynch[8,9], Michael Levin[6,7], Jeffery Haessler[10], Krishna Aragam[2,3], Mark Chaffin[3], Mary Haas[3], Sara Lindström[10,11], Themistocles L. Assimes[12,13], Jie Huang[14], Kyung Min Lee[8,15,16], Qing Shao[15], Jennifer E. Huffman[14], Christopher Kabrhel[17,18], Yunfeng Huang[19,20], Yan V. Sun[19,20], Marijana Vujkovic[6,21], Danish Saleheen[6,21], Donald R. Miller[15,16], Peter Reaven[22], Scott DuVall[8,23], William E. Boden[14], Saiju Pyarajan[14,24], Alex P. Reiner[10], David-Alexandre Trégouët[25], Peter Henke[26], Charles Kooperberg[10], J. Michael Gaziano[14,24], John Concato[27,28], Daniel J. Rader[7], Kelly Cho[14,24], Kyong-Mi Chang[6,7], Peter W. F. Wilson[20,29], Nicholas L. Smith[11,30,31], Christopher J. O'Donnell[1,14,32], Philip S. Tsao[12,13], Sekar Kathiresan[2,3,33], Andrea Obi[26], Scott M. Damrauer[6,34,35], Pradeep Natarajan[1,2,3,5,35*], INVENT Consortium and Veterans Affairs' Million Veteran Program[36]

Venous thromboembolism is a significant cause of mortality[1], yet its genetic determinants are incompletely defined. We performed a discovery genome-wide association study in the Million Veteran Program and UK Biobank, with testing of approximately 13 million DNA sequence variants for association with venous thromboembolism (26,066 cases and 624,053 controls) and meta-analyzed both studies, followed by independent replication with up to 17,672 venous thromboembolism cases and 167,295 controls. We identified 22 previously unknown loci, bringing the total number of venous thromboembolism-associated loci to 33, and subsequently fine-mapped these associations. We developed a genome-wide polygenic risk score for venous thromboembolism that identifies 5% of the population at an equivalent incident venous thromboembolism risk to carriers of the established factor V Leiden p.R506Q and prothrombin G20210A mutations. Our data provide mechanistic insights into the genetic epidemiology of venous thromboembolism and suggest a greater overlap among venous and arterial cardiovascular disease than previously thought.

Venous thromboembolism (VTE) is a complex disease impacted by both environmental[1] and genetic determinants[2,3]. The narrow-sense heritability of VTE has been estimated to be approximately 30% (ref. [4]) At the time of analysis, genome-wide association studies (GWAS) revealed only 11 loci reaching genome-wide significance[4–10], leaving a substantial portion of VTE heritability unknown.

Large-scale biobanks linking genetic and diverse phenotypic data in the electronic health record (EHR) are being developed throughout the world[11,12]. Leveraging two large-scale biobanks—the UK Biobank and the Million Veteran Program (MVP)—we sought to: (1) perform a genetic discovery analysis for VTE; (2) evaluate the causal role of blood lipids in VTE; (3) further characterize the role of plasminogen activator inhibitor 1 (PAI-1) in VTE; and (4) develop and evaluate a genome-wide polygenic risk score (PRS) for VTE.

We designed a two-phased VTE discovery GWAS (Fig. 1 and Supplementary Fig. 1). In phase 1, we used MVP v.2.1 data and performed testing for association separately among individuals of European (white), African (black) and Hispanic ancestry and meta-analyzed the results across ancestral groups. In the UK Biobank, association testing was performed in individuals of European ancestry. We combined statistical evidence across the MVP and UK Biobank and set a significance threshold of $P < 5 \times 10^{-8}$ (genome-wide significance); we also required an internal replication $P < 0.01$ in each of the individual MVP and UK Biobank analyses, with concordant directions of effect, to minimize false positives. In phase 2, an additional round of external replication was performed for lead variants using summary data of up to 15,572 VTE cases and 113,430 disease-free controls from the INVENT (International Network on Venous Thrombosis) Consortium[13] combined with 2,100 VTE cases and 53,865 controls from MVP 3.0 data, requiring $P < 0.05$ with consistent direction of effect for successful replication.

In the MVP, the discovery analysis consisted of 11,844 VTE cases (8,929 white, 2,261 black, 654 Hispanic) and 211,753 controls from the MVP v.2.1 data. In the UK Biobank, we identified 14,222 VTE cases and 372,102 controls. The baseline characteristics for both cohorts are presented in Supplementary Tables 1 and 2. VTE cases tended to be older, had a history of smoking, and had a higher body mass index and type 2 diabetes. Following *trans*-ancestry meta-analysis across the MVP and UK Biobank, a total of 2,706 variants at 39 loci met a genome-wide significance threshold, with $P < 0.01$ and concordant effect directions in both datasets (Supplementary Fig. 2–5). The factor V Leiden (*F5*) variant rs6025 (p.R506Q, NC_000001.10:g.169519049T>C), was the top association result (2.5% frequency for the T allele; odds ratio (OR) = 2.53; 95%

A full list of affiliations appears at the end of the paper.

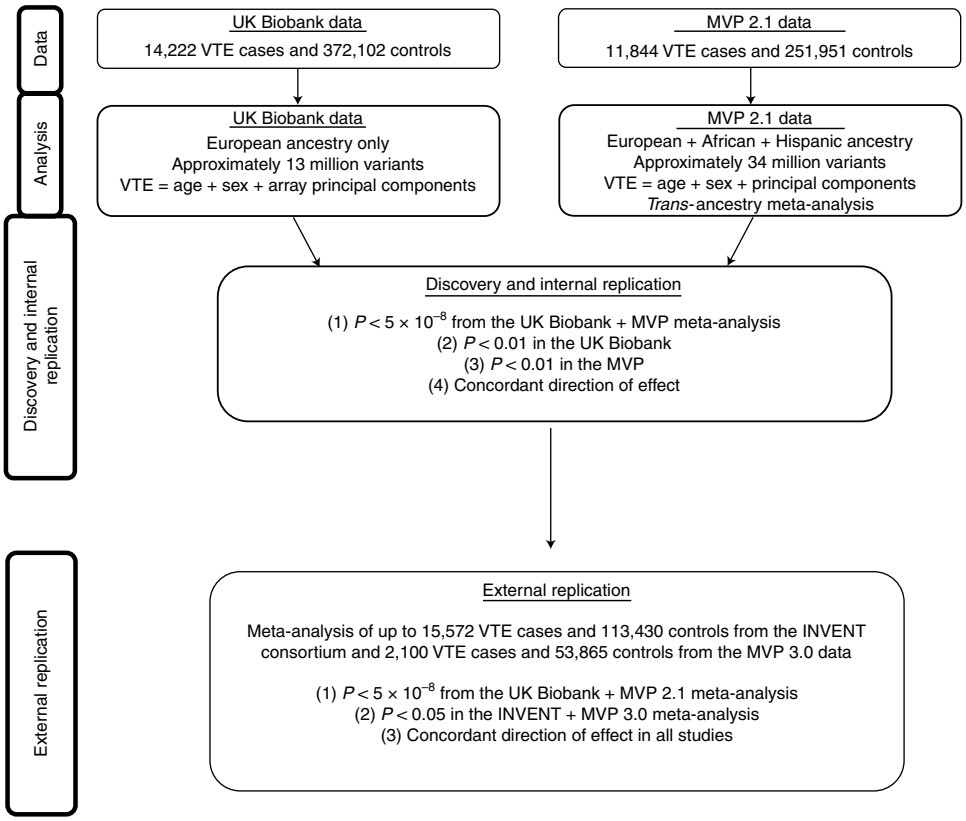


Fig. 1 | Venous thromboembolic disease genetic discovery and replication study design. In the UK Biobank, we performed an association analysis for DNA sequence variants in 14,222 VTE cases and 372,102 controls of European ancestry using logistic regression. These results were combined with association statistics from DNA sequence variants across 3 mutually exclusive ancestry groups in the MVP v.2.1 data representing 11,844 VTE cases and 251,951 controls. Data from the UK Biobank and MVP were meta-analyzed using an inverse variance-weighted fixed effects method. We set a significance threshold of two-sided $P<5\times10^{-8}$ (genome-wide significance) and also required an internal replication two-sided $P<0.01$ in each of the MVP and UK Biobank analyses, with concordant direction of effect, to minimize false positives. We subsequently performed external replication using summary data from the INVENT consortium (up to 15,572 VTE cases and 113,430 controls) meta-analyzed with data from MVP 3.0 (2,100 VTE cases and 53,865 controls), requiring an external replication $P<0.05$ with a consistent direction of effect.

confidence interval (CI) = 2.43–2.64; $P<1.0\times10^{-300}$). We replicated all 11 previously described genome-wide VTE loci and identified 28 candidate, previously unknown VTE loci brought forward for external replication (Supplementary Tables 3 and 4). Of the 28 candidate loci, 22 successfully replicated in an independent set of up to 17,672 VTE cases and 167,295 controls (Supplementary Tables 5 and 6).

One large randomized controlled trial showed that low-density lipoprotein (LDL) cholesterol lowering with a statin versus placebo led to a reduced risk of venous thromboembolic events[14]. We sought to explore the potential causal relationships of blood lipids with VTE development by performing a multivariate Mendelian randomization analysis using a weighted polygenic score of 222 lipid-associated variants from the Global Lipids Genetics Consortium and summary data from the MVP v.2.1 and UK Biobank VTE GWAS restricted to Europeans (Supplementary Table 7) (ref. [15]). We observed that a 1 s.d. of genetically elevated LDL cholesterol was associated with an increased risk of VTE ($OR_{LDL}=1.17$, 95% CI = 1.05–1.29, $P_{LDL}=0.003$). In contrast, both a 1 s.d. of genetically elevated high-density lipoprotein (HDL) cholesterol and a 1 s.d. of genetically elevated triglycerides were not associated with a risk of VTE ($OR_{HDL}=1.01$, 95% CI = 0.91–1.13, $P_{HDL}=0.82$; $OR_{Triglycerides}=0.88$, 95% CI = 0.77–1.00, $P_{Triglycerides}=0.04$) after Bonferroni correction ($P<0.016$ = (0.05/3 lipid fractions). A Mendelian randomization-Egger (MR-Egger) analysis[16] indicated no pleiotropic biases of our lipid genetic instruments (MR-Egger intercept $P>0.05$ for all 3 lipid fractions (Supplementary Table 8 and Fig. 2). The following are provided in the Supplementary Results: (1) a phenome-wide association study; (2) an analysis of how DNA sequence variants differ in their contribution to vascular disease risk in the arterial and venous territories; (3) an examination of VTE risk variant-protein quantitative trait loci (pQTL) associations; (4) and results of a VTE fine-mapping analysis including a 99% credible set of 4 variants at the *ZFPM2* locus, which were genome-wide *trans*-pQTL associations with plasma PAI-1 concentration (Supplementary Table 9).

Given the known role of PAI-1 in venous thrombosis and fibrinolysis in model systems[17], we hypothesized that the *ZFPM2* VTE GWAS and the PAI-1 *trans*-pQTL associations may represent colocalizing signals at the *ZFPM2* locus. We used a recently described colocalization analysis pipeline[18] to compute the colocalization posterior probability (CLPP) for the *ZFPM2* locus. Using European MVP v.2.1 and UK Biobank European VTE meta-analyzed summary statistics, PAI-1 pQTL results in human plasma from the INTERVAL study[19], and reference linkage disequilibrium information of 503 European participants from the 1000 Genomes Project[20] phase 3 whole-genome sequencing data, we calculated a CLPP of 0.203 at this locus. Previous work suggests that a CLPP > 0.01 is indicative of a 'reasonably high' probability of colocalization[18,21]; the LocusCompare plot at this site further indicates that the *ZFPM2* VTE GWAS and PAI-1 pQTL associations probably represent a true colocalization event (Supplementary Fig. 6).

PAI-1 influences thrombosis by directly inhibiting conversion of plasminogen to plasmin and indirectly via disrupting the interaction

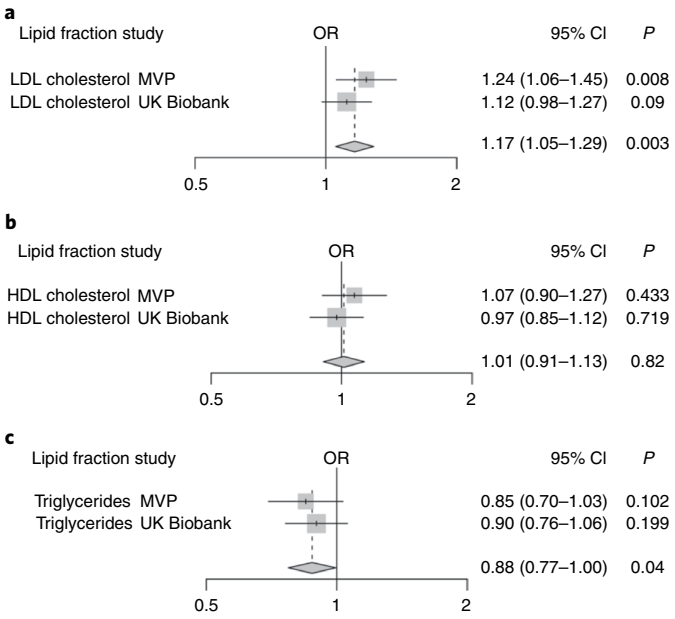


Fig. 2 | Blood lipids and VTE risk. a–c, Association of the 222 variant lipid genetic risk scores with VTE in a multivariable Mendelian randomization analysis. Logistic regression odds ratios (ORs) are displayed per 1 s.d. genetically increased LDL cholesterol (**a**), HDL cholesterol (**b**) and triglycerides (**c**). Wald statistic two-sided values of *P* are displayed. Summary-level lipids data from up to 319,677 participants of the Global Lipids Genetics Consortium[15] and VTE association data from the MVP ($n=8{,}929$ cases, 181,337 controls) and UK Biobank ($n=14{,}222$ cases, 372,102 controls) were used for this analysis. The gray boxes reflect the inverse variance weight for each study.

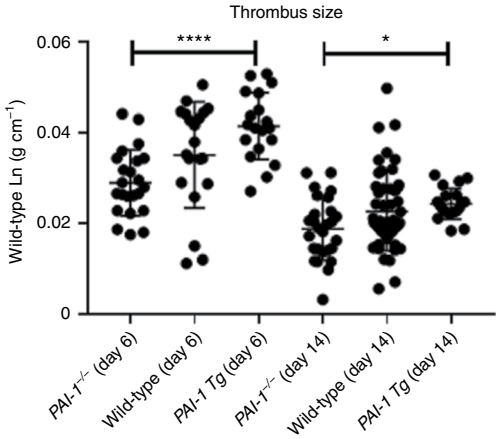


Fig. 3 | Functional assessment of PAI-1 in murine models. The inferior vena cava venous thrombus size was measured at days 6 and 14 after inferior vena cava ligation in *PAI-1* Tg (day 6, $n=19$; day 14, $n=20$), wild-type (day 6, $n=20$; day 14, $n=49$) and *PAI-1*$^{-/-}$ mice (day 6, $n=23$; day 14, $n=27$). Thrombus size was larger in *PAI-1* Tg compared to *PAI-1*$^{-/-}$ mice (one-way ANOVA followed by Tukey's multiple comparisons post-hoc test, $^{*}P=0.02$, $^{****}P<0.0001$). A scatterplot depicting mean thrombus size ± s.d. is shown.

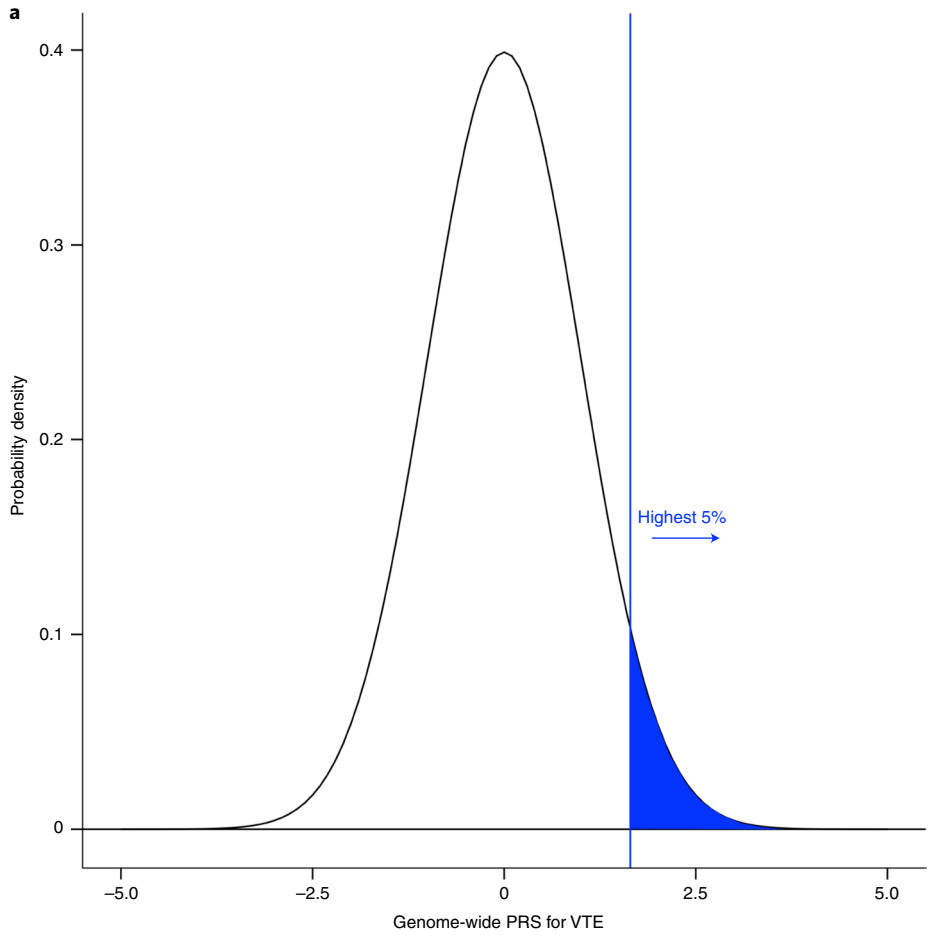
of circulating monocytes with the glycoprotein vitronectin within the thrombus and adjacent vein wall[22]. Monocytes are a key source of factor III (tissue factor) as well as matrix metalloproteinases during thrombus clearance[23,24]. Given the colocalization between PAI-1 concentration and human VTE, we sought to determine the impact of PAI-1 levels on venous thrombus size in an experimental deep vein thrombosis model using transgenic mice. *PAI-1*$^{-/-}$ mice have no circulating active PAI-1, whereas those overexpressing PAI-1 (*PAI-1 Tg*) have levels approximately 137-fold greater than wild-type C57B/L6 mice[25]. Six days following inferior vena cava (IVC) occlusion with generation of thrombus, PAI-1-overexpressing mice had 1.5-fold larger thrombus size compared to *PAI-1*$^{-/-}$ mice, with wild-type mice demonstrating an intermediate phenotype. This difference persisted during late thrombus resolution at day 14 (Fig. 3), demonstrating progressive impairment in thrombus clearance in the setting of increasing PAI-1 protein levels.

Finally, we sought to examine the contribution of polygenic inheritance on VTE risk. Currently, the *F5* (p.R506Q) and *F2* (prothrombin) G20210A mutations, low-frequency variants that confer a two to threefold risk of VTE, are frequently tested in clinical settings to evaluate the role of inherited thrombophilia predisposing to acute thrombotic syndromes. Given the individual associations of common genetic variants with VTE, heritable VTE risk may also be explained by an aggregate of common variant VTE risk alleles[26]. We hypothesized that those at the right tail of the normally distributed VTE PRS (highest 5%) would be at significantly increased VTE risk (Fig. 4a).

We generated a 297-variant VTE PRS using a pruning and thresholding method ($R^2<0.2$, $P<1\times10^{-5}$) from European MVP v.2.1 and UK Biobank European VTE meta-analyzed summary statistics (Supplementary Table 10). Notably, we excluded the linkage disequilibrium blocks ($R^2>0.2$) containing the *F5* p.R506Q and *F2* G20210A variants from the PRS. We first assessed the associated VTE risk for the 5% of individuals with the highest PRS_{VTE} relative to the rest of the population using prevalent data from the MVP v.3.0, a set of 2,100 VTE cases and 53,865 VTE controls entirely independent from the individuals in the MVP discovery GWAS. We observed that the 2,798 individuals in MVP v.3.0 with the 5% highest PRS_{VTE} had 2.89-fold increased risk of VTE relative to the rest of the population ($OR_{PRS}=2.89$, 95% CI = 2.52–3.30, $P_{PRS}=7.2\times10^{-53}$). This effect estimate was similar in magnitude to those observed for *F5* p.R506Q ($OR_{F5}=2.97$, 95% CI = 2.63–3.36, $P_{F5}=3.4\times10^{-67}$) and *F2* G20210A ($OR_{F2}=2.61$, 95% CI = 2.19–3.12, $P_{F2}=5.2\times10^{-27}$) (Fig. 4b). In addition, we observed that this risk was further compounded for individuals among the top 5% with increased polygenic VTE risk who were also *F5* p.R506Q or *F2* G20210A carriers.

We sought replication of our PRS findings using incident VTE data from the prospective Women's Health Initiative (WHI) Hormone Trial. In total, among 10,975 European women prospectively followed for up to 25 years in the WHI Hormone Trial, 690 incident VTE events were identified in participants with genetic data. The demographic and clinical characteristics of the WHI participants in our VTE incident event analysis are shown in Supplementary Table 11. We estimated the risk for carriers of *F5* p.R506Q and *F2* G20210A mutations as well as those among the 5% highest PRS_{VTE} through Cox proportional hazards models. We observed that *F5* p.R506Q carriers were at greater than twofold risk of developing VTE (hazard ratio $(HR)_{F5}=2.34$, 95% CI = 1.86–3.35, $P_{F5}=2.8\times10^{-13}$), and the *F2* G20210A mutation was nominally associated with increased VTE risk ($HR_{F2}=3.35$, 95% CI = 1.10–10.23, $P_{F2}=0.033$). The 549 individuals in the WHI with the 5% highest PRS_{VTE} had a 2.51-fold risk of incident VTE relative to the rest of the population ($HR_{PRS}=2.51$, 95% CI = 1.97–3.19, $P_{PRS}=4.4\times10^{-14}$) as shown in Fig. 5. Much like in the MVP, the risk among the 5% of the population with the highest PRS_{VTE} in the WHI was comparable in effect size to that of large-effect, monogenic mutations in *F5* and *F2*.

These findings allow several conclusions. First, our results lend human genetic support to LDL cholesterol lowering as a preventive strategy for VTE. In the JUPITER (Justification for the Use of statins in Prevention: an Interventional Trial Evaluating Rosuvastatin) trial, administration of 20 mg of rosuvastatin in asymptomatic participants resulted in a reduced occurrence of symptomatic VTE[14].



b

Gene	Variants	Chromosome	Carriers (%)	Cases	Controls	OR	95% CI	*P*
F5	p.R506Q	1	3,048 (5.4%)	2,100	53,865		2.97 (2.63–3.36)	3.4×10^{-67}
F2	G20210A	11	1,581 (2.8%)	2,100	53,865		2.61 (2.19–3.12)	5.2×10^{-27}
Multiple	297-variant PRS	Many	2,798 (5%)	2,100	53,865		2.89 (2.52–3.30)	7.2×10^{-53}
Multiple	G20210A + 297-variant PRS	Many	131 (0.2%)	2,100	53,865		4.90 (3.06–7.85)	3.1×10^{-11}
Multiple	p.R506Q + 297-variant PRS	Many	908 (1.6%)	2,100	53,865		4.62 (3.82–5.59)	3.9×10^{-56}

Fig. 4 | Genome-wide PRS for VTE. a, Distribution of the PRS_{VTE} in the MVP 3.0 dataset ($n = 55{,}965$). The *x* axis represents the PRS with values transformed to have a mean of 0 and s.d. of 1. The region shaded in blue represents those with the highest 5% PRS_{VTE} values. **b**, VTE ORs in the MVP 3.0 data for carriers of the *F5* p.R506Q and *F2* G20210A mutations. In addition, the OR for individuals with the highest 5% PRS_{VTE} compared to individuals in the lower 95% of PRS_{VTE}, as well as carriers of the *F5* p.R506Q and *F2* G20210A mutations within the highest 5% PRS_{VTE}, are depicted. Wald statistic two-sided values of *P* are displayed.

Gene	Variants	Chromosome	Allele frequency (%)	Incident events	Total *n*	HR	95% CI	*P*
F5	p.R506Q	1	2.4	690	10,975		2.34 (1.86–2.95)	2.8×10^{-13}
F2	G20210A	11	0.45	690	10,975		3.35 (1.10–10.23)	0.033
Multiple	297-variant PRS	Many	Highest percentage	690	10,975		2.51 (1.97–3.19)	4.4×10^{-14}

Fig. 5 | Genome-wide PRS and incident VTE events. Hazard ratios calculated from the Cox proportional hazards model for incident VTE events in the WHI study for carriers of the *F5* p.R506Q and *F2* G20210A mutations. The hazard ratio for individuals with the highest 5% PRS_{VTE} compared to individuals in the lower 95% of PRS_{VTE} is also depicted. Two-sided values of *P* are displayed.

This implies that the apparent VTE risk reduction from statins may be due to on-target lowering of lipoproteins, much like the benefits observed for multiple atherosclerotic syndromes[27,28]. Second, partial antagonism of PAI-1 as a preventive treatment for VTE deserves further consideration. In our analysis, we noted colocalizing *ZFPM2* VTE GWAS and PAI-1 pQTL associations and observed that PAI-1-overexpressing mice had 1.5-fold larger thrombus size compared to *PAI-1*$^{-/-}$ mice in an IVC ligation model. These results suggest that

imbalance in the thrombosis-fibrinolysis pendulum in the human condition may lead to development of pathological VTE, whereas lower active PAI-1 levels may allow for resolution of incidental venous thrombosis before becoming clinically relevant. Third, our data provide further evidence for the usefulness of polygenic risk prediction in the clinical realm. In a recent publication by Khera et al.[29], the authors generated expanded PRS and demonstrated that those within the right tail of the distribution had a greater than threefold increased risk of developing the disease, akin to carriers of monogenic mutations. We build on these findings by extending polygenic scoring to incident VTE events, where we observed similar magnitudes of effect for our PRS_{VTE} and *F5* p.R506Q/*F2* G20210A mutations. Our data suggest that extending current thrombophilia genetic panels to include testing for polygenic VTE risk would significantly increase the yield of current genetic testing and may be warranted.

Our study should be interpreted within the context of its limitations. First, our VTE phenotype is based on EHR data and may result in misclassification of case status. However, such misclassification should reduce the statistical power for discovery and bias results toward the null. Second, while our colocalization analysis and murine functional data support the role of PAI-1 in VTE, further research is needed to fully understand the causal variant at the *ZFPM2* locus and its underlying mechanism. Lastly, while those with the highest PRS_{VTE} are at increased risk for VTE, the PRS mechanism of action represents a combination of many causal risk factors, rather than one single pathway that leads to disease. However, assessment of individual risk may help identify a subpopulation that may benefit from thromboprophylaxis during periods of increased risk, for instance, perioperatively[30] or during hospitalizations for acute, medical illness[31].

In conclusion, our data provide mechanistic insights into the genetic epidemiology of VTE and suggest a greater intersection between blood lipids, VTE and arterial vascular disease than previously understood.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of code and data availability are available at https://doi.org/10.1038/s41588-019-0519-3.

Received: 5 March 2019; Accepted: 24 September 2019;
Published online: 1 November 2019

References

1. Heit, J. A. Epidemiology of venous thromboembolism. *Nat. Rev. Cardiol.* **12**, 464–474 (2015).
2. Bertina, R. M. et al. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369**, 64–67 (1994).
3. Poort, S. R., Rosendaal, F. R., Reitsma, P. H. & Bertina, R. M. A common genetic variation in the 3′-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis. *Blood* **88**, 3698–3703 (1996).
4. Klarin, D., Emdin, C. A., Natarajan, P., Conrad, M. F. & Kathiresan, S. Genetic analysis of venous thromboembolism in UK biobank identifies the ZFPM2 locus and implicates obesity as a causal risk factor. *Circ. Cardiovasc. Genet.* **10**, e001643 (2017).
5. 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).
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. 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).
8. Hernandez, W. et al. Novel genetic predictors of venous thromboembolism risk in African Americans. *Blood* **127**, 1923–1929 (2016).
9. 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).
10. Trégouët, D. A. et al. Common susceptibility alleles are unlikely to contribute as strongly as the *FV* and *ABO* loci to VTE risk: results from a GWAS approach. *Blood* **113**, 5298–5303 (2009).
11. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).
12. Gaziano, J. M. et al. Million veteran program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
13. Lindstrom, S.et al. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood* https://doi.org/10.1182/blood.2019000435 (2019).
14. Glynn, R. J. et al. A randomized trial of rosuvastatin in the prevention of venous thromboembolism. *N. Engl. J. Med.* **360**, 1851–1861 (2009).
15. Liu, D. J. et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
16. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
17. Eitzman, D. T., Westrick, R. J., Nabel, E. G. & Ginsburg, D. Plasminogen activator inhibitor-1 and vitronectin promote vascular thrombosis in mice. *Blood* **95**, 577–580 (2000).
18. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).
19. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
20. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
21. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
22. Fogo, A. B. Renal fibrosis: not just PAI-1 in the sky. *J. Clin. Invest.* **112**, 326–328 (2003).
23. Pawlinski, R. & Mackman, N. Cellular sources of tissue factor in endotoxemia and sepsis. *Thromb. Res.* **125** (Suppl. 1), S70–S73 (2010).
24. Henke, P. K. et al. Deep vein thrombosis resolution is modulated by monocyte CXCR2-mediated activity in a mouse model. *Arterioscler. Thromb. Vasc. Biol.* **24**, 1130–1137 (2004).
25. Obi, A. T. et al. Plasminogen activator-1 overexpression decreases experimental postthrombotic vein wall fibrosis by a non-vitronectin-dependent mechanism. *J. Thromb. Haemost.* **12**, 1353–1363 (2014).
26. Wassel, C. L. et al. A genetic risk score comprising known venous thromboembolism loci is associated with chronic venous disease in a multi-ethnic cohort. *Thromb. Res.* **136**, 966–973 (2015).
27. Ridker, P. M. et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N. Engl. J. Med.* **359**, 2195–2207 (2008).
28. Mihaylova, B. et al. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *Lancet* **380**, 581–590 (2012).
29. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
30. Bahl, V. et al. A validation study of a retrospective venous thromboembolism risk scoring method. *Ann. Surg.* **251**, 344–350 (2010).
31. Anderson, F. A. Jr. & Spencer, F. A. Risk factors for venous thromboembolism. *Circulation* **107**, I9–I16 (2003).

[1]Veterans Affairs Boston Healthcare System, Boston, MA, USA. [2]Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Division of Vascular Surgery and Endovascular Therapy, University of Florida School of Medicine, Gainesville, FL, USA. [5]Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [6]Corporal Michael Crescenz Veterans Affairs Medical Center, Philadelphia, PA, USA. [7]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [8]Veterans Affairs Informatics and Computing Infrastructure, Veterans Affairs Salt Lake City Health Care System, Salt Lake City, UT, USA. [9]University of Massachusetts College of Nursing & Health Sciences, Boston, MA, USA. [10]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [11]Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA. [12]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA. [13]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [14]Massachusetts Veterans Epidemiology Research and Information Center, Veterans Affairs Boston Healthcare System, Boston, MA, USA. [15]Center for Healthcare Organization and Implementation Research, Edith Nourse Rogers Memorial Veterans Hospital, Bedford, MA, USA. [16]Boston University School of Public Health, Department of Health Law, Policy & Management, Boston, MA, USA. [17]Center for Vascular Emergencies, Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, USA. [18]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [19]Department of Epidemiology, Emory University Rollins School of Public Health, Department of Biomedical Informatics Emory University School of Medicine, Atlanta, GA, USA. [20]Atlanta Veterans Affairs Health Care System, Decatur, GA, USA. [21]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, Philadelphia, PA, USA. [22]Phoenix Veterans Affairs Health Care System, Phoenix, AZ, USA. [23]Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA. [24]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [25]Bordeaux Population Health Research Center (INSERM UMR S 1219), University of Bordeaux, Bordeaux, France. [26]Section of Vascular Surgery, Department of Surgery, University of Michigan, Ann Arbor, MI, USA. [27]Clinical Epidemiology Research Center, Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA. [28]Department of Medicine, Yale University School of Medicine, New Haven, CT, USA. [29]Emory Clinical Cardiovascular Research Institute, Atlanta, GA, USA. [30]Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, WA, USA. [31]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA. [32]Cardiovascular Medicine Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [33]Verve Therapeutics, Cambridge, MA, USA. [34]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [35]These authors contributed equally: Scott M. Damrauer, Pradeep Natarajan. [36]A full list of authors appears in the Supplementary Note. *e-mail: pnatarajan@mgh.harvard.edu

Methods

Study populations. We conducted genetic association analyses using DNA samples and phenotypic data from two cohorts: the MVP and UK Biobank. In the MVP, individuals aged 19 to over 100 years were recruited from 63 Veterans Affairs Medical Centers across the USA. In our initial MVP analysis, we evaluated 11,844 VTE cases (8,929 white, 2,261 black, 654 Hispanic) and 211,753 VTE-free controls.

In the UK Biobank, individuals aged 45–69 years old were recruited from across the UK for participation. In this study, we identified 14,222 VTE cases and 372,102 controls of European ancestry. Further details of cohort descriptions and disease definitions are described in the Supplementary Note. All studies received ethical and study protocol approval by their appropriate institutional review boards and informed consent was obtained from all participants. Additional information regarding the experimental design and participants is provided in the Nature Research Reporting Summary.

In addition, we examined incident VTE data from the WHI randomized clinical trial of hormone therapy for our PRS analysis. The overall design of the WHI study has been described previously[32]. In brief, at the inception of the WHI study (1993–1998), 161,808 postmenopausal women between the ages of 50 and 79 years were eligible for inclusion in multiple clinical trials. Exclusion criteria related to the presence of medical conditions predisposing to shortened survival or safety concerns. The protocol and consent forms were approved by institutional review committees and all participants provided written informed consent. The WHI Hormone Trial initially comprised 27,347 postmenopausal women who were randomized to receive either estrogen plus progestin or estrogen alone versus placebo until the trials were stopped early in July 2002 and March 2004, respectively. All WHI Hormone Trial participants subsequently continued to be followed without intervention until closeout. Of the various components of the WHI, VTE was adjudicated by physician adjudicators for participants who were enrolled in the hormone therapy trials.

Genetic data and quality control for association analysis. DNA extracted from whole blood was genotyped in the MVP using a customized Affymetrix Axiom biobank array, the MVP 1.0 Genotyping Array. Veterans (US military personnel) of three mutually exclusive ancestry groups were identified for analysis: (1) non-Hispanic white population (European ancestry); (2) non-Hispanic black population (African ancestry); and (3) self-identified Hispanic population. After pre-phasing using EAGLE[33] v.2, genotypes from the 1000 Genomes Project[20] phase 3, v.5 reference panel were imputed into MVP participants using the Minimac3 software[34]. Ancestry-specific principal component analysis was performed using the EIGENSOFT v.6 software[35]. Additional details of the quality control procedures used to assign ancestry and perform genotype imputation are described in the Supplementary Note.

In the MVP, sample and variant quality control was performed as described previously[36]. In brief, duplicate samples, samples with more heterozygosity than expected, an excess (>2.5%) of missing genotype calls or discordance between genetically inferred sex and phenotypic sex were excluded. In addition, one individual from each pair of related individuals (kinship >0.0884 as measured by the KING v.2.0 software[37]) were removed. In total, we identified 312,571 multi-ancestry participants passing quality control from the MVP v.2.1 data (used in the association analysis) and another 69,578 from the MVP v.3.0 data used for the PRS analysis.

Following imputation, variant-level quality control was performed using the EasyQC v.9.2 R package[38] and exclusion metrics included: ancestry-specific Hardy–Weinberg equilibrium $P < 1 \times 10^{-20}$; posterior call probability < 0.9; imputation quality < 0.3; minor allele frequency (MAF) < 0.0003; call rate < 97.5% for common variants (MAF > 1%); and call rate <99% for rare variants (MAF < 1%). Variants were also excluded if they deviated >10% from their expected allele frequency based on reference data from the 1000 Genomes Project[20]. Following variant-level quality control, we obtained 19.9 million, 31.9 million and 28.1 million DNA sequence variants for analysis in white, black and Hispanic participants, respectively.

In the UK Biobank, analysis was performed separately in white individuals after genotyping using either the UK Biobank Lung Exome Variant Evaluation or UK Biobank Axiom Arrays. Approximately 500,000 individuals were genotyped and subsequently imputed to the Haplotype Reference Consortium and UK10K reference panels (UK Biobank v.3 imputation release). Details of these procedures are described elsewhere[39]. We performed genome-wide association testing for VTE in the UK Biobank using all variants in the v.3 release with a MAF > 0.3% and an imputation quality INFO > 0.4. To avoid potential population stratification, only European ancestry samples were included in the analysis. This subset was selected based on self-reported white ancestry that was subsequently confirmed using genetic principal component analysis. Outliers within the self-reported white ancestry samples in the first six principal components of ancestry were detected and subsequently removed using the R package *aberrant* v.1.0 (ref. [40]). In addition, individuals with sex chromosome aneuploidy (neither XX or XY), discordant self-reported and genetic sex or excessive heterozygosity or missingness, as defined centrally by the UK Biobank, were removed. Finally, one individual from each pair of second-degree or closer relatives (kinship > 0.0884) was removed, selectively retaining VTE cases when possible.

VTE discovery association analysis. In the MVP, genotyped and imputed DNA sequence variants were tested for association with VTE through logistic regression adjusting for age, sex and five principal components of ancestry assuming an additive model using the SNPTEST v.2.5.4 (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) statistical software program. In our discovery analysis, we performed association analyses using MVP v.2.1 data separately for each ancestral group (white, black and Hispanic populations) and then meta-analyses using an inverse variance-weighted fixed effects method implemented in the METAL software (release 2011) program[41]. We excluded variants with a high level of heterogeneity (I^2 statistic > 75%) across the three ancestries. In the UK Biobank, association testing was performed using a logistic regression model adjusted for age at baseline, sex, genotyping array and the first five principal components of ancestry. All testing was performed in PLINK 2.00 alpha (https://www.cog-genomics.org/plink/2.0/).

We combined results across the MVP v.2.1 and UK Biobank cohorts using inverse variance-weighted fixed effects meta-analysis and set a significance threshold of $P < 5 \times 10^{-8}$ (genome-wide significance). In addition, we also required an internal replication $P < 0.01$ in each of the MVP and UK Biobank analyses (for example, MVP discovery and subsequent UK Biobank replication, and vice versa), with concordant direction of effect, to minimize false positives. Previously unknown loci were defined as being greater than 500,000 base pairs away from a known VTE genome-wide associated lead variant. Additionally, European linkage disequilibrium information from the 1000 Genomes Project[20] was used to determine independent variants where a locus extended beyond 500,000 base pairs. All logistic regression *P* values were two-sided. For the X chromosome analyses, male genotypes were coded as if they were homozygous diploid for the observed allele.

Replication. In phase 2, an additional round of external replication was performed for lead variants using summary data of up to 15,572 VTE cases and 113,430 disease-free controls from the INVENT Consortium's current VTE meta-analysis[13] combined with 2,100 VTE cases and 53,865 controls from the MVP v.3.0 data. Of note, the UK Biobank data were excluded from the summary statistics provided by INVENT. We defined significant previously unknown associations as those that were at least nominally significant in replication ($P < 0.05$) with consistent direction of effect and had an overall $P < 5 \times 10^{-8}$ (genome-wide significance) in the discovery and replication cohorts combined.

VTE disease definitions. From the 312,571 multi-ancestry participants in MVP v.2.1 and 69,578 European participants in MVP v.3.0, individuals were defined as having VTE based on possessing at least two of the International Classification of Diseases 9/10 codes outlined in Supplementary Table 12 in their EHR. Individuals were defined as controls if they did not meet the definition of a VTE case and their EHR reflected two or more separate encounters in the Veterans Affairs Healthcare System in each of the 2 years before enrollment in the MVP. In the UK Biobank, individuals were defined as having VTE based on the definition by Klarin et al.[4] as described previously. All other individuals were defined as controls.

Lipids and VTE Mendelian randomization analysis. Summary-level data for 222 genome-wide lipid-associated variants were obtained from publicly available data from the Global Lipids Genetics Consortium[15] using a previously described genetic risk score instrument[36]. As described previously, cohorts either excluded participants on statins or adjusted total cholesterol and LDL cholesterol (by dividing by 0.8 or 0.7, respectively) if a statin was prescribed. One variant, rs77375493, was excluded from the current analysis after not passing quality control. We then used the results from the MVP and UK Biobank GWAS meta-analysis restricted to Europeans. The effect alleles were matched with all lipid and VTE summary data and three different Mendelian randomization analyses were performed: (1) inverse variance-weighted; (2) multivariable; (3) MR-Egger to account for pleiotropic bias. First, we performed inverse variance-weighted Mendelian randomization using each set of variants for each lipid trait as the instrumental variables. However, this method does not account for possible pleiotropic bias. Therefore, we next performed inverse variance-weighted multivariable Mendelian randomization. This method adjusts for possible pleiotropic effects across the included lipid traits in our analyses, using effect estimates from the variant-VTE outcome and effect estimates from variant-LDL cholesterol, variant-HDL cholesterol and variant-triglycerides as predictors in one multivariable model. We additionally performed MR-Egger analysis as described previously[16]. This technique can be used to detect bias secondary to unbalanced pleiotropy in Mendelian randomization studies. In contrast to inverse variance-weighted analysis, the regression line is unconstrained and the intercept represents the average pleiotropic effects across all variants. Bonferroni-corrected two-sided *P* values ($P = 0.016$; 0.05/3) for 3 tests were used to declare statistical significance. Analysis was performed using R v.3.2.1.

Colocalization of *ZFPM2* VTE GWAS and PAI-1 plasma pQTL signals. To evaluate whether there was evidence of colocalization across the VTE GWAS and PAI-1 pQTL studies, we used the European MVP v.2.1 and European UK Biobank VTE meta-analyzed summary statistics and PAI-1 pQTL results from the

INTERVAL study[19]. For the 2,178 variants within the 1-Mb region surrounding the *ZFPM2* lead VTE GWAS variant, we performed a locus-wide colocalization analysis using FINEMAP v.1.0 (ref. [42]) to generate posterior causal probabilities for each of these variants in the GWAS and pQTL analyses. We used the European superpopulation subset of the 1000 Genomes Project[20] phase 3 whole-genome sequence data as a reference for the linkage disequilibrium statistics and assumed only 1 causal variant at the locus. We then analyzed these posterior probabilities with a publicly available pipeline[18] to compute the CLPP for the entire locus as described previously[21]. The R package LocusCompareR v.1.0 was used to visualize the colocalizing signals.

Functional assessment of PAI-1 in murine models. Male C57BL/6 (wild-type) mice (The Jackson Laboratory), *PAI-1*$^{-/-}$ (backcrossed 5–10 generations on a C57BL/6 background) and PAI-1-overexpressing mice (*PAI-1 Tg*, backcrossed 5–10 generations on a C57BL/6 background) were used in this study[25,43,44]. Previous data comparing homozygous littermates to wild-type C57BL/6 controls demonstrated an identical phenotype with regard to venous thrombosis (size and cellular composition)[25,44]. Therefore, in the interest of humane and responsible animal use, wild-type C57BL/6 mice were used as controls. Animals underwent a well-characterized deep vein thrombosis model, stasis IVC thrombosis, at 8–10 weeks of age and with a body weight of 20–25 g (refs. [24,25,45–47]). Isoflurane 2% was administered as inhaled anesthetic. A midline laparotomy was performed, the retroperitoneum exposed and the dorsal IVC branches were interrupted with electrocautery. The infrarenal IVC and any accompanying side branches caudal to the left renal vein were ligated with 7-0 Prolene (Ethicon) to generate blood stasis. A running continuous 5-0 Vicryl suture was used to close the fascia and Vetbond Tissue Adhesive was applied for skin closure (3M Animal Care Products). Mice were euthanized at 6 and 14 d post-thrombosis. The IVC and its associated thrombus were weighed (g) and measured (cm) for weight-to-length analysis of thrombus size[24,48,49]. Prism 6.0 (GraphPad Software) was used to analyze thrombus size. Data are presented as the mean ± s.d. Statistical significance among multiple groups was determined using one-way analysis of variance (ANOVA) followed by Tukey's multiple comparisons post-hoc test. A value of $P < 0.05$ was considered significant. All work was approved by the University of Michigan, University Committee on Use and Care of Animals and was performed in compliance with the Guide for the Care and Use of Laboratory Animals published by the National Institutes of Health.

VTE PRS generation. A PRS represents an individual's risk of a given disease conferred by the cumulative impact of many common DNA sequence variants. A weight is assigned to each genetic variant based on its strength of association with disease risk (β). Individuals are then additively scored in a weighted fashion based on the number of risk alleles they carry for each variant in the PRS.

To generate our score, we used summary statistics from the combined MVP v.2.1 and UK Biobank VTE summary statistics restricted to Europeans (23,151 VTE cases, 553,439 controls) and a linkage disequilibrium panel of 20,000 randomly selected European ancestry samples from the UK Biobank. We restricted variants to those present in both MVP v.2.1 and UK Biobank VTE summary statistics with a consistent direction of effect. To increase the number of independent variants included in our score, we performed a pruning and thresholding analysis using the linkage disequilibrium-driven clumping procedure in PLINK v.1.90b (-clump). In brief, this algorithm formed 'clumps' around variants with a VTE association $P < 1 \times 10^{-5}$ and with an $R^2 > 0.2$ based on the linkage disequilibrium reference. From our initial set of summary statistics, the algorithm selects only one associated variant from each clump below our prespecified *P* value threshold. The final output from this procedure generated a score of 299 independent ($R^2 < 0.2$), VTE-associated ($P < 1 \times 10^{-5}$) variants, representing the strongest disease-associated variant for each linkage disequilibrium-based clump across the genome. From this 299-variant PRS, we then removed the clumps containing the *F5* p.R506Q and *F2* G20210A variants, resulting in a 297-variant PRS_{VTE} for downstream analysis.

VTE PRS analysis. From the 69,578 MVP v.3.0 participants (none of whom were included in the VTE discovery analysis), we identified 2,100 prevalent VTE cases and 53,865 controls. We first assessed the associated VTE risk for the 5% of individuals with the highest PRS_{VTE} relative to the rest of the population using logistic regression adjusting for age, sex and 5 principal components of ancestry. We then tested the association of the *F5* p.R506Q and *F2* G20210A variants among the 5% of individuals with the highest PRS_{VTE} relative to the rest of the population in the MVP v.3.0 data using an identical logistic regression model.

We replicated our findings using incident VTE data from the WHI. Data used in this analysis included genetic data from WHI Hormone Trial participants derived from three separate GWAS substudies: (1) the WHI Genomics and Randomized Trials Network study (WHI-GARNET, 457 incident VTE events among 4,233 participants); (2) the WHI Memory Study (WHIMS, 180 incident VTE events among 5,637 participants); and (3) the WHI Long Life Study (WHILLS, 53 incident VTE events among 1,105 participants). All individuals included in the incident event analysis were of European ancestry. Specific details of each WHI substudy including genotyping, study design and imputation are included in the Supplementary Note. Cox proportional hazards models were used to estimate HRs and 95% CIs for the associations of the *F5* p.R506Q and *F2* G20210A mutations with VTE adjusting for age, 10 principal components of ancestry and hormone therapy intervention status during the active phase of the WHI Hormone Trial. We then tested the associated VTE risk for the 5% of individuals with the highest PRS_{VTE} relative to the rest of the population using Cox proportional hazards models adjusting for age, 10 principal components of ancestry and hormone therapy intervention status during the active phase of the WHI Hormone Trial. Results from the WHIMS, WHILLS and WHI-GARNET were combined using an inverse variance-weighted fixed effects meta-analysis. Bonferroni-corrected two-sided *P* values ($P = 0.016$; 0.05/2 variants + 1 PRS_{VTE}) for 3 tests were used to declare statistical significance. Analyses were performed using R v.3.2.1.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The full summary-level association data from the MVP *trans*-ancestry VTE meta-analysis from this report are available on request through dbGAP, accession code no. phs001672.v2.p1. Data contributed by the CARDIoGRAMplusC4D investigators are available at http://www.CARDIOGRAMPLUSC4D.org/. Data on large artery stroke have been contributed by the MEGASTROKE investigators and are available at http://www.megastroke.org/. The genetic and phenotypic UK Biobank data are available on application to the UK Biobank.

References

32. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. *Control. Clin. Trials* **19**, 61–109 (1998).
33. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
34. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
35. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
36. Klarin, D. et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
37. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
38. Winkler, T. W. et al. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
39. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
40. Bellenguez, C., Strange, A., Freeman, C., Donnelly, P. & Spencer, C. C. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134–135 (2012).
41. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
42. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
43. Eitzman, D. T. et al. Bleomycin-induced pulmonary fibrosis in transgenic mice that either lack or overexpress the murine plasminogen activator inhibitor-1 gene. *J. Clin. Invest.* **97**, 232–237 (1996).
44. Baldwin, J. F. et al. The role of urokinase plasminogen activator and plasmin activator inhibitor-1 on vein wall remodeling in experimental deep vein thrombosis. *J. Vasc. Surg.* **56**, 1089–1097 (2012).
45. Wojcik, B. M. et al. Interleukin-6: a potential target for post-thrombotic syndrome. *Ann. Vasc. Surg.* **25**, 229–239 (2011).
46. Diaz, J. A. et al. Critical review of mouse models of venous thrombosis. *Arterioscler. Thromb. Vasc. Biol.* **32**, 556–562 (2012).
47. Obi, A. T. et al. Endotoxaemia-augmented murine venous thrombosis is dependent on TLR-4 and ICAM-1, and potentiated by neutropenia. *Thromb. Haemost.* **117**, 339–348 (2017).
48. Henke, P. K. et al. Targeted deletion of CCR2 impairs deep vein thrombosis resolution in a mouse model. *J. Immunol.* **177**, 3388–3397 (2006).
49. Laser, A. et al. Deletion of cysteine-cysteine receptor 7 promotes fibrotic injury in experimental post-thrombotic vein wall remodeling. *Arterioscler. Thromb. Vasc. Biol.* **34**, 377–385 (2014).

Acknowledgements

Funding was received from the Department of Veterans Affairs Office of Research and Development, Million Veteran Program (grant no. MVP000). This publication does not represent the views of the Department of Veterans Affairs or the US Government. This research was also supported by three additional Department of Veterans Affairs awards (no. I01-01BX03340 to K.C. and P.W.; no I01-BX003362 to P.T. and K.M.C.; and

no. I01-CX001025 to P.W.) and used the resources and facilities at the VA Informatics and Computing Infrastructure (no. VA HSR RES 13-457). S.M.D. is supported by the Veterans Administration (no. IK2-CX001780). S.K. is supported by a Research Scholar award from the Massachusetts General Hospital, the Donovan Family Foundation and the National Institutes of Health (NIH) (no. R01HL127564). P.N. is supported by the NIH/National Heart, Lung, and Blood Institute (NHLBI) (nos. K08HL140203 and R01HL142711). D.T. was financially supported by the EPIDEMIOM-VTE Senior Chair from the Initiative of Excellence of the University of Bordeaux. C.K. is supported by the NIH (grant no. HL116854). Data on coronary artery disease have been contributed by the CARDIoGRAMplusC4D investigators. Data on large artery stroke have been contributed by the MEGASTROKE investigators. The MEGASTROKE project received funding from sources specified at http://www.megastroke.org/acknowledgements.html. The WHI program is funded by the NHLBI, NIH and the US Department of Health and Human Services (contract nos. HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C and HHSN268201600004C). For a list of all the investigators who have contributed to WHI science, see https://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf. This research has been conducted using the UK Biobank resource, application no. 7089.

Author contributions

D.K., E.B., R.J., J.L., M.L., J. Haessler, K.A., M.C., M.H., T.L.A., J. Huang, S.P., A.P.R., P.H., C. Kooperberg, J.M.G., J.C., D.J.R., K.C., K.-M.C., P.W.F.W., C.J.O., P.S.T., S.K., A.O., S.M.D. and P.N. conceived and designed the study. D.K., E.B., R.J., J.L., M.L., J. Haessler, K.A., M.C., M.H., S.L., T.L.A., J. Huang, K.M.L., Q.S., J.E.H., C. Kabrhel, Y.H., Y.V.S., M.V., D.S., D.R.M., P.R., S.D., W.E.B., S.P., A.P.R., D.A.T., P.H., C. Kooperberg, J.M.G., J.C., D.J.R., K.C., K.-M.C., P.W.F.W., N.L.S., C.J.O., P.S.T., S.K., A.O., S.M.D. and P.N. acquired, analyzed or interpreted the data. D.K., S.K., S.M.D. and P.N. drafted the manuscript. D.K., E.B., R.J., J.L., M.L., J. Haessler, K.A., M.C., M.H., S.L., T.L.A., J. Huang, K.M.L., Q.S., J.E.H., C. Kabrhel, Y.H., Y.V.S., M.V., D.S., D.R.M., P.R., S.D., W.E.B., S.P., A.P.R., D.A.T., P.H., C. Kooperberg, J.M.G., J.C., D.J.R., K.C., K.-M.C., P.W.F.W., N.L.S., C.J.O., P.S.T., S.K., A.O., S.M.D. and P.N. critically revised the manuscript for key intellectual content. D.K., Y.V.S, J.C., J.M.G., D.J.R., K.C., K.-M.C., C.J.O., P.W.F.W., S.K., P.S.T., S.M.D. and P.N. provided administrative, technical or material support.

Competing interests

P.N. reports grant support from Amgen, Apple and Boston Scientific, and consulting income from Apple, all unrelated to the submitted work. S.K. reports grant support from Regeneron Pharmaceuticals and Bayer, grant support and personal fees from Aegerion Pharmaceuticals, personal fees from the Regeneron Genetics Center, Merck, Celera Corporation, Novartis, Bristol-Myers Squibb, Sanofi, AstraZeneca, Alnylam Pharmaceuticals, Eli Lilly and Company and Leerink Partners, personal fees and other support from Catabasis and other support from San Therapeutics outside the submitted work. He is also the chair of the scientific advisory board at Genomics plc and the Chief Executive Officer of Verve Therapeutics. S.D. reports grants to his institution in the last 3 years outside the submitted work: AbbVie Inc.; Anolinx LLC; Astellas Pharma Inc.; AstraZeneca Pharmaceuticals LP; Boehringer Ingelheim International GmbH; Celgene Corporation; Eli Lilly and Company; Genentech, Inc.; Genomic Health, Inc.; Gilead Sciences, Inc.; GlaxoSmithKline; Innocrin Pharmaceuticals Inc.; Janssen Pharmaceuticals; Kantar Health; Myriad Genetic Laboratories, Inc.; Novartis International AG; and Parexel International Corporation. C.K. reports grants to his institution from Janssen Pharmaceuticals, Diagnostica Stago and Siemens Healthcare Diagnostics for research related to VTE, but not related to the current work. J.C. is now with the US Food and Drug Administration.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0519-3.

Correspondence and requests for materials should be addressed to P.N.

nature research

Corresponding author(s): Pradeep Natarajan

Last updated by author(s): Aug 21, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Phenotypic data was collected from the electronic health record and genetic data using the Million Veteran Program (MVP) Axiom array.
Data analysis	Imputation was performed using MiniMac3/EAGLE v2, and data was collected and cleaned using the EasyQC package, SNPTESTv2.5.4, EIGENSOFT v6, METAL (released 2011), KING 2.0, ADMIXTUREv1.3, GraphPad Prism v6.0, and PLINK2 software programs as outlined in the online methods. Code for the colocalization analysis is available at: https://bitbucket.org/mgloud/production_coloc_pipeline. For the other software programs, clear code for analysis is available at their associated website (see text). Additional analyses were performed in R-3.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The full summary level association data from the MVP trans-ancestry PAD meta-analysis from this report are available through dbGAP, accession code phs001672.v2.p1. Data contributed by CARDIoGRAMplusC4D investigators are available online (http://www.CARDIOGRAMPLUSC4D.org/). Data on large artery stroke have been contributed by the MEGASTROKE investigators and are available online (http://www.megastroke.org/). The genetic and phenotypic UK Biobank data are available upon application to the UK Biobank.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	disclose on these points even when the disclosure is negative. Sample size All samples available of three ancestries (European, African, Hispanic) were used for analysis (after quality control, see Supplementary Table 1 for full details). Sample size was determined based on using all genetic data available from MVP/UK Biobank. Participants were excluded if they failed to meet case or control definitions.
Data exclusions	Data were excluded if they did not pass our pre-established quality control metrics, or if they did not fall within the three main ancestries used for analysis.
Replication	In Phase 2 of the discovery GWAS, an additional round of external replication was performed for lead variants using summary data of up to 15,572 VTE cases and 113,430 disease-free controls from the INVENT consortium combined with 2,100 VTE cases and 53,865 controls from MVP 3.0 data, requiring $P < 0.05$ with consistent direction of effect for successful replication. The majority of replication efforts were successful, a few were unsuccessful likely secondary to power considerations.
Randomization	Randomization is not applicable, as this is a population based case-control analysis of prevalent data.
Blinding	Blinding is not applicable, as this is a population based case-control analysis of prevalent data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Male C57BL/6 (WT) mice (Jackson Laboratory, Farmington, CT), PAI-1-/- (backcrossed >10 generations on C57BL/6 mice) and PAI-1 over-expressing mice (PAI-1 Tg, backcrossed > 10 generations on C57BL/6 background) were utilized in this study. Previous data comparing homozygous littermates to wild type C57BL/6 controls demonstrated identical phenotype with regards to venous thrombosis with regards to size and cellular composition. Therefore, in the interest of humane and responsible animal use, wild type C57BL/6 mice (WT) were utilized as controls. Animals underwent a well-characterized DVT model, stasis inferior vena cava (IVC) thrombosis, at 8-10 weeks of age and 20-25 grams body weight.
Wild animals	This study did not include wild animals
Field-collected samples	This study did not include field-collected animals
Ethics oversight	All work was approved by the University of Michigan, University Committee on Use and Care of Animals and was performed in compliance with the Guide for the Care and Use of Laboratory Animals published by the US National Institutes of Health. The WHI-HT trial was approved by the local Institutional Review Board at the Fred Hutchinson Cancer Research Center.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Demographics and participant counts for the European, African, and Hispanic MVP participants and UK Biobank participants that passed our quality control and were included in the analysis are depicted in Supplementary Table 1-2. In MVP, European - Mean Age VTE cases = 71.0 years, 95.0% male, Mean Age VTE controls = 68.0 years, 92.7% male; African - Mean Age VTE cases = 66.1 years, 91.6% male, Mean Age VTE controls = 61.5 years, 86.5% male; Hispanic - Mean Age VTE cases = 66.8 years, 93.7% male, Mean Age VTE controls = 60.5 years, 90.8% male).
Recruitment	Individuals aged 19 to 104 years have been recruited voluntarily from more than 50 VA Medical Centers nationwide for participation in the Million Veteran Program biobank study. Recruitment is currently occurring in person at selected sites in the VHA health care system. Every Veteran is assigned a study ID number, which is used to track them throughout the entire process of recruitment, enrollment, sample collection and use; this approach also provides a level of protection for personal identifiers from the outset. Given that study enrollment is voluntary, biases of this study are similar to those of any mega-biobank with voluntary enrollment, including survivorship bias. A complete description of the entire MVP Biobank study including recruitment can be found at PMID: 26441289.
Ethics oversight	All studies received ethical and study protocol approval by their appropriate Institutional Review Boards and informed consent was obtained from all participants (Partners Healthcare IRB, VA Central IRB, UK Research Ethics Committee).

Note that full information on the approval of the study protocol must also be provided in the manuscript.