Genetic determinants of venous thrombosis

Haan, H.G. de

Genetic determinants of venous thrombosis

Hugoline de Haan

Colofon

Genetic determinants of venous thrombosis

Layout and design: Eduard Boxem | www.persoonlijkproefschrift.nl
Printing: Ridderprint BV | www.ridderprint.nl

Genetic determinants of venous thrombosis

proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 8 januari 2020
klokke 11:15 uur

door

Hugoline Georgette de Haan
Geboren te De Bilt,
in 1987

Promotor

Prof. dr F.R. Rosendaal

Co-promotor

dr A. van Hylckama Vlieg

Leden promotiecommissie

Prof. dr S.C. Cannegieter
Prof. dr C. Wijmenga (Rijksuniversiteit Groningen)
Prof. dr P.E. Slagboom
Prof. dr F.W.G. Leebeek (Erasmus Medisch Centrum)

The work described in this thesis was performed at the department of Clinical Epidemiology of the Leiden University Medical Center, the Netherlands. Research described in this thesis was supported by a grant of the Rembrandt Institute for Cardiovascular Science.

TABLE OF CONTENTS

CHAPTER 1

General introduction and outline

GENERAL INTRODUCTION

Venous thrombosis (VT), the occlusion of the venous system by a blood clot, is a multicausal disorder affecting 1-2 per 1000 individuals annually.[1,2] The most common manifestations are deep vein thrombosis (DVT) of the lower extremities and pulmonary embolism (PE). Mortality and morbidity after a thrombotic event are considerable: PE has a case-fatality rate of about 10% within the first month,[1-4] whereas 20 to 60% of the DVT patients develop the post-thrombotic syndrome.[5-7] In addition, VT recurs in 20 to 30% of the patients within five years of the first event.[5,8] The risk of VT and its complications is not equal for all individuals. For example, the incidence of first events increases exponentially with age and men have approximately a twofold higher risk of recurrence than women.[9-11] The established risk factors for VT are often present concurrently and include recent immobilization, surgery, cancer, pregnancy or postpartum period, and hormone use (see Table for a short overview).[12,13] Most, if not all, risk factors relate to hypercoagulability, vascular endothelial injury, or stasis, also known as Virchow's triad, and trigger a shift in the hemostatic balance towards clotting.

Table. Main risk factors for venous thrombosis

Factor	Relation with venous thrombosis*
Increasing age	Weak to strong
Male sex	Weak
Genetic factors	Weak to strong
Active cancer	Strong
Surgery, trauma, immobilization	Strong
Long-haul (air) travel	Moderately strong
Oral contraceptive use	Moderately strong
Hormone replacement therapy	Moderately strong
Pregnancy and postpartum period	Moderately strong
Overweight or obesity	Moderately strong

*Strong denotes a relative risk >5; moderately strong: relative risk 2-5; weak: relative risk <2
For an extensive review on risk factors for venous thrombosis see Lijfering *et al*.[40]

Many individuals who develop VT do not have any of the established risk factors,[2] which suggests that as yet unrecognized factors must play a role in VT pathophysiology. This is also in line with the observation that patients whose first thrombotic event

is not provoked by any of the established risk factors have a 2- to 3-fold increased recurrence risk.[5,8,14-16] Furthermore, (prophylactic) treatment of VT by anticoagulant use is not without risks, as all currently available anticoagulants are associated with bleeding complications.[17] In order to have better prevention and treatment strategies, we need to advance our knowledge on risk factors for VT and their underlying biological mechanisms.

In addition to clinical or acquired risk factors, genetic variation contributes to the risk of VT. Individuals with a positive family history of VT have an increased risk of developing VT compared with individuals with a negative family history,[18] with the risk being proportional to the degree of relatedness to the affected family member.[19] Overall, VT has a strong genetic basis with heritability estimates between 50 and 60% based on family and twin studies.[20-22] To identify genes and specific genetic variants contributing to VT pathophysiology, different strategies have been employed including linkage analysis, candidate gene studies, genome-wide association studies, and (next-generation) DNA sequencing. Variants in seventeen genes have so far been identified as well-established genetic risk factors for VT.[23]

Among the first identified genetic risk factors for VT are the deficiencies in the natural anticoagulant proteins, i.e., antithrombin, protein C, and protein S (encoded by *SERPINC1, PROC*, and *PROS1,* respectively).[24-26] These deficiencies are mainly caused by rare or even family-specific variants and have a large effect on VT risk. Other major genetic risk variants for VT include factor V (FV) Leiden (in *F5*, rs6025) and prothrombin (PT) G20210A (in *F2,* rs1799963), which reach an average population frequency of 5% and 2% in Northwest Europe, respectively.[23,27,28] FV Leiden was identified in individuals with activated protein C (APC) resistance, as the missense variant demolishes one of the APC cleavage sites in activated FV.[27,29] The absence of this cleavage site also hampers the cofactor function of FV in degrading activated factor VIII by APC and protein S.[29] As a result, FV Leiden carriers have a 3-fold increased risk of VT,[23,27] which can be further increased in combination with other risk factors such as oral contraception use.[30] The 2-fold increased VT risk observed in carriers of PT G20210A is due to a substitution in the 3' untranslated region of *F2,* which affects the post-transcriptional regulation of PT mRNA and thereby increases PT plasma levels.[23,28,31] The remaining established genetic risk factors are common variants associated with modest effects on VT risk. Similar to FV Leiden and PT G20210A, most risk variants are located in or near genes coding

for proteins involved in hemostasis.[23] However, for some of the identified genetic loci, such as the locus in *TSPAN15*,[32] the causal variant and biological mechanism remain unknown. In addition, the established genetic risk factors explain around 5% of the phenotypic variance,[33] suggesting that there exist genetic risk factors for VT that have not yet been identified.

For recurrence, previous studies have mainly focussed on genetic variants associated with a first thrombotic event. For most variants no association with recurrence or much smaller effect sizes have been observed.[34-37] For example, carriers of FV Leiden have a 1.4-fold increased risk of recurrent VT compared with non-carriers, whereas PT G20210A is associated with a risk increase of recurrence of around 20 to 70 %.[34,35] In part, these findings can be explained by the difference in absolute risks of first and recurrent VT, resulting in incomparability of effects on a relative risk scale.[38] In addition, research into risk factors for recurrence risk may be hindered by index event bias, although this could lead to both under- and overestimation of the risk estimate.[39] This all assumes that the risk factors and underlying biological mechanisms for a first and recurrent VT are the same, whereas different genetic mechanisms may be involved in recurrence. For example, genetic variants that control the response to damaged vessels and valves after a thrombotic event could play a role in recurrence pathophysiology, but few studies have investigated recurrence-specific variants.

The main aim of the research conducted for this thesis was to identify novel genetic risk factors for a first and recurrent VT. This will not only advance our understanding of the genetic architecture of (recurrent) VT, but also aid in unravelling the biological mechanisms, improve risk stratification, and help to identify potential drug targets. In addition, we aim to show potential applications of genetic risk variants in risk stratification and causal inference.

OUTLINE

In **chapter 2**, we aim to identify novel genetic risk factors for a first VT by studying common and rare genetic variants in mainly coding regions of over 700 genes involved in hemostasis and related pathways using targeted next-sequencing. A more agnostic approach is used in **chapter 3**, where we conduct a genome-wide association study to

uncover common genetic variants associated with recurrent VT. To explore whether the difference in (recurrent) VT risk between men and women can be explained by variations on the Y chromosome, we study in **chapter 4** the association between common European Y haplogroups and the association with the risk of a first and recurrent VT. In **chapter 5**, our aim is to validate the synergistic effect of variation in *CADM1* and protein C deficiency which was previously observed in a family with thrombophilia. For this, we study the joint effects on VT risk of over 300 common variants in *CADM1* and abnormalities in the protein C pathway. The discriminative value of a risk score based on genetic risk factors for a first VT is assessed and compared with a clinical risk model in **chapter 6**. In addition, in **chapter 7**, we discuss the basic concepts of Mendelian randomisation analyses and their use in causal inference.

REFERENCES

1. Anderson FA Jr, Wheeler HB, Goldberg RJ, Hosmer DW, Patwardhan NA, Jovanovic B, et al. A population-based perspective of the hospital incidence and case-fatality rates of deep vein thrombosis and pulmonary embolism. The Worcester DVT Study. *Arch Intern Med*. 1991;151(5):933-8.
2. Naess IA, Christiansen SC, Romundstad P, Cannegieter SC, Rosendaal FR, Hammerstrøm J. Incidence and mortality of venous thrombosis: a population-based study. *J Thromb Haemost*. 2007;5(4):692-9.
3. Cushman M, Tsai AW, White RH, Heckbert SR, Rosamond WD, Enright P, et al. Deep vein thrombosis and pulmonary embolism in two cohorts: the longitudinal investigation of thromboembolism etiology. *Am J Med*. 2004;117(1):19-25.
4. White RH. The epidemiology of venous thromboembolism. *Circulation*. 2003;107(23 Suppl 1):I4-8. Review.
5. Prandoni P, Lensing AW, Cogo A, Cuppini S, Villalta S, Carta M, et al. The long-term clinical course of acute deep venous thrombosis. *Ann Intern Med*. 1996;125(1):1-7.
6. Ashrani AA, Heit JA. Incidence and cost burden of post-thrombotic syndrome. *J Thromb Thrombolysis*. 2009;28(4):465-76.
7. Kahn SR. The post thrombotic syndrome. *Thromb Res*. 2011;127 Suppl 3:S89-92.
8. Hansson PO, Sörbo J, Eriksson H. Recurrent venous thromboembolism after deep vein thrombosis: incidence and risk factors. *Arch Intern Med*. 2000;160(6):769-74.
9. Kyrle PA, Minar E, Bialonczyk C, Hirschl M, Weltermann A, Eichinger S. The risk of recurrent venous thromboembolism in men and women. *N Engl J Med.* 2004;350:2558–63.
10. Christiansen SC, Cannegieter SC, Koster T, Vandenbroucke JP, Rosendaal FR. Thrombophilia, clinical factors, and recurrent venous thrombotic events. *JAMA*. 2005;293:2352-61.
11. Douketis J, Tosetto A, Marcucci M, Baglin T, Cosmi B, Cushman M, et al. Risk of recurrence after venous thromboembolism in men and women: patient level meta-analysis. *BMJ*. 2011;342:d813.
12. Rosendaal FR. Venous thrombosis: a multicausal disease. Lancet. 1999;353(9159):1167-73.
13. Lijfering WM, Rosendaal FR, Cannegieter SC. Risk factors for venous thrombosis - current understanding from an epidemiological point of view. *Br J Haematol*. 2010;149(6):824-33.
14. Baglin T, Luddington R, Brown K, Baglin C. Incidence of recurrent venous thromboembolism in relation to clinical and thrombophilic risk factors: prospective cohort study. *Lancet*. 2003;362(9383):523-6.
15. Christiansen SC, Cannegieter SC, Koster T, Vandenbroucke JP, Rosendaal FR. Thrombophilia, clinical factors, and recurrent venous thrombotic events. *JAMA*. 2005;293(19):2352-61.

16. Prandoni P, Noventa F, Ghirarduzzi A, Pengo V, Bernardi E, Pesavento R, et al. The risk of recurrent venous thromboembolism after discontinuing anticoagulation in patients with acute proximal deep vein thrombosis or pulmonary embolism. A prospective cohort study in 1,626 patients. *Haematologica*. 2007;92(2):199-205.
17. Schulman S, Beyth RJ, Kearon C, Levine MN. Hemorrhagic complications of anticoagulant and thrombolytic treatment: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). *Chest*. 2008;133(6 Suppl):257S-298S.
18. Bezemer ID, van der Meer FJ, Eikenboom JC, Rosendaal FR, Doggen CJ. The value of family history as a risk indicator for venous thrombosis. *Arch Intern Med*. 2009;169(6):610-5.
19. Zöller B, Ohlsson H, Sundquist J, Sundquist K. Familial risk of venous thromboembolism in first-, second- and third-degree relatives: a nationwide family study in Sweden. *Thromb Haemost.* 2013;109(3):458-63.
20. Souto J, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria J, et al. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. *Am J Hum Genet*. 2000;67(6):1452–9.
21. Larsen T, Sorensen H, Skytthe A, Johnsen S, Vaupel J, Christensen K. Major genetic susceptibility for venous thromboembolism in men: a study of Danish twins. *Epidemiology*. 2003;14(3):328–32.
22. Heit J, Phelps M, Ward S, Slusser J, Petterson T, De Andrade M. Familial segregation of venous thromboembolism. *J Thromb Haemost*. 2004;2(5):731–6. Ho WK, Hankey GJ, Quinlan DJ, Eikelboom JW. Risk of recurrent venous thromboembolism in patients with common thrombophilia: a systematic review. *Arch Intern Med*. 2006;166(7):729-36.
23. Trégouët DA, Morange PE. What is currently known about the genetics of venous thromboembolism at the dawn of next generation sequencing technologies. *Br J Haematol*. 2018;180(3):335-345.
24. Egeberg O. Inherited antithrombin deficiency causing thrombophilia. *Thromb Diath Haemorrh*. 1965;13:516–30.
25. Griffin JH, Evatt B, Zimmerman TS, Kleiss AJ, Wideman C. Deficiency of protein C in congenital thrombotic disease. *J Clin Invest*. 1981;68(5):1370–3.
26. Schwarz HP, Fischer M, Hopmeier P, Batard MA, Griffin JH. Plasma protein S deficiency in familial thrombotic disease. *Blood*. 1984;64(6):1297–300.
27. Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, et al. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature*. 1994;369(6475):64-7.
28. Poort SR, Rosendaal FR, Reitsma PH, Bertina RM. A common genetic variation in the 3'-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis. *Blood*. 1996;88(10):3698-703.

29. Nicolaes GA, Dahlbäck B. Factor V and thrombotic disease: description of a janus-faced protein. *Arterioscler Thromb Vasc Biol*. 2002;22(4):530-8.
30. Vandenbroucke JP, Koster T, Briët E, Reitsma PH, Bertina RM, Rosendaal FR. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet.* 1994;344(8935):1453-7.
31. Pollak ES, Lam HS, Russell JE. The G20210A mutation does not affect the stability of prothrombin mRNA in vivo. *Blood*. 2002;100(1):359-62.
32. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng LC, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet*. 2015;96(4):532-42.
33. Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, et al. Genetics of venous thrombosis: insights from a new genome wide association study. *PLoS One*. 2011;6(9):e25581.
34. Ho WK, Hankey GJ, Quinlan DJ, Eikelboom JW. Risk of recurrent venous thromboembolism in patients with common thrombophilia: a systematic review. *Arch Intern Med*. 2006;166(7):729-36.
35. Marchiori A, Mosena L, Prins MH, Prandoni P. The risk of recurrent venous thromboembolism among heterozygous carriers of factor V Leiden or prothrombin G20210A mutation. A systematic review of prospective studies. *Haematologica*. 2007;92(8):1107-14.
36. van Hylckama Vlieg A, Flinterman LE, Bare LA, Cannegieter SC, Reitsma PH, Arellano AR, et al. Genetic variations associated with recurrent venous thrombosis. *Circ Cardiovasc Genet*. 2014;7(6):806-13.
37. Bruzelius M, Ljungqvist M, Bottai M, Bergendal A, Strawbridge RJ, Holmström M, et al. F11 is associated with recurrent VTE in women. A prospective cohort study. *Thromb Haemost*. 2016;115(2):406-14.
38. Cannegieter SC, van Hylckama Vlieg A. Venous thrombosis: understanding the paradoxes of recurrence. *J Thromb Haemost*. 2013;11 Suppl 1:161-9.
39. Dahabreh IJ, Kent DM. Index event bias as an explanation for the paradoxes of recurrence risk research. *JAMA*. 2011;305(8):822-3.
40. Lijfering WM, Rosendaal FR, Cannegieter SC. Risk factors for venous thrombosis - current understanding from an epidemiological point of view. *Br J Haematol*. 2010;149:824-33.

CHAPTER 2

Targeted sequencing to identify novel genetic risk factors for deep vein thrombosis: a study of 734 genes

de Haan HG, van Hylckama Vlieg A, Lotta LA, Gorski MM, Bucciarelli P, Martinelli I, Baglin TP, Peyvandi F, Rosendaal FR, INVENT consortium.

J Thromb Haemost. 2018;16(12):2432-2441.

ABSTRACT

Background

Although several genetic risk factors for deep vein thrombosis (DVT) are known, almost all related to hemostasis, a large genetic component remains unexplained.

Objectives

We aimed to identify novel genetic determinants using targeted DNA sequencing.

Patients/Methods

We included 899 DVT patients and 599 controls from three case-control studies (DVT-Milan, MEGA, and THE-VTE) for sequencing of the coding regions of 734 genes involved in hemostasis or related pathways. We performed single-variant association tests for common variants (minor allele frequency [MAF]≥1%) and gene-based tests for rare variants (MAF≤1%), accounting for multiple testing by the false discovery rate (FDR).

Results

Sixty-two out of 3,617 common variants were associated with DVT risk (FDR<0.10). Most of these mapped to *F5, ABO, FGA-FGG,* and *CYP4V2-KLKB1-F11*. Lead variant at *F5* was rs6672595 (odds ratio [OR] 1.58, 95% confidence interval [CI] 1.29-1.92), in moderate linkage with known variant rs4524. Reciprocal conditional analyses suggested that intronic variation might drive this association. We also observed a secondary association at the *F11* region: missense *KLKB1* variant rs3733402 remained associated conditional on known variants rs2039614 and rs2289252 (OR 1.36, 95% CI 1.10-1.69). Two novel variant associations were observed, in *CBS* and *MASP1*, but these did not replicate in the meta-analysis data from the INVENT consortium. There was no support for a burden of rare variants contributing to DVT risk (FDR>0.2).

Conclusions

We confirmed associations between DVT and common variants in *F5, ABO, FGA-FGG,* and *CYP4V2-KLKB1-F11* and observed secondary signals in *F5* and *CYP4V2-KLKB1-F11* that warrant replication and fine-mapping in larger studies.

INTRODUCTION

The hemostatic system ensures the delicate balance between clotting and bleeding. Disturbance of this balance towards clotting may lead to venous thrombosis (VT), mainly manifested as pulmonary embolism (PE) or deep vein thrombosis (DVT).[1,2] Abnormal levels of both fibrinolytic and coagulation factors have been associated with VT risk.[3-6] The role of platelets as risk factor is less well studied, with conflicting results being reported for associations between VT and several platelet markers.[7,8] In addition, genetic variants predominantly in genes encoding proteins of the hemostatic system have been linked to VT risk.[9] Deficiencies of the natural anticoagulants, antithrombin, protein C and protein S, were among the first identified genetic causes of VT, and by now hundreds of (mainly rare) mutations have been reported.[10] Two recent meta-analyses of genome-wide association studies (GWAS), each including over 6000 patients and a multifold of controls, confirmed the association of six loci and identified three novel loci.[11,12] The established loci all map to genes related to hemostasis, specifically: *F5, FGG, F11, ABO, F2,* and *PROCR*.[9-12] Two of the novel loci (*TSPAN15* and *SLC44A2*), and potentially a third locus at *HIVEP1* identified in an earlier GWAS[13] but not confirmed in the latest meta-analyses,[11,12] are the only replicated loci not directly connected to the hemostatic system. This suggests that genes regulating (components of) the hemostatic system are the main genetic contributors to VT risk.

While VT has a strong genetic basis, with heritability estimates of 50-60%,[14-16] the established genetic risk factors only explain a small proportion of the phenotypic variance.[17] In addition, the genetic component remains unknown in 30% of families with multiple family members affected by VT.[18] GWAS efforts have had limited success in identifying novel genetic risk factors, which were mainly common variants in hemostatic-related genes conferring small effects on VT risk. Therefore, a focus on rare and low-frequency variants in coding regions of the genome, may help to discover novel determinants of VT. As such, we have previously shown that a burden of rare coding *ADAMTS13* variants is associated with a 4.8-fold increased DVT risk.[19]

To extend the GWAS efforts, we performed targeted DNA sequencing of the coding regions of 734 genes that were or could be related to the hemostatic system in 899 DVT patients and 599 controls. We subsequently sought replication for associated variants using meta-analysis data from the International Network against Thrombosis (INVENT) collaboration.[11]

PATIENTS AND METHODS

Study population

We set up the Milan Leiden Sequencing study (MILES), in which we included patients with a first VT and controls without a history of VT from three population-based case-control studies: DVT-Milan, Multiple Environmental and Genetic Assessment of risk factors for venous thrombosis (MEGA), and the Thrombophilia, Hypercoagulability and Environmental Risks in Venous Thromboembolism (THE-VTE) study. All studies have been previously described in detail.[19-21] Briefly, DVT-Milan recruited 2,139 consecutive patients with a first DVT at the Angelo Bianchi Bonomi Hemophilia and Thrombosis Center in Milan (Italy) between 1995 and 2010.[19] As controls served non-consanguineous relatives, partners or friends who accompanied patients to center visits. In MEGA, 4,956 consecutive patients with a first DVT or PE were recruited at six anticoagulation clinics in the Netherlands between 1999 and 2004.[20] Partners of patients were invited to participate as a control subject. Additional controls were recruited from the general population using random digit dialling. Patients and controls were invited to provide a blood sample until 2002, after which we switched, for logistical reasons, to buccal swabs. THE-VTE is a two-center case-control study, with a similar design as MEGA, in which 796 consecutive patients with a first DVT or PE and 531 controls were enrolled in Leiden (the Netherlands) and Cambridge (United Kingdom) between 2003 and 2008.[21] Again, partners of eligible patients were invited to participate as control subject.

From each study we included patients and controls based on the following criteria: high-quality DNA sample available from blood, European ancestry as defined by self-reported country of birth of the parents, no major surgery or cancer diagnosis related to the index date, and no deficiency of the natural anticoagulant proteins defined as having normal levels of protein C, protein S, and antithrombin. To eliminate two major genetic causes of VT, we included patients and controls who did not carry factor V (FV) Leiden (rs6025) or prothrombin (PT) G20210A (rs1799963). In addition, we oversampled patients who had a recurrence during the follow-up studies of MEGA and THE-VTE (N=241), as these are more likely to carry genetic risk factors for VT. To ensure a sufficient sample size, we allowed recurrent VT patients to carry FV Leiden or PT G20210A (N=94). In total, 899 DVT patients and 599 controls were selected for sequencing. An overview of the participants per study is presented in Supplemental Table 1.

All participants provided written informed consent. DVT-Milan was approved by the Institutional Review Board of the Fondazione IRCCS Ca' Granda–Ospedale Maggiore Policlinico, whereas MEGA and THE-VTE were approved by the Medical Ethics Committee of the Leiden University Medical Center. THE-VTE was also approved by the NHS Research Ethics Committee in Cambridge.

Targeted DNA sequencing

We selected pathways involved in thrombosis and hemostasis, including the coagulation system, fibrinolysis, platelet function, inflammation, and the complement system. Using literature and gene ontology databases, we extracted genes belonging to these pathways. From the ThromboGenomics database,[22] we included additional genes that have been linked to inherited clotting, platelets or bleeding disorders. In total, we included 734 genes, of which we sequenced the coding regions plus 10 base pairs flanking the exons to cover the splice junctions. For a subset of 48 genes, we additionally sequenced the 3' and 5' untranslated regions (UTR). In addition, we performed whole gene sequencing including 10 kilo base pairs promoter area of three genes, that is *F5, VWF*, and *F8*, which are of particular interest for VT. *F5* harbours the strongest genetic risk factor for VT, that is FV Leiden, in the general population. Von Willebrand factor and factor VIII, encoded by *VWF* and *F8*, are tightly interconnected proteins of which levels are strongly associated with first and recurrent VT risk.[5,23] We also targeted 179 single nucleotide variants, consisting of 28 variants previously associated with VT and 151 ancestry-informative markers. To facilitate the capture, we allowed some 200 base pairs of target region surrounding each variant. A list of the targeted genes and variants can be found in Supplemental Table 2.

The target area was designed with the Reference Sequence (RefSeq) Database using tools in the UCSC Genome Browser[24] and sent to NimbleGen (Roche NimbleGen, Madison, WI, USA) for probe design. Next-generation DNA sequencing was subsequently performed at the Human Genome Sequencing Center (HGSC), Baylor College of Medicine (Houston, USA). A complete sequencing protocol can be accessed on the HGSC website (https://www.hgsc.bcm.edu/content/protocols-sequencing-library-construction). Briefly, DNA samples were constructed into Illumina paired-end pre-capture libraries according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D) with some minor modifications. We multiplexed 24 samples per capture and included

two capture pools per HiSeq lane. Enriched samples were sequenced using the HiSeq 2000 platform (Illumina, San Diego, CA, USA).

Sequence analysis was performed using the Mercury analysis pipeline.[25] In short, sequence reads and base-call confidence values were generated for de-multiplexed pools using the vendor's primary analysis software (CASAVA). Next, reads and qualities were mapped to reference genome hg19 using the Burrows-Wheeler aligner,[26] resulting in BAM files per sample.[27] Realignment around insertions and deletions (indels), and recalibration of quality scores was performed with the Genome Analysis Toolkit.[28] Variant calling was conducted using the Atlas2suite,[29] followed by variant annotation as implemented in the Cassandra annotation suite. Individual variant files were subsequently merged into a project-level file to generate a genotype matrix of all identified variants.

Initial exclusion criteria for variant calls were as follows: variant posterior probability <0.95, number of variant reads <3, variant read ratio <0.1, variant reads in a single strand direction, total coverage <6 or >1024 reads. Called variants that passed quality control in at least one individual were included in the project-level variant file. In total, 31,540 variants were identified in 1495 individuals with sequencing data available (897 DVT patients and 598 controls). We subsequently performed additional filtering using VCFtools[30] to identify high-quality variants, requiring a sequencing depth ≥10 reads, call rate ≥80%, Phred score ≥30, and Hardy-Weinberg equilibrium $P>1.0x10^{-4}$ in the controls separately per study. A total of 20,054 variants passed quality control.

Statistical analysis

We conducted single-variant association analyses for 3,617 low-frequency and common variants, defined as a minor allele frequency (MAF) ≥1%, using logistic regression as implemented in PLINK.[31] We calculated effect estimates as odds ratios (OR) with corresponding 95% confidence intervals (95% CI) per risk allele copy and adjusted for sex, age, (study) origin, carriership of FV Leiden per allele copy, and carriership of PT G20210A. We assumed that X-chromosomal loci undergo complete inactivation. Linkage disequilibrium (LD) between variants was assessed in Europeans from the 1000 Genomes Project.[32] To identify secondary associations, we performed conditional analyses by adjusting for the lead variant at a locus (defined as region within 1 Mb of the lead variant). The Bonferroni threshold for significance was set at $1.38 x10^{-5}$ (0.05

divided by 3,617 variants) to account for multiple testing. We additionally calculated false discovery rates (FDR) and variants with a FDR <0.10 were carried forward for replication.

Rare variants (MAF ≤1%) were collapsed per gene and analysed with the T1 burden test and the Sequence Kernel Association Test (SKAT),[33] the latter allowing differential effect directions. In total, we analysed 16,188 variants in 647 genes with a cumulative minor allele count (cMAC) ≥5. Analyses were adjusted for sex, age, (study) origin, carriership of FV Leiden, and PT G20210A. In the burden test, we used adaptive permutations to calculate empirical P-values, which were stratified by Northwest versus South European origin. We calculated FDRs to take multiple testing into account. To identify which rare variant contributed to an association signal, we excluded one variant at a time and repeated the analyses. The gene-based association tests were performed with the PLINK/SEQ suite.

Replication

Novel associations between common and low-frequency variants and DVT (FDR <0.10) were examined in meta-analysis data from INVENT. Details on the meta-analysis and the included studies are provided elsewhere.[11] In short, GWAS data from 12 studies, totalling 7,507 VT patients and 52,632 controls, were meta-analysed using an inverse-variance weighting fixed-effects model. Of note, there was a small amount of overlap in VT patients (N=384) between the discovery and the replication analyses, as some patients were also included in the meta-analysis of INVENT.

RESULTS

Targeted DNA sequencing was successfully performed in 897 DVT patients and 598 controls. The study population characteristics are presented in Table 1. In total, 20,054 high-quality variants were identified, of which 11,268 were singletons (median of 7 singletons per person, interquartile range 4-10). An overview of the functional classes and the MAF distribution is shown in Supplemental Figure 1. The majority of the variants was rare and mapped to protein-coding sequence (N=10,131), including several stop-loss and -gain variants. We also observed 168 indels and 530 splice variants. In addition, we identified a total of 5,210 variants which had not been reported in any database.

Table 1. Study population characteristics

	DVT patients	Controls
N	897	598
Age in years, mean (SD)	48.1 (13.7)	47.1 (13.3)
male sex, N (%)	449 (50.1)	277 (46.3)
North-west European origin, N (%)	599 (67.8)	300 (50.2)
DVT only, N (%)	755 (84.2)	NA
*Carriers PT, N (%)	15 (1.67)	NA
*Carriers FVL, N (%)	75 (8.36)	NA

DVT deep vein thrombosis; SD standard deviation; FVL factor V Leiden; PT prothrombin G20210A; NA not applicable

* These were part of a subgroup of 241 DVT patients who had a recurrence during follow-up in MEGA and THE-VTE (prevalence of FVL and PT in that subgroup of 31.1% and 6.2%, respectively).

Single variant association analyses

We tested 3,617 low-frequency and common variants for an association with DVT risk. The quantile-quantile plot of the observed P-values versus the expected distribution is shown in Supplemental Figure 2. Statistically significant associations at the Bonferroni threshold were observed for 12 variants in four loci: *ABO, FGA-FGG, CYP4V2-KLKB1-F11*, and *F5* (Table 2). All four loci harbour established genetic risk factors for VT. Interestingly, only three of the 12 variants mapped to coding sequence. Exclusion of recurrent VT patients in a sensitivity analysis resulted in similar associations with DVT risk (Supplemental Table 3). Lead variant in *ABO* was the well-known risk variant rs8176719 (frameshift variant, risk allele frequency (RAF) 45%), encoding non-O blood groups. C-carriers had a 1.9-fold (95% CI 1.61-2.24) increased DVT risk per allele copy. The intronic *ABO* variant rs4962040 also reached statistical significance (RAF 59%, OR 1.53, 95% CI 1.28-1.83), though this association was diminished upon conditioning on rs8176719 ($OR_{adjusted}$ 1.12, 95% CI 0.88-1.41). Likewise, none of the other 22 *ABO* variants were associated with DVT risk conditional on rs8176719 (Supplemental Table 4). In *CYP4V2-KLKB1-F11*, lead variant was intronic *F11* variant rs2036914 (RAF 60%, OR 1.65, 95% CI 1.38-1.97), which has been linked to increased FXI levels and VT.[34,35] Three additional variants were associated with DVT risk at the Bonferroni threshold, of which one remained associated upon conditioning on rs2036914 (rs3733402 in *KLKB1*, $OR_{adjusted}$ 1.33, 95% CI 1.08-1.64). Conditioning on a second known *F11* risk variant (rs2289252), did not materially change this association ($OR_{adjusted}$ 1.36, 95% CI 1.10-1.69).

The *KLKB1* missense variant (p.Ser143Asn) leads to reduced binding of prekallikrein to its cofactor high-molecular weight kininogen,[36] affecting the initiation of the intrinsic coagulation cascade. In the *FGA-FGG* locus, the association with DVT was driven by missense *FGA* variant rs6050 (RAF 39%, OR 1.66, 95% CI 1.37-2.02) and downstream *FGG* variant rs2066865 (RAF 35%, OR 1.60, 95% CI 1.33-1.92), which have both been linked to increased γ' fibrinogen levels and VT risk.[37,38] rs6050 and rs2066865 were in high LD (r^2 0.90) and reciprocal conditional analysis showed that they represented the same association signal (Supplemental Table 5). We did not identify additional associations after conditioning on the lead variants (Supplemental Table 4). Four intronic *F5* variants were associated with DVT risk at the Bonferroni threshold, which were in almost complete LD (lowest r^2 between any pair was 0.90) and represented the same association signal. Carriers of the lead variant (rs6672595, RAF 76%) had a 1.6-fold increased DVT risk (95% CI 1.29-1.92) per risk allele. The variants were also in high LD (r^2 0.77) with *F5* missense variant rs4524, for which an association with VT independent of FV Leiden has been reported.[39] In our study, carriers of rs4524 (RAF 73%) had a 1.3-fold higher DVT risk (95% CI 1.11-1.60) per allele copy, which attenuated with adjustment for lead variant rs6672595 ($OR_{adjusted}$ 1.10, 95% CI 0.74-1.63). On the other hand, the association between rs6672595 (and its proxies) and DVT risk remained, albeit with wider confidence intervals, with adjustment for rs4524 (Supplemental Table 6). No secondary association signals were observed in the *F5* region (Supplemental Figure 3).

Table 2. Associations between common variants and first deep vein thrombosis ($P < 1.38 \times 10^{-5}$)

							Discovery analysis		*Conditional analysis	
rsID	**Chr.**	**Position**	**Class**	**Gene**	**A_1/A_2**	**RAF**	**OR (95%CI)**	**P**	***OR (95%CI)**	**P**
rs3766110	1	169515183	intronic	*F5*	C/A	0.774	1.54 (1.27-1.86)	1.07×10^{-5}	NA	NA
rs3766111	1	169515204	intronic	*F5*	C/T	0.773	1.57 (1.29-1.91)	6.82×10^{-6}	NA	NA
rs3766113	1	169515307	intronic	*F5*	G/A	0.770	1.55 (1.28-1.88)	8.59×10^{-6}	NA	NA
rs6672595	1	169515536	intronic	*F5*	T/C	0.757	1.58 (1.29-1.92)	6.11×10^{-6}	NA	NA
rs6050	4	155507590	missense	*FGA*	T/C	0.393	1.66 (1.37-2.02)	2.33×10^{-7}	NA	NA
rs2066865	4	155525276	downstream	*FGG*	G/A	0.352	1.60 (1.33-1.92)	4.86×10^{-7}	1.36 (0.81-2.31)	0.245
rs3733402	4	187158034	missense	*KLKB1*	G/A	0.573	1.55 (1.30-1.86)	1.27×10^{-6}	1.33 (1.08-1.64)	0.006
rs4253399	4	187188094	intronic	*F11*	T/G	0.458	1.50 (1.27-1.76)	8.34×10^{-7}	1.16 (0.90-1.49)	0.246
rs3822057	4	187188152	intronic	*F11*	A/C	0.545	1.44 (1.23-1.70)	6.74×10^{-6}	0.91 (0.62-1.35)	0.642
rs2036914	4	187192481	intronic	*F11*	T/C	0.602	1.65 (1.38-1.97)	2.47×10^{-8}	NA	NA
rs8176719	9	136132908	frameshift	*ABO*	T/TC	0.451	1.90 (1.61-2.24)	1.39×10^{-14}	NA	NA
rs4962040	9	136133531	intronic	*ABO*	G/A	0.594	1.53 (1.28-1.83)	3.67×10^{-6}	1.12 (0.88-1.41)	0.355

Chr. chromosome; A_1 reference allele; A_2 risk allele; RAF risk allele frequency; OR odds ratio; CI confidence interval; P P value; NA not applicable

Single variants association analyses for 3,617 low-frequency and common variants (MAF > 1%) were conducted using logistic regression assuming an additive mode of inheritance. Analyses were adjusted for sex, age, (study) origin, carriership of FV Leiden per copy of the risk allele, and carriership of PT G20210A.

*We conducted conditional logistic regression analyses in which we adjusted for the lead variant per locus (highlighted in bold, i.e. *F5* rs6672595, *FGA* rs6050, *F11* rs2036914, and *ABO* rs8176719).

In addition, we observed 50 variants that did not exceed the Bonferroni threshold for statistical significance, but did have low FDR (<0.10). Almost all of these mapped to the four main loci and did not represent new association signals (Supplemental Table 7). We additionally identified two novel, suggestive variant associations with DVT risk (Table 3). In *MASP1*, we observed an association with DVT for 3' UTR variant rs72549167 (RAF 1.6%, FDR 9%). Carriers of the risk allele had a 3.5-fold increased DVT risk (95% CI 1.62-7.67) per allele copy. The *MASP1* gene encodes mannan-binding lecture serine peptidase 1, which is involved in the lectin pathway of complement activation and has crosslinks with the clotting cascade.[40,41] In particular, activated by thrombin and activated platelets,[42] MASP1 can cleave several coagulation factors, including prothrombin, thrombin-activatable fibrinolysis inhibitor, and factor XIII.[41] Of the other 16 *MASP1* variants, one was also associated with DVT risk (Supplemental Table 8), which was in complete LD with rs72549167. The other novel variant association mapped to a synonymous variant in *CBS*, encoding cystathionine beta-synthase, associated with DVT risk with an allelic OR of 1.31 (95% CI 1.11-1.55, FDR 9%). Cystathionine beta-synthase catalyses the conversion of homocysteine to cystathionine and specific genetic defects in *CBS* lead to homocystinuria, a disorder which has been linked to increased VT risk.[43] We observed two additional common variants in *CBS*, all not associated with rs1801181, and none of these were associated with DVT risk (Supplemental Table 9). We next aimed to replicate the two novel variant associations using the meta-analysis data from INVENT, which included 7,507 VT patients and 52,632 controls (Table 3). There was no clear evidence for an association of DVT with rs72549167 in *MASP1* (OR 1.21, 95% CI 0.96-1.52), nor with rs1801181 in *CBS* (OR 1.00, 95% CI 0.96-1.05).

Table 3. Novel variant associations with deep vein thrombosis (FDR < 0.10) and replication effort

						Discovery analysis				*Replication			
rs ID	**Chr.**	**Position**	**Class**	**Gene**	**A_1/A_2**	**RAF**	**OR (95% CI)**	**P**	**FDR**	**RAF**	**OR (95% CI)**	**P**	**Info (SD)**
rs1801181	21	44480616	synonymous	*CBS*	G/A	0.370	1.31 (1.11-1.55)	0.002	0.09	0.364	1.00 (0.95-1.05)	0.926	0.96 (0.02)
rs72549167	3	186952375	3' UTR	*MASP1*	C/G	0.016	3.52 (1.62-7.67)	0.002	0.09	0.010	1.21 (0.96-1.52)	0.102	0.86 (0.12)

Chr. chromosome; A_1 reference allele; A_2 risk allele; RAF risk allele frequency; OR odds ratio; CI confidence interval; P P value; info mean imputation quality score; SD standard deviation; UTR untranslated region

Discovery analysis was performed using logistic regression assuming an additive mode of inheritance. Analyses were adjusted for sex, age, (study) origin, carriership of FV Leiden per copy of the risk allele, and carriership of PT G20210A.

*Replication was performed in data from the INVENT consortium. GWAS results from 12 studies were meta-analysed using a fixed-effect meta-analysis model based on inverse-variance weighting. Heterogeneity was assessed by the Cochran's Q statistic and the I^2 index. For rs1801181 we observed a Q 8.69, I^2 0.00, P-value 0.65. For rs72549167, we observed a Q of 9.06, I^2 0.00, P-value 0.62.

Gene-based association analyses

The impact of 16,188 rare variants mapping to 647 genes (cMAC ≥5) on DVT risk was assessed with aggregation tests. The results from the SKAT-based joint analyses of all rare variants per gene did not provide support for an association between rare variants and DVT risk. The most suggestive association signal was observed for *F2RL2* (P 0.0013, FDR 60%), encoding proteinase-activated receptor-3 (PAR-3). The burden tests identified one gene suggestive of an association with DVT risk. DVT patients had a burden of rare variants in *KLK5* (P 0.0003, FDR 21%), which encodes a serine protease named kallikrein related peptidase 5 and is involved in inflammatory responses through the PAR-2 system.[44] Of the 10 rare variants identified in *KLK5*, including five singletons, 26 variant alleles were observed in DVT patients compared with three alleles in controls. All 10 variants mapped to protein-coding sequence. None of the variants was solely driving the association signal (data not shown).

DISCUSSION

To identify novel genetic risk factors for DVT which have been missed by GWAS, we sequenced the coding regions of 734 genes related to hemostasis in 899 DVT patients and 599 controls. Our targeted sequencing approach confirmed several established risk loci. Specifically, lead variants at *ABO*, *FGA-FGG*, and *CYP4V2-KLKB1-F11* have all previously been implicated in VT risk, both directly or via proxy variants.[11-13,19,34-36] The effect sizes observed in our study were slightly higher than in earlier reports, which may in part be explained by our selection of individuals without a cancer diagnosis or recent surgery. Differences in genetic effects on PE versus DVT could also have played a role, in line with the so-called 'FV Leiden paradox'.[45] Although we did not discover novel risk loci, the secondary risk loci identified at *F5* and *CYP4V2-KLKB1-F11* may provide leads for a better understanding of the biological mechanism underlying these loci.

Interestingly, almost all associated variants mapped to non-coding sequence, while our sequencing design mainly targeted coding variation. In *F5* and *CYP4V2-KLKB1-F11*, there was little evidence that the (lead) associations could be explained by linkage to common, coding variants. This may point to non-coding variation as causal risk factor, potentially influencing DVT risk by affecting gene regulation. Four co-inherited intronic variants in *F5* were associated with DVT risk at the Bonferroni threshold, which have

not been implicated in VT risk. Missense *F5* variant rs4524, an established risk variant independent of FV Leiden[38] and in moderate LD with the associated *F5* variants, did not attain a high level of statistical significance in our study. Furthermore, its effect on DVT risk was strongly diminished when adjusting for our lead *F5* variant (rs6672595). Both variants are part of a large, strongly-linked cluster of variants, which spans across several introns and exons of *F5*. Additional fine-mapping in a large study is necessary to uncover the most likely causal variant. Another notable finding was the suggestive, secondary association signal at *CYP4V2-KLKB1-F11*, missense *KLKB1* variant rs3733402, which remained associated with DVT risk with an allelic odds ratio of 1.4 upon adjusting for rs2036914 and rs2289252. We are not the first to report an association signal at *CYP4V2-KLKB1-F11* secondary to rs2289252 and rs2036914,[11,34] although the previously reported variants are not in LD with rs3733402, suggesting that this locus may indeed harbour multiple causal variants. In addition, we were unable to disentangle the effects of *FGA*-rs6050 and *FGG*-rs2068865 on DVT risk due to their strong, though imperfect, linkage. However, a previously reported haplotype analysis did not show an independent association with VT for the haplotype carrying *FGA*-rs6050.[36]

In addition to the associations at the known loci, we identified two variants, which have not been linked to VT risk, with low FDR but association tests that did not pass the Bonferroni threshold. These were a synonymous variant in *CBS* and a 3' UTR variant in *MASP1*. Both variants did, however, not replicate in the meta-analysis data from INVENT. Imputation quality was sufficient and there was no evidence of statistical heterogeneity. We cannot rule out that differences in the discovery and the replication study populations, for example due to the inclusion of DVT patients versus patients with any VT event, could have explained the lack of replication. Alternatively, the associations in the discovery analysis might have been chance findings, taking into account the FDR of 9% for both variants.

The gene-based analyses did not support the hypothesis of a burden of rare, mainly coding variants in hemostasis-related genes contributing to DVT risk. We observed a potential association for a burden of rare variants in *KLK5* with 26 alleles observed in DVT patients compared with 3 alleles in controls, though the FDR was relatively high (21%). The lack of significant gene associations may be explained by our limited sample size. Gene-based analyses for complex diseases generally require large study sizes given the likely modest effect sizes and the expected proportion of causal variants.[46]

Therefore, we might have missed associations between genes with rare variants and DVT risk. We also did not distinguish between rare variants with or without a predicted deleterious consequence, as advocated by some,[46,47] since this would have further increased the multiple testing burden and lowered cMAC counts. As the effects of VT on fitness are limited, we also did not expect strong purifying selection on deleterious variants. In addition, our group has previously reported an association between DVT and a burden of rare coding variants in *ADAMTS13* (17 alleles in DVT patients compared with 4 alleles in controls, N=192 individuals).[19] In the present study, we observed a nominal association for a burden of rare variants in *ADAMTS13* with DVT risk (P 0.048, 84 alleles in DVT patients compared with 42 alleles in controls). Although the majority of studied rare *ADAMTS13* variants mapped to coding sequence (75%), the inclusion of noncoding variants may explain the difference in the results of the burden analyses. However, when only focusing on rare coding variation in *ADAMTS13*, we observed a similar association with DVT risk (P 0.066, 55 alleles in DVT patients compared with 27 alleles in controls). Larger studies are needed to elucidate the role of rare coding and noncoding variants in *ADAMTS13* on DVT risk.

The major limitation of our study is its limited sample size, which did not allow us to detect associations across the entire allele frequency spectrum. Given the multicausal nature of DVT, genetic effect estimates on DVT risk are expected to be modest, requiring an even larger sample size. We attempted to maximize our statistical power by studying genetic variation in biologically plausible genes in a well-characterized study population. Specifically, we selected genetically enriched DVT patients, without some of the major clinical risk factors. In addition, we oversampled VT patients who had developed a recurrence and are therefore more likely to carry genetic risk variants. Except for a small number of patients with recurrent VT, we selected individuals not carrying FV Leiden and PT G20210A, and, therefore, we could not study these variants or those in strong LD. Another limitation is the lack of generalizability of our findings to non-European populations. In addition, by design, our targeted sequencing approach did not allow us to study variation in regulatory regions outside our target area nor variation in genes not previously linked to the hemostatic system. Therefore, we were unable to identify variants in untargeted regions of the candidate genes, novel DVT-associating genes outside the hemostatic system, and to assess variation in the recently identified risk loci *SLC44A2* and *TSPAN15*.[11]

In conclusion, our targeted sequencing approach confirmed the association of several of the established VT risk loci. The secondary loci identified at *F5* and *CYP4V2-KLKB1-F11* suggest that the underlying biological mechanism might be more complex than initially thought. In addition, we did not find evidence of a burden of rare variants in hemostasis-related genes affecting DVT risk.

REFERENCES

1. Reitsma PH, Versteeg HH, Middeldorp S. Mechanistic view of risk factors for venous thromboembolism. Arterioscler Thromb Vasc Biol. 2012;32:563-8.
2. Reitter EM, Kaider A, Ay C, Quehenberger P, Marosi C, Zielinski C, Pabinger I. Longitudinal analysis of hemostasis biomarkers in cancer patients during antitumor treatment. J Thromb Haemost. 2016;14:294-305.
3. van Tilburg NH, Rosendaal FR, Bertina RM. Thrombin activatable fibrinolysis inhibitor and the risk for deep vein thrombosis. Blood. 2000;95:2855-9.
4. Meltzer ME, Lisman T, de Groot PG, Meijers JC, le Cessie S, Doggen CJ, Rosendaal FR. Venous thrombosis risk associated with plasma hypofibrinolysis is explained by elevated plasma levels of TAFI and PAI-1. Blood. 2010;116:113-21.
5. Tsai AW, Cushman M, Rosamond WD, Heckbert SR, Tracy RP, Aleksic N, Folsom AR. Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE). Am J Med. 2002;113:636-42.
6. Cushman M, O'Meara ES, Folsom AR, Heckbert SR. Coagulation factors IX through XIII and the risk of future venous thrombosis: the Longitudinal Investigation of Thromboembolism Etiology. Blood. 2009;114:2878-83.
7. Braekkan SK, Mathiesen EB, Njølstad I, Wilsgaard T, Størmer J, Hansen JB. Mean platelet volume is a risk factor for venous thromboembolism: the Tromsø Study, Tromsø, Norway. J Thromb Haemost. 2010;8:157-62.
8. Puurunen MK, Hwang SJ, O'Donnell CJ, Tofler G, Johnson AD. Platelet function as a risk factor for venous thromboembolism in the Framingham Heart Study. Thromb Res. 2017;151:57-62.
9. Rosendaal FR, Reitsma PH. Genetics of venous thrombosis. J Thromb Haemost. 2009; 7 Suppl 1:301–304.
10. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet;136:665-677.
11. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng LC, de Andrade M, de Visser MC, Wiggins KL, Suchon P, Saut N, Smadja DM, Le Gal G, van Hylckama Vlieg A, Di Narzo A, Hao K, Nelson CP, Rocanin-Arjo A, Folkersen L, Monajemi R, et al. Meta-analysis of 65,734 Individuals Identifies TSPAN15 and SLC44A2 as Two Susceptibility Loci for Venous Thromboembolism. Am J Hum Genet. 2015;96:532-42.

12. Hinds DA, Buil A, Ziemek D, Martinez-Perez A, Malik R, Folkersen L, Germain M, Mälarstig A, Brown A, Soria JM, Dichgans M, Bing N, Franco-Cereceda A, Souto JC, Dermitzakis ET, Hamsten A, Worrall BB, Tung JY; METASTROKE Consortium, INVENT Consortium., Sabater-Lleal M. Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. Hum Mol Genet. 2016;25:1867-74.
13. Morange PE, Bezemer I, Saut N, Bare L, Burgos G, Brocheton J, Durand H, Biron-Andreani C, Schved JF, Pernod G, Galan P, Drouet L, Zelenika D, Germain M, Nicaud V, Heath S, Ninio E, Delluc A, Münzel T, Zeller T, et al. A follow-up study of a genome-wide association scan identifies a susceptibility locus for venous thrombosis on chromosome 6p24.1. Am J Hum Genet. 2010;86:592-5.
14. Souto J, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria J, Coll I, Felices R, Stone W, Fontcuberta J, Blangero J. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. Am J Hum Genet. 2000;67:1452–9.
15. Larsen T, Sorensen H, Skytthe A, Johnsen S, Vaupel J, Christensen K. Major genetic susceptibility for venous thromboembolism in men: a study of Danish twins. Epidemiology. 2003;14(3):328–32.
16. Heit J, Phelps M, Ward S, Slusser J, Petterson T, De Andrade M. Familial segregation of venous thromboembolism. J Thromb Haemost. 2004;2:731–6.
17. Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, Cohen W, Oudot-Mellakh T, Antoni G, Alessi MC, Zelenika D, Cambien F, Tiret L, Bertrand M, Dupuy AM, Letenneur L, Lathrop M, Emmerich J, Amouyel P, Trégouët DA, Morange PE. Genetics of venous thrombosis: insights from a new genome wide association study. PLoS One. 2011;6:e25581.
18. Bezemer ID, van der Meer FJ, Eikenboom JC, Rosendaal FR, Doggen CJ. The value of family history as a risk indicator for venous thrombosis. Arch Intern Med. 2009;169(6):610-5.
19. Lotta LA, Tuana G, Yu J, Martinelli I, Wang M, Yu F, Passamonti SM, Pappalardo E, Valsecchi C, Scherer SE, Hale W 4th, Muzny DM, Randi G, Rosendaal FR, Gibbs RA, Peyvandi F. Next-generation sequencing study finds an excess of rare, coding single-nucleotide variants of ADAMTS13 in patients with deep vein thrombosis. J Thromb Haemost. 2013;11:1228-39.
20. Blom JW, Doggen CJ, Osanto S, Rosendaal FR. Malignancies, prothrombotic mutations, and the risk of venous thrombosis. JAMA. 2005;293:715–722.
21. van Hylckama Vlieg A, Baglin CA, Luddington R, MacDonald S, Rosendaal FR, Baglin TP. The risk of a first and a recurrent venous thrombosis associated with an elevated D-dimer level and an elevated thrombin potential: results of the THE-VTE study. J Thromb Haemost. 2015;13:1642-52.

22. Simeoni I, Stephens JC, Hu F, Deevi SV, Megy K, Bariana TK, Lentaigne C, Schulman S, Sivapalaratnam S, Vries MJ, Westbury SK, Greene D, Papadia S, Alessi MC, Attwood AP, Ballmaier M, Baynam G, Bermejo E, Bertoli M, Bray PF, et al. A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorders. Blood. 2016;127:2791-803.
23. Timp JF, Lijfering WM, Flinterman LE, van Hylckama Vlieg A, le Cessie S, Rosendaal FR, Cannegieter SC. Predictive value of factor VIII levels for recurrent venous thrombosis: results from the MEGA follow-up study. J Thromb Haemost. 2015;13:1823-32.
24. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12:996-1006.
25. Reid JG, Carroll A, Veeraraghavan N, Dahdouli M, Sundqvist A, English A, Bainbridge M, White S, Salerno W, Buhay C, Yu F, Muzny D, Daly R, Duyk G, Gibbs RA, Boerwinkle E. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics. 2014;15,30.
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
28. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43,491–8.
29. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs RA, Yu F. An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics. 2012;13:8.
30. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011;27:2156-8.
31. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.
32. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68-74.
33. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82-93.

34. Li Y, Bezemer ID, Rowland CM, Tong CH, Arellano AR, Catanese JJ, Devlin JJ, Reitsma PH, Bare LA, Rosendaal FR. Genetic variants associated with deep vein thrombosis: the F11 locus. J Thromb Haemost. 2009;7:1802-8.
35. Bezemer ID, Bare LA, Doggen CJ, Arellano AR, Tong C, Rowland CM, Catanese J, Young BA, Reitsma PH, Devlin JJ, Rosendaal FR. Gene variants associated with deep vein thrombosis. JAMA. 2008;299:1306-14.
36. Katsuda I, Maruyama F, Ezaki K, Sawamura T, Ichihara Y. A new type of plasma prekallikrein deficiency associated with homozygosity for Gly104Arg and Asn124Ser in apple domain 2 of the heavy-chain region. Eur J Haematol. 2007;79:59–68.
37. Uitte de Willige S, de Visser MC, Houwing-Duistermaat JJ, Rosendaal FR, Vos HL, Bertina RM. Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen gamma' levels. Blood. 2005;106:4176-83.
38. Lovely RS, Yang Q, Massaro JM, Wang J, D'Agostino RB Sr, O'Donnell CJ, Shannon J, Farrell DH. Assessment of genetic determinants of the association of γ' fibrinogen in relation to cardiovascular disease. Arterioscler Thromb Vasc Biol. 2011;31:2345-52.
39. Smith NL, Hindorff LA, Heckbert SR, Lemaitre RN, Marciante KD, Rice K, Lumley T, Bis JC, Wiggins KL, Rosendaal FR, Psaty BM. Association of genetic variations with nonfatal venous thrombosis in postmenopausal women. JAMA. 2007;297:489–98.
40. Dobó J, Schroeder V, Jenny L, Cervenak L, Závodszky P, Gál P. Multiple roles of complement MASP-1 at the interface of innate immune response and coagulation. Mol Immunol. 2014;61:69-78.
41. Hess K, Ajjan R, Phoenix F, Dobó J, Gál P, Schroeder V. Effects of MASP-1 of the complement system on activation of coagulation factors and plasma clot formation. PLoS One. 2012;7:e35690.
42. Kozarcanin H, Lood C, Munthe-Fog L, Sandholm K, Hamad OA, Bengtsson AA, Skjoedt MO, Huber-Lang M, Garred P, Ekdahl KN, Nilsson B. The lectin complement pathway serine proteases (MASPs) represent a possible crossroad between the coagulation and complement systems in thromboinflammation. J Thromb Haemost. 2016;14:531-45.
43. Mudd SH, Skovby F, Levy HL, Pettigrew KD, Wilcken B, Pyeritz RE, Andria G, Boers GH, Bromberg IL, Cerone R, Fowler B, Gröbe H, Schmidt H, Schweitzeret L. The natural history of homocystinuria due to cystathionine β-synthase deficiency. Am J Hum Genet. 1997;37:1-31.
44. Oikonomopoulou K, Hansen KK, Saifeddine M, Vergnolle N, Tea I, Blaber M, Blaber SI, Scarisbrick I, Diamandis EP, Hollenberg MD. Kallikrein-mediated cell signalling: targeting proteinase-activated receptors (PARs). Biol Chem. 2006;387:817-24.
45. Desmarais S, de Moerloose P, Reber G, Minazio P, Perrier A, Bounameaux H. Resistance to activated protein C in an unselected population of patients with pulmonary embolism. Lancet. 1996;347:1374-75.

46. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, Albers PK, GoT2D Consortium, McVean G, Boehnke M, Altshuler D, McCarthy MI. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. 2015;11:e1005165.
47. Richardson TG, Campbell C, Timpson NJ, Gaunt TR. Incorporating Non-Coding Annotations into Rare Variant Analysis. PLoS One. 2016;11:e0154181.

SUPPLEMENTAL INFORMATION

Supplemental Table 1. Included participants per study

Supplemental Table 2. Targeted genes and variants

Supplemental Table 3. Sensitivity analysis excluding recurrent VT patients

Supplemental Table 4. Discovery and conditional association analyses of variants in *ABO*, *CYP4V2-KLKB1-F11*, and *FGA-FGG* ($P > 1.38 \times 10^{-5}$)

Supplemental Table 5. Reciprocal conditional association analyses rs6050 and rs2066865

Supplemental Table 6. Association analysis at the *F5* association locus conditional on rs4524

Supplemental Table 7. Suggestive single variant associations (FDR <0.10)

Supplemental Table 8. Single variant association analyses of common *MASP1* variants

Supplemental Table 9. Single variant association analyses of common *CBS* variants

Supplemental Figure 1. Minor allele frequency distribution of identified variants (left) and overview of functional classes (right)

Supplemental Figure 2. Quantile-quantile plot of single variant association analyses

Supplemental Figure 3. Regional association plots for single variant associations in the *F5* region

Supplemental Table 1. Included participants per study

	MEGA		THE-VTE		DVT-Milan	
	DVT patients	Control subjects	DVT patients	Control subjects	DVT patients	Control subjects
N	459	230	140	70	298	298
Men, N (%)	246 (53.6)	130 (56.5)	86 (61.4)	27 (38.6)	117 (39.3)	120 (40.3)
Age in years, mean (SD)	49.0 (12.4)	49.6 (11.5)	54.4 (12.9)	55.7 (10.9)	43.6 (14.7)	43.2 (13.8)
FV Leiden carriers, N	69	0	10	0	0	0
PT G20210A carriers, N	11	0	2	0	2	0

DVT, deep vein thrombosis; SD, standard deviation; PT, prothrombin

Supplemental Table 2. Targeted genes and variants

Target area	Gene names
Exons and intron-exon boundaries (N=670)	*ABCB6, ABCG2, ABO, ACTB, ACTG1, ACVRL1, ADAM17, ADIPOQ, ADRA1A, ADRA1B, ADRA1D, ADRA2A, ADRA2B, ADRA2C, ADRB1, ADRB2, ADRB3, ADRBK1, ADRBK2, AGPAT1, ALOX5, ALOX5AP, ANKRD26, ANO6, ANXA1, ANXA10, ANXA11, ANXA13, ANXA2, ANXA3, ANXA4, ANXA5, ANXA6, ANXA7, ANXA8, ANXA8L1, ANXA8L2, ANXA9, AP3B1, APCS, APOA1, APOA2, APOA4, APOA5, APOB, APOC1, APOC2, APOC3, APOC4, APOD, APOE, APOF, APOL1, APOL2, APOL3, APOL4, APOL5, APOL6, ARHGDIB, ARHGEF1, ARHGEF3, ARNT, ARNT2, ARNT3, ASIC2, ATP2A3, AXL, B2M, BAI3, BAZ1B, BIRC5, BLOC1S3, C1GALT1, C1orf114, C1QA, C1QB, C1QC, C1R, C1S, C2, C21orf7, C2orf88, C3, C4A, C4B, C5, C6, C6orf25, C7, C8A, C8B, C8G, C9, CA2, CADM1, CADM2, CADM3, CADM4, CALM1, CALM2, CALM3, CALR, CASK, CASP8AP2, CBS, CCL5, CD34, CD36, CD4, CD40, CD40LG, CD46, CD55, CD59, CDH1, CDKN1A, CDKN2D, CFB, CFD, CFH, CFI, CFP, CHST12, CHST14, CLEC4M, CLU, CMTM5, COL10A1, COL11A1, COL11A2, COL12A1, COL13A1, COL14A1, COL15A1, COL16A1, COL17A1, COL18A1, COL19A1, COL1A1, COL1A2, COL20A1, COL21A1, COL22A1, COL23A1, COL24A1, COL25A1, COL27A1, COL28A1, COL2A1, COL3A1, COL4A1, COL4A2, COL4A3, COL4A3BP, COL4A4, COL4A5, COL4A6, COL5A1, COL5A2, COL5A3, COL6A1, COL6A2, COL6A3, COL6A5, COL7A1, COL8A1, COL8A2, COL9A1, COL9A2, COL9A3, COMP,CR1, CR2, CRP, CRTAM, CTSA, CTSG, CYCS, CYP1A2, CYP2A6, CYP2C9, CYP3A5, CYP4V2, DTNBP1, EBI3, EDEM2, EDIL3, ELANE, EMID2, ENG, ENTPD1, EPAS1, EPB41L1, EPB41L2, EPB41L3, EPR1, EPS8L2, ERBB2, ERBB3, ESR1, FCGR2A, FCGR2B, FERMT3, FLI1, FLNA, FOXA1, FOXA2, FOXA3, FTH1, FTO, GAS6, GATA1, GBGT1, GCKR, GFI1B, GGCX, GLT6D1, GNA12, GNA13, GNAQ, GNAS, GNB2L1, GNG11, GP1BA, GP1BB, GP5, GP6, GP9, GPR30, GPX1, GPX2, GPX3, GPX4, GPX5, GPX6, GPX7, GPX8, GRAP2, HIF1A, HIF3A, HIST1H2AC, HIST1H2BH, HIST1H2BJ, HIST1, H2BK, HIST1H3H, HIVEP1, HLA-A, HLA-B, HLA-C, HOXA11, HPS1, HPS3, HPS4, HPS5, HPS6, HRG, HRH1, HRH2, HRH3, HRH4, HS6ST2, HTR1A, HTR1B, HTR1D, HTR1E, HTR1F, HTR2A, HTR2B, HTR2C, HTR3A, HTR3B, HTR3C, HTR3D, HTR3E, HTR4, HTR5A, HTR6, HTR7, ICAM1, ICAM2, ICAM3, ICAM4, ICAM5, IL10, IL10RA, IL10RB, IL11, IL11RA, IL12A, IL12B, IL12RB1, IL12RB2, IL13, IL13R IL15, IL15RA, IL16, IL17A, IL17B, IL17C, IL17D, IL17F, IL17RA, IL17RB, IL18, IL18R1, IL19, IL1A, IL1B, IL1R1, IL1R2, IL1RN, IL2, IL20, IL20RA, IL20RB, IL21, IL21R, IL22, IL22RA1, IL22RA2, IL23A, IL23R, IL24, IL25, IL26, IL27, IL27RA, IL28A, IL28B, IL28RA, IL29, IL2RA, IL2RB, IL2RG, IL3, IL31, IL31RA, IL32, IL33, IL3RA, IL3RB, IL4, IL4R, IL5, IL5RA, IL6, IL6RA, IL6RB, IL7, IL7RA, IL8, IL8RA, IL8RB, IL9, IL9R, IRAK1, IRAK2, IRAK3, IRAK4, ITFG2, ITGA1, ITGA10, ITGA11, ITGA2, ITGA2B, ITGA3, ITGA4, ITGA5, ITGA6, ITGA7, ITGA8, ITGA9, ITGAD, ITGAE, ITGAL, ITGAM, ITGAV, ITGAX, ITGB1, ITGB2, ITGB3, ITGB4, ITGB5, ITGB6, ITGB7, ITGB8, ITM2B, JAK1, JAK2, JAK3, KLF2, KLK1, KLK10, KLK11, KLK12, KLK13, KLK14, KLK15, KLK2, KLK3, KLK4, KLK5, KLK6, KLK7, KLK8, KLK9, LDLR, LFNG, LOC401913, LPA, LPAR4, LPAR6, LY6E, LY6G6F, LYST, MARCKS, MASP1, MASP2, MASTL, MBL2, MBTPS1, MERTK, MET, MFGE8, MMP2, MMP24, MMP9, MMRN1, MMRN2, MPL, MPP1, MPP3, MPP6, MTHFR, MYBPC3, MYH9, MYL6, MYL9, NAT8B, NBEA, NBEAL2, NCOA1, NFKB1, NFKB2, NNMT, NOS1, NOS2, NOS3, NQO1, NR1I2, NRGN, OAZ1, ODZ1, OS9, OST4, P2RX1, P2RX2, P2RX3, P2RX4, P2RX5, P2RX6, P2RX7, P2RY1, P2RY10,P2RY11, P2RY12, P2RY13, P2RY14, P2RY2, P2RY4, P2RY6, P2RY8, PCSK9, PDGFA, PDIA2, PDLIM1, PDZK1IP1, PECAM1, PF4, PGRMC1, PIGM, PKM2, PKN1, PKN2, PKN3, PLA1A, PLA2G10, PLA2G12A, PLA2G12B, PLA2G1B, PLA2G2A, PLA2G2C, PLA2G2D, PLA2G2E, PLA2G2F, PLA2G3, PLA2G4A, PLA2G4B, PLA2G4C, PLA2G4D,*

Supplemental Table 2. Continued

Target area	Gene names
Exons and intron-exon boundaries (N=670)	*PLA2G4E, PLA2G4F, PLA2G5, PLA2G6, PLA2G7, PLB1, PLCB1, PLCB2, PLCB3, PLCB4, PLCD1, PLCD3, PLCD4, PLCE1, PLCG1, PLCG2, PLCH1, PLCH2, PLCL1, PLCL2, PLCZ1, PLD1, PLD2, PPBP, PPDPF, PPP1CA, PPP1CB, PPP2CA, PPP2CB, PPP3CA, PPP3CB, PRKAR2B, PRKCA, PRKCB, PRKCD, PRKCE, PRKCG, PRKCH, PRKCI, PRKCQ, PRKCZ, PRKD1, PRKD2, PRKD3, PRTN3, PSEN1, PTAFR, PTGDR, PTGDS, PTGER1, PTGER2, PTGER3, PTGER4, PTGES, PTGES2, PTGES3, PTGFR, PTGIR, PTGIS, PTGS1, PTGS2, PTK2B, PTX3, PVRL3, RAC1, RAP1B, RASGRP2, RBM8A, RGS10, RGS18, RGS7, RND2, RUNX1, SAA1, SAA2, SAA4, SCARA5, SCARB1, SCIN, SDCBP, SDPR, SELE, SELL, SELP, SERF2, SERPINA1, SERPINA11, SERPINA12, SERPINA13, SERPINA2, SERPINA3, SERPINA4, SERPINA6, SERPINA7, SERPINA8, SERPINB1, SERPINB10, SERPINB11, SERPINB12, SERPINB13, SERPINB2, SERPINB3, SERPINB4, SERPINB5, SERPINB6, SERPINB7, SERPINB8, SERPINB9, SERPINE2, SERPINF1, SERPING1, SERPINH1, SERPINI1, SERPINI2, SH3BGRL2, SH3BGRL3, SHBG, SMAD4, SMOX, SNCA, SPARC, SPINT1, SPINT2, SPINT3, SPINT4, SRGN, STAB2, STX2, STXBP3, STXBP5, SULT1A1, SULT1E1, TACR1, TAGLN2, TAOK1, TBX2, TBXA2R, TBXAS1, TC2N, TFPI2, TGFB1, THBS1, THBS2, THBS3, THBS4, THPO, TLN1, TLR1, TLR10, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TNF, TNFRSF1A, TPH1, TREML1, TRIM58, TUBA4A, TUBB1, TXNLA,TYK2, TYRO3, UGT1A1, UGT1A3, UGT1A9, UGT2B7, USF1, VASP, VCAM1, VHL, VHLL, VKORC1, VTN, WAS, WASF2, WDR66, WIPF1, ZNF185, ZNF544*
Exons, intron-exon boundaries, and UTRS (N=48)	*A2M, ADAMTS13, APOH, C4BPA, C4BPB, CALU, CPB2, F10, F11, F12, F13A1, F13B, F2, F2R, F2RL1, F2RL2, F2RL3, F3, F7, F9, FGA, FGB, FGG, HABP2, HNF1A, HNF4A, KLKB1, KNG1, LMAN1, LRP1, MCFD2, PLAT, PLAU, PLAUR, PLG, PROC, PROCR, PROS1, PROZ, SERPINA10, SERPINA5, SERPINA9, SERPINC1, SERPIND1, SERPINE1, SERPINF2, TFPI, THBD*
Entire gene including promoter (N=3)	*F5, F8, VWF*
Variants (N=178)	rs988436, rs942793, rs1368136, rs2060983, rs1404402, rs1016120, rs1414411, rs2014303, rs1030626, rs1517661, rs764138, rs2218497, rs725379, rs1377724, rs1406121, rs869538, rs1905471, rs764681, rs1280100, rs723211, rs1368666, rs1896974, rs1366578, rs262981, rs262980, rs262982, rs2876712, rs2366267, rs234693, rs32524, rs3845765, rs746600, rs2227139, rs2132055, rs869537, rs2389530, rs1964428, rs1377725, rs710616, rs2312211, rs1343747, rs719985, rs768131, rs1428555, rs1073197, rs1398064, rs31096, rs718255, rs725105, rs1937147, rs2351002, rs1383279, rs1393358, rs1376264, rs1854673, rs1585217, rs727864, rs461081, rs1573275, rs1343809, rs1986710, rs959284, rs956377, rs448041, rs222948, rs952993, rs959100, rs743337, rs1074342, rs1395479, rs1476873, rs950668, rs1846466, rs1531124, rs1401535, rs953177, rs722626, rs952784, rs1327806, rs2225251, rs1931621, rs1412917, rs950228, rs718092, rs1825015, rs728337, rs2383188, rs754798, rs721544, rs361263, rs958967,

Supplemental Table 2. Continued

Target area	Gene names
Variants (N=178)	rs1416464, rs726364, rs1526952, rs1390049, rs1485254, rs1368666, rs829669, rs1362489, rs703698, rs10108270, rs10236187, rs1040045, rs1040404, rs10496971, rs10510228, rs10512572, rs10513300, rs10839880, rs11227699, rs11652805, rs12130799, rs12439433, rs12544346, rs12629908, rs12657828, rs1296819, rs1325502, rs13400937, rs1369093, rs1407434, rs1471939, rs1513181, rs1760921, rs1837606, rs1871428, rs1950993, rs200354, rs2030763, rs2073821, rs2125345, rs214678, rs2306040, rs2330442, rs2357442, rs2397060, rs2416791, rs2504853, rs260690, rs2627037, rs2702414, rs2946788, rs2986742, rs3118378, rs316598, rs316873, rs32314, rs3737576, rs3745099, rs3784230, rs1799963, rs8176719, rs2066865, rs2036914, rs2069951, rs2289252, rs4149755, rs2069952, rs2227589, rs169713, rs3136520, rs1799809, rs867186, rs1613662, rs3136516, rs1039084, rs2001490, rs6003, rs670659, rs6048, rs5985, rs8176592, rs3822057, rs1523127, rs3742264, rs710446, rs3813948. rs12941510

Supplemental Table 3. Sensitivity analysis excluding recurrent VT patients

							Discovery analysis		Sensitivity analysis	
rsID	**Chr.**	**Position**	**A_1/A_2**	**RAF**	**Class**	**Gene**	**OR (95%CI)**	**P**	***OR (95%CI)**	**P**
rs3766110	1	169515183	C/A	0.774	intronic	*F5*	1.54 (1.27-1.86)	$1.07x10^{-5}$	1.49 (1.22-1.81)	$9.87x10^{-5}$
rs3766111	1	169515204	C/T	0.773	intronic	*F5*	1.57 (1.29-1.91)	$6.82x10^{-6}$	1.52 (1.24-1.86)	$6.27x10^{-5}$
rs3766113	1	169515307	G/A	0.77	intronic	*F5*	1.55 (1.28-1.88)	$8.59x10^{-6}$	1.51 (1.24-1.85)	$5.28x10^{-5}$
rs6672595	1	169515536	T/C	0.757	intronic	*F5*	1.58 (1.29-1.92)	$6.11x10^{-6}$	1.54 (1.25-1.89)	$3.67x10^{-5}$
rs6050	4	155507590	T/C	0.393	missense	*FGA*	1.66 (1.37-2.02)	$2.33x10^{-7}$	1.55 (1.27-1.90)	$1.47x10^{-5}$
rs2066865	4	155525276	G/A	0.352	downstream	*FGG*	1.60 (1.33-1.92)	$4.86x10^{-7}$	1.56 (1.29-1.89)	$5.43x10^{-6}$
rs3733402	4	187158034	G/A	0.573	missense	*KLKB1*	1.55 (1.30-1.86)	$1.27x10^{-6}$	1.49 (1.24-1.79)	$2.78x10^{-5}$
rs4253399	4	187188094	T/G	0.458	intronic	*F11*	1.50 (1.27-1.76)	$8.34x10^{-7}$	1.43 (1.21-1.69)	$3.48x10^{-5}$
rs3822057	4	187188152	A/C	0.545	intronic	*F11*	1.44 (1.23-1.70)	$6.74x10^{-6}$	1.35 (1.15-1.60)	$3.80x10^{-4}$
rs2036914	4	187192481	T/C	0.602	intronic	*F11*	1.65 (1.38-1.97)	$2.47x10^{-8}$	1.55 (1.29-1.86)	$3.02x10^{-6}$
rs8176719	9	136132908	T/TC	0.451	Frameshift	*ABO*	1.90 (1.61-2.24)	$1.39x10^{-14}$	1.88 (1.59-2.23)	$2.28x10^{-13}$
rs4962040	9	136133531	G/A	0.594	intronic	*ABO*	1.53 (1.28-1.83)	$3.67x10^{-6}$	1.51 (1.26-1.82)	$1.21x10^{-5}$

Chr chromosome; A1 reference allele; A2 risk allele; RAF risk allele frequency; OR odds ratio; CI confidence interval
After exclusion of recurrent VT patients, single-variant association analyses were performed in 656 patients and 598 controls, adjusting for age, sex, and (study) origin.

Supplemental Table 4. Discovery and conditional association analyses of variants in *ABO*, *CYP4V2-KLKB1-F11*, and *FGA-FGG* (P >1.38x10^{-5})

							Discovery analyses		$^{\$}$Conditional analyses	
rsID	**Chr**	**Position**	**A1/A2**	**RAF**	**Functional class**	**Gene**	**OR (95%CI)**	**P**	**OR (95%CI)**	**P**
rs2070022	4	155504948	A/G	0.828	3' UTR	*FGA*	1.31 (1.06-1.62)	0.013	1.06 (0.81-1.39)	0.68
rs2070011	4	155511897	C/T	0.406	5' UTR	*FGA*	1.37 (1.16-1.62)	2.18x10^{-4}	0.90 (0.67-1.21)	0.478
rs1049636	4	155525970	G/A	0.761	3' UTR	*FGG*	1.30 (1.08-1.56)	0.005	1.03 (0.82-1.31)	0.778
rs13146272	4	187120211	C/A	0.692	missense	*CYP4V2*	1.35 (1.12-1.61)	0.001	1.20 (0.99-1.46)	0.07
rs3817184	4	187122304	C/T	0.456	intronic	*CYP4V2*	1.43 (1.21-1.69)	2.24x10^{-5}	1.21 (1.00-1.46)	0.054
rs3736455	4	187122319	T/G	0.706	synonymous	*CYP4V2*	1.31 (1.09-1.56)	0.003	1.18 (0.97-1.44)	0.104
rs34745240	4	187122332	G/A	0.046	missense	*CYP4V2*	1.42 (0.95-2.11)	0.083	1.18 (0.79-1.78)	0.42
rs4253301	4	187173012	G/T	0.868	missense	*KLKB1*	1.31 (1.03-1.66)	0.029	1.00 (0.75-1.33)	0.978
rs925453	4	187179210	T/C	0.715	synonymous	*KLKB1*	1.25 (1.05-1.49)	0.012	1.14 (0.94-1.39)	0.187
rs3087505	4	187179486	A/G	0.912	3' UTR	*KLKB1*	1.20 (0.92-1.56)	0.188	0.97 (0.71-1.33)	0.854
rs3733403	4	187187135	G/C	0.898	5' UTR	*F11*	1.05 (0.79-1.38)	0.752	1.28 (0.96-1.72)	0.093
rs4253398	4	187188061	C/T	0.705	intronic	*F11*	1.16 (0.97-1.39)	0.115	0.87 (0.69-1.09)	0.222
rs35709976	4	187188141	G/GAT	0.898	intronic	*F11*	1.33 (1.03-1.72)	0.027	1.12 (0.83-1.51)	0.469
rs2289252*	4	187207381	C/T	0.528	intronic	*F11*	1.39 (1.17-1.65)	1.94x10^{-4}	1.15 (0.92-1.44)	0.211
rs5976	4	187209729	A/G	0.949	synonymous	*F11*	1.21 (0.84-1.76)	0.307	0.89 (0.58-1.35)	0.579
rs4253429*	4	187210033	G/A	0.803	3' UTR	*F11*	1.28 (1.04-1.57)	0.021	1.01 (0.78-1.31)	0.958
rs4253430*	4	187210064	C/G	0.590	3' UTR	*F11*	1.38 (1.16-1.63)	2.50x10^{-4}	1.05 (0.83-1.31)	0.694
rs4253865*	4	187210090	A/G	0.950	3' UTR	*F11*	1.20 (0.83-1.73)	0.334	0.87 (0.57-1.32)	0.499
rs1062547*	4	187210247	T/A	0.605	3' UTR	*F11*	1.33 (1.13-1.57)	7.97x10^{-4}	1.01 (0.80-1.26)	0.954

Supplemental Table 4. Continued

							Discovery analyses		[$]Conditional analyses	
rsID	**Chr**	**Position**	**A1/A2**	**RAF**	**Functional class**	**Gene**	**OR (95%CI)**	**P**	**OR (95%CI)**	**P**
rs186377697	4	187210319	T/A	0.987	3' UTR	*F11*	1.39 (0.69-2.83)	0.359	1.32 (0.60-2.92)	0.495
rs8176749	9	136131188	C/T	0.087	ncRNA	*ABO*	1.53 (1.15-2.04)	0.004	1.09 (0.80-1.49)	0.569
rs8176748	9	136131289	T/C	0.774	ncRNA	*ABO*	1.37 (1.13-1.66)	0.001	1.01 (0.82-1.26)	0.898
rs8176747	9	136131315	C/G	0.092	ncRNA	*ABO*	1.60 (1.19-2.14)	0.002	1.13 (0.82-1.55)	0.452
rs41302905	9	136131316	T/C	0.982	ncRNA	*ABO*	1.38 (0.77-2.48)	0.276	2.09 (1.14-3.81)	0.017
rs8176746	9	136131322	G/T	0.096	ncRNA	*ABO*	1.56 (1.18-2.08)	0.002	1.10 (0.81-1.50)	0.527
rs8176745	9	136131347	A/G	0.766	ncRNA	*ABO*	1.37 (1.13-1.66)	0.001	1.02 (0.82-1.26)	0.882
rs8176744	9	136131350	T/G	0.967	ncRNA	*ABO*	1.72 (1.10-2.67)	0.017	1.37 (0.86-2.19)	0.188
rs8176743	9	136131415	C/T	0.099	ncRNA	*ABO*	1.52 (1.14-2.02)	0.004	1.07 (0.79-1.46)	0.654
rs8176742	9	136131437	T/C	0.760	ncRNA	*ABO*	1.39 (1.15-1.68)	$8.42x10^{-4}$	1.01 (0.81-1.25)	0.959
rs8176741	9	136131461	G/A	0.094	ncRNA	*ABO*	1.57 (1.18-2.08)	0.002	1.14 (0.84-1.54)	0.399
rs8176740	9	136131472	T/A	0.759	ncRNA	*ABO*	1.34 (1.10-1.63)	0.003	0.98 (0.79-1.23)	0.882
rs8176739	9	136131523	A/G	0.983	ncRNA	*ABO*	1.63 (0.86-3.08)	0.136	1.25 (0.65-2.41)	0.506
rs7853989	9	136131592	G/C	0.093	ncRNA	*ABO*	1.37 (1.03-1.81)	0.029	0.96 (0.71-1.31)	0.808
rs1053878	9	136131651	G/A	0.046	ncRNA	*ABO*	1.24 (0.85-1.82)	0.266	0.89 (0.60-1.34)	0.584
rs8176720	9	136132873	C/T	0.618	ncRNA	*ABO*	1.08 (0.91-1.29)	0.374	1.00 (0.84-1.20)	0.979
rs75179845	9	136132954	T/C	0.098	intronic	*ABO*	1.49 (1.11-2.00)	0.008	1.07 (0.78-1.46)	0.684
rs8176718	9	136132957	T/C	0.766	intronic	*ABO*	1.39 (1.15-1.68)	$7.46x10^{-4}$	1.03 (0.83-1.27)	0.821
rs512770	9	136133506	A/G	0.819	ncRNA	*ABO*	1.36 (1.11-1.67)	0.003	1.11 (0.89-1.38)	0.348

Supplemental Table 4. Continued

							Discovery analyses		$Conditional analyses	
rsID	**Chr**	**Position**	**A1/A2**	**RAF**	**Functional class**	**Gene**	**OR (95%CI)**	**P**	**OR (95%CI)**	**P**
rs549443	9	136135237	A/G	0.789	ncRNA	*ABO*	1.32 (1.08-1.61)	0.006	0.97 (0.78-1.21)	0.792
rs549446	9	136135238	T/C	0.785	ncRNA	*ABO*	1.38 (1.14-1.67)	0.001	1.00 (0.81-1.25)	0.968
rs688976	9	136136770	A/C	0.803	ncRNA	*ABO*	1.31 (1.06-1.61)	0.012	0.97 (0.77-1.22)	0.807
rs8176696	9	136136773	T/C	0.980	ncRNA	*ABO*	1.65 (0.91-2.97)	0.098	1.15 (0.62-2.13)	0.656

Chr chromosome; A1 reference allele; A2 risk allele; RAF risk allele frequency; UTR untranslated region; ncRNA non-coding RNA; OR odds ratio; CI confidence interval; NA not applicable

** F11* variants rs2289252, rs4253429, rs4253430, rs4253865, and rs1062547 also map to a non-coding RNA transcript of *LOC285441 (F11-AS1).*

$ We conducted conditional logistic regression analyses in which we adjusted for the lead variant per locus (i.e. *F5* rs6672595, *FGA* rs6050, *F11* rs2036914, and *ABO* rs8176719).

Supplemental Table 5. Reciprocal conditional association analyses rs6050 and rs2066865

rsID	**Chr**	**Position**	**A1/A2**	**RAF**	**Class**	**Gene**	***LD R^2 (D')**	**OR (95% CI)**	**P**	**OR_{cond} (95% CI)**	**P_{cond}**
rs6050	4	155507590	T/C	0.393	missense	*FGA*	0.90 (0.99)	1.66 (1.37-2.02)	$2.33x10^{-7}$	1.24 (0.73-2.12)	0.430
rs2066865	4	155525276	G/A	0.352	downstream	*FGG*	0.90 (0.99)	1.60 (1.33-1.92)	$4.86x10^{-7}$	1.37 (0.81-2.31)	0.245

Chr chromosome; A1 reference allele; A2 risk allele; RAF risk allele frequency; LD linkage disequilibrium; OR odds ratio; CI confidence interval; cond conditional

* Linkage disequilibrium based on data of European population of 1000 Genomes Project.

Supplemental Table 6. Association analysis at the *F5* association locus conditional on rs4524

								Discovery analyses		Conditional analyses[$]	
rsID	Chr	Position	A1/A2	RAF	Functional class	[*]LD R^2 (D')	Gene	OR (95% CI)	P	OR (95% CI)	P
rs2009814	1	169471917	T/C	0.721	intergenic	0.96 (0.99)	.	1.42 (1.18-1.71)	$2.29x10^{-4}$	1.91 (0.73-5.00)	0.186
rs974793	1	169478654	T/C	0.712	intergenic	0.96 (0.99)	.	1.36 (1.13-1.64)	$1.44x10^{-3}$	1.32 (0.47-3.69)	0.596
rs2187952	1	169481950	A/G	0.697	3' UTR	0.96 (0.99)	*F5*	1.39 (1.14-1.68)	$9.44x10^{-4}$	1.21 (0.30-4.98)	0.789
rs4656685	1	169483844	T/C	0.746	intronic	0.96 (0.99)	*F5*	1.40 (1.17-1.68)	$3.02x10^{-4}$	0.74 (0.20-2.67)	0.645
rs2227244	1	169489358	C/T	0.728	intronic	0.96 (0.99)	*F5*	1.36 (1.13-1.63)	$1.12x10^{-3}$	0.77 (0.13-4.72)	0.778
rs2213867	1	169489585	C/T	0.727	intronic	0.96 (0.99)	*F5*	1.38 (1.15-1.66)	$6.34x10^{-4}$	0.59 (0.10-3.29)	0.544
rs9332655	1	169490592	G/A	0.730	intronic	0.97 (1.00)	*F5*	1.36 (1.13-1.64)	$1.03x10^{-3}$	0.29 (0.03-2.68)	0.277
rs9332652	1	169491021	AT/A	0.706	intronic	0.97 (1.00)	*F5*	1.35 (1.14-1.60)	$6.30x10^{-4}$	1.06 (0.61-1.83)	0.847
rs9332627	1	169497820	A/G	0.720	intronic	0.97 (1.00)	*F5*	1.41 (1.17-1.71)	$2.80x10^{-4}$	0.39 (0.04-3.78)	0.414
rs2420373	1	169498181	T/C	0.709	intronic	0.97 (1.00)	*F5*	1.44 (1.19-1.75)	$1.70x10^{-4}$	NA	NA
rs2187953	1	169499381	C/A	0.732	intronic	0.97 (1.00)	*F5*	1.40 (1.17-1.68)	$3.25x10^{-4}$	0.38 (0.04-3.73)	0.407
rs9332620	1	169499951	T/C	0.717	intronic	0.98 (1.00)	*F5*	1.39 (1.15-1.68)	$5.88x10^{-4}$	NA	NA
rs9332619	1	169500348	A/G	0.696	intronic	0.98 (1.00)	*F5*	1.37 (1.13-1.66)	$1.43x10^{-3}$	0 (0-∞)	0.968
rs6670393	1	169502533	C/A	0.723	intronic	0.98 (1.00)	*F5*	1.35 (1.12-1.62)	$1.50x10^{-3}$	NA	NA
rs10800453	1	169507076	A/T	0.772	intronic	1.00 (1.00)	*F5*	1.41 (1.16-1.71)	$5.24x10^{-4}$	NA	NA
rs9287090	1	169510380	A/G	0.714	synonymous	1.00 (1.00)	*F5*	1.36 (1.12-1.64)	$1.65x10^{-3}$	1.48 (0.09-24)	0.782
rs6032	1	169511555	C/T	0.725	missense	1.00 (1.00)	*F5*	1.37 (1.14-1.65)	$7.90x10^{-4}$	NA	NA
rs6021	1	169512027	C/T	0.729	synonymous	1.00 (1.00)	*F5*	1.36 (1.13-1.63)	$1.35x10^{-3}$	NA	NA

Supplemental Table 6. Continued

								Discovery analyses		Conditional analyses[$]	
rsID	**Chr**	**Position**	**A1/A2**	**RAF**	**Functional class**	**[*]LD R^2 (D')**	**Gene**	**OR (95% CI)**	**P**	**OR (95% CI)**	**P**
rs6016	1	169512120	A/G	0.710	synonymous	1.00 (1.00)	*F5*	1.40 (1.16-1.70)	$4.71x10^{-4}$	NA	NA
rs2239851	1	169512497	A/C	0.721	intronic	1.00 (1.00)	*F5*	1.38 (1.14-1.66)	$7.59x10^{-4}$	NA	NA
rs6675244	1	169512562	C/T	0.731	intronic	0.99 (1.00)	*F5*	1.37 (1.14-1.64)	$8.40x10^{-4}$	NA	NA
rs6662593	1	169512594	A/G	0.724	intronic	1.00 (1.00)	*F5*	1.36 (1.13-1.63)	$1.21x10^{-3}$	NA	NA
rs6662696	1	169512651	A/G	0.725	intronic	1.00 (1.00)	*F5*	1.35 (1.12-1.62)	$1.49x10^{-3}$	NA	NA
rs9332600	1	169512913	T/C	0.727	intronic	1.00 (1.00)	*F5*	1.43 (1.19-1.72)	$1.50x10^{-4}$	1.94 (0.17-22)	0.592
rs9287092	1	169513436	A/C	0.747	intronic	0.86 (0.94)	*F5*	1.41 (1.17-1.71)	$2.95x10^{-4}$	1.44 (0.85-2.43)	0.171
rs9332595	1	169514355	C/G	0.758	intronic	0.77 (0.93)	*F5*	1.51 (1.24-1.83)	$4.10x10^{-5}$	1.81 (1.21-2.72)	$4.17x10^{-3}$
rs929130	1	169514779	A/G	0.760	intronic	0.77 (0.93)	*F5*	1.55 (1.27-1.89)	$1.42x10^{-5}$	1.59 (1.07-2.37)	0.022
rs3766110	1	169515183	C/A	0.774	intronic	0.77 (0.93)	*F5*	1.54 (1.27-1.86)	$1.07x10^{-5}$	1.60 (1.09-2.36)	0.017
rs3766111	1	169515204	C/T	0.773	intronic	0.77 (0.93)	*F5*	1.57 (1.29-1.91)	$6.82x10^{-6}$	1.52 (1.02-2.27)	0.041
rs3766112	1	169515296	G/C	0.748	intronic	0.77 (0.93)	*F5*	1.55 (1.27-1.90)	$1.53x10^{-5}$	1.77 (1.17-2.69)	0.007
rs3766113	1	169515307	G/A	0.770	intronic	0.77 (0.93)	*F5*	1.55 (1.28-1.88)	$8.59x10^{-6}$	1.57 (1.06-2.32)	0.024
rs6672595	1	169515536	T/C	0.757	intronic	0.77 (0.93)	*F5*	1.58 (1.29-1.92)	$6.11x10^{-6}$	1.65 (1.10-2.48)	0.016
rs13306345	1	169515874	A/T	0.775	intronic	0.75 (0.90)	*F5*	1.46 (1.20-1.77)	$1.33x10^{-4}$	1.63 (1.09-2.43)	0.016
rs1894695	1	169517833	G/C	0.763	intronic	0.75 (0.90)	*F5*	1.47 (1.21-1.78)	$9.68x10^{-5}$	1.72 (1.15-2.56)	0.008
rs1894696	1	169517975	T/C	0.784	intronic	0.75 (0.90)	*F5*	1.44 (1.18-1.76)	$2.71x10^{-4}$	1.47 (1.00-2.15)	0.051
rs72248387	1	169518819	T/TCA	0.766	intronic	0.75 (0.90)	*F5*	1.52 (1.25-1.84)	$2.56x10^{-5}$	1.47 (1.00-2.17)	0.052
rs10158595	1	169520364	T/C	0.765	intronic	0.74 (0.89)	*F5*	1.49 (1.23-1.81)	$5.22x10^{-5}$	1.50 (1.02-2.20)	0.038

Supplemental Table 6. Continued

rsID	Chr	Position	A1/A2	RAF	Functional class	*LD R^2 (D')	Gene	Discovery analyses OR (95% CI)	Discovery analyses P	Conditional analyses$ OR (95% CI)	Conditional analyses$ P
rs2420375	1	169520459	C/G	0.796	intronic	0.74 (0.89)	*F5*	1.41 (1.16-1.72)	$5.98x10^{-4}$	1.43 (0.98-2.08)	0.065
rs2420376	1	169520549	A/G	0.788	intronic	0.74 (0.89)	*F5*	1.45 (1.20-1.76)	$1.29x10^{-4}$	1.52 (1.04-2.21)	0.03
rs2420377	1	169520592	A/G	0.788	intronic	0.74 (0.89)	*F5*	1.42 (1.17-1.72)	$4.12x10^{-4}$	1.46 (0.99-2.15)	0.055

Chr chromosome; A1 reference allele; A2 risk allele; RAF risk allele frequency; LD linkage disequilibrium; OR odds ratio; CI confidence interval; cond conditional; UTR untranslated region; NA not applicable

* Linkage disequilibrium with rs4524, based on data of European population of 1000 Genomes Project.

$ Conditional analyses on rs4525

Supplemental Table 7. Suggestive single variant associations (FDR <0.10)

rsID	Chr.	Position	A1/A2	RAF	Functional class	Gene	Discovery analyses OR (95% CI)	Discovery analyses P	FDR	Conditional analyses$ OR (95% CI)	Conditional analyses$ P
rs2009814	1	169471917	T/C	0.721	intergenic	.	1.42 (1.18-1.71)	$2.29x10^{-4}$	0.032	1.12 (0.78-1.61)	0.544
rs974793	1	169478654	T/C	0.712	intergenic	.	1.36 (1.13-1.64)	0.001	0.090	1.04 (0.71-1.51)	0.846
rs2187952	1	169481950	A/G	0.697	3' UTR	*F5*	1.39 (1.14-1.68)	$9.44x10^{-4}$	0.072	0.97 (0.66-1.43)	0.88
rs4656685	1	169483844	T/C	0.746	intronic	*F5*	1.40 (1.17-1.68)	$3.02x1^{-4}$	0.035	0.95 (0.65-1.40)	0.80
rs2227244	1	169489358	C/T	0.728	intronic	*F5*	1.36 (1.13-1.63)	0.001	0.08	0.94 (0.63-1.39)	0.746
rs2213867	1	169489585	C/T	0.727	intronic	*F5*	1.38 (1.15-1.66)	$6.34x10^{-4}$	0.058	0.93 (0.63-1.37)	0.711
rs9332655	1	169490592	G/A	0.730	intronic	*F5*	1.36 (1.13-1.64)	0.001	0.075	0.91 (0.62-1.34)	0.629
rs9332652	1	169491021	AT/A	0.706	intronic	*F5*	1.35 (1.14-1.60)	$6.30x10^{-4}$	0.058	1.01 (0.73-1.38)	0.964

Supplemental Table 7. Continued

							Discovery analyses			Conditional analyses[$]	
rsID	Chr.	Position	A1/A2	RAF	Functional class	Gene	OR (95% CI)	P	FDR	OR (95% CI)	P
rs9332627	1	169497820	A/G	0.720	intronic	*F5*	1.41 (1.17-1.71)	$2.80x10^{-4}$	0.035	0.94 (0.64-1.39)	0.77
rs2420373	1	169498181	T/C	0.709	intronic	*F5*	1.44 (1.19-1.75)	$1.70x10^{-4}$	0.027	0.97 (0.65-1.44)	0.888
rs2187953	1	169499381	C/A	0.732	intronic	*F5*	1.40 (1.17-1.68)	$3.25x10^{-4}$	0.037	0.94 (0.64-1.39)	0.765
rs9332620	1	169499951	T/C	0.717	intronic	*F5*	1.39 (1.15-1.68)	$5.88x10^{-4}$	0.058	0.90 (0.60-1.33)	0.59
rs9332619	1	169500348	A/G	0.696	intronic	*F5*	1.37 (1.13-1.66)	0.001	0.09	0.96 (0.64-1.43)	0.841
rs6670393	1	169502533	C/A	0.723	intronic	*F5*	1.35 (1.12-1.62)	0.002	0.09	0.89 (0.60-1.31)	0.554
rs10800453	1	169507076	A/T	0.772	intronic	*F5*	1.41 (1.16-1.71)	$5.24x10^{-4}$	0.054	1.00 (0.65-1.55)	0.992
rs9287090	1	169510380	A/G	0.714	synonymous	*F5*	1.36 (1.12-1.64)	0.002	0.096	0.97 (0.66-1.43)	0.872
rs6032	1	169511555	C/T	0.725	missense	*F5*	1.37 (1.14-1.65)	$7.90x10^{-4}$	0.065	0.99 (0.66-1.46)	0.939
rs4525	1	169511734	C/T	0.736	missense	*F5*	1.38 (1.15-1.66)	$6.40x10^{-4}$	0.058	1.02 (0.69-1.51)	0.93
rs6021	1	169512027	C/T	0.729	synonymous	*F5*	1.36 (1.13-1.63)	0.001	0.088	0.95 (0.64-1.40)	0.787
rs6016	1	169512120	A/G	0.710	synonymous	*F5*	1.40 (1.16-1.70)	$4.71x10^{-4}$	0.05	0.97 (0.65-1.44)	0.875
rs2239851	1	169512497	A/C	0.721	intronic	*F5*	1.38 (1.14-1.66)	$7.59x10^{-4}$	0.065	0.97 (0.66-1.45)	0.90
rs6675244	1	169512562	C/T	0.731	intronic	*F5*	1.37 (1.14-1.64)	$8.40x10^{-4}$	0.066	0.95 (0.65-1.41)	0.815
rs6662593	1	169512594	A/G	0.724	intronic	*F5*	1.36 (1.13-1.63)	0.001	0.082	0.96 (0.64-1.42)	0.822
rs6662696	1	169512651	A/G	0.725	intronic	*F5*	1.35 (1.12-1.62)	0.001	0.09	0.95 (0.64-1.40)	0.788
rs9332600	1	169512913	T/C	0.727	intronic	*F5*	1.43 (1.19-1.72)	$1.50x10^{-4}$	0.025	0.99 (0.67-1.47)	0.96
rs9287092	1	169513436	A/C	0.747	intronic	*F5*	1.41 (1.17-1.71)	$2.95x10^{-4}$	0.035	0.69 (0.40-1.18)	0.175
rs9332595	1	169514355	C/G	0.758	intronic	*F5*	1.51 (1.24-1.83)	$4.10x10^{-5}$	0.009	NA	NA

Supplemental Table 7. Continued

rsID	Chr.	Position	A1/A2	RAF	Functional class	Gene	Discovery analyses			Conditional analyses[$]	
							OR (95% CI)	P	FDR	OR (95% CI)	P
rs929130	1	169514779	A/G	0.760	intronic	*F5*	1.55 (1.27-1.89)	$1.42x10^{-5}$	0.004	NA	NA
rs3766112	1	169515296	G/C	0.748	intronic	*F5*	1.55 (1.27-1.90)	$1.53x10^{-5}$	0.004	NA	NA
rs13306345	1	169515874	A/T	0.775	intronic	*F5*	1.46 (1.20-1.77)	$1.33x10^{-4}$	0.023	1.05 (0.09-11.7)	0.967
rs1894695	1	169517833	G/C	0.763	intronic	*F5*	1.47 (1.21-1.78)	$9.68x10^{-5}$	0.018	0.59 (0.06-5.78)	0.647
rs1894696	1	169517975	T/C	0.784	intronic	*F5*	1.44 (1.18-1.76)	$2.71x10^{-4}$	0.035	0.56 (0.06-5.49)	0.616
rs72248387	1	169518819	T/TCA	0.766	intronic	*F5*	1.52 (1.25-1.84)	$2.56x10^{-5}$	0.006	0.61 (0.06-6.01)	0.673
rs10158595	1	169520364	T/C	0.765	intronic	*F5*	1.49 (1.23-1.81)	$5.22x10^{-5}$	0.01	0.62 (0.12-3.23)	0.567
rs2420375	1	169520459	C/G	0.796	intronic	*F5*	1.41 (1.16-1.72)	$5.98x10^{-4}$	0.058	0.59 (0.11-3.07)	0.528
rs2420376	1	169520549	A/G	0.788	intronic	*F5*	1.45 (1.20-1.76)	$1.29x10^{-4}$	0.023	0.60 (0.11-3.13)	0.543
rs2420377	1	169520592	A/G	0.788	intronic	*F5*	1.42 (1.17-1.72)	$4.12x10^{-4}$	0.045	0.31 (0.04-2.65)	0.282
rs2070011	4	155511897	C/T	0.406	5' UTR	*FGA*	1.37 (1.16-1.62)	$2.19x10^{-4}$	0.031	0.90 (0.67-1.21)	0.478
rs13146272	4	187120211	C/A	0.692	missense	*CYP4V2*	1.35 (1.12-1.61)	0.001	0.083	1.20 (0.99-1.46)	0.07
rs3817184	4	187122304	C/T	0.456	intronic	*CYP4V2*	1.43 (1.21-1.69)	$2.24x10^{-5}$	0.005	1.21 (1.00-1.46)	0.054
rs2289252*	4	187207381	C/T	0.528	intronic	*F11*	1.39 (1.17-1.65)	$1.94x10^{-4}$	0.029	1.15 (0.92-1.44)	0.211
rs4253430*	4	187210064	C/G	0.590	3' UTR	*F11*	1.38 (1.16-1.63)	$2.50x10^{-4}$	0.033	1.05 (0.83-1.31)	0.694
rs1062547*	4	187210247	T/A	0.605	3' UTR	*F11*	1.33 (1.13-1.57)	$7.97x10^{-4}$	0.065	1.01 (0.80-1.26)	0.954
rs8176748	9	136131289	T/C	0.774	ncRNA	*ABO*	1.37 (1.13-1.66)	0.001	0.081	1.01 (0.82-1.26)	0.898
rs8176745	9	136131347	A/G	0.766	ncRNA	*ABO*	1.37 (1.13-1.66)	0.001	0.081	1.02 (0.82-1.26)	0.882
rs8176742	9	136131437	T/C	0.760	ncRNA	*ABO*	1.39 (1.15-1.68)	$8.42x10^{-4}$	0.066	1.01 (0.81-1.25)	0.959

Supplemental Table 7. Continued

rsID	Chr.	Position	A1/A2	RAF	Functional class	Gene	Discovery analyses OR (95% CI)	P	FDR	Conditional analyses[$] OR (95% CI)	P
rs8176718	9	136132957	T/C	0.766	intronic	*ABO*	1.39 (1.15-1.68)	$7.46x10^{-4}$	0.065	1.03 (0.83-1.27)	0.821
rs549446	9	136135238	T/C	0.785	unknown	*ABO*	1.38 (1.14-1.67)	0.001	0.075	1.00 (0.81-1.25)	0.968

Chr. chromosome; A1 reference allele; A2 risk allele; RAF risk allele frequency; UTR untranslated region; ncRNA; non-coding RNA; OR odds ratio; CI confidence interval; FDR false discovery rate; NA not applicable

** F11* variants rs2289252, rs4253430, and rs1062547 also map to a non-coding RNA transcript of *LOC285441 (F11-AS1).*

[$] We conducted conditional logistic regression analyses in which we adjusted for the lead variant per locus (i.e. *F5* rs6672595, *FGA* rs6050, *F11* rs2036914, and *ABO* rs8176719).

Supplemental Table 8. Single-variant association analyses of common *MASP1* variants

rsID	Chr.	Position	A1/A2	RAF	Functional class	Gene	[*]LD R^2 (D')	OR (95%CI)	P
rs62294422	3	186952037	G/A	0.067	3' UTR	*MASP1*		1.06 (0.78-1.45)	0.707
rs12489890	3	186952415	A/G	0.821	3' UTR	*MASP1*	0.00 (1.00)	1.04 (0.84-1.28)	0.730
rs72549168	3	186952569	C/T	0.016	3' UTR	*MASP1*	1.00 (1.00)	3.11 (1.42-6.81)	0.005
rs850314	3	186952587	T/C	0.622	3' UTR	*MASP1*	0.01 (1.00)	1.03 (0.87-1.21)	0.766
rs1109452	3	186952588	A/G	0.699	3' UTR	*MASP1*	0.01 (1.00)	1.15 (0.97-1.38)	0.117
rs72549262	3	186952914	C/G	0.905	3' UTR	*MASP1*	0.00 (1.00)	1.12 (0.85-1.49)	0.424
rs874603	3	186953037	A/G	0.059	3' UTR	*MASP1*	0.00 (1.00)	1.10 (0.79-1.53)	0.582
rs850313	3	186953226	C/G	0.746	3' UTR	*MASP1*	0.01 (1.00)	1.17 (0.98-1.40)	0.083
rs78393224	3	186953244	T/A	0.976	3' UTR	*MASP1*	0.00 (1.00)	1.53 (0.90-2.60)	0.118

Supplemental Table 8. Continued

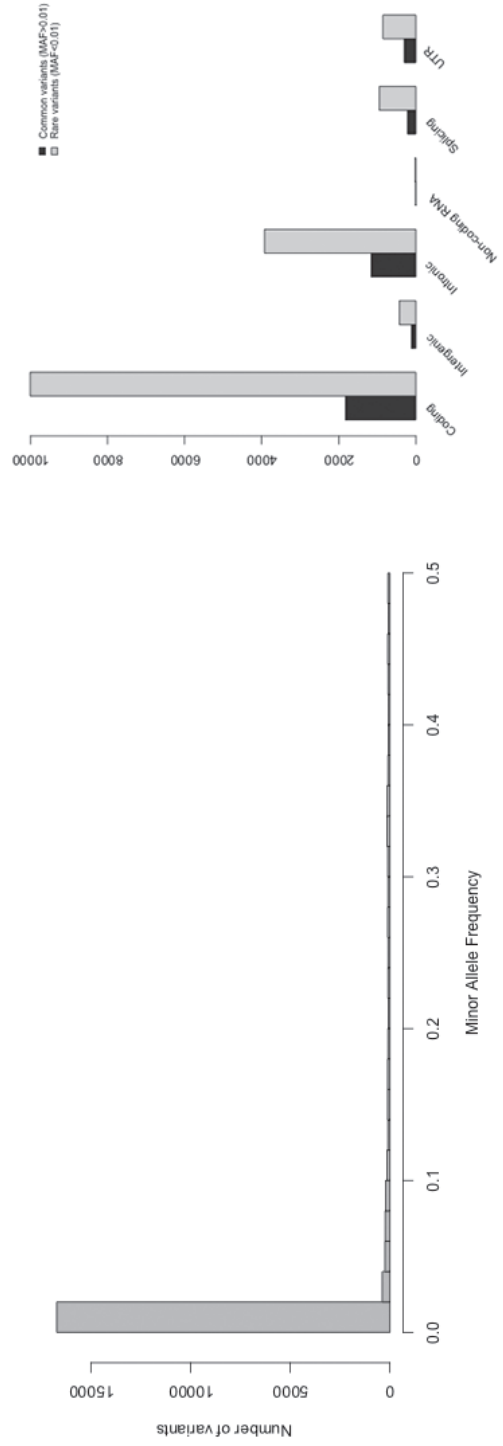
rsID	Chr.	Position	A1/A2	RAF	Functional class	Gene	*LD R^2 (D')	OR (95%CI)	P
rs67143992	3	186953321	T/C	0.816	3' UTR	*MASP1*	0.00 (1.00)	1.14 (0.93-1.40)	0.196
rs850312	3	186953808	C/T	0.364	synonymous	*MASP1*	0.00 (0.27)	1.05 (0.89-1.24)	0.548
rs72549154	3	186953932	C/A	0.028	missense	*MASP1*	0.00 (1.00)	1.04 (0.66-1.65)	0.861
rs3774268	3	186954324	G/A	0.133	synonymous	*MASP1*	0.01 (0.38)	1.04 (0.82-1.32)	0.775
rs698090	3	186964300	C/T	0.681	3' UTR	*MASP1*	0.01 (0.85)	1.01 (0.85-1.21)	0.877
rs16848736	3	186964312	T/C	0.023	3' UTR	*MASP1*	0.00 (1.00)	1.23 (0.71-2.12)	0.455
rs72549254	3	187009412	A/G	0.812	Splicing	*MASP1*	0.00 (0.62)	1.17 (0.93-1.46)	0.186

Chr. chromosome; A1 reference allele; A2 risk allele; RAF risk allele frequency; UTR untranslated region; LD linkage disequilibrium; OR odds ratio; CI confidence interval
* Linkage disequilibrium with rs72549167, based on data of European population of 1000 Genomes Project.

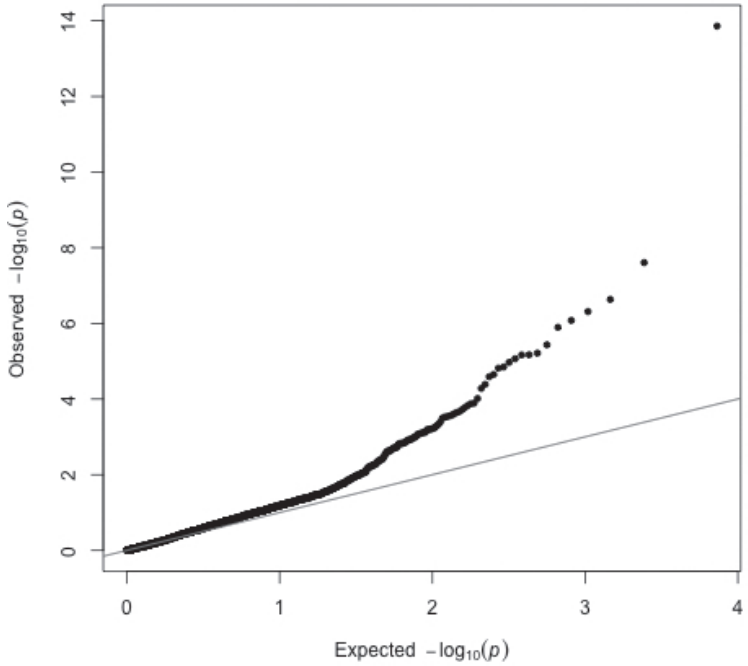
Supplemental Table 9. Single-variant association analyses of common CBS variants

rsID	Chr.	Position	A1/A2	RAF	Functional class	Gene	*LD R^2 (D')	OR (95%CI)	P
rs9978104	21	44473980	T/G	0.894	3' UTR	*CBS*	0.06 (1.00)	1.16 (0.88-1.52)	0.295
rs234706	21	44485350	A/G	0.640	synonymous	*CBS*	0.17 (0.79)	1.15 (0.97-1.36)	0.110

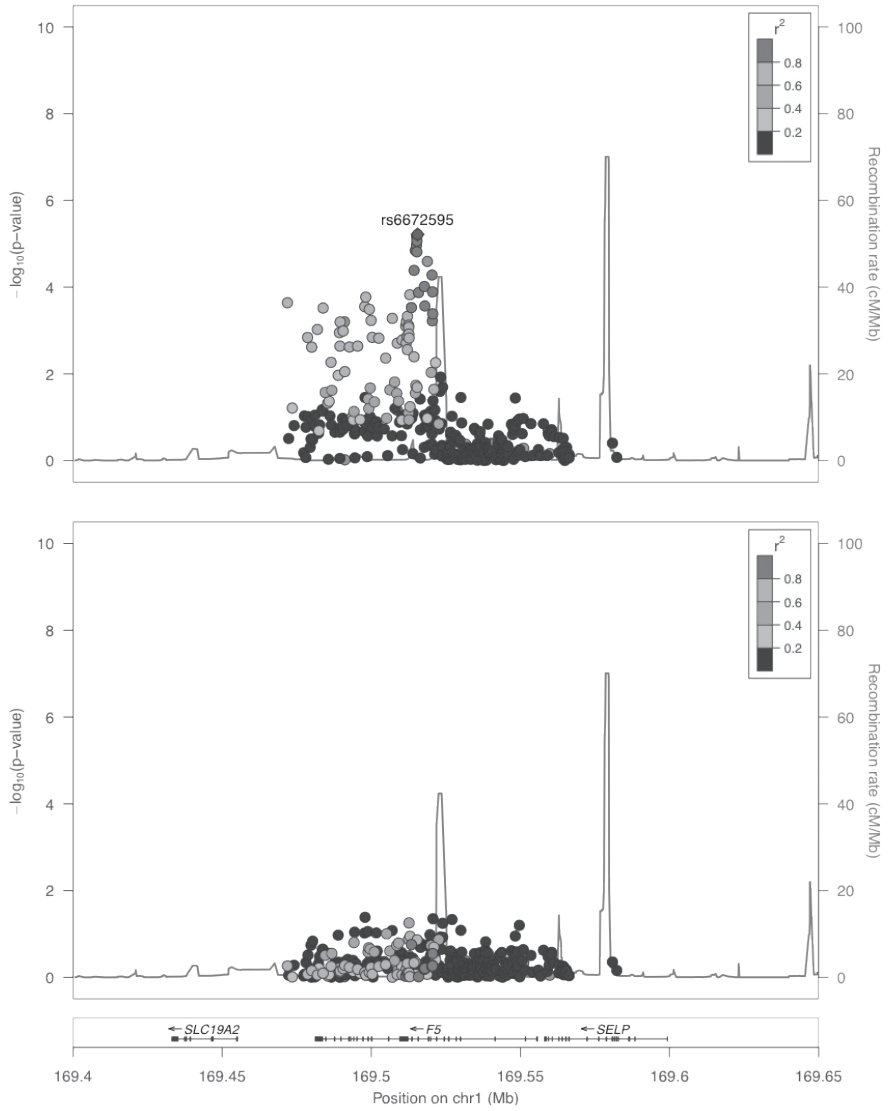
Chr. chromosome; A1 reference allele; A2 risk allele; RAF risk allele frequency; UTR untranslated region; LD linkage disequilibrium; OR odds ratio; CI confidence interval
* Linkage disequilibrium with rs1801181, based on data of European population of 1000 Genomes Population.



Supplemental Figure 1. Minor allele frequency distribution of identified variants (left) and overview of functional classes (right)



Supplemental Figure 2. Quantile-Quantile plot of the single variant association analyses



Supplemental Figure 3. Regional association plots for single variants in the *F5* region

Regional association plots showing single variant association results between common variants in the *F5* region and DVT risk before (upper panel) and after conditioning (lower panel) on lead variant rs6672595.

CHAPTER 3

Genome-wide association study identifies a novel genetic risk factor for recurrent venous thrombosis

de Haan HG, van Hylckama Vlieg A, Germain M, Baglin TP, Deleuze JF, Trégouët DA, Rosendaal FR.

Circ Genom Precis Med. 2018;11:e001827.

ABSTRACT

Background

Genetic risk factors for a first venous thrombosis (VT) seem to have little effect on recurrence risk. Therefore, we aimed specifically to identify novel genetic determinants of recurrent VT. So far, genome-wide association studies are lacking.

Methods and Results

We performed a genome-wide association scan in 1279 patients from the MEGA follow-up study; 832 patients with a first VT only and 447 recurrent VT patients. We analysed genotype probabilities of about 8.6 million variants, imputed to the Genome of the Netherlands project reference panel, with a minor allele frequency ≥ 1% for an association with recurrent VT. One region exceeded genome-wide significance (P-value ≤ $5x10^{-8}$), mapping to the well-known FV Leiden locus. Conditional association analyses on FV Leiden did not yield any secondary association signals. We also identified 52 suggestive association signals (P-value < $1x10^{-5}$) at 17 additional loci. None of these loci were previously implicated in VT risk. Replication analyses for 17 lead variants were performed in 350 recurrent VT patients and 1866 patients with a single VT event from the MEGA follow-up study, THE-VTE study, and LETS study. We observed an association with recurrence for an intergenic variant at 18q22.1 with an odds ratio of 1.7 (95% CI 1.2-2.6) per copy of the minor allele.

Conclusions

We confirmed the association of FV Leiden and identified a novel risk locus at 18q22.1 in the first large genetic study on recurrent VT.

INTRODUCTION

Approximately 20 to 30% of patients with a first venous thrombosis (VT) develop a recurrence within five years of the first event,[1,2] and therefore predicting and preventing recurrence is of crucial importance. However, risk factors for a first event do no predict recurrence well and hence risk profiling is difficult.[3-6] Recurrence risk is the highest amongst patients whose thrombotic event was not provoked by transient risk factors such as surgery and immobilization.[1,2,7-9] In particular, previous studies have shown that patients with a first unprovoked event have a two to three-fold increased risk of recurrence compared with patients with a first provoked event.[7-9] This suggests that patients with recurrent VT are enriched for genetic risk factors. The minor effects of determinants of first events on recurrence on the relative risk scale can be explained by the difference in absolute risks of first and recurrent VT, and index event bias.[10,11]

In addition, different genetic variants may play a role in recurrence than in first thrombosis, for example factors that affect clot lysis or the recanalization of the vein after a thrombotic event. So far, few studies have focused on recurrence-specific genetic risk factors. Zee and colleagues studied a panel of 86 variants in 56 candidate genes and observed suggestive associations with recurrent VT for four variants.[12] In addition, homozygosity of Ser128Arg in the E-selectin gene and length of a GT-dinucleotide repeat in the promoter of the gene encoding heme oxygenase 1 have been linked to recurrent VT in an Austrian study.[13,14] However, none of these findings have sofar been confirmed in large independent studies.

In order to identify novel genetic determinants of recurrent VT, we performed the first genome-wide association study (GWAS) on recurrence in 447 patients with recurrent VT and a sample of 832 patients who remained recurrence-free in the Multiple Environmental and Genetic Assessment of risk factors for venous thrombosis (MEGA) follow-up study.[15] To validate our findings, we additionally performed a replication study of the newly identified risk variants among 350 recurrent VT patients and a sample of 1866 patients with a single event only from three cohort studies.

MATERIAL AND METHODS

GWAS analysis

Study population

We included patients from the MEGA follow-up study, a large population-based cohort study on risk factors for recurrent VT. Details of this study have been described elsewhere.[15] In short, 4956 patients with a first deep vein thrombosis (DVT) of the leg or pulmonary embolism (PE), who were enrolled in the MEGA case-control study between 1999 and 2004[16], were invited to participate. Follow-up started at the date of the first event. Between 2008 and 2009, questionnaires related to recurrent VT were sent to the patients. Occurrence of recurrent VT was determined by information from patients, anticoagulation clinics, and treating physicians according to a decision rule.[15] Follow-up ended when a recurrent VT occurred, the patient died or migrated, or when the questionnaire was returned, whichever occurred first. For the patients who died, information on the cause of death was retrieved from the national registry of death certificates. If no questionnaire was returned, patients were considered lost to follow-up.

For the GWAS analysis, 1499 patients were selected according to the following process (Flow diagram is shown in Supplemental Figure 1). First, patients who had not provided a high-quality blood sample or buccal swap for DNA analysis were excluded (667 out of 4956 eligible patients). In addition, we excluded all patients who had been diagnosed with cancer (N=457). We then selected all patients for whom a recurrent VT event was reported at time of sample selection for the current analysis (N=542). Of these, 16 recurrences were classified as uncertain recurrences according to the decision rule,[15] and these patients were subsequently analyzed as recurrence-free patients. In addition, we randomly sampled 957 patients, totaling 973 patients who remained without a recurrent event during a median period of 7.1 years (interquartile range [IQR] 5.5-8.4). Follow-up was incomplete for 19.5% of these patients, as some died without recurrence (N=11), whereas others were last seen at the anticoagulation clinic (N=77) or at time of blood sampling for the MEGA case-control study (N=102). Patients with incomplete follow-up were followed for a median period of 312 days (range 60 days to 9.7 years). As these patients did not or no longer visit the anticoagulation clinic, which monitor anticoagulant treatment, it is unlikely that these patients suffered from a recurrent

VT and, therefore, these patients were considered as recurrence-free patients in the GWAS analysis. We performed a sensitivity analysis for the top GWAS findings in which we excluded patients with incomplete follow-up and patients who had an uncertain recurrent event.

This study was approved by the Medial Ethics Committee of the Leiden University Medical Center, and all participants gave written informed consent.

GWAS quality-control and imputation

Genome-wide genotyping was performed with the Illumina Human660-Quad v.1 BeadChip (Illumina Inc., San Diego, USA) at Centre National de Génotypage (Institut de Génomique, Evry, France). Genotyping was successfully completed for 1461 patients, of whom 1426 had a call rate of at least 98%. Additional exclusions at the individual level included discrepancy between self-reported and genotypic sex, abnormal level of autosomal heterozygosity (false discovery rate <1%), and ethnic outliers based on multidimensional scaling analysis of the identity-by-state matrix. Furthermore, 32 patients withdrew their consent for the MEGA follow-up study, leaving a total of 1279 patients for imputation and association analyses (447 patients with a recurrence during follow-up and 832 recurrence-free patients). The following exclusions were applied to identify a final set of 497,563 high-quality variants: minor allele frequency (MAF) below 1%, genotyping call rate below 98%, significant deviation from Hardy-Weinberg equilibrium (P-value $<1x10^{-6}$) in patients with a first event only. All quality-control procedures were performed with the R-package GenABEL.[17]

Following the conversion of the genomic positions from hg18 to hg19 using the UCSC Genome Browser LiftOver tool, imputation of 19.6 million autosomal variants was performed using IMPUTE2 software[18] according to the Genome of the Netherlands reference panel (GoNL release 4).[19] Prior to the association analyses, we excluded variants with a MAF below 1% or an imputation quality score I below 0.5.

Statistical analysis

Imputed genotypes of 8.6 million variants were tested for an association with recurrent VT using SNPTEST version 2[20] by means of logistic regression with the missing data likelihood score test, which takes the uncertainty of the imputed genotypes into account. All analyses were adjusted for age and sex. We assumed an additive mode of

inheritance. The level of genome-wide significance was set at P-value $<5x10^{-8}$, whereas the threshold for highly suggestive association signals was set at P-value $<1x10^{-5}$. In order to identify independent secondary association signals at a locus, we performed conditional analyses on the lead variant or the previously reported VT risk variant. In addition, we grouped associated variants in clumps based on linkage disequilibrium (LD) and genomic distance according to standard settings in PLINK.[21] Regional association plots were created with LocusZoom[22] and functional annotation of the variants was performed with AnnoVar.[23]

The quantile-quantile plot of the genome-wide test statistics against the expected null distribution showed no appreciable evidence of inflation due to population stratification or genotyping artefacts (Supplemental Figure 2). Likewise, the genomic inflation factor (lambda[24]) before and after imputation was 1.033 and 1.001, respectively. None of the first four principal components were associated with recurrent VT, and these were therefore not included as covariates in the association analyses.

Look-up of previously reported risk variants

In order to validate previously reported genetic associations with (recurrent) venous thrombosis that may not have attained genome-wide significance in our study, we specifically explored the association results for 17 variants. Selected variants were either previously shown to be associated with recurrence only[12-14,25] or reached genome-wide significance in one of the two recent GWAS studies on first VT.[26,27] Effects were calculated per copy of the risk allele based on the reporting in the original studies. Additional information on the selected variants is provided in Supplemental Table 1. Two variants (rs3025058 and rs3074372) could not be studied due to the absence of (tagging) variants in the GWAS, one variant (rs114209171) could not be studied as it was located on the X chromosome.

Replication analysis

Study population

The replication analysis was conducted in 350 patients with recurrent VT and a sample of 1866 patients with a single event only. These individuals were included from three European studies into VT risk, that is the MEGA follow-up study, the Leiden Thrombophilia Study (LETS) study[4], and the Thrombophilia, Hypercoagulability and

Environmental Risks in Venous Thromboembolism (THE-VTE) study[28]. From the MEGA follow-up study, we included 155 recurrent VT patients who had not been included in the original GWAS or who were excluded during the quality-control procedures of the GWAS. In addition, we randomly sampled 929 patients with a single VT event only, of whom 72.9% had complete follow-up.

LETS and THE-VTE study are both population-based case-control studies into risk factors for VT with subsequent follow-up of the VT patients. The study designs are similar to that of the MEGA study and have been described in detail previously.[4,28] In LETS, 474 consecutive patients with a first DVT in the leg or arm were recruited at three anticoagulation clinics in or near Leiden. Patients were subsequently followed for recurrence until 2000 using repeated questionnaires. Follow-up started 90 days after the date of the first event and ended at the date of recurrence, date of death, date of emigration, or the end of the study, whichever occurred first.[4] A total of 471 patients had a DNA sample available for genotyping. Of these, 90 patients developed a recurrence during a median follow-up of 8.0 years (IQR 6.8-9.0). Follow-up was complete for 88.2% of the recurrence-free patients. THE-VTE is a two-center case-control study, in which 796 consecutive patients with a first VT were enrolled in Leiden and Cambridge (UK).[28] Patients were subsequently followed for recurrence starting at the date of the first event. In Leiden, follow-up ended when a recurrent event occurred, when a patient died or migrated, or when patients were untraceable, whichever occurred first. For patients included in Cambridge, recurrence status was checked on 1 July 2013 using hospital records. In the absence of recurrence or death, this date was registered as the end of follow-up. For the current analysis, we excluded patients who did not have a DNA sample available (N=135). During a median follow-up of 5.4 years (IQR 4.2-6.6), 105 of the 661 patients experienced a recurrent VT event. Follow-up was complete for 88.5% of the patients with a single VT event only. In both LETS and THE-VTE, individuals with a recent cancer diagnosis were not enrolled.

All participants gave written informed consent. The THE-VTE and LETS study were both approved by the Medical Ethics Committee of the Leiden University Medical Center. In addition, THE-VTE was also approved by the NHS Research Ethics Committee in Cambridge, UK.

Genotyping

For each novel locus that showed a highly significant association with recurrent VT in the discovery GWAS, we selected the lead variant or the variant with the largest functional impact. These variants were genotyped with predesigned or custom-made TaqMan assays (Life Technologies, Thermo Fisher Scientific, USA) according to manufacturer's specifications. Primer design failed for three variants (rs9834479 in *ROBO1*, rs61504683 in *LPPR3*, and rs111750150 in *TSPEAR*), which were subsequently replaced by variants in high LD (r^2>0.8) in our GWAS study population or based on the CEU 1000 Genomes population using SNAP software[29].

Statistical analysis

Association with recurrent VT was assessed using logistic regression analyses adjusting for age, sex, study, and study center in case of THE-VTE. Patients who were lost to follow-up were analyzed as recurrence-free. These patients remained without a recurrent event during a median follow-up period of 1.2 years (IQR 0.7-3.4). To account for multiple hypothesis testing, the threshold for statistical significance was set at 0.05 divided by the number of variants tested in the replication analyses. We also calculated the false discover rate (FDR). In addition, we performed a sub-analysis including only the patients from LETS and THE-VTE in a Cox regression model to calculate hazard ratios with 95% confidence intervals (95% CI). In this analysis, patients who were lost to follow-up were censored at the last date known to be recurrence-free. To ensure comparability of follow-up time between the LETS and THE-VTE study in the Cox regression analysis, we recalculated the follow-up time in THE-VTE to start 90 days after the date of the first event.

For the variant that replicated, we performed a meta-analysis of the results obtained in the replication cohorts and in the original GWAS in order to obtain the most robust estimate of its effect size. For this, we used a fixed-effects model based on inverse-variance weighting as implemented in the METAL software.[30] Heterogeneity was assessed by the Cochran's Q statistic and the I^2 index.

Discriminative value

To explore the potential clinical value of the two identified and validated genetic risk loci, we assessed the discriminative accuracy of two prediction models: a clinical model

and a combined model to which we added dosages of two genetic variants (rs6025 and rs9946608). The clinical model included sex, age, event type (DVT only versus PE with or without a DVT), and provoking status (recent surgery, trauma, immobilization, hormone use, pregnancy, and travel). We fitted both models in the GWAS population, which had complete clinical information for 1260 individuals (443 recurrence patients and 817 patients with a first VT only). Areas under the receiver-operating characteristic curves (AUC) were constructed using the predicted risks derived from logistic regression models. We calculated and compared the AUCs of the two prediction models using DeLong's test for correlated ROC curves as implemented in R package "pROC".[31]

RESULTS

GWAS analysis

Population characteristics

After quality-control assessments, 447 patients with a recurrent VT and 832 patients with a single VT event were included in the genome-wide association analyses. Overall, these patients had been followed for a median period of 6.1 years (IQR 2.2-7.9). Seventeen percent of the recurrence-free patients did not complete follow-up, as some died without recurrence (N=9) or had an uncertain recurrent event (N=10), whereas others were last seen at the anticoagulation clinic (N=46) or at time of blood sampling for the MEGA case-control study (N=75). The mean age at time of the first event was 48.1 years (standard deviation [sd] 12.9) and 49% of the patients was a man. Sixty-one percent of the patients had a first DVT of the leg, whereas 29% had a PE and 10% of the patients were diagnosed with both. Compared with patients with a single VT event, patients who experienced a recurrence were more often men and had more often a first unprovoked event (Table 1).

Table 1. Characteristics of GWAS study population

	Patients with a first VT only N=832	**Patients with a recurrent VT** N=447
Age at first event, mean years (SD)	47.0 (12.8)	50.2 (12.7)
Male sex, N (%)	339 (40.7)	287 (64.2)
Body mass index, kg/m^2	26.8 (4.7)	27.1 (4.5)
Smoking, N (%)	297 (35.7)	144 (32.9)
First event was unprovoked*, N (%)	248 (29.8)	220 (49.2)
Duration of anticoagulant therapy, median days (IQR)	183 (110-213)	185 (111-212)
Type of first event:		
DVT, N (%)	497 (59.7)	283 (63.3)
PE, N (%)	265 (31.9)	102 (22.8)
DVT and PE, N (%)	70 (8.4)	61 (13.8)

VT venous thrombosis, DVT deep vein thrombosis, PE pulmonary embolism, SD standard deviation, IQR interquartile range

*Provoking factors: recent surgery, immobilization (plaster cast, bedridden at home, hospitalization), hormone use, pregnancy or post-partum, and travel.

Association analyses

We assessed the association between 8.6 million variants and recurrent VT. The Manhattan plot of the GWAS results is shown in Supplemental Figure 3. Nineteen variants, all mapping to the *F5* region, were associated with recurrent VT at genome-wide significance (Supplemental Table 2). The lead variant mapped to a non-coding sequence in *F5* (rs2213868, MAF 14%, P-value 2.67x10^{-9}). The *F5* locus also included the established VT-associated variant FV Leiden (rs6025, MAF 9.6%, P-value 1.28x10^{-8}), of which the T-allele was associated with a 2.4-fold increased risk of recurrent VT (95% CI 1.75-3.15). Conditional analyses on rs6025 did not reveal any secondary association signals at the locus (Supplemental Figure 4). Of the genome-wide significant variants, the lowest remaining P-value was 0.02 for rs2213868 (Supplemental Table 2).

We additionally identified 52 variants that showed suggestive evidence of an association (P-value <1.0x10^{-5}) with recurrent VT (Supplemental Table 3). Of these, nine variants

were part of the *F5* locus and were no longer associated with recurrent VT when conditioning on FV Leiden. The other 43 variants mapped to 17 loci, mainly at non-coding sequence. None of the variants or gene regions have previously been implicated in the risk of recurrent or a first VT. We did not identify independent association signals at any of these loci when conditioning on the lead variant of each locus (data not shown). The effect estimates of the lead variants did not materially change in a sensitivity analysis excluding patients who were lost to follow-up, although confidence intervals became wider due to the smaller sample size (Supplemental Table 4). Likewise, all lead variants remained associated with recurrence risk, with similar effect sizes, in a sensitivity analysis adjusting for provoking status (Supplemental Table 5).

Furthermore, we aimed to replicate previous genetic associations with recurrent VT and to explore associations for variants recently reported in GWAS analyses on first VT. Results are reported in Table 2. We assessed the association of eight variants that reached genome-wide significance in two recent GWAS studies. Besides the association with FV Leiden, we observed a nominal association with recurrent VT for *FGG* rs2066865 (OR 1.30, 95% CI 1.09-1.56) and *F5* rs4524 (OR 1.25, 95% CI 1.02-1.54). The recently identified risk variants in *SCL44A2* and *TSPAN15* showed no evidence of an association with the risk of recurrence (rs2288904, OR 1.14, 95% CI 0.90-1.44 and rs78707713, OR 1.14, 95% CI 0.85-1.54, respectively). In addition, five variants that have previously been linked to recurrent VT risk were not associated with recurrence in the present GWAS analysis (Table 2).

Table 2. GWAS look-up of genetic variants previously associated with first or recurrent venous thrombosis

Literature	rs ID	Chr.	Position	Gene	A1/A2	EAF	Info	OR (95% CI)	P-value
First VT	rs4524	1	169511755	*F5*	C/T	0.800	1.00	1.25 (1.02-1.54)	0.032
	rs6025	1	169519049	*F5*	C/T	0.096	0.94	2.35 (1.75-3.15)	$1.28x10^{-8}$
	rs2066865	4	155525276	*FGG*	G/A	0.339	1.00	1.30 (1.09-1.56)	0.003
	rs4253417	4	187199005	*F11*	T/C	0.483	0.95	1.17 (0.99-1.39)	0.068
	rs529565	9	136149500	*ABO*	T/C	0.447	1.00	1.19 (1.00-1.42)	0.055
	rs78707713	10	71245276	*TSPAN15*	C/T	0.915	0.98	1.14 (0.85-1.54)	0.381
	rs1799963	11	46761055	*F2*	G/A	0.021	0.77	1.25 (0.64-2.42)	0.516
	rs2288904	19	10742170	*SLC44A2*	A/G	0.839	1.00	1.14 (0.90-1.44)	0.265
	rs6087685	20	33777612	*PROCR*	G/C	0.381	0.98	1.05 (0.89-1.25)	0.543
Recurrence	rs5361	1	169701060	*SELE*	T/G	0.121	1.00	1.14 (0.89-1.47)	0.296
	rs1799864	3	46399208	*CCR5*	G/A	0.067	0.83	0.91 (0.63-1.32)	0.622
	rs805297	6	31622606	*APOM*	C/A	0.313	0.98	1.07 (0.90-1.28)	0.447
	rs662	7	94937446	*PON1*	T/C	0.298	1.00	0.93 (0.77-1.11)	0.405
	rs1800775	16	56995236	*CETP*	C/A	0.493	1.00	1.03 (0.87-1.22)	0.718

Chr. chromosome, A1 reference allele, A2 effect allele, EAF effect allele frequency, Info imputation quality info score, OR odds ratio, CI confidence interval, VT venous thrombosis

Effects were calculated per copy of the risk allele, as reported in the original study, and adjusted for age and sex assuming an additive mode of inheritance.

Table 3. Main findings of GWAS and replication study for lead variants at highly suggestive loci of GWAS

					GWAS				Replication		
rs ID	**Chr**	**Position**	**(Nearest) Gene**	**A1/A2**	**info**	**MAF**	**OR (95% CI)**	**P-value**	**MAF**	**OR (95% CI)**	**P-value**
rs112349920	1	159933483	*LINC01133*	C/T	0.98	0.011	6.90 (3.06-15.6)	$3.36x10^{-6}$	0.006	0.21 (0.03-1.55)	0.125
rs144482539	2	164295170	*FIGN*	A/G	0.60	0.015	8.11 (3.33-19.8)	$4.14x10^{-6}$	0.011	1.85 (0.92-3.71)	0.084
rs34029315	3	10571102	*ATP2B2*	A/G	0.95	0.081	2.16 (1.57-2.97)	$2.48x10^{-6}$	0.101	1.01 (0.76-1.34)	0.931
rs41499647	3	50525154	*CACNA2D2*	C/T	0.91	0.205	1.61 (1.31-1.99)	$8.90x10^{-6}$	0.204	0.81 (0.65-1.01)	0.063
rs6548639*	3	79687975	*ROBO1*	C/T	0.70	0.380	1.59 (1.29-1.96)	$1.39x10^{-5}$	0.545	1.03 (0.87-1.23)	0.699
rs114497105	5	13759735	*DNAH5*	C/T	0.70	0.014	6.91 (2.97-16.1)	$7.55x10^{-6}$	0.012	1.26 (0.60-2.62)	0.547
rs142454359	5	135637432	*TRPC7*	G/A	0.58	0.036	4.06 (2.28-7.25)	$2.07x10^{-6}$	0.000	∞ (0.00-∞)	1.000
rs79438589	5	158397065	*EBF1*	G/T	0.80	0.022	4.32 (2.30-8.11)	$5.21x10^{-6}$	0.018	0.93 (0.49-1.77)	0.820
rs78069640	6	8859837	*RP11-314C16.1*	C/T	0.63	0.014	7.69 (3.18-18.6)	$6.01x10^{-6}$	0.020	0.86 (0.45-1.64)	0.639
rs2334321†	6	110567409	*METTL24*	G/A	0.97	0.084	0.50 (0.37-0.68)	$7.86x10^{-6}$	0.077	0.82 (0.59-1.15)	0.256
rs142720518	9	39156170	*CNTNAP3*	T/C	0.69	0.074	0.41 (0.28-0.60)	$6.11x10^{-6}$	0.065	0.88 (0.61-1.28)	0.515
rs4766986	12	113076475	*PTPN11*	C/T	0.85	0.057	2.42 (1.64-3.58)	$8.75x10^{-6}$	0.054	0.74 (0.49-1.13)	0.159
rs9946608	18	65817281	*RP11-638L3.1*	T/C	0.94	0.037	2.91 (1.83-4.61)	$5.76x10^{-6}$	0.033	1.73 (1.16-2.59)	0.008
rs351995*	19	809732	*PTBP1*	C/A	0.60	0.449	0.62 (0.50-0.78)	$2.05x10^{-5}$	0.472	0.97 (0.82-1.14)	0.676
rs203551	20	1192766	*C20orf202*	T/G	0.94	0.237	1.61 (1.31-1.97)	$5.26x10^{-6}$	0.246	1.00 (0.83-1.21)	0.987
rs78571420	21	36377390	*RUNX1*	T/A	0.99	0.042	2.58 (1.70-3.93)	$9.88x10^{-6}$	0.041	0.79 (0.50-1.25)	0.311
rs117161628*	21	46138322	*TSPEAR*	C/T	NA	NA	NA	NA	0.000	∞ (0.00-∞)	1.000

Chr. chromosome, A1 major allele, A2 minor allele, info imputation quality info score, MAF minor allele frequency, OR odds ratio, CI confidence interval, NA not applicable

* Primer design failed for rs9834479 in *ROBO1*, rs61504683 in *LPPR3*, and rs111750150 in *TSPEAR* and each of these were replaced by tagging variants.

For rs111750150, a tagging variant was selected based on 1000G CEU population using SNAP software[28], as there was no tagging variant available in the GWAS.

† This variant was not the lead variant at this locus, but it was selected on its functional impact (LD with lead variant r^2 0.93)

GWAS: effects calculated per copy of the minor allele, adjusted for age and sex

Replication: effects calculated per copy of the minor allele using a logistic regression model, adjusted for age, sex, study and, country of study

Replication analyses

To eliminate false-positive findings, we next performed a replication study in 350 patients with recurrent VT and 1866 patients with a single event only from three population-based cohorts. Overall, patients were followed for recurrence for a median period of 6.1 years (IQR 3.8-7.8), albeit follow-up started at different moments in time (see Material and Methods). Follow-up was complete for 83.7 percent of the patients.

For each of the 17 loci, we genotyped either the lead variant or the variant with substantial functional impact, and tested these for an association with recurrent VT in the replication cohorts. Results of the replication analyses are presented in Table 3. For two variants, rs142454359 and rs117161628, we observed only one carrier and, therefore, these variants could not be studied in detail. We observed an association with recurrent VT for one variant, whereas the remaining variants showed no evidence of an association with recurrent VT. Variant rs9946608 is located in an intergenic region at 18q22.1 and was associated with a 1.7-fold (95% CI 1.16-2.59, P-value 0.008, FDR 0.136) increased recurrence risk per copy of the minor allele. Similarly, we observed a hazard ratio of 1.69 (95% CI 1.18-2.42) per copy of the minor allele of rs9946608 for recurrence risk in a sub-analysis of patients from the LETS and THE-VTE cohorts. When we meta-analyzed the results obtained in the replication cohorts and the discovery GWAS, the minor allele of rs9946608 was associated with a 2.2-fold increased recurrence risk (Table 4, 95% CI 1.62-2.98, P-value 4.83×10^{-7}). There was no evidence for heterogeneity across the three replication cohorts (Q-statistic 1.12, I^2 0.00, P-value 0.57), nor across the replication cohorts and the discovery GWAS (Q-statistic 3.66, I^2 18.1, P-value 0.30).

We subsequently interrogated several publicly available databases for potential mechanistic information on rs9946608. No significant expression quantitative trait loci have been reported in GTEx[32] for rs9946608 or any of the linked variants (r^2>0.8). We used RegulomeDB[33], which integrates information from the ENCODE[34] and Roadmap Epigenomic[35] projects, to assess whether rs9946608 or linked variants may have a regulatory function. There is minimal evidence that several variants at this locus, including rs9946608, may affect transcription factor binding affinity. In some cell lines, DNase peaks in the chromatin structure have been identified using DNase-sequencing. Genes located nearby, which could be potential target genes, are two long intergenic non-coding RNA (lincRNA) genes *(RPH11-526H11-1* and *RP11-638L3.1)* and protein-coding gene *TMX3*. The latter encodes thioredoxin-related transmembrane protein

3 (TMX3), which has been detected in human megakaryocytes, platelets, and at the platelet surface of both resting and stimulated platelets.[36]

Table 4. Association results of rs9946608 in three replication cohorts

	TT	TC	CC	MAF	OR (95% CI)
MEGA					
recurrent VT patients	132	10	0	0.035	1.43 (0.71-2.87)
first VT patients	850	43	1	0.025	reference
LETS					
recurrent VT patients	66	10	1	0.078	2.40 (1.17-4.90)
first VT patients	330	25	1	0.038	reference
THE-VTE					
recurrent VT patients	95	9	1	0.052	1.61 (0.78-3.29)
first VT patients	519	36	0	0.032	reference
Meta-analysis					1.76 (1.17-2.65)
Combined with GWAS					2.20 (1.62-2.98)

MAF minor allele frequency, OR odds ratio, CI confidence interval, VT venous thrombosis, GWAS genome-wide association study

Results were meta-analyzed using a fixed-effect meta-analysis model based on inverse-variance weighting. Heterogeneity was assessed by the Cochran's Q statistic and the I^2 index. Across the three replication cohorts the heterogeneity measures were as follows: Q 1.12, I^2 0.00, P-value 0.57. For the three replication studies and the discovery GWAS, we observed a Q of 3.66, I^2 18.1, P-value 0.30. In the GWAS, the MAF of rs9946608 was 0.583 in recurrence patients and 0.256 in patients with a first VT only.

Discriminative value

In a preliminary analysis, we explored the added discriminative value of FV Leiden and rs9946608 to a prediction model with clinical risk factors alone. The AUC of the clinical prediction model, which included sex, age, event type, and provoking status, was 0.65 (95% CI 0.61-0.68). Predictive accuracy of recurrence risk significantly improved when adding the two genetic risk variants to the model (AUC 0.68, 95% CI 0.65-0.71).

DISCUSSION

This GWAS is the first large-scale genetic discovery effort for recurrent VT. Previous studies were either small or focussed on candidate gene variants, such as FV Leiden and prothrombin G20210A. The high recurrence rate of VT, especially in patients with a first unprovoked event, and the subsequent lifelong treatment with anticoagulants make it important to uncover the genetic and biological architecture of recurrent VT. Here, we confirm the association of FV Leiden with recurrence and identify a novel potential risk locus at chromosome 18q22.1.

Genome-wide significance was attained by several variants at the *F5* locus, which included the well-known risk variant FV Leiden. We observed a 2.4-fold increased risk of recurrence per copy of the T-allele of FV Leiden, which is slightly higher than previously reported,[3,4] albeit still lower than the risk estimates observed for a first VT.[26,27] There were no secondary association signals observed at the *F5* locus. Known VT risk variant rs4524, which has been shown to affect the risk of a first thrombotic event independent of FV Leiden,[26,37] was only nominally associated with recurrent VT. This may suggest that FV Leiden is the key determinant at the *F5* locus of recurrence risk.

We additionally identified 43 variants at 17 novel loci associated with recurrent VT at suggestive significance (P-value $<1.0\times10^{-5}$). We sought to replicate these findings in independent samples from three studies. Our results suggest that carriers of rs9946608-C have a 1.7-fold increased recurrence risk compared with non-carriers. We observed little evidence for statistical heterogeneity between the replication studies which could explain our findings. Formal replication is needed to confirm the association between rs9946608 and recurrent VT, as the meta-analysis of the GWAS and the replication studies did not reach genome-wide significance. From a clinical perspective, it would also be interesting to evaluate whether this variant has a differential effect on recurrent DVT or PE, which was now impossible to study due to low number of patients.

Variant rs9946608 and proxies map to noncoding sequence at chromosome 18q22.1 and have not been implicated in disease risk before. If the association with recurrence risk is true, this intergenic locus has most likely a regulatory function. We observed some evidence of transcription factor binding affinity and DNase peaks in the chromatin structure of some cell lines. Additional work, including fine-mapping of the GWAS signal to identify the functional variant, is needed to unravel the potential underlying

mechanism. Candidate genes could be nearby lincRNA genes *RPH11-526H11-1* and *RP11-638L3.*1. Increasing evidence suggests that lincRNAs may play an important role in epigenetic and post-transcriptional regulation in health and (cardiovascular) disease.[38,39] However, the characteristics and function of the majority of these RNAs are currently not known. Interrogation of several publicly available databases, such as GTEx[32] and several long noncoding RNA databases, did not yield additional information. The nearest protein-coding gene, *TMX3*, lies over 500Kb away, but could also be a target given its biological function. As TMX3 has been detected at the platelet surface,[36] it may play a role in platelet functioning, in line with other members of the protein disulphide isomerase family. Functional follow-up experiments could help to identify and characterize the potential role of these genes in recurrent VT. In addition, long-range chromatin interaction analyses using chromosome conformation capture technologies, such as 4C and Hi-C, might aid to identify other potential target genes.

Another notable finding is that almost all variants, which have previously been linked to a first VT at genome-wide significance[26,27] including the novel risk variants at *TSPAN15* and *SLC44A2*, were not or only nominally associated with the risk of recurrent VT. This is in line with previous reports on the risk variants which have been studied for recurrence risk.[3-6] Several explanations for this discrepancy have been proposed. To some extent, this can be explained by the difference in absolute risks for first and recurrent VT, resulting in the incomparability of effects on a relative risk scale between first and recurrent VT.[11] In addition, research into risk factors for recurrence risk may be hindered by index event bias, although this could lead to both under- and overestimation of the risk estimate.[10] Of note, as all candidate risk variants had effects in the expected direction and three out of nine variants were associated with recurrence risk at a significance level of 0.05, which is more than expected by chance, our results provide some evidence that these variants may also impact recurrence risk. In particular, *FGG* rs2066865 might be promising, as earlier studies have also reported some evidence of an association.[5,6]

The main limitation of this study is the small sample size with 447 and 345 recurrent VT patients in the discovery GWAS and the combined replication studies, respectively. As a result, we may have missed associations between recurrent VT and variants with a small effect or a low MAF. The small sample size may also explain why we failed to replicate most suggestively associated variants identified in our GWAS. We therefore emphasise

the need of a large international collaborative effort to substantially increase the sample size for recurrent VT analyses. Of note, mainly patients of Northwest-European origin were included in our analyses and, therefore, caution is needed in generalizing our results to other populations. In addition, the X chromosome was not interrogated in the discovery GWAS.

In both the GWAS and the replication analyses, patients who were lost to follow-up or who experienced an uncertain recurrent VT were considered to be recurrence-free. This could have affected our results, as we cannot rule out that these patients experienced a recurrent thrombotic event. However, this is unlikely, since these patients did not visit the anticoagulation clinics, which monitor anticoagulant treatment. In addition, the results of the sensitivity GWAS, in which these patients were excluded, did not materially differ from the discovery GWAS. Likewise, we obtained a similar effect estimate for rs9946608 in the logistic regression model and the time-to-event analysis, in which patients who were lost to follow-up were censored. Together, this suggests that the impact of misclassification in our study was probably low.

Our findings could lead to a better understanding of the biological mechanism underlying recurrent VT. In addition, we have previously shown the potential clinical value of genetic risk factors in the risk stratification of first and recurrent VT.[5,40] In a preliminary analysis, we showed that adding FV Leiden and rs9946608 to a clinical prediction model slightly improved the risk discrimination of recurrence. Identification of novel risk variants may further improve risk prediction of recurrent VT. Although additional replication and functional analyses are required, we identified a potential risk locus at chromosome 18q22.1 and confirmed the role of FV Leiden in recurrent VT pathophysiology.

REFERENCES

1. Prandoni P, Lensing AW, Cogo A, Cuppini S, Villalta S, Carta M, et al. The long-term clinical course of acute deep venous thrombosis. *Ann Intern Med*. 1996;125(1):1-7.
2. Hansson PO, Sörbo J, Eriksson H. Recurrent venous thromboembolism after deep vein thrombosis: incidence and risk factors. *Arch Intern Med*. 2000;160(6):769-74.
3. Ho WK, Hankey GJ, Quinlan DJ, Eikelboom JW. Risk of recurrent venous thromboembolism in patients with common thrombophilia: a systematic review. *Arch Intern Med*. 2006;166(7):729-36.
4. Marchiori A, Mosena L, Prins MH, Prandoni P. The risk of recurrent venous thromboembolism among heterozygous carriers of factor V Leiden or prothrombin G20210A mutation. A systematic review of prospective studies. *Haematologica*. 2007;92(8):1107-14.
5. van Hylckama Vlieg A, Flinterman LE, Bare LA, Cannegieter SC, Reitsma PH, Arellano AR, et al. Genetic variations associated with recurrent venous thrombosis. *Circ Cardiovasc Genet*. 2014;7(6):806-13.
6. Bruzelius M, Ljungqvist M, Bottai M, Bergendal A, Strawbridge RJ, Holmström M, et al. F11 is associated with recurrent VTE in women. A prospective cohort study. *Thromb Haemost*. 2016;115(2):406-14.
7. Baglin T, Luddington R, Brown K, Baglin C. Incidence of recurrent venous thromboembolism in relation to clinical and thrombophilic risk factors: prospective cohort study. *Lancet*. 2003;362(9383):523-6.
8. Christiansen SC, Cannegieter SC, Koster T, Vandenbroucke JP, Rosendaal FR. Thrombophilia, clinical factors, and recurrent venous thrombotic events. *JAMA*. 2005;293(19):2352-61.
9. Prandoni P, Noventa F, Ghirarduzzi A, Pengo V, Bernardi E, Pesavento R, et al. The risk of recurrent venous thromboembolism after discontinuing anticoagulation in patients with acute proximal deep vein thrombosis or pulmonary embolism. A prospective cohort study in 1,626 patients. *Haematologica*. 2007;92(2):199-205.
10. Dahabreh IJ, Kent DM. Index event bias as an explanation for the paradoxes of recurrence risk research. *JAMA*. 2011;305(8):822-3.
11. Cannegieter SC, van Hylckama Vlieg A. Venous thrombosis: understanding the paradoxes of recurrence. *J Thromb Haemost*. 2013;11 Suppl 1:161-9.
12. Zee RY, Bubes V, Shrivastava S, Ridker PM, Glynn RJ. Genetic risk factors in recurrent venous thromboembolism: A multilocus, population-based, prospective approach. *Clin Chim Acta*. 2009;402(1-2):189-92.
13. Jilma B, Kovar FM, Hron G, Endler G, Marsik CL, Eichinger S, et al. Homozygosity in the single nucleotide polymorphism Ser128Arg in the E-selectin gene associated with recurrent venous thromboembolism. *Arch Intern Med*. 2006;166(15):1655-9.

14. Mustafa S, Weltermann A, Fritsche R, Marsik C, Wagner O, Kyrle PA, et al. Genetic variation in heme oxygenase 1 (HMOX1) and the risk of recurrent venous thromboembolism. *J Vasc Surg*. 2008;47(3):566-70.
15. Timp JF, Lijfering WM, Flinterman LE, van Hylckama Vlieg A, le Cessie S, Rosendaal FR, et al. Predictive value of factor VIII levels for recurrent venous thrombosis: results from the MEGA follow-up study. *J Thromb Haemost*. 2015;13(10):1823-32.
16. Blom JW, Doggen CJ, Osanto S, Rosendaal FR. Malignancies, prothrombotic mutations, and the risk of venous thrombosis. *JAMA*. 2005;293(6):715-22.
17. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: a R library for genome-wide association analysis. *Bioinformatics*. 2007 23(10):1294-6.
18. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009;5(6):e1000529.
19. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46(8):818-25.
20. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007;39(7):906-13.
21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-75.
22. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010;26(18):2336-7.
23. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
24. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55(4):997-1004.
25. Ahmad A, Sundquist K, Zöller B, Dahlbäck B, Svensson PJ, Sundquist J, et al. Identification of polymorphisms in Apolipoprotein M gene and their relationship with risk of recurrent venous thromboembolism. *Thromb Haemost*. 2016;116(3):432-41.
26. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng LC, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet*. 2015 Apr 2;96(4):532-42.
27. Hinds DA, Buil A, Ziemek D, Martinez-Perez A, Malik R, Folkersen L, 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*. 2016;25(9):1867-74.
28. van Hylckama Vlieg A, Baglin CA, Luddington R, MacDonald S, Rosendaal FR, Baglin TP. The risk of a first and a recurrent venous thrombosis associated with an elevated D-dimer level and an elevated thrombin potential: results of the THE-VTE study. *J Thromb Haemost*. 2015;13(9):1642-52.

29. Johnson AD, Handsaker RE, Pulit S, Nizzari MM, O'Donnell CJ, de Bakker P I. SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008;24(24):2938-9.
30. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26:2190-1.
31. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
32. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648-60.
33. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-7.
34. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
35. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*. 2010;28(10):1045-8.
36. Holbrook LM, Watkins NA, Simmonds AD, Jones CI, Ouwehand WH, Gibbins JM. Platelets release novel thiol isomerase enzymes which are recruited to the cell surface following activation. *Br J Haematol*. 2010;148(4):627-37.
37. Smith NL, Hindorff LA, Heckbert SR, Lemaitre RN, Marciante KD, Rice K, et al. Association of genetic variations with nonfatal venous thrombosis in postmenopausal women. *JAMA*. 2007;297: 489–98.
38. Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet*. 2011;12(12):861-74.
39. Uchida S, Dimmeler S. Long noncoding RNAs in cardiovascular diseases. *Circ Res*. 2015;116(4):737-50.
40. de Haan HG, Bezemer ID, Doggen CJ, Le Cessie S, Reitsma PH, Arellano AR, et al. Multiple SNP testing improves risk prediction of first venous thrombosis. *Blood*. 2012;120(3):656-63.

SUPPLEMENTAL FIGURES AND TABLES

Supplemental Table 1. Previously reported associations with first or recurrent VT

Supplemental Table 2. GWAS associations with recurrent VT at genome-wide significance

Supplemental Table 3. GWAS associations with recurrent VT at significance threshold of $P<1x10^{-5}$

Supplemental Table 4. Association results for lead variants in sensitivity analysis excluding patients who were lost to follow-up

Supplemental Table 5. Association results for lead variants in sensitivity analysis adjusting on provoking status

Supplemental Figure 1. Flow diagram of patients included and excluded from GWAS analyses

Supplemental Figure 2. Quantile-quantile plot of the genome-wide test statistics

Supplemental Figure 3. Manhattan plot of the GWAS association results

Supplemental Figure 4. Regional association plots at the *F5* locus before and after conditioning on FV Leiden

Supplemental Table 1. Previously reported associations with first or recurrent VT

A. Previously reported genome wide significant associations with first VT

rs ID	Chr.	Position	Gene	Effect	Alleles	A2	Analysis	OR (95% CI)	P-value	Study
rs4524	1	169511755	*F5*	missense	T/C	T	additive	1.20 (1.14-1.26)	$2.65x10^{-11}$	Germain *et al.*
rs6025	1	169519049	*F5*	missense	C/T	T	additive	3.25 (2.91-3.64)	$1.10x10^{-96}$	Germain *et al.*
rs2066865	4	155525276	*FGG*	3'UTR	G/A	A	additive	1.24 (1.18-1.31)	$1.03x10^{-16}$	Germain *et al.*
rs4253417	4	187199005	*F11*	intronic	T/C	C	additive	1.27 (1.22-1.34)	$1.21x10^{-23}$	Germain *et al.*
rs529565	9	136149500	*ABO*	intronic	T/C	C	additive	1.55 (1.48-1.63)	$4.23x10^{-75}$	Germain *et al.*
rs78707713	10	71245276	*TSPAN15*	intronic	T/C	T	additive	1.28 (1.19-1.39)	$5.74x10^{-11}$	Germain *et al.*
rs1799963	11	46761055	*F2*	intronic	G/A	A	additive	2.29 (1.75-2.99)	$1.73x10^{-9}$	Germain *et al.*
rs2288904	19	10742170	*SLC44A2*	missense	G/A	G	additive	1.19 (1.12-1.26)	$1.07x10^{-9}$	Germain *et al.*
rs6087685	20	33777612	*PROCR*	intronic	G/C	C	additive	1.15 (1.10-1.21)	$1.65x10^{-8}$	Germain *et al.*
rs114209171	X	154278797	*FUNCD2**	intronic	T/C	T	additive	1.15 (1.11-1.20)	$7.0x10^{-13}$	Hinds *et al.*

Chr. Chromosome, A2 effect allele, OR odds ratio, CI confidence interval, UTR untranslated region *near *F8*

B. Previously reported associations with recurrent VT specifically

rs ID	Chr.	Position	Gene	Effect	Alleles	A2	Analysis	HR (95% CI)	P-value	Study
rs5361	1	169701060	*SELE*	missense	T/G	G	homozygous	4.1 (1.5-11.4)	0.01	Jilma *et al.*
rs1799864	3	46399208	*CCR2*	missense	G/A	A	additive	2.00 (1.15-3.48)	0.014	Zee *et al.*
rs805297	6	31622606	*APOM*	intronic	C/A	A	additive, men	1.72 (1.03-2.88)	0.038	Ahmad *et al.*
rs662	7	94937446	*PON1*	missense	T/C	C	additive	1.79 (1.08-2.95)	0.023	Zee *et al.*
rs3025058	11	102715948-102715949	*MMP3*	upstream	5A/6A	6A	additive	1.66 (1.10-2.49)	0.015	Zee *et al.*
rs1800775	16	56995236	*CETP*	upstream	C/A	A	additive	0.63 (0.40-0.98)	0.041	Zee *et al.*
rs3074372	22	35776887-35776888	*HMOX1*	5'UTR	GT-Repeat	Long	heterozygous	2.2 (1.4-3.4)	0.001	Mustafa *et al.*

Chr. Chromosome, A2 effect allele, HR hazard ratio, CI confidence interval, UTR untranslated region

Supplemental Table 2. GWAS associations with recurrent VT at genome-wide significance

rs ID	Chr	Position	Effect	(Nearest) Gene	A1/ A2	MAF overall	Info	OR (95%CI)	P-value	OR_{cond} (95%CI)	$P\text{-value}_{cond}$
rs1894692	1	169467654	intergenic	*SLC19A2*	A/G	0.094	0.93	2.38 (1.76-3.20)	$1.18x10^{-8}$	2.39 (0.39-14.7)	0.349
rs4264045	1	169470748	intergenic	*F5*	G/T	0.137	0.98	2.07 (1.62-2.65)	$6.38x10^{-9}$	1.57 (1.01-2.44)	0.045
rs6670848	1	169472899	intergenic	*F5*	G/A	0.137	0.98	2.07 (1.62-2.64)	$6.99x10^{-9}$	1.56 (1.00-2.42)	0.048
rs10737547	1	169476052	intergenic	*F5*	G/A	0.137	0.98	2.06 (1.61-2.63)	$7.56x10^{-9}$	1.55 (1.00-2.41)	0.050
rs6687813	1	169477574	intergenic	*F5*	C/A	0.137	0.98	2.06 (1.61-2.63)	$7.80x10^{-9}$	1.55 (1.00-2.40)	0.051
rs970740	1	169479974	intergenic	*F5*	T/C	0.137	0.98	2.06 (1.61-2.63)	$7.91x10^{-9}$	1.55 (1.00-2.40)	0.052
rs6427194	1	169481121	downstream	*F5*	A/T	0.137	0.98	2.06 (1.61-2.63)	$7.92x10^{-9}$	1.55 (1.00-2.40)	0.052
rs6427195	1	169481176	downstream	*F5*	T/A	0.137	0.98	2.06 (1.61-2.63)	$7.92x10^{-9}$	1.55 (1.00-2.40)	0.052
rs6427196	1	169481223	3' UTR	*F5*	G/C	0.137	0.98	2.06 (1.61-2.63)	$7.92x10^{-9}$	1.55 (1.00-2.40)	0.052
rs9332666	1	169486641	intronic	*F5*	C/G	0.137	0.98	2.05 (1.61-2.62)	$8.24x10^{-9}$	1.54 (0.99-2.39)	0.054
rs2420370	1	169490392	intronic	*F5*	C/G	0.137	0.99	2.05 (1.60-2.62)	$9.05x10^{-9}$	1.53 (0.99-2.37)	0.058
rs6682179	1	169490401	intronic	*F5*	C/T	0.137	0.99	2.05 (1.60-2.62)	$9.05x10^{-9}$	1.53 (0.99-2.37)	0.058
rs2420371	1	169491555	intronic	*F5*	A/G	0.138	0.99	2.03 (1.59-2.59)	$1.16x10^{-8}$	1.50 (0.97-2.32)	0.066
rs2420372	1	169498056	intronic	*F5*	G/A	0.138	0.99	2.02 (1.58-2.57)	$1.37x10^{-8}$	1.48 (0.96-2.29)	0.075
rs6009	1	169498834	intronic	*F5*	C/T	0.138	0.99	2.02 (1.58-2.57)	$1.37x10^{-8}$	1.48 (0.96-2.29)	0.075
rs6427197	1	169500590	intronic	*F5*	A/C	0.138	0.99	2.02 (1.58-2.57)	$1.43x10^{-8}$	1.48 (0.96-2.29)	0.077
rs1018827	1	169514006	intronic	*F5*	G/A	0.140	0.98	2.07 (1.62-2.64)	$4.34x10^{-9}$	1.63 (1.07-2.50)	0.024
rs6025	1	169519049	missense	*F5*	C/T	0.096	0.94	2.35 (1.75-3.15)	$1.28x10^{-8}$	NA	NA
rs2213868	1	169521553	intronic	*F5*	A/G	0.138	0.93	2.14 (1.67-2.75)	$2.67x10^{-9}$	1.71 (1.10-2.68)	0.018

Chr Chromosome, A1 major allele, A2 minor allele, MAF minor allele frequency, OR odds ratio, CI confidence interval, UTR untranslated region, NA not applicable, cond conditional, info imputation quality info measure

GWAS analyses were adjusted for age and sex, assuming an additive mode of inheritance. Subsequently, we performed a conditional analysis on FV Leiden (rs6025).

Supplemental Table 3. GWAS associations with recurrent VT at significance threshold of $P<1x10^{-5}$

rs ID	Chr.	Position	Effect	(Nearest) Gene	A1/A2	MAF overall	info	OR (95%CI)	P-value
rs111438240	1	159902335	synonymous	*IGSF9*	G/A	0.011	0.82	7.59 (3.18-18.1)	$5.07x10^{-6}$
rs112349920	1	159933483	intronic	*LINC01133*	C/T	0.011	0.98	6.90 (3.06-15.6)	$3.36x10^{-6}$
rs111272082	1	159940634	intronic	*LINC01133*	A/G	0.011	0.98	6.92 (3.06-15.6)	$3.37x10^{-6}$
rs145163454	1	169090748	intronic	*ATP1B1*	T/C	0.091	0.95	2.19 (1.63-2.95)	$2.38x10^{-7}$
rs77979353	1	169115022	intronic	*NME7*	T/C	0.119	0.96	1.81 (1.39-2.35)	$8.97x10^{-6}$
rs144737447	1	169160458	intronic	*NME7*	C/T	0.093	0.97	2.15 (1.60-2.87)	$2.72x10^{-7}$
rs2227246	1	169208179	intronic	*NME7*	T/C	0.091	0.96	2.16 (1.61-2.91)	$3.03x10^{-7}$
rs2040445	1	169216412	intronic	*NME7*	G/C	0.093	0.97	2.15 (1.61-2.87)	$2.58x10^{-7}$
rs6692824	1	169226218	intronic	*NME7*	G/C	0.094	0.96	2.16 (1.62-2.90)	$2.07x10^{-7}$
rs1209731	1	169324793	intronic	*NME7*	C/T	0.094	0.97	2.18 (1.63-2.92)	$1.42x10^{-7}$
rs2678166	1	169435027	3' UTR	*SLC19A2*	T/C	0.091	0.96	2.21 (1.64-2.96)	$1.43x10^{-7}$
rs6696217	1	169460726	intergenic	*SLC19A2*	G/A	0.147	0.97	1.80 (1.42-2.28)	$9.80x10^{-7}$
rs144482539	2	164295170	intergenic	*FIGN*	A/G	0.015	0.60	8.11 (3.33-19.8)	$4.14x10^{-6}$
rs34029315	3	10571102	intergenic	*ATP2B2*	A/G	0.081	0.95	2.16 (1.57-2.97)	$2.48x10^{-6}$
rs41499647	3	50525154	intronic	*CACNA2D2*	C/T	0.205	0.91	1.61 (1.31-1.99)	$8.90x10^{-6}$
rs9834479	3	79687062	intronic	*ROBO1*	T/A	0.411	0.73	1.61 (1.32-1.96)	$3.10x10^{-6}$
rs114497105	5	13759735	intronic	*DNAH5*	C/T	0.014	0.70	6.91 (2.97-16.1)	$7.55x10^{-6}$
rs142454359	5	135637432	intronic	*TRPC7*	G/A	0.036	0.58	4.06 (2.28-7.25)	$2.07x10^{-6}$
rs77962281	5	135710076	intergenic	*TRPC7*	T/C	0.036	0.59	3.87 (2.19-6.86)	$3.37x10^{-6}$
rs79438589	5	158397065	intronic	*EBF1*	G/T	0.022	0.80	4.32 (2.30-8.11)	$5.21x10^{-6}$

Supplemental Table 3. Continued

rs ID	Chr.	Position	Effect	(Nearest) Gene	A1/A2	MAF overall	info	OR (95%CI)	P-value
rs78069640	6	8859837	intergenic	*RP11-314C16.1*	C/T	0.014	0.63	7.69 (3.18-18.6)	$6.01x10^{-6}$
rs2334321	6	110567409	missense	*METTL24*	G/A	0.084	0.97	0.50 (0.37-0.68)	$7.86x10^{-6}$
rs72935918	6	110579775	intronic	*METTL24*	A/G	0.085	0.98	0.50 (0.37-0.68)	$7.07x10^{-6}$
rs72935927	6	110616710	intronic	*METTL24*	C/T	0.085	0.95	0.50 (0.37-0.68)	$9.86x10^{-6}$
rs10974147	9	39113963	intronic	*CNTNAP3*	C/A	0.073	0.67	0.41 (0.28-0.61)	$9.60x10^{-6}$
rs150155153	9	39130703	intronic	*CNTNAP3*	A/G	0.069	0.72	0.41 (0.28-0.61)	$9.74x10^{-6}$
rs1692979	9	39145910	intronic	*CNTNAP3*	T/C	0.071	0.70	0.41 (0.28-0.61)	$7.96x10^{-6}$
rs115460012	9	39146124	intronic	*CNTNAP3*	G/C	0.095	0.60	0.43 (0.30-0.63)	$9.22x10^{-6}$
rs142720518	9	39156170	intronic	*CNTNAP3*	T/C	0.074	0.69	0.41 (0.28-0.60)	$6.11x10^{-6}$
rs115361037	9	39158211	intronic	*CNTNAP3*	A/C	0.071	0.71	0.42 (0.28-0.61)	$9.29x10^{-6}$
rs233713	12	113049776	intergenic	*PTPN11*	G/A	0.057	0.85	2.43 (1.64-3.59)	$8.85x10^{-6}$
rs4766986	12	113076475	intergenic	*PTPN11*	C/T	0.057	0.97	2.42 (1.64-3.58)	$8.75x10^{-6}$
rs58823953	18	65794561	intergenic	*RP11-638L3.1*	A/G	0.036	0.97	2.87 (1.81-4.54)	$6.82x10^{-6}$
rs78684150	18	65797020	intergenic	*RP11-638L3.1*	A/G	0.036	0.97	2.87 (1.81-4.54)	$6.84x10^{-6}$
rs116330040	18	65799997	intergenic	*RP11-638L3.1*	T/C	0.036	0.97	2.87 (1.81-4.54)	$6.82x10^{-6}$
rs58636937	18	65801415	intergenic	*RP11-638L3.1*	C/T	0.036	0.97	2.87 (1.81-4.54)	$6.85x10^{-6}$
rs115198055	18	65804403	intergenic	*RP11-638L3.1*	A/C	0.036	0.97	2.87 (1.81-4.54)	$6.91x10^{-6}$
rs150366483	18	65804614	intergenic	*RP11-638L3.1*	G/T	0.036	0.97	2.87 (1.81-4.54)	$6.91x10^{-6}$
rs114858887	18	65809262	intergenic	*RP11-638L3.1*	A/G	0.036	0.97	2.87 (1.81-4.55)	$6.82x10^{-6}$
rs77684223	18	65810005	intergenic	*RP11-638L3.1*	T/A	0.036	0.97	2.87 (1.81-4.54)	$6.83x10^{-6}$

Supplemental Table 3. Continued

rs ID	Chr.	Position	Effect	(Nearest) Gene	A1/A2	MAF overall	info	OR (95%CI)	P-value
rs74443278	18	65810476	intergenic	*RP11-638L3.1*	T/C	0.036	0.97	2.87 (1.81-4.55)	$6.82x10^{-6}$
rs77238268	18	65813987	intergenic	*RP11-638L3.1*	C/G	0.036	0.96	2.88 (1.82-4.56)	$6.57x10^{-6}$
rs7228982	18	65814097	intergenic	*RP11-638L3.1*	A/G	0.036	0.96	2.88 (1.82-4.56)	$6.55x10^{-6}$
rs9946608	18	65817281	intergenic	*RP11-638L3.1*	T/C	0.037	0.94	2.91 (1.83-4.61)	$5.76x10^{-6}$
rs145772467	18	65825166	intergenic	*RP11-638L3.1*	T/C	0.037	0.88	2.99 (1.85-4.83)	$7.21x10^{-6}$
rs118036929	18	65825184	intergenic	*RP11-638L3.1*	T/C	0.037	0.88	2.99 (1.85-4.83)	$7.21x10^{-6}$
rs61504683	19	816919	intronic	*LPPR3*	C/T	0.468	0.58	0.60 (0.48-0.74)	$3.80x10^{-6}$
rs7269259	20	1179082	intergenic	*C20orf202*	C/T	0.237	1.00	1.57 (1.29-1.92)	$7.45x10^{-6}$
rs126622	20	1191140	intergenic	*C20orf202*	T/C	0.238	0.95	1.60 (1.30-1.96)	$5.86x10^{-6}$
rs203551	20	1192766	intergenic	*C20orf202*	T/G	0.237	0.94	1.61 (1.31-1.97)	$5.26x10^{-6}$
rs78571420	21	36377390	intronic	*RUNX1*	T/A	0.042	0.99	2.58 (1.70-3.93)	$9.88x10^{-6}$
rs111750150	21	45960880	intronic	*TSPEAR*	C/T	0.020	0.52	8.25 (3.56-19.1)	$8.46x10^{-7}$

Chr. Chromosome, A1 major allele, A2 minor allele, MAF minor allele frequency, OR odds ratio, CI confidence interval, UTR untranslated region
GWAS analyses were adjusted for age and sex, assuming an additive mode of inheritance. Variants in or near *ATP1B1*, *NME7*, or *SLC19A2* mapped to the F5 locus.

Supplemental Table 4. Association results for lead variants in sensitivity analysis excluding patients who were lost to follow-up

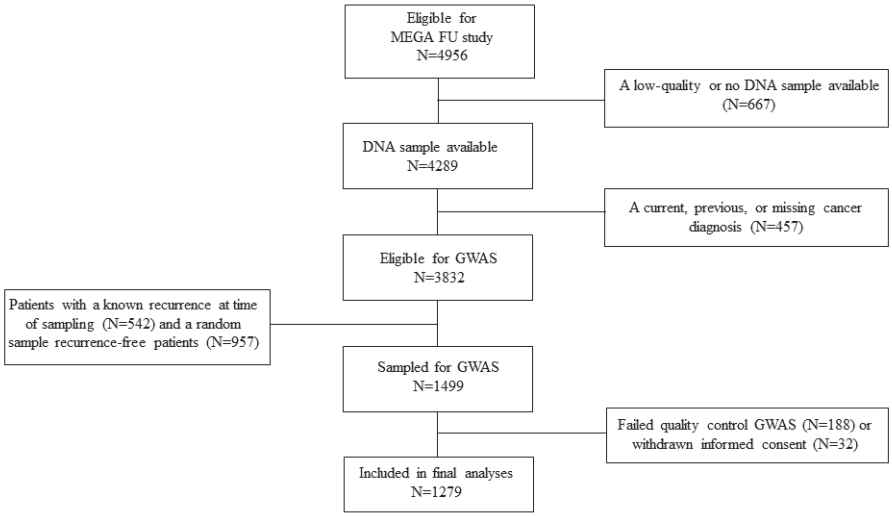
rs ID	Chr.	Position	Effect	(Nearest) Gene	A1/A2	MAF overall	OR (95% CI)	P-value
rs112349920	1	159933483	intronic	*LINC01133*	C/T	0.011	6.99 (3.05-16.0)	$4.47x10^{-6}$
rs2213868	1	169521553	intronic	*F5*	A/G	0.143	2.05 (1.59-2.66)	$4.07x10^{-8}$
rs144482539	2	164295170	intergenic	*FIGN*	A/G	0.016	6.74 (2.79-16.2)	$2.17x10^{-5}$
rs34029315	3	10571102	intergenic	*ATP2B2*	A/G	0.084	2.10 (1.52-2.92)	$7.66x10^{-6}$
rs41499647	3	50525154	intronic	*CACNA2D2*	C/T	0.209	1.59 (1.28-1.97)	$3.05x10^{-5}$
rs9834479	3	79687062	intronic	*ROBO1*	T/A	0.416	1.59 (1.29-1.95)	$9.82x10^{-6}$
rs114497105	5	13759735	intronic	*DNAH5*	C/T	0.015	5.39 (2.35-12.4)	$7.04x10^{-5}$
rs142454359	5	135637432	intronic	*TRPC7*	G/A	0.037	3.71 (2.07-6.62)	$9.65x10^{-6}$
rs79438589	5	158397065	intronic	*EBF1*	G/T	0.022	4.94 (2.57-9.49)	$1.70x10^{-6}$
rs78069640	6	8859837	intergenic	*RP11-314C16.1*	C/T	0.015	6.65 (2.75-16.1)	$2.70x10^{-5}$
rs72935918	6	110579775	intronic	*METTL24*	A/G	0.079	0.53 (0.38-0.73)	$1.04x10^{-4}$
rs142720518	9	39156170	intronic	*CNTNAP3*	T/C	0.073	0.37 (0.25-0.56)	$1.84x10^{-6}$
rs4766986	12	113076475	intergenic	*PTPN11*	C/T	0.060	2.20 (1.49-3.27)	$8.51x10^{-5}$
rs9946608	18	65817281	intergenic	*RP11-638L3.1*	T/C	0.038	2.78 (1.74-4.44)	$1.97x10^{-5}$
rs61504683	19	816919	intronic	*LPPR3*	C/T	0.464	0.60 (0.48-0.75)	$1.08x10^{-5}$
rs203551	20	1192766	intergenic	*C20orf202*	T/G	0.239	1.62 (1.31-2.00)	$8.27x10^{-6}$
rs78571420	21	36377390	intronic	*RUNX1*	T/A	0.043	2.62 (1.70-4.04)	$1.21x10^{-5}$
rs111750150	21	45960880	intronic	*TSPEAR*	C/T	0.020	9.09 (3.84-21.5)	$5.00x10^{-7}$

Chr. Chromosome, A1 major allele, A2 minor allele, MAF minor allele frequency, OR odds ratio, CI confidence interval
GWAS analyses were adjusted for age and sex, assuming an additive mode of inheritance.

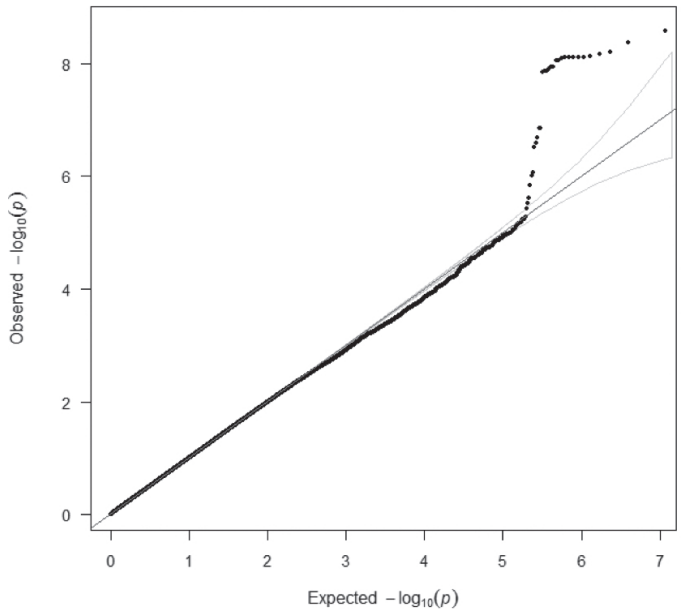
Supplemental Table 5. Association results for lead variants in sensitivity analysis adjusting on provoking status

rs ID	Chr.	Position	Effect	(Nearest) Gene	A1/A2	MAF overall	OR (95% CI)	P-value
rs112349920	1	159933483	intronic	*LINC01133*	C/T	0.011	6.59 (2.87-15.1)	$8.85x10^{-6}$
rs2213868	1	169521553	intronic	*F5*	A/G	0.139	2.14 (1.66-2.76)	$3.67x10^{-9}$
rs144482539	2	164295170	intergenic	*FIGN*	A/G	0.015	8.72 (3.56-21.4)	$2.20x10^{-6}$
rs34029315	3	10571102	intergenic	*ATP2B2*	A/G	0.081	2.24 (1.62-3.10)	$1.05x10^{-6}$
rs41499647	3	50525154	intronic	*CACNA2D2*	C/T	0.206	1.58 (1.28-1.96)	$2.24x10^{-5}$
rs9834479	3	79687062	intronic	*ROBO1*	T/A	0.413	1.58 (1.29-1.93)	$9.49x10^{-6}$
rs114497105	5	13759735	intronic	*DNAH5*	C/T	0.014	7.62 (3.22-18.1)	$3.85x10^{-6}$
rs142454359	5	135637432	intronic	*TRPC7*	G/A	0.036	4.05 (2.27-7.24)	$2.25x10^{-6}$
rs79438589	5	158397065	intronic	*EBF1*	G/T	0.022	4.48 (2.37-8.46)	$3.93x10^{-6}$
rs78069640	6	8859837	intergenic	*RP11-314C16.1*	C/T	0.014	7.17 (2.94-17.5)	$1.54x10^{-5}$
rs72935918	6	110579775	intronic	*METTL24*	A/G	0.084	0.49 (0.36-0.67)	$6.13x10^{-6}$
rs142720518	9	39156170	intronic	*CNTNAP3*	T/C	0.074	0.40 (0.27-0.58)	$3.25x10^{-6}$
rs4766986	12	113076475	intergenic	*PTPN11*	C/T	0.058	2.44 (1.65-3.61)	$8.85x10^{-6}$
rs9946608	18	65817281	intergenic	*RP11-638L3.1*	T/C	0.037	2.93 (1.84-4.65)	$5.32x10^{-6}$
rs61504683	19	816919	intronic	*LPPR3*	C/T	0.469	0.59 (0.47-0.77)	$2.92x10^{-6}$
rs203551	20	1192766	intergenic	*C20orf202*	T/G	0.236	1.58 (1.28-1.94)	$1.43x10^{-5}$
rs78571420	21	36377390	intronic	*RUNX1*	T/A	0.043	2.47 (1.62-3.78)	$2.80x10^{-5}$
rs111750150	21	45960880	intronic	*TSPEAR*	C/T	0.019	8.70 (3.64-20.8)	$1.14x10^{-6}$

Chr. Chromosome, A1 major allele, A2 minor allele, MAF minor allele frequency, OR odds ratio, CI confidence interval
GWAS analyses were adjusted for age, sex, and provoking status, assuming an additive mode of inheritance.
Provoking status at time of the first event was defined by the following factors: recent surgery, immobilization (plaster cast, bedridden at home, hospitalization), hormone use, pregnancy or post-partum, and travel.

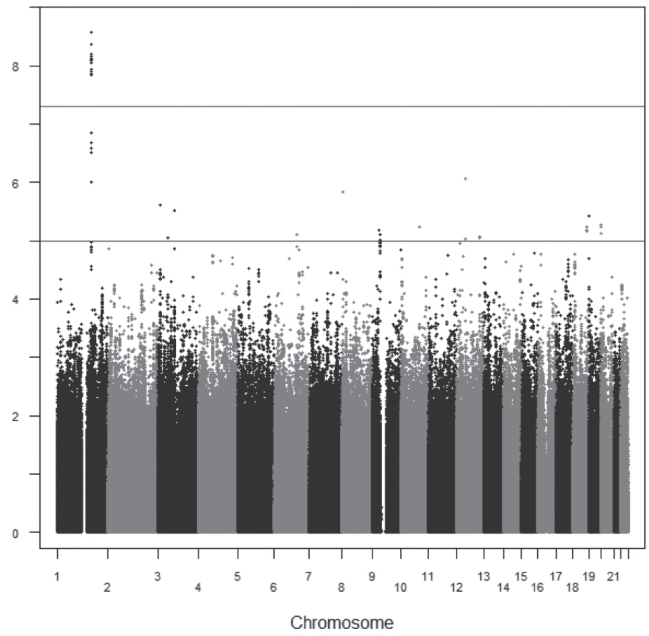


Supplemental Figure 1. Flow diagram of patients included and excluded from GWAS analyses



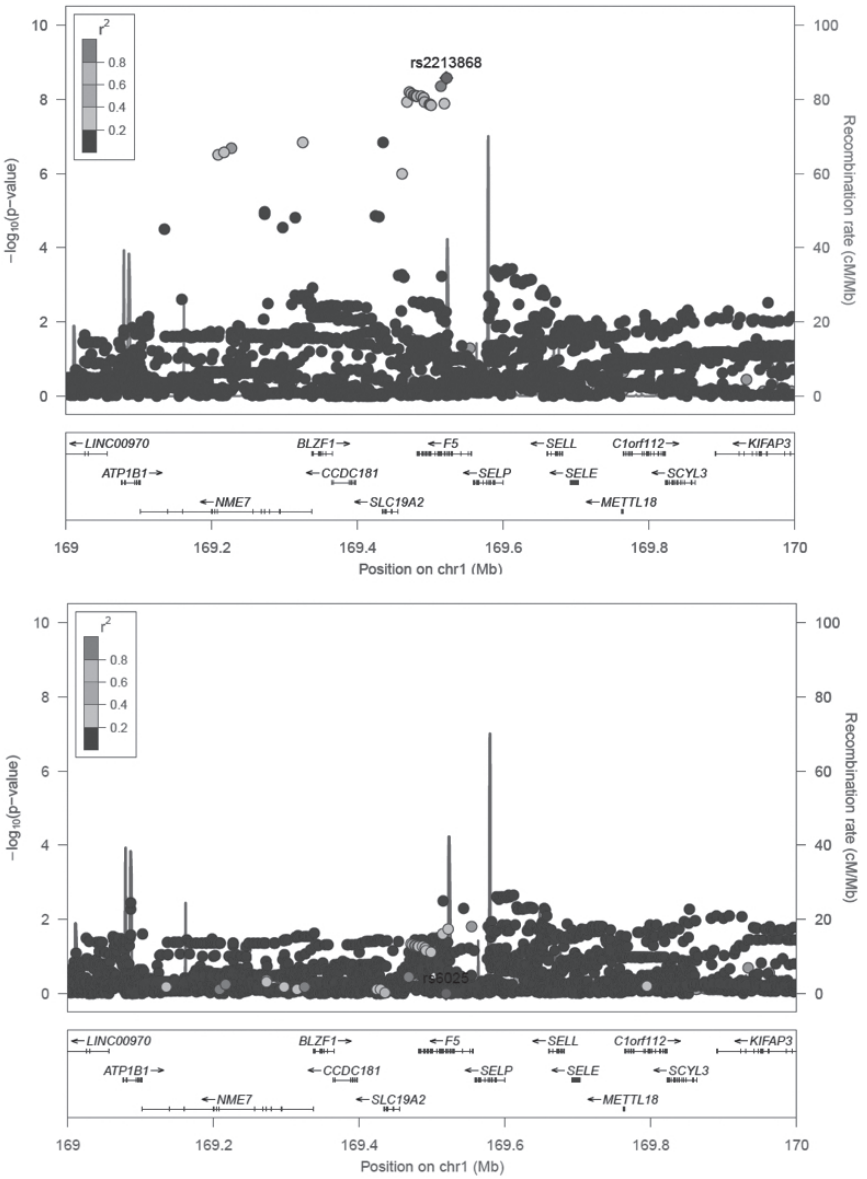
Supplemental Figure 2. Quantile-quantile plot of the genome-wide test statistics

The test statistics of the GWAS are plotted against the expected null distribution. Results are shown as $-\log_{10}$(P-values).



Supplemental Figure 3. Manhattan plot of the GWAS association results

Manhattan plot of $-\log_{10}$(P-values) for the associations between genotyped and imputed variants with recurrent venous thrombosis. We used logistic regression models to calculate the effects per copy of the minor allele, adjusted for age and sex. A total of 8.6 million autosomal variants were tested for an association with recurrent VT. The upper horizontal line at $5x10^{-8}$ represents the genome-wide significance threshold, whereas the lower line at $1x10^{-5}$ indicates the highly suggestive threshold.



Supplemental Figure 4. Regional association plots at the *F5* locus before and after conditioning on FV Leiden.

Results are shown as $-\log_{10}$(P-values) for both genotyped and imputed variants. The most associated variant in the discovery GWAS is shown as a triangle (rs2213868, upper panel). The colors of the other variants reflect the extent of linkage disequilibrium with the lead variant. The lower panel shows the association plot for recurrent VT after conditioning on the well-known FV Leiden variant (rs6025). The plots were generated using LocusZoom software.

CHAPTER 4

Male-specific risk of first and recurrent venous thrombosis: a phylogenetic analysis of the Y chromosome

de Haan HG, van Hylckama Vlieg A, van der Gaag KJ, de Knijff P, Rosendaal FR.

J Thromb Haemost. 2016;14(10):1971-1977.

ABSTRACT

Background

Recurrence risk of venous thrombosis (VT) is higher in men than in women. When excluding reproductive risk factors, this sex difference is also apparent for first VT. Current explanations for this difference are insufficient.

Objectives

We aimed to study the association between chromosome Y haplogroups and the risk of first and recurrent VT.

Methods

Y chromosomes of 3742 men (1729 patients; 2013 controls) from the MEGA case-control study were tracked into haplogroups according to the phylogenetic tree. We calculated the risk of first VT by comparing the major haplogroups with the most frequent haplogroup. For recurrence risk, 1645 patients were followed for a mean of five years, during which 350 developed a recurrence (21%, MEGA follow-up study). We calculated recurrence rates for the major haplogroups and compared groups by calculating hazard ratios.

Results

We observed 13 haplogroups, of which R1b was the most frequent (59%). The major haplogroups were not associated with first VT with odds ratios ranging from 1.01 to 1.15. Haplogroup E-carriers had the highest recurrence rate (53.5 per 1000 person-years, 95% confidence interval (CI) 33.3-86.1), whereas R1a-carriers had the lowest recurrence rate (24.3 per 1000 person-years, 95% CI 12.6-46.6). Compared with R1b-carriers, both haplogroups were not significantly associated with recurrence risk.

Conclusions

In contrast to a study on coronary artery disease, our results do not show a clear predisposing effect of Y haplogroups on first and recurrent VT risk in men. It is therefore unlikely that Y variation can explain the sex difference in VT risk.

INTRODUCTION

Venous thrombosis (VT), a common and complex disease, recurs in 20-30% of patients within five years of the first episode.[1,2] Interestingly, the risk of recurrence differs between men and women. Kyrle and colleagues observed a 5-year cumulative incidence of recurrence of 30.7% among men compared with 8.5% among women.[3] Overall, previous studies have reported a 1.5- to 3.6-fold higher recurrence risk in men than in women.[3-7] Our group was the first to suggest that the disparity by sex may not only concern recurrence risk, as we showed that men had a two-fold increased risk of a first thrombotic event compared with women when controlling for reproductive risk factors.[8]

Several explanations for the sex difference in VT risk have been proposed and, so far, only body height could explain a modest proportion.[7,9] However, almost all research has focused on recurrence risk and environmental factors. Analyses in biological and adoptive families from a nationwide Swedish registry showed stronger familial clustering in men than in women.[10,11] Similarly, the Danish twin registry reported high heritability of VT among male twins but not among female twins, providing evidence for a potential role for Y- or X-linked genetic factors.[12] Plausible candidates would be the X-chromosomal *F8* and *F9* genes, which encode coagulation factors VIII and IX. However, no sex difference in the heritability of either factor has been observed.[13] In addition, women have higher factor VIII levels than men do,[14,15] whereas factor IX levels are similar.[16] Recently, Roach et al. did not observe a difference in risk of recurrence between carriers and non-carriers of *F9* Malmö in four pooled European cohorts.[17]

Accumulating evidence suggest that genes on the male-specific region of the Y chromosome (MSY) are not only involved in sex determination and development but also in basic cellular processes.[18,19] Genetic variation on the MSY is highly conserved due to limited recombination, making traditional analysis of genetic variation almost impossible. Due to this high conservation, however, Y chromosomes can be grouped into haplogroups forming a phylogenetic tree.[20,21] Phylogenetic analyses have identified associations between Y haplogroups and several diseases, including atherosclerosis and AIDS progression.[22,23] Recently, a 50% increased risk of coronary artery disease was reported in carriers of haplogroup I compared with non-carriers.[24] The role of the Y chromosome in VT risk has not been studied before.

We hypothesized that the sex difference in first and recurrent VT risk could in part be explained by Y-linked genetic variation. We therefore studied the association between Y haplogroups and the risk of a first and recurrent VT in men of Northwestern European origin.

MATERIAL AND METHODS

Study population

We included all men with a DNA sample available from the Multiple Environmental and Genetic Assessment of risk factors for venous thrombosis (MEGA) study, which is a large population-based case-control study. Collection and ascertainment of patients have previously been described in detail.[25] Patients with a first episode of deep vein thrombosis (DVT) or a pulmonary embolism (PE) were identified at six anticoagulation clinics, which monitor outpatient treatment with vitamin K antagonists, within the Netherlands between 1999 and 2004. Control subjects were recruited by random-digit dialling and by invitation of partners of the patients.

Participants provided a blood sample or buccal swap for DNA analysis and several well-known genetic risk factors for venous thrombosis have previously been genotyped, including Factor V Leiden (rs6025), prothrombin G20210A (rs1799963) and ABO non-O blood type (rs8176719).[25] Self-reported country of birth of the patients and their parents was used to determine continental origin of the participants, and the present analyses were restricted to men of Northwestern European origin. We defined provoked venous thrombosis as recent (within 3 months before the index date) surgery, minor injury to the leg,[26] immobilization (i.e., plaster cast, bedridden at home, hospitalization), travel for more than 4 hours in 2 months before the index date, and a cancer diagnosis between 5 years before and 6 months after the index date. For the current study, we included 1811 male patients and 2037 male control subjects.

Subsequently, 1655 male patients with a first VT gave their consent to be followed for recurrence in the MEGA follow-up study. We have reported on the design and methods in detail elsewhere.[27] In brief, start of follow-up was defined as the date of the first event. Between 2007 and 2009, we retrieved the vital status of all patients from the central Dutch population register and sent questionnaires concerning recurrent VT to

all patients who were alive. Diagnosis of a recurrent event was verified by information from patients, anticoagulation clinics and treating physicians. We classified the reported recurrences into certain and uncertain recurrences according to a decision rule previously described.[28] For the current analyses, only the certain recurrences (N=350) were used as end point, and patients with an uncertain recurrence (N=80) were censored at time of their uncertain recurrent event. For the end of follow-up, we used the date of the recurrence or the date of filling in the questionnaire when no recurrence had occurred. If patients did not fill in the questionnaire, they were censored at the last date known to be recurrence free, that is, the last visit to the anticoagulant clinic (N=109), date of death (N=36) or emigration (N=0), or the last time the patient was known to be recurrence free from information of the MEGA case-control study (N=117).

In addition, we performed a sensitivity analysis for the incidence rate calculations in which start of follow-up was defined as the date of stopping anticoagulant therapy. If patients restarted anticoagulant therapy during follow-up for other reasons than a recurrent event (for example, atrial fibrillation), we considered them not at risk during these periods. Out of 1645 patients with a first VT, 176 patients left the study before stopping anticoagulant therapy, of which 10 patients developed a recurrence in this period. These patients were excluded in the sensitivity analyses. A total of 136 patients restarted anticoagulant therapy at some point during follow-up, of which four patients developed a recurrent event while using anticoagulants. If any patient left the study before stopping the anticoagulant therapy for a second time, they were censored at time of restarting the anticoagulant therapy.

Both studies were approved by the Medial Ethics Committee of the Leiden University Medical Center, and all participants gave written informed consent.

Phylogenetic analysis

To classify all participants into the major clades of the phylogenetic tree (Figure 1), we determined 26 single nucleotide variants in MSY (i.e., SRY10831, M91, M181, M145, M174, M96, P143, M213, M201, M69, M170, M304, M9, M20, P256, M214, M231, M175, M45, M242, M207, M173, M343, M124, P202, and M70) in a multiplex reaction using the SNaPshot kit (Applied Biosystems, California, USA).[21] Sequences of PCR primers used for amplification of the genomic DNA samples (1.5 ng/μl) are available upon request. After amplification, samples were treated with shrimp alkaline phosphatase (SAP,

Affymetrix, Cleveland, USA) and exonuclease I (EXOSAP-IT, Affymetrix) to eliminate remaining primers and dNTPs. Next, we performed SNaPshot minisequencing, which is a fluorescent-based primer extension method. Purified extension products were analyzed using ABI Prism 3100 Genetic Analyzer (Applied Biosystems) and evaluated with GeneMarker software (Softgenetics, State College, USA). Participants who could not be classified into one of the major haplogroups due to missing genotype data were excluded (N=34, 10 patients and 24 control subjects), leaving 1729 patients and 2013 control subjects for association analysis for first VT risk and 1645 patients for association analysis for recurrent VT risk.

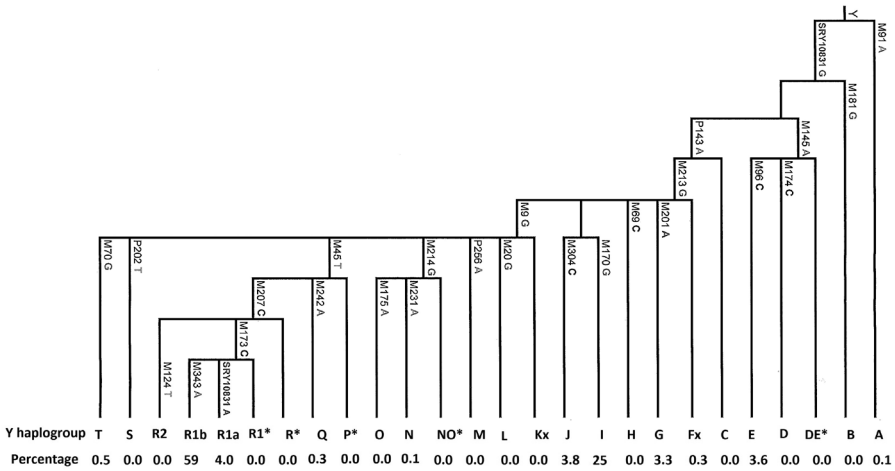


Figure 1. Phylogenetic tree of the Y chromosome and overall haplogroup distribution in MEGA.

We genotyped 26 variants in MSY to categorize Y chromosomes into lineages of the phylogenetic tree. Variants are depicted as terminal markers of the haplogroups. * and x define deeper branches unifying multiple (subclades of) haplogroups.

Statistical analysis

In order to determine the association between Y variation and the risk of a first thrombotic event, we compared carriers of the most common haplogroup with carriers of each of the other major haplogroups in the MEGA case-control study. We calculated odds ratios (OR) and corresponding 95% confidence intervals (CI) using logistic regression models, which were adjusted for age. In addition, we performed a

subanalysis in which patients were stratified on the type of the first venous thrombosis (deep venous thrombosis of the leg and pulmonary embolism).

Recurrence risk was determined by calculation of cumulative incidences and incidence rates for each of the major haplogroups in the MEGA follow-up study. For evaluation of recurrence risk, we calculated hazard ratios (HR) using age-adjusted Cox regression models with the most common haplogroup as reference group. We verified the proportional hazard assumption by evaluating the curves of the log-log survivor function.

For both the risk of a first and the risk of a recurrent event, we performed sensitivity analyses in which we adjusted for established common genetic risk factors (i.e., FV Leiden, F2 G20210A, and ABO non-O) and restricted to unprovoked VT. Analyses were carried out using statistical software packages SPSS (version 20, IBM, Armonk, NY, USA) and STATA (version 12, StataCorp, Texas, USA).

We performed a power calculation based on the results of the study on coronary artery disease by Charchar and colleagues.[24] Assuming a prevalence of 20% of haplogroup I, we had a 99.6% power level with an alpha of 0.05 to observe a 50% increased risk of a first thrombotic event in haplogroup I carriers compared with R1b-carriers. Based on our sample size, the minimum odds ratio we could have detected with a power level of 80% and an alpha of 0.05 was 1.26.

RESULTS

Y haplogroups and risk of first VT – case-control study

We included 3742 men of Northwestern European ancestry, of whom there were 1729 patients with a first thrombotic event and 2013 control subjects. Patients were slightly older than control subjects (mean age patients: 53.1 years, standard deviation (SD): 11.4 versus mean age controls: 48.2 years, SD: 12.4). VT diagnoses were as follows: 1020 (59%) patients had a first DVT in the legs only, 464 (27%) patients had a first PE only, and 245 (14%) patients had both a DVT and a PE. The thrombotic event was not precipitated by provoking risk factors in 748 (44%) patients.

For the phylogenetic analysis, we genotyped 26 biallelic Y variants that allow partitioning into the major European Y haplogroups. We observed 13 Y haplogroups among the 3742 men, of which six groups (i.e., R1b, I, R1a, J, E, and G) accounted for more than 98% of the Y lineages (Figure 1). R1b and I were the most common haplogroups, which were carried by 59% and 25% of the participants, respectively. We compared VT risk between carriers of haplogroup R1b and carriers of each of the other major haplogroups. No clear associations with VT were observed and the results did not change when restricting to unprovoked VT risk or when adjusted for the established genetic risk factors (Table 1). Although not significant, haplogroup E carriers had a weak increase in risk of unprovoked VT compared with R1b carriers (OR: 1.49, 95% CI 0.96-2.30). A subanalysis stratifying on the risk of a DVT only and risk of a PE only, did not identify any associations with the main Y haplogroups (Supplemental Table 1). If anything, carriers of haplogroup E had a higher risk of PE compared with R1b carriers (OR: 1.41, 95% CI 0.83-2.38). However, the confidence interval was wide due to low number of patients carrying this haplogroup.

Y haplogroups and risk of recurrent VT – follow-up study

A total of 1645 male VT patients gave their consent to be followed for recurrence. During a mean follow-up of 5 years (SD: 2.93), recurrent VT was confirmed in 350 men, corresponding to an incidence rate of 41.5 (95% CI 37.4-46.1) per 1000 person-years and a 5-year cumulative incidence of 20% (95% CI 18.2-22.4). Incidence rates and 5-year cumulative incidences for the six most common haplogroups are reported in Table 2. Haplogroup E carriers had the highest risk of recurrent VT with an incidence rate of 53.5 (95% CI 33.3-86.1) per 1000 person-years and a 5-year cumulative incidence of 26.3% (95% CI 16.5-40.5). The incidence rate of recurrence for carriers of haplogroup R1a was lowest at 24.3 (95% CI 12.6-46.6) per 1000 person-years and the 5-year cumulative incidence was 14.5% (95% CI 7.78-26.0) suggesting that these men were at lower risk of developing a recurrent event. Sensitivity analyses using time of stopping anticoagulant therapy as start of follow-up resulted in somewhat higher incidence rates, but did not change the overall results (Supplemental Table 2).

Table 1. Risk of a first venous thrombotic event in carriers of each of the major European Y haplogroups compared with R1b-carriers.

Haplogroup	Overall VT risk						Unprovoked VT risk		
	Patients, N	Controls, N	$OR	$95% CI	*OR	*95% CI	Patients, N	$OR	$95% CI
R1b	1022	1194	Ref	-	Ref	-	439	Ref	-
I	431	494	1.02	0.87-1.19	1.06	0.89-1.25	190	1.03	0.84-1.26
R1a	68	82	1.01	0.72-1.42	1.00	0.70-1.44	29	1.03	0.66-1.62
J	64	77	1.02	0.72-1.45	0.96	0.66-1.40	24	0.92	0.57-1.50
E	64	71	1.15	0.81-1.65	1.23	0.84-1.80	34	1.49	0.96-2.30
G	58	66	1.07	0.74-1.55	1.19	0.81-1.75	20	0.62	0.52-1.48

OR, odds ratio; CI, confidence interval; Ref, reference group; $Analyses were adjusted for age; *Analyses were additionally adjusted for the established genetic risk factors for VT (i.e., FV Leiden, F2 G20210A and ABO non-O).

Table 2. Incidence rates and cumulative incidences of recurrent VT for all men and for carriers of the six major Y haplogroups.

Haplogroup	Men, N	Recurrences, N	Sum FU in years	Incidence rate per 1000 pys (95% CI)	5-year cumulative incidence (95% CI)
All men	1645	350	8439	41.5 (37.4-46.1)	20.2 (18.2-22.4)
1st provoked VT	920	160	4697	34.1 (29.2-39.8)	16.4 (13.9-19.2)
1st unprovoked VT	713	187	3701	50.5 (43.8-58.3)	24.6 (21.4-28.2)
R1b	967	211	4935	42.8 (37.4-48.9)	20.0 (17.4-22.9)
I	409	86	2096	41.0 (33.2-50.7)	20.4 (16.6-25.0)
R1a	66	9	371	24.3 (12.6-46.6)	14.5 (7.78-26.0)
J	63	13	309	42.1 (24.5-72.5)	24.4 (14.6-39.2)
E	63	17	317	53.5 (33.3-86.1)	26.3 (16.5-40.5)
G	56	11	306	35.9 (19.9-64.8)	19.6 (10.7-34.4)

FU, follow-up since first VT event; pys, person-years; CI, confidence interval.

We calculated hazard ratios of time to recurrence for carriers of the major Y haplogroups compared with haplogroup R1b carriers (Table 3). Although not significant, inheritance of haplogroup R1a reduced the risk of a recurrence on average by 42% whereas carrying E increased the recurrence risk by 25%. We observed similar results when restricting to men with a first unprovoked event or when adjusting for the established genetic risk factors, albeit with wider confidence intervals due to the low number of individuals (Table 3). When we compared carriers of the haplogroup with the highest recurrence risk with carriers of the haplogroup with the lowest recurrence risk, we observed that haplogroup E carriers had a 2.2-fold increased risk of recurrence (95% CI 0.97-4.90) compared with men carrying haplogroup R1a, albeit the confidence interval was wide and crossed unity.

Table 3. Risk of recurrent VT in carriers of each of the major European Y haplogroups compared with R1b-carriers.

	Overall recurrence risk						Recurrence risk after first unprovoked VT			
Haplogroup	**Men**	**Recurrences, N**	**$HR**	**$95% CI**	***HR**	***95% CI**	**Men**	**Recurrences, N**	**$HR**	**$95% CI**
R1b	967	211	Ref	-	Ref	-	415	108	Ref	-
I	409	86	0.96	0.74-1.23	0.97	0.74-1.26	182	47	0.99	0.71-1.40
R1a	66	9	0.58	0.30-1.13	0.54	0.27-1.10	28	4	0.47	0.17-1.27
J	63	13	0.98	0.56-1.72	0.82	0.43-1.55	24	6	1.09	0.48-2.49
E	63	17	1.25	0.76-2.04	1.21	0.72-2.05	34	14	1.79	1.02-3.12
G	56	11	0.86	0.47-1.57	0.90	0.49-1.66	19	5	0.98	0.40-2.41

HR, hazard ratio; CI, confidence interval; Ref, reference group; $ Analyses were adjusted for age; * analyses were additionally adjusted for FV Leiden, F2 G20210A and ABO non-O.

DISCUSSION

So far, none of the proposed explanations for the sex difference in VT risk have proven to be sufficient. We hypothesized that male predisposition to venous thrombotic events may be determined by the Y chromosome. This is the first study to explore the association between genetic variation in MSY and the risk of a first and recurrent venous thrombosis. Identification of a male-specific risk factor for venous thrombosis would aid in risk stratification and unraveling the pathophysiology of VT.

We did not observe a clear association between any of the major European Y haplogroups and risk of a first VT, as almost all risk estimates were close to unity. For risk of unprovoked VT, carriers of haplogroup E had a mild increased risk. In contrast, Charchar and colleagues reported a 1.5-fold increased risk of coronary artery disease in carriers of haplogroup I compared with non-carriers.[24] The lack of association between haplogroup I and VT could be explained by differences in disease mechanism. Although several links between arterial and venous thrombosis have been described, they are generally regarded as separate diseases with shared risk factors.[29] Our results suggest that the proposed mechanism of haplogroup I, i.e., down regulation of two MSY genes (*UTY* and *PRKY*) in macrophages,[30] does not play an important role in venous thrombosis.

For the risk of recurrent VT, we also did not observe a strong association with any of the major Y haplogroups, although carriers of haplogroup R1a had a somewhat decreased risk of recurrence. In addition, in line with our findings for risk of first VT, recurrence risk was highest for carriers of haplogroup E. The recurrence rate was similar to that for men with a first unprovoked VT event. Both findings were consistent when restricting to unprovoked VT risk or when adjusting for the established genetic risk factors. This suggests that our results were not influenced by differences in the major risk factors for VT. The prevalence of haplogroup R1a and E in our study population were 5.0% and 4.5%, respectively. To confirm that carriers of haplogroups R1a and E have differential risk of recurrent VT, follow-up in a large and well-characterized study population with a higher prevalence of these haplogroups would be needed. R1a is a wide-spread Y haplogroup with branches both in Europe and Asia. The haplogroup is estimated to have arrived in Europe over 20,000 years ago.[31-33] Nowadays, the European clade of R1a is most frequent in East-Europe, with different branches exceeding a frequency of 20% in the population.[31] Haplogroup E is the predominant haplogroup on the African continent. However, a subclade (E1b1b) entered Europe via the Middle East more than 10,000

years ago during the Neolithization of Europe.[32,33] This subclade reaches frequencies up to 25% in Europe with a distinct South to North gradient.[32]

Our study has several limitations. Possibly, due to a limited sample size, we may have missed associations between haplogroups and the risk of venous thrombosis. However, it is unlikely that we have missed an association between haplogroup I (which was associated with coronary artery disease[24]) and venous thrombosis as our study was adequately powered to detect a similar association. As the prevalence of the other haplogroups was much smaller, we can therefore not rule out that we have missed an association with VT. Sample size was even smaller for the analyses of recurrence risk, which was reflected by the wide confidence intervals, and, therefore caution is needed in the interpretation of our findings both regarding an association or the lack thereof.

As Y haplogroups are highly geographically differentiated, a further limitation of our study is the inability to rule out the presence of population stratification. To limit the possibility that our data reflects recent admixture, we excluded all men who reported that their parents were born outside Northwest Europe. We did not observe an association between the major haplogroups and any of the established genetic risk factors, which are known to vary in allele frequency between populations of different origin.[34] In addition, the haplogroup distribution in the controls was in range with what has previously been reported for The Netherlands.[33,35,36] For example, a study of men with a confirmed paternal ancestor born in the Dutch province Noord-Brabant before 1800, reported the following Y haplogroup distribution: 3.8% E, 3.0% G, 16% I, 7.6% J, 3.0% R1a and 65% R1b.[34] Of note, the estimates are often based on small sample sizes and show spatial and temporal differences.

A potential source of bias could be survival bias, as we included patients who survived a first venous thrombotic event. However, the impact of survival bias on our results is probably limited, as it is unlikely that survival differed between the carriers of the Y haplogroups.

Among the strengths of this study are the long follow-up period and the objectively confirmed recurrent VT events. Furthermore, this is the first study to explore variation in the Y chromosome as a male-specific risk factor for VT.

Even if carriers of haplogroups R1a and E have a slightly different recurrence risk, our results do not show a clear predisposing effect of variation in MSY on recurrence risk which can explain the inequity by sex. For comparison, 212 out of 1868 female patients from the MEGA study developed a recurrence during follow-up, corresponding to an incidence rate of 18.4 (95% CI 15.9-20.9) per 1000 person-years. This rate is still lower than the recurrence rate in haplogroup R1a carriers. However, it is possible that we missed minor Y-linked contributions to VT risk by rare Y haplogroups or subgroups. Alternative explanations could be X-linked factors or differential gene expression of autosomal genes. In conclusion, our data suggest that Y-linked variation plays a limited role in risk of venous thrombosis.

REFERENCES

1. Hansson PO, Sörbo J, Eriksson H. Recurrent venous thromboembolism after deep vein thrombosis: incidence and risk factors. *Arch Intern Med.* 2000;160:769-774.
2. Prandoni P, Noventa F, Ghirarduzzi A, Pengo V, Bernardi E, Pesavento R, Iotti M, Tormene D, Simioni P, Pagnan A. The risk of recurrent venous thromboembolism after discontinuing anticoagulation in patients with acute proximal deep vein thrombosis or pulmonary embolism. A prospective cohort study in 1,626 patients. *Haematologica.* 2007;92:199-205.
3. Kyrle PA, Minar E, Bialonczyk C, Hirschl M, Weltermann A, Eichinger S. The risk of recurrent venous thromboembolism in men and women. *N Engl J Med.* 2004;350:2558–63.
4. Christiansen SC, Cannegieter SC, Koster T, Vandenbroucke JP, Rosendaal FR. Thrombophilia, clinical factors, and recurrent venous thrombotic events. *JAMA.* 2005;293:2352-61.
5. McRae S, Tran H, Schulman S, Ginsberg J, Kearon C. Effect of patient's sex on risk of recurrent venous thromboembolism: a meta-analysis. *Lancet.* 2006;368:371-8.
6. Rodger MA, Kahn SR, Wells PS, Anderson DA, Chagnon I, Le Gal G, Solymoss S, Crowther M, Perrier A, White R, Vickars L, Ramsay T, Betancourt MT, Kovacs MJ. Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy. *CMAJ.* 2008;179:417-426.
7. Douketis J, Tosetto A, Marcucci M, Baglin T, Cosmi B, Cushman M, Kyrle P, Poli D, Tait RC, Iorio A. Risk of recurrence after venous thromboembolism in men and women: patient level meta-analysis. *BMJ.* 2011;342:d813.
8. Roach RE, Lijfering WM, Rosendaal FR, Cannegieter SC, le Cessie S. Sex difference in risk of second but not of first venous thrombosis: paradox explained. *Circulation.* 2014;129:51-6.
9. Roach RE, Cannegieter SC, Lijfering WM. Differential risks in men and women for first and recurrent venous thrombosis: the role of genes and environment. *J Thromb Haemost.* 2014;12:1593-600.
10. Zöller B, Ohlsson H, Sundquist J, Sundquist K. Familial risk of venous thromboembolism in first-, second- and third-degree relatives: a nationwide family study in Sweden. *Thromb Haemost.* 2013;109:458-63.
11. Zöller B, Li X, Sundquist J, Sundquist K. Familial transmission of venous thromboembolism: a cohort study of 80 214 Swedish adoptees linked to their biological and adoptive parents. *Circ Cardiovasc Genet.* 2014;7:296-303.
12. Larsen TB, Sørensen HT, Skytthe A, Johnsen SP, Vaupel, JW, Christensen K. Major genetic susceptibility for venous thromboembolism in men: a study of Danish twins. *Epidemiology. 2003;14:328-32.*

13. Orstavik KH, Magnus P, Reisner H, Berg K, Graham JB, Nance W. Factor VIII and factor IX in a twin population: evidence for a major effect of ABO locus on factor VIII level. *Am J Hum Genet.* 1985;37:89–101.
14. Tracy RP, Bovill EG, Fried LP, Heiss G, Lee MH, Polak JF, Psaty BM, Savage PJ. The distribution of coagulation factors VII and VIII and fibrinogen in adults over 65 years. Results from the Cardiovascular Health Study. *Ann Epidemiol.* 1992;2:509-19.
15. Conlan MG, Folsom AR, Finch A, Davis CE, Sorlie P, Marcucci G, Wu KK. Associations of factor VIII and von Willebrand factor with age, race, sex, and risk factors for atherosclerosis. The Atherosclerosis Risk in Communities (ARIC) Study. *Thromb Haemost.* 1993;70:380-5.
16. van Hylckama Vlieg A, van der Linden IK, Bertina RM, Rosendaal FR. High levels of factor IX increase the risk of venous thrombosis. *Blood.* 2000;95:3678-82.
17. Roach RE, Lijfering WM, Tait RC, Baglin T, Kyrle PA, Cannegieter SC, Rosendaal FR. Sex difference in the risk of recurrent venous thrombosis: a detailed analysis in four European cohorts. *J Thromb Haemost*. 2015;13:1815-22.
18. Lahn BT, Page DC. Functional coherence of the human Y chromosome. *Science.* 1997;278:675-80.
19. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature.* 2003;423:825-37.
20. Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* 2002;12:339–48.
21. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 2008;18:830-8.
22. Voskarides K, Hadjipanagi D, Papazachariou L, Griffin M, Panayiotou AG. Evidence for contribution of the y chromosome in atherosclerotic plaque occurrence in men. *Genet Test Mol Biomarkers.* 2014;18:552-6.
23. Sezgin E, Lind JM, Shrestha S, Hendrickson S, Goedert JJ, Donfield S, Kirk GD, Phair JP, Troyer JL, O'Brien SJ, Smith MW. Association of Y chromosome haplogroup I with HIV progression, and HAART outcome. *Hum Genet.* 2009;125:281-94.
24. Charchar FJ, Bloomer LD, Barnes TA, Cowley MJ, Nelson CP, Wang Y, Denniff M, Debiec R, Christofidou P, Nankervis S, Dominiczak AF, Bani-Mustafa A, Balmforth AJ, Hall AS, Erdmann J, Cambien F, Deloukas P, Hengstenberg C, Packard C, Schunkert H, et al. Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. *Lancet.* 2012;379:915-22.
25. Blom JW, Doggen CJ, Osanto S, Rosendaal FR. Malignancies, prothrombotic mutations, and the risk of venous thrombosis. *JAMA.* 2005;293:715–22.

26. van Stralen KJ, Rosendaal FR, Doggen CJ. Minor injuries as a risk factor for venous thrombosis. *Arch Intern Med*. 2008;168:21-6.
27. van Hylckama Vlieg A, Flinterman LE, Bare LA, Cannegieter SC, Reitsma PH, Arellano AR, Tong CH, Devlin JJ, Rosendaal FR. Genetic variations associated with recurrent venous thrombosis. *Circ Cardiovasc Genet.* 2014;7:806-13.
28. Timp JF, Lijfering WM, Flinterman LE, van Hylckama Vlieg A, le Cessie S, Rosendaal FR, Cannegieter SC. Predictive value of factor VIII levels for recurrent venous thrombosis: results from the MEGA follow-up study. *J Thromb Haemost*. 2015;13:1823-32.
29. Lijfering WM, Flinterman LE, Vandenbroucke JP, Rosendaal FR, Cannegieter SC. Relationship between venous and arterial thrombosis: a review of the literature from a causal perspective. *Semin Thromb Hemost.* 2011;37:885-96.
30. Bloomer LD, Nelson CP, Eales J, Denniff M, Christofidou P, Debiec R, Moore J, Consortium C, Zukowska-Szczechowska E, Goodall AH, Thompson J, Samani NJ, Charchar FJ, Tomaszewski M. Male-specific region of the Y chromosome and cardiovascular risk: phylogenetic analysis and gene expression studies. *Arterioscler Thromb Vasc Biol*. 2013;33:1722-7.
31. Underhill PA, Poznik GD, Rootsi S, Järve M, Lin AA, Wang J, Passarelli B, Kanbar J, Myres NM, King RJ, Di Cristofaro J, Sahakyan H, Behar DM, Kushniarevich A, Šarac J, Šarec T, Rudan P, Kumar Pathak A, Chaubey G, Grugni V, et al. The phylogenetic and geographic structure of Y-chromosome haplpogroup R1a. *Eur J Hum Genet.* 2015;23:124-31.
32. Wiik, K. Where did European men come from? *J Genet Genealogy.* 2008;4:35-85.
33. Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science.* 2000;290:1155-9.
34. Tang L, Hu Y. Ethnic diversity in the genetics of venous thromboembolism. *Thromb Haemost.* 2015;114:901-9.
35. Larmuseau MH, Vanderheyden N, Jacobs M, Coomans M, Larno L, Decorte R. Micro-geographic distribution of Y-chromosomal variation in the central-western European region Brabant. *Forensic Sci Int Genet.* 2011;5:95-9.
36. Larmuseau MH, Ottoni C, Raeymaekers JA, Vanderheyden N, Larmuseau HF, Decorte R. Temporal differentiation across a West-European Y-chromosomal cline: genealogy as a tool in human population genetics. *Eur J Hum Genet.* 2012;20:434-40.

Supplemental Table 1. Risk of a first deep vein thrombosis and pulmonary embolism in carriers of each of the major European Y haplogroups compared with R1b-carriers.

		Risk of a first DVT only		Risk of a first PE only	
Haplogroups	**[$]Controls, N (%)**	**[$]Patients, N (%)**	***OR (95% CI)**	**[$]Patients, N (%)**	***OR (95% CI)**
R1b	1194 (60)	600 (59)	Ref	276 (60)	Ref
I	494 (25)	254 (25)	1.01 (0.84-1.22)	119 (26)	1.04 (0.82-1.32)
R1a	82 (4.1)	44 (4.3)	1.12 (0.76-1.65)	15 (3.2)	0.82 (0.46-1.46)
J	77 (3.9)	40 (3.9)	1.08 (0.72-1.61)	14 (3.0)	0.87 (0.48-1.57)
E	71 (3.6)	35 (3.4)	1.07 (0.70-1.64)	20 (4.3)	1.41 (0.83-2.38)
G	66 (3.3)	37 (3.6)	1.14 (0.75-1.74)	13 (2.8)	0.87 (0.47-1.61)

DVT, deep vein thrombosis; PE, pulmonary embolism; OR, odds ratio; CI, confidence interval; Ref, reference group; [$]Percentages calculated within the carriers of the six major European Y haplogroups. *Analyses were adjusted for age.

Supplemental Table 2. Sensitivity analyses for the incidence rates of recurrent VT.

Haplogroup	Men, N	Recurrences, N	Sum FU in years	Incidence rate per 1000 pys (95% CI)
All men	1469	336	6880	48.8 (43.9-54.4)
1st provoked VT	814	155	3859	40.2 (34.3-47.0)
1st unprovoked VT	648	178	2993	59.5 (51.3-68.9)
R1b	858	200	4057	49.3 (42.9-56.6)
I	367	83	1676	49.5 (40.0-61.4)
R1a	63	9	312	28.8 (15.0-55.4)
J	60	13	257	50.6 (29.4-87.1)
E	53	17	258	66.0 (41.0-106)
G	50	11	256	42.9 (23.8-77.5)

FU, follow-up since the date of stopping anticoagulant therapy; pys, person-years; CI, confidence interval.

CHAPTER 5

Genetic variants in Cell Adhesion Molecule 1 (CADM1): a validation study of a novel endothelial cell venous thrombosis risk factor

de Haan HG, Bezemer ID, Vossen CY, van Hylckama Vlieg A, Böehringer S, Hasstedt SJ, Levy S, Rosendaal FR, Bovill EG.

Thromb Res. 2014;134(6):1186-92.

ABSTRACT

Introduction

In a protein C deficient family, we recently identified a candidate gene, *CADM1*, which interacted with protein C deficiency in increasing the risk of venous thrombosis (VT). This study aimed to determine whether *CADM1* variants also interact with protein C pathway abnormalities in increasing VT risk outside this family.

Materials and methods

We genotyped over 300 *CADM1* variants in the population-based MEGA case-control study. We compared VT risks between cases with low protein C activity (N=194), low protein S levels (N=23), high factor VIII activity (N=165) or factor V Leiden carriers (N=580), and all 4004 controls. Positive associations were repeated in all 3496 cases and 4004 controls.

Results

We found 22 variants which were associated with VT in one of the protein C pathway risk groups. After mutual adjustment, six variants remained associated with VT. The strongest evidence was found for rs220842 and rs11608105. For rs220842, the odds ratio (OR) for VT was 3.2 (95% CI 1.2-9.0) for cases with high factor VIII activity compared with controls. In addition, this variant was associated with an increased risk of VT in the overall study population (OR 1.5, 95% CI 1.0-2.2). The other variant, rs11608105, was not associated with VT in the overall study population (OR 1.0, 95% CI 0.8-1.1), but showed a strong effect on VT risk (OR 21, 95% CI 5.1-88) when combined with low protein C or S levels.

Conclusions

In a population-based association study, we confirm a role for *CADM1* variants in increasing the risk of VT by interaction with protein C pathway abnormalities.

INTRODUCTION

We have identified a candidate gene, cell adhesion molecule 1 (*CADM1*), which appears to interact with protein C deficiency to increase the risk of venous thrombosis in an extended French Canadian family with type I protein C deficiency due to a *PROC* 3363C insertion ("Vermont family").[1] The 300kb *CADM1* gene is also known as nectin-like protein 2 (*NECL2*), tumor suppressor in lung cancer 1 (*TSLC1*), synapse cell adhesion molecule (*SynCAM1*), spermatogenic immunoglobulin super family (*SgIGSF*), and immunoglobulin super family 4 (*IGSF4*).[2-6] CADM1, an immunoglobulin cell adhesion molecule involved in binding interactions supporting intercellular adhesion, has been best characterized as a constitutive cell-cell adhesion molecule in epithelial cells and at neuronal synapses.[4,5]

In the Vermont family study, several single nucleotide variants (SNVs) in *CADM1* showed a strong association with venous thrombosis in interaction with protein C deficiency.[1] For example, among protein C deficient family members, carriers of the rs6589488 minor allele had a 17-fold increased risk of venous thrombosis (OR 17, 95% CI 13.5-21.4) compared with homozygous major allele carriers. Subsequent *CADM1* gene expression assays, using blood outgrowth endothelial cells cultured from family members, showed a decreased expression compared with controls, lending phenotypic support to the SNV associations. We also demonstrated *CADM1* in endothelial cells, where it appears to be selectively involved in endothelial cell migration, suggesting a role in maintenance of endothelial barrier function.[1,7]

Activated Protein C, bound to the endothelial protein C receptor (APC-EPCR) on the endothelial membrane, mediates endothelial barrier enhancement through activation of protease activated receptor 1 (PAR-1) and the sphingosine-1-phosphate-receptor-1 (S_1P_1) pathways.[8-12] This APC-EPCR mediated activation of PAR-1 and S_1P_1 leads to activation of endothelial Rac1 and the cytoskeletal rearrangements associated with endothelial barrier enhancement.[10,11,13] The *CADM1* pathway,[14] which is associated with migration and adhesion in epithelial cells, appears to mediate this epithelial cell behavior, in part, through regulating small Rho-GTPases including Rac1.[15,16] This suggests that our observation of a strong interaction between the *CADM1* and protein C genes in increasing thrombosis risk in the Vermont family may be related to a shared common signalling pathway involving the small Rho-GTPases. Thus, the *CADM1* pathway interaction with the protein C system may represent a novel biological pathway

conferring increased risk for venous thrombosis at the level of the vessel wall due to impaired maintenance of endothelial barrier function.

In order to validate the association between *CADM1* and thrombosis observed in the Vermont family study, we investigated *CADM1* gene variants in the Multiple Environmental and Genetic Assessment of risk factors for venous thrombosis (MEGA study), a case-control study on venous thrombosis including over 4000 patients and 4000 controls. To study the effect of *CADM1* variants on thrombosis risk, we primarily focused on subsets of thrombosis patients with protein C pathway abnormalities (i.e. low levels of protein C or S, high factor VIII levels, and the factor V Leiden variant) as *CADM1* variants were found to interact with protein C deficiency in the Vermont family study.[1] Protein S interacts closely with protein C in the inactivation of the procoagulant factors Va and VIIIa,[17] and synergistic effects of *CADM1* with protein C deficiency might therefore also occur with protein S deficiency, high levels of factor VIII, or activated protein C resistance due to factor V Leiden (*F5*, rs6025).

MATERIAL AND METHODS

Study population

The MEGA study is a population-based case-control study.[18,19] Consecutive patients aged 18 to 70 years with a first venous thrombosis of the leg or arm, or with a pulmonary embolism were recruited from 6 anticoagulation clinics in the western part of the Netherlands between 1999 and 2004. Partners of patients, as well as additional individuals recruited by random digit dialling and frequency-matched on age and sex, were invited as control subjects. All participants received a standardized questionnaire about risk factors for venous thrombosis. A blood sample was taken approximately 3 months after discontinuation of anticoagulant therapy (usually 3-12 months after the diagnosis of venous thrombosis), or after a year when patients continued their anticoagulant therapy, and from control subjects. Participants who refused to or were unable to provide a blood sample and patients and their partners included after June 1, 2002 were offered the option of providing a buccal swab sample for DNA. Exclusion criteria were previous venous thrombosis (patients and controls), no venous thrombosis (patients, after checking hospital records), age younger than 18 or older than 70, severe psychiatric problems, inability to speak Dutch and, for genetic and

blood sample analysis, poor sample quality. For the present analysis, we only included individuals from North- or Western European origin (90%), which was assessed by self-reported country of birth of the parents, in order to avoid population stratification. This left 1970 patients and 2490 control subjects (N=4460) with a plasma and DNA sample and another 1526 patients and 1514 control subjects (N=3040) with only a DNA sample eligible for analysis.

Protein C pathway abnormalities

We selected individuals with protein C pathway abnormalities, i.e., low protein C activity, low protein S levels, high factor VIII activity levels, or factor V Leiden carriership. The protein C, protein S and factor VIII abnormalities were not individually diagnosed, but instead we used clinically relevant cut-off levels to categorize individuals as abnormal. Low protein C activity was defined by taking the lower limit of normal (67% of normal in our laboratory) as cut-off point. When individuals were on oral anticoagulant therapy at time of blood draw, we calculated the expected protein C activity relative to factor VII activity by linear regression according to a method described by O'Brien et al.[20] The observed levels were classified as "low" when the observed/expected ratio was below the geometric mean minus 2 standard deviations as calculated among control subjects. Of 1959 patients and 2471 control subjects with protein C (and factor VII) measurements, 194 patients (10%; mean protein C activity 43% of normal; range 19-66) and 28 control subjects (1%; mean protein C activity 42% of normal; range 30-62) had low protein C activity. Of these 194 patients and 28 control subjects, 178 patients and 21 controls were on oral anticoagulant therapy at the time of the blood draw.

Similarly to the selection of individuals with low protein C, we selected low protein S individuals by selecting total protein S levels below the lower limit of normal (67% of normal) for individuals not on oral anticoagulant therapy at the time of the blood draw and calculated protein S levels relative to factor II for patients using oral anticoagulant therapy at the time of the blood draw. Of the 1828 patients and 2252 control subjects with protein S (and factor II) measurements, 23 patients (1%; mean protein S level 58% of normal; range 32-66) and 26 controls (1%; mean protein S level 60% of normal; range 45-67) had low protein S levels. Of these 33 patients and 28 controls, 3 patients and none of the controls were on oral anticoagulant therapy at the time of the blood draw.

High factor VIII was defined as activity levels higher than the geometric mean plus 2 standard deviations as calculated among control subjects, which was 204 IU/ml. In total, 165 (8%) of 1969 patients and 51 (2%) of 2488 control subjects with factor VIII levels available had high factor VIII activity levels.

For the factor V Leiden subgroup analysis, we selected 580 (17%) patients and 219 (5%) control subjects who carried the variant from among 3493 patients and 4000 control subjects with factor V Leiden genotypes available.

Laboratory analysis

Collection and processing of blood and buccal swab samples, subsequent DNA isolation and genotyping of factor V Leiden variant have been described previously.[18] Measurements of protein C activity were performed with a chromogenic assay and factor II, VII and VIII activity measurements were based on clotting time assays using immune-depleted plasma, deficient for the factor under study. These measurements were performed on a STA-R coagulation analyzer following the instructions of the manufacturer (Diagnostica Stago, Asnières, France). Total protein S levels were measured by an enzyme-linked immunosorbent assay (ELISA, Diagnostica Stago, Asnières, France). The mean intra- and inter-assay coefficients of variation in our laboratory were 1.4% and 3.5%, respectively, for protein C, 2.7% and 4.2% for factor II, 3.4% and 4.0% for factor VII, 3.6% and 8.9% for factor VIII and 5.0% and 3.5% for protein S. All measurements were performed on a single blood draw.

SNV Selection

We selected 364 SNVs throughout *CADM1* and 2kb downstream and 10kb upstream of the gene in order to include conserved elements which may play a regulatory role (chr11:114,543,000-114,893,000, NCBI B36 assembly). From the *CADM1* SNVs that were genotyped in the European HapMap population, we chose 86 tagging SNVs with minor allele frequency (MAF)>0.01 by pairwise tagging (r^2>0.8) as implemented in Haploview.[21] From the HapMap list we added 42 SNVs from blocks with multiple SNVs for redundancy and 29 SNVs in regions where the distance between adjacent SNVs was largest. In addition, we selected 99 SNVs that had not been genotyped by HapMap but were validated in dbSNP and 108 SNVs that we identified by resequencing the region in the Vermont family. Of 364 SNVs selected for genotyping, 47 were excluded because of poor assay performance, 3 SNV assays were excluded because of atypical clustering

and 30 were not polymorphic in the MEGA study population, which left 284 SNVs for statistical analysis. Genotyping was performed at the Johns Hopkins University through the NHLBI Genotyping and Resequencing Service. Genotyping quality was assessed by establishing the call rate (>99%) and the Hardy-Weinberg equilibrium of each SNV.

Statistical analysis

The primary analysis was to compare allele frequencies between patients with specific abnormalities in the protein C pathway (i.e. low protein C, low protein S, high factor VIII or factor V Leiden) and all control subjects. The choice for taking all control subjects as a reference group was made because few control subjects had low protein C activity or low protein S levels.

Odds ratios (OR) and 95% confidence intervals (95% CI) were computed using logistic regression for an additive genetic model. The reference allele was the most prevalent (major) allele in the total study population and the OR was calculated per additional minor allele copy. Variants that were associated with venous thrombosis in the primary analysis (one of the subgroups of protein C pathway abnormalities versus all controls) with p-value <0.05 were further studied. Next, linkage disequilibrium (LD) between SNVs of interest was studied in Haploview.[21] Of the variants that were in strong linkage disequilibrium, defined as r^2 of 0.7 or higher, we selected the variant with the highest allele frequency in controls for follow-up. To assess the causal effects of the SNVs, we mutually adjusted the associations by entering all positive variants into a conditional logistic regression model. Positive associations were repeated in the overall MEGA study (3496 cases and 4004 controls) and studied for the joint effects of the variants and the protein C pathway abnormality under study.

With more than 250 variants tested for association with venous thrombosis in each subgroup, the chance of false positive findings is substantial. In order to decrease the chance of false-positive reporting, we calculated an FDR-adjusted q-value.[22]

RESULTS

Characteristics of the study population are presented in Table 1. We studied 284 variants in four subgroups of venous thrombosis patients with a protein C pathway abnormality,

i.e., patients with low protein C activity (N=194), patients with low protein S levels (N=23), patients with high FVIII activity (N=165) and patients carrying the FV Leiden polymorphism (N=580), and all controls (N=4004). The subgroups were not mutually exclusive, i.e., 72 patients (12%) had multiple abnormalities in the protein C pathway.

Table 1. Characteristics of the MEGA study population.

	Patients (N=3496)	**Controls (N=4004)**
Men (%)	1633 (46.7)	1892 (47.3)
Mean age (SD)	49.18 (12.81)	48.40 (12.36)
FVL carrier (%)	580 (16.60)	219 (5.48)
Plasma available	1970	2490
Low protein C (%)	194 (9.90)	28 (1.13)
Low protein S (%)	23 (1.26)	26 (1.15)
High factor VIII (%)	165 (8.38)	51 (2.05)

SD standard deviation; FVL Factor V Leiden
Low protein C was defined as activity levels below 67% of normal or when on anticoagulant treatment relative to factor VII (see Methods). Similarly, low protein S was defined as activity levels below 67% of normal or when on anticoagulant treatment relative to factor II (see Methods). High factor VIII was defined as activity levels higher than the geometric mean plus two standard deviations among controls (see Methods).

Associations between CADM1 variants and VT within protein C pathway subgroups

For all 284 variants, allele frequencies among all MEGA study patients and all MEGA study controls are listed in Supplemental Table 1. Twelve of the 284 variants were monomorphic among control subjects and eight were monomorphic among patients of the overall MEGA study. In addition, several variants were monomorphic in one of the subgroups of patients with a protein C pathway abnormality: 16 variants among patients with low protein C activity, 46 variants among patients with low protein S levels, 14 variants among patients with high factor VIII activity and 17 variants among patients carrying factor V Leiden. These variants could not be studied.

During the first stage of the analysis, we identified 22 *CADM1* variants that were associated with venous thrombosis (p-value<0.05) in one of the subgroups of patients with a protein C pathway abnormality and all controls (Table 2). One variant was associated with venous thrombosis in the low protein C subgroup, nine variants in the low protein S subgroup, six variants in the high factor VIII subgroup, and seven variants in the factor V Leiden subgroup (Table 2). Only one variant (rs11608105) was associated

with venous thrombosis in multiple subgroups, i.e. the low protein C subgroup (OR 1.57, 95% CI 1.05-2.34) and the low protein S subgroup (OR 2.98, 95% CI 1.27-7.02). To correct for multiple testing, we calculated FDR-adjusted q-values after which none of the variants remained associated with venous thrombosis (Table 2).

Table 2. Associations with venous thrombosis in the different subgroups.

	Risk allele frequency, %					FDR
	Patients	Controls	OR	95% CI	p-value	q-value
Low protein C patients						
rs11608105	7.22	4.72	1.57	1.05-2.34	0.026	1
Low protein S patients						
rs4938182	32.6	19.8	1.95	1.05-3.63	0.034	0.756
rs4450197	8.70	2.04	4.95	1.67-14.7	0.004	0.333
rs10128746	13.0	3.63	4.40	1.75-11.1	0.002	0.333
rs11215418	10.9	3.62	3.37	1.29-8.83	0.013	0.371
rs45595941	4.35	0.70	6.71	1.54-29.3	0.011	0.371
rs45616036	4.35	0.84	5.03	1.25-20.3	0.023	0.575
rs11608105	13.0	4.72	2.98	1.27-7.02	0.013	0.371
rs45520832	2.17	0.11	20.1	2.45-166	0.005	0.333
rs45583332	4.35	0.70	6.71	1.54-29.3	0.011	0.371
High factor VIII patients						
rs10891823	9.47	6.48	1.48	1.02-2.16	0.040	0.999
rs11215504	7.58	4.35	1.79	1.18-2.73	0.006	0.750
rs11215515	7.10	4.26	1.75	1.14-2.68	0.010	0.833
rs11215458	5.62	3.65	1.61	1.00-2.60	0.050	0.999
rs220842	1.52	0.51	3.02	1.18-7.74	0.022	0.999
rs10891856	9.47	5.75	1.75	1.21-2.55	0.003	0.750
Factor V Leiden patients						
rs12577709	15.9	13.6	1.19	1.01-1.42	0.041	0.988
rs45545346	1.73	3.42	0.50	0.32-0.79	0.003	0.741
rs45608938	3.89	5.42	0.71	0.52-0.97	0.032	0.988
rs17443832	3.97	5.38	0.73	0.54-0.99	0.045	0.988
rs45578937	5.10	7.07	0.71	0.54-0.93	0.014	0.865
rs45458294	4.84	6.92	0.68	0.52-0.91	0.008	0.741
rs314497	7.84	5.86	1.37	1.08-1.73	0.009	0.741

OR odds ratio; CI confidence interval; FDR false discovery rate
In the univariable analysis, 22 variants were associated with venous thrombosis. The risk allele frequency was calculated in the subgroup of cases in which the variant was identified and in the overall controls.

Next, we studied linkage disequilibrium between the positive variants. Of the 22 variants, four pairs of variants were in strong linkage disequilibrium (Figure 1; $r^2 \geq 0.7$). Of each pair of variants, the variant having the highest risk allele frequency among controls was selected for the remaining analyses. To study the causal effects of the positive variants on venous thrombosis, we entered all positive variants within each subgroup in a logistic regression model. In the subgroup of protein S, two variants remained associated with venous thrombosis, i.e., rs11608105 and rs45520832 (Table 3; OR 3.54, 95% CI 1.46-8.60 and OR 22.1, 95% CI 2.35-208, respectively). In addition, two variants, i.e., rs11215504 and rs220842, remained associated with venous thrombosis in patients with high factor VIII activity (OR 1.89, 95% CI 1.24-2.88 and OR 3.23, 95% CI 1.17-8.97, respectively). In the patients that carried FV Leiden, another two variants, i.e., rs45608938 and rs45545346, remained associated with a decreased risk of venous thrombosis (Table 3; OR 0.71, 95% CI 0.52-0.97 and OR 0.53, 95% CI 0.30-0.93 respectively).

Table 3. Mutually adjusted associations with venous thrombosis in the different subgroups.

	Risk allele frequency, %			
	Patients	**Controls**	**OR**	**95% CI**
Low protein C patients				
rs11608105	7.22	4.72	1.57	1.05-2.34
Low protein S patients				
rs4938182	32.6	19.8	1.60	0.79-3.22
rs4450197	8.70	2.04	1.22	0.21-7.25
rs10128746	13.0	3.63	3.04	0.72-13.0
rs45616036	4.35	0.84	1.92	0.07-51.9
rs11608105	13.0	4.72	3.54	1.46-8.60
rs45520832	2.17	0.11	22.1	2.35-208
rs45583332	4.35	0.70	4.27	0.15-124
High factor VIII patients				
rs10891823	9.47	6.48	1.13	0.59-2.14
rs11215504	7.58	4.35	1.89	1.24-2.88
rs11215515	7.10	4.26	1.16	0.56-2.40
rs11215458	5.62	3.65	1.03	0.42-2.50
rs220842	1.52	0.51	3.23	1.17-8.97
rs10891856	9.47	5.75	1.60	0.89-2.85

Table 3. Continued

	Risk allele frequency, %			
	Patients	**Controls**	**OR**	**95% CI**
Factor V Leiden patients				
rs12577709	15.9	13.6	1.09	0.90-1.32
rs45545346	1.73	3.42	0.53	0.30-0.93
rs45608938	3.89	5.42	0.71	0.52-0.97
rs45578937	5.10	7.07	0.93	0.66-1.30
rs314497	7.84	5.86	1.27	0.98-1.65

OR odds ratio; CI confidence interval
When including the positive associations per subgroup together in a logistic regression model, six variants remained associated with venous thrombosis.

Associations to venous thrombosis in overall MEGA study

We further investigated the six variants, which remained associated with venous thrombosis after mutual adjustment, in the overall MEGA study population in order to study the effect on venous thrombosis independently of the protein C pathway abnormalities. We observed a weak association between rs220842 and venous thrombosis (OR 1.49, 95% CI 0.99-2.24) and between rs11215504 and venous thrombosis (OR 1.14, 95% CI 0.98-1.33). The other four variants were not associated with venous thrombosis in the overall MEGA study population (Table 4).

Table 4. Associations with venous thrombosis in MEGA overall study population.

	Risk allele frequency, %			
***CADM1* variants**	**Patients**	**Controls**	**OR**	**95% CI**
rs11608105	4.57	4.72	0.97	0.83-1.13
rs45520832	0.14	0.11	1.27	0.52-3.13
rs11215504	4.94	4.35	1.14	0.98-1.33
rs220842	0.76	0.51	1.49	0.99-2.24
rs45608938	5.33	5.42	0.98	0.85-1.13
rs45545346	3.13	3.42	0.92	0.77-1.09

OR odds ratio; CI confidence interval

Joint effect of CADM1 variants and protein C pathway abnormalities

We studied the joint effect of the thrombosis associated variants and the protein C pathway abnormalities by using homozygous major allele carriers without the protein C pathway abnormality under study as a reference for the odds ratio (Table 5). The combination of carrying variant rs11608105 and having low protein C or protein S levels was associated with a 21-fold increased risk (95% CI 5.08-88.8) of venous thrombosis. Compared with non-carriers having low protein C or S levels, the risk of venous thrombosis was a 4-fold increased (95% CI 1.00-18.7) in carriers of the risk allele with low protein C or S levels.

Similar to findings in the overall MEGA study population, variant rs220842 was associated with an increased risk of venous thrombosis (OR 1.88, 95% CI 1.07-3.31; Table 5) in individuals without high factor VIII activity. The joint effect of the variant and high factor VIII activity could not be studied as only patients and no controls with high factor VIII activity carried the variant (N=5; Table 5). Furthermore, having high factor VIII activity and carrying the risk allele of variant rs11215504 was associated with a 6.5-fold increased risk of venous thrombosis. This exceeded the risk for rs11215504 or the defect alone (Table 5) albeit with a wide confidence interval due to the small number of carriers with also a defect (95% CI 2.48-17.1). For the other positive variants, no clear joint effect with a protein C pathway abnormality could be calculated (rs45520832) or was observed (rs45608938, rs45545346) (Table 5).

Table 5. Combined associations for *CADM1* SNVs with protein C pathway abnormalities and venous thrombosis.

***CADM1* variants**		**Pathway defect**		**Patients, N**	**Controls, N**	**OR**	**95% CI**
rs11608105	No	PC/PS	No	1503	1996	1(REF)	
rs11608105	Yes	PC/PS	No	123	206	0.79	0.63-1.00
rs11608105	No	PC/PS	Yes	181	49	4.91	3.55-6.77
rs11608105	Yes	PC/PS	Yes	32	2	21.3	5.08-88.8
rs45520832	No	PS	No	1798	2219	1(REF)	
rs45520832	Yes	PS	No	5	3	2.06	0.49-8.62
rs45520832	No	PS	Yes	22	26	1.04	0.59-1.85
rs45520832	Yes	PS	Yes	1	0	NA	NA
rs220842	No	FVIII	No	1775	2416	1(REF)	
rs220842	Yes	FVIII	No	29	21	1.88	1.07-3.31
rs220842	No	FVIII	Yes	160	51	4.27	3.10-5.89
rs220842	Yes	FVIII	Yes	5	0	NA	NA
rs11215504	No	FVIII	No	1638	2222	1(REF)	
rs11215504	Yes	FVIII	No	166	215	1.05	0.85-1.30
rs11215504	No	FVIII	Yes	141	46	4.16	2.96-5.84
rs11215504	Yes	FVIII	Yes	24	5	6.51	2.48-17.1
rs45608938	No	FVL	No	2591	3379	1(REF)	
rs45608938	Yes	FVL	No	318	397	1.04	0.89-1.22
rs45608938	No	FVL	Yes	533	198	3.51	2.96-4.17
rs45608938	Yes	FVL	Yes	45	19	3.09	1.80-5.29
rs45545346	No	FVL	No	2721	3524	1(REF)	
rs45545346	Yes	FVL	No	192	255	0.98	0.80-1.18
rs45545346	No	FVL	Yes	559	210	3.45	2.92-4.07
rs45545346	Yes	FVL	Yes	20	9	2.88	1.31-6.33

OR odds ratio; CI confidence interval; PC protein C; PS protein S; FVL Factor V Leiden; REF reference.

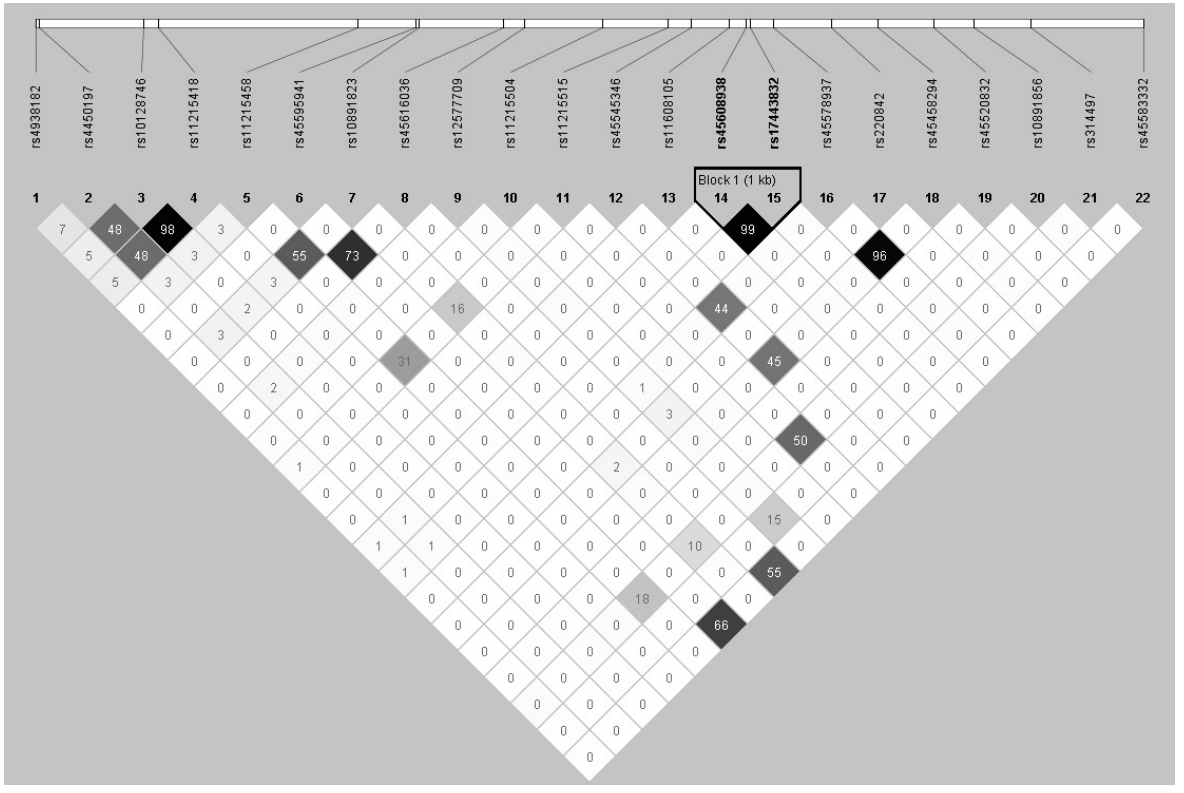


Figure 1. Linkage disequilibrium between *CADM1* variants which were associated with venous thrombosis in the univariable analysis of specific protein C pathway abnormality groups. Linkage equilibrium between the variants, as calculated in the controls, is expressed as r squared.

DISCUSSION

In this study we aimed to validate the *CADM1* gene, encoding cell adhesion molecule 1, as a gene involved in the etiology of venous thrombosis. We identified this gene as a candidate risk gene in the Vermont family.[1] The thrombosis association was most pronounced among individuals in this family with both variation in *CADM1* and protein C deficiency. To confirm the interaction of protein C deficiency and *CADM1* variants in increasing the risk of thrombosis, we studied 284 variants in *CADM1* in the population-based MEGA study. We performed analyses mainly by comparing thrombosis cases with protein C pathway abnormalities, i.e. low protein C or S levels, high factor VIII activity or factor V Leiden, with all controls.

For six variants in the *CADM1* gene, a consistent association with venous thrombosis was observed in one of the subgroups of protein C pathway abnormalities. Within individuals with low protein C or S levels, rs11608105 showed a 21-fold increased risk of venous thrombosis. Another variant (rs220842) was associated with venous thrombosis in the overall MEGA population and was only present in patients, and not in control subjects, with high factor VIII activity. Whether the variants are causal or are in linkage disequilibrium with unmeasured causal variants is not known. Our results suggest independent effects for the two variants. Both variants lie in intron 1, which comprises 240 kB of the 300 kB *CADM1* gene. There are a number of transcription factor binding sites and regulatory elements in intron 1. Examination of the 500 bp sequence flanking the variants revealed the occurrence of conserved elements (across 37 mammals) and open chromatin regions (DNase I hypersensitivity assay).[23] This suggests that epigenetic control may be the underlying functional mechanism by which these variants exert their effect on venous thrombosis.

One of the drawbacks of our study is the relatively low number of individuals per protein C pathway abnormality subgroup, which decreased our power to detect effects for *CADM1* variants. In addition, testing multiple SNVs for association with venous thrombosis increases the chance of false-positive associations. We therefore calculated FDR-adjusted q-values, after which we were no longer able to detect an association between the *CADM1* variants and venous thrombosis. We sought support for our hypothesis through addressing the association between venous thrombosis and *CADM1* variants in not only patients with low protein C levels, but also in other subgroups of patients with protein C pathway abnormalities. Although there was some overlap in

patients within the protein C pathway subgroups, we observed almost no overlap in the thrombosis-associated *CADM1* variants across the subgroups of protein C pathway abnormalities. Only one variant (rs11608105) was found to be associated with venous thrombosis in multiple subgroups, in this case in patients with low protein C and S levels. In some cases, the direction of the odds ratio for venous thrombosis risk of the positive *CADM1* variant differed across the protein C pathway abnormalities (Supplemental Table 2). Taken together, this may suggest that genetic variation in *CADM1* interacts only with single or specific factors within the protein C pathway.

Another drawback of our study is that the protein C and protein S deficiencies were not individually diagnosed, but we determined levels below clinical cut-offs using a single test. Therefore, the prevalence of the protein C pathway abnormalities may vary and some misclassification may have occurred. It is unlikely though to have affected the comparisons on a group level. In addition, as in all case-control studies, we cannot rule out that the thrombotic event itself influenced the coagulation factor levels, in particular the levels of the acute phase reactant factor VIII. However, the median time between blood draw and thrombotic event was 10 months and we did not observe any difference between the mean FVIII levels of blood samples drawn less than 6 months after the thrombotic event and blood samples drawn 6 or more months after the thrombotic event (mean levels of 134.9 and 132.7 IU/ml, respectively).

We identified several variants of which the risk allele was carried by patients or control subjects only. These might be involved as risk or protective alleles for venous thrombosis when co-occurring with a protein C pathway abnormality. However, since these variants were rare and the number of individuals was low, we are not able to draw conclusions about these variants.

The variant that was most strongly associated in the French Canadian family study, rs6589488, was not associated in the overall MEGA study (OR 1.07, 95% CI 0.98-1.17) nor in one of the subgroups of protein C pathway abnormalities (Supplemental Table 3). Linkage disequilibrium, as determined by r^2, with the variants consistently associated with venous thrombosis in our analysis (listed in Table 2) was low (<0.15). One explanation for the lack of a clear effect of rs6589488 in the current study is that the variants in the family study are rare mutations, private to this family or the French Canadian population. The results found in the current case-control study for

a joint thrombophilic effect of *CADM1* variants with protein C deficiency, protein S deficiency, or high factor VIII levels does suggest though that the *CADM1* pathway might play a role in the biology of hemostasis in the general population as well. The *CADM1* pathway links to the actin cytoskeleton and in the cancer literature its oncogenic effect is due to variants in *CADM1* as well as downstream proteins.[24-27] Analysis of genes of downstream members of the *CADM1* pathway might identify additional novel risk factors for venous thrombosis. Another possibility is that mutations in the gene for protein C (*PROC*) itself affect the interaction between CADM1 and protein C pathway. However, this would involve an indirect interaction between the downstream pathways associated respectively with the Endothelial Cell Protein C receptor and CADM1, as there is no evidence for a direct interaction of protein C with CADM1.

In conclusion, this study found some evidence of a joint effect of genetic variation in *CADM1* and protein C pathway abnormalities on the risk of venous thrombosis. This study aimed to validate a previous genetic study in a large thrombophilic family study, but could not replicate the specific associations observed in the family. Therefore, further study of the *CADM1* pathway is needed to determine whether abnormalities of the *CADM1* pathway link the risk for venous thrombosis to the vessel wall.

REFERENCES

1. Hasstedt SJ, Bezemer ID, Callas PW, Vossen CY, Trotman W, Hebbel RP, et al. Cell adhesion molecule 1: a novel risk factor for venous thrombosis. *Blood*. 2009;114:3084-91.
2. Giangreco A, Jensen KB, Takai Y, Miyoshi J, Watt FM. Necl2 regulates epidermal adhesion and wound repair. *Development*. 2009;136:3505-14.
3. Gomyo H, Arai Y, Tanigami A, Murakami Y, Hattori M, Hosoda F, et al. A 2-Mb sequence-ready contig map and a novel immunoglobulin superfamily gene IGSF4 in the LOH region of chromosome 11q23.2. *Genomics*. 1999;62:139-46.
4. Ito A, Okada M, Uchino K, Wakayama T, Koma Y, Iseki S, et al. Expression of the TSLC1 adhesion molecule in pulmonary epithelium and its down-regulation in pulmonary adenocarcinoma other than bronchioloalveolar carcinoma. *Lab Invest*. 2003;83:1175-83.
5. Biederer T, Sara Y, Mozhayeva M, Atasoy D, Liu X, Kavalali ET, et al. SynCAM, a synaptic adhesion molecule that drives synapse assembly. *Science*. 2002;297:1525-31.
6. Van der Weyden L, Arends MJ, Chausiaux OE, Ellis PJ, Lange UC, Surani MA, et al. Loss of TSLC1 causes male infertility due to a defect at the spermatid stage of spermatogenesis. *Mol Cell Biol*. 2006;26:3595-609.
7. Tatsumi K, Taatjes DJ, Wadsworth MP, Bouchard BA, Bovill EG. Cell adhesion molecule 1 (CADM1) is ubiquitously present in the endothelium and smooth muscle cells of the human macro- and micro-vasculature. *Histochem Cell Biol*. 2012;138:815-20.
8. Feistritzer C, Riewald M. Endothelial barrier protection by activated protein C through PAR1-dependent sphingosine 1-phosphate receptor-1 crossactivation. *Blood*. 2005;105:3178-84.
9. Niessen F, Furlan-Freguia C, Fernandez JA, Mosnier LO, Castellino FJ, Weiler H, et al. Endogenous EPCR/aPC-PAR1 signaling prevents inflammation-induced vascular leakage and lethality. *Blood*. 2009;113: 2859-66.
10. Finigan JH, Dudek SM, Singleton PA, Chiang ET, Jacobson JR, Camp SM, et al. Activated protein C mediates novel lung endothelial barrier enhancement: role of sphingosine 1-phosphate receptor transactivation. *J Biol Chem*. 2005;280:17286-93.
11. Bae JS, Rezaie AR. Thrombin inhibits nuclear factor kappaB and RhoA pathways in cytokine-stimulated vascular endothelial cells when EPCR is occupied by protein C. *Thromb Haemost*. 2009;101:513-20.
12. McVerry BJ, Garcia JG. In vitro and in vivo modulation of vascular barrier integrity by sphingosine 1-phosphate: mechanistic insights. *Cell Signal*. 2005;17: 131-9.
13. Wojciak-Stothard B, Ridley AJ. Shear stress-induced endothelial cell polarization is mediated by Rho and Rac but not Cdc42 or PI 3-kinases. *J Cell Biol*. 2003;161:429-39.
14. Murakami Y. Involvement of a cell adhesion molecule, TSLC1/IGSF4, in human oncogenesis. *Cancer Sci*. 2005;96:543-52.

15. Masuda M, Kikuchi S, Maruyama T, Sakurai-Yageta M, Williams YN, Ghosh HP, et al. Tumor suppressor in lung cancer (TSLC)1 suppresses epithelial cell scattering and tubulogenesis. *J Biol Chem*. 2005;280:42164-71.
16. Kawano S, Ikeda W, Kishimoto M, Ogita H, Takai Y. Silencing of ErbB3/ErbB2 signaling by immunoglobulin-like Necl-2. *J Biol Chem*. 2009;284:23793-805.
17. Martinelli I, De Stefano V, Mannucci PM. Inherited risk factors for venous thromboembolism. *Nat Rev Cardiol*. 2014;11:140-56.
18. Blom JW, Doggen CJ, Osanto S, Rosendaal FR. Malignancies, prothrombotic mutations, and the risk of venous thrombosis. *JAMA*. 2005;293:715-22.
19. Van Stralen KJ, Rosendaal FR, Doggen CJ. Minor injuries as a risk factor for venous thrombosis. *Arch Intern Med*. 2008;168:21-6.
20. O'Brien AE, Tate GM, Shiach C. Evaluation of protein C and protein S levels during oral anticoagulant therapy. *Clin Lab Haematol*. 1998;20:245-52.
21. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21:263-5.
22. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;57:289-300.
23. University of California SC: Genome browser version NCBI 36/hg18
24. Sakurai-Yageta M, Masuda M, Tsuboi Y, Ito A, Murakami Y. Tumor suppressor CADM1 is involved in epithelial cell structure. *Biochem Biophys Res Commun.* 2009;390:977-82.
25. Yageta M, Kuramochi M, Masuda M, Fukami T, Fukuhara H, Maruyama T, et al. Direct association of TSLC1 and DAL-1, two distinct tumor suppressor proteins in lung cancer. *Cancer Res*. 2002;62:5129-33.
26. Fukuhara H, Masuda M, Yageta M, Fukami T, Kuramochi M, Maruyama T, et al. Association of a lung tumor suppressor TSLC1 with MPP3, a human homologue of Drosophila tumor suppressor Dlg. *Oncogene*. 2003;22:6160-5.
27. Shingai T, Ikeda W, Kakunaga S, Morimoto K, Takekuni K, Itoh S, et al. Implications of nectin-like molecule 2/IGSF4/RA175/SgIGSF/TSLC1/SynCAM1 in cell-cell adhesion and transmembrane protein localization in epithelial cells. *J Biol Chem*. 2003;278:35421-7.

SUPPLEMENTAL TABLES

Table S1. Minor allele frequencies of *CADM1* variants in overall MEGA study population

		Minor allele frequency, %	
***CADM1* variant**	**Position**	**Patients**	**Controls**
rs11215392	114543618	2.65	2.75
rs34157656	114544511	44.3	45.0
rs10444329	114544893	18.1	17.0
rs17118020	114545350	1.50	1.33
rs17118023	114546173	18.1	16.9
rs17649730	114546639	15.0	14.1
rs4936321	114546799	47.6	46.2
rs11606837	114548047	49.4	48.4
rs4938182	114548246	21.0	19.8
rs45460594	114548330	3.09	2.75
rs45486791	114548565	0.53	0.44
rs45539744	114548882	0.01	0.02
rs4450197	114549421	2.41	2.04
rs1048932	114550060	43.9	43.0
rs45483591	114551963	0.04	0.03
rs45445298	114554121	0.36	0.45
rs17304149	114554390	48.3	49.1
rs17118046	114554937	3.84	3.80
rs45508098	114555249	16.6	15.8
rs7928746	114556120	2.03	2.36
rs4938183	114556779	4.10	4.09
rs45479795	114557630	4.41	4.27
rs11215400	114557845	27.3	27.4
rs45483594	114558449	16.6	15.9
rs12807135	114558718	49.6	50.4
rs45594631	114559767	0.00	0.02
rs11215403	114563795	25.9	25.8
rs45604639	114565259	0.33	0.28
rs7937380	114565377	26.0	26.0
rs45614835	114565529	16.3	15.6
rs4936322	114566743	45.1	43.8
rs45605138	114567521	1.70	1.78
rs4245160	114567760	0.01	0.00
rs45625839	114568381	0.01	0.01
rs7101437	114568851	49.9	50.4
rs45628237	114569486	0.33	0.26
rs11215406	114570292	27.3	27.3
rs11215407	114570503	6.13	6.38

Table S1. Continued

		Minor allele frequency, %	
***CADM1* variant**	**Position**	**Patients**	**Controls**
rs10891805	114571691	3.61	3.72
rs45456599	114571885	16.3	15.5
rs45617644	114571999	0.01	0.02
rs45574838	114572238	0.04	0.00
rs6589484	114576024	3.69	3.76
rs45529533	114576096	11.0	11.1
rs45479100	114577512	16.6	15.9
rs12226198	114579444	5.74	5.86
rs10128746	114580646	3.71	3.63
rs11215415	114580742	2.18	2.44
rs45505693	114583362	0.64	0.89
rs3802858	114583702	45.0	44.2
rs3802857	114583828	35.1	35.4
rs11215418	114585104	3.70	3.62
rs7125361	114585252	44.8	43.7
rs9645660	114586773	49.2	48.1
rs11215419	114587020	49.4	50.6
rs45516099	114587093	16.8	15.8
rs7482812	114588382	3.00	2.81
rs6589486	114589507	45.8	46.8
rs12281523	114589876	5.39	5.26
rs45525440	114590677	5.37	5.18
rs45489793	114592265	18.2	17.2
rs11215424	114592631	28.5	28.9
rs4938190	114592960	47.9	47.0
rs7106961	114593510	1.57	1.66
rs7947402	114593630	49.3	48.0
rs45593334	114594650	28.5	28.7
rs45583736	114595117	0.03	0.04
rs4245161	114595636	0.03	0.00
rs7479259	114595925	45.1	45.9
rs45614535	114596076	44.9	45.7
rs11825649	114597503	1.63	1.51
rs45460202	114597825	1.63	1.76
rs1938736	114598207	18.2	17.2
rs11215427	114598648	28.5	29.0
rs12575340	114600534	17.2	16.4
rs11215430	114601206	5.30	5.14
rs10891812	114601641	46.8	46.1
rs6589488	114602166	15.2	14.3

Table S1. Continued

		Minor allele frequency, %	
***CADM1* variant**	**Position**	**Patients**	**Controls**
rs12284489	114602367	5.32	5.20
rs12280033	114603084	7.04	6.83
rs12417740	114603646	45.2	46.0
rs11215431	114604893	0.03	0.00
rs11602686	114605848	45.3	46.2
rs11215433	114606504	7.01	6.89
rs10458967	114608081	5.32	5.18
rs10458969	114608403	16.6	15.5
rs11215437	114609382	24.8	25.0
rs10891814	114609820	38.6	37.8
rs10502200	114610942	3.34	3.52
rs45593037	114612214	4.64	4.36
rs947802	114613194	38.9	38.2
rs12283904	114614312	0.00	0.03
rs2269737	114616515	19.2	19.1
rs11215439	114617425	19.1	18.2
rs12421121	114617518	19.1	19.7
rs17118125	114619942	19.0	19.0
rs11215445	114620383	22.6	22.5
rs9633941	114621837	19.5	18.8
rs12225639	114622453	16.1	15.2
rs45624531	114622551	19.0	19.1
rs10502199	114625825	15.7	15.2
rs1892773	114627836	20.5	20.7
rs7127390	114627937	20.4	20.4
rs4936325	114630329	15.4	15.1
rs17118149	114630440	0.09	0.09
rs45604331	114632418	0.00	0.02
rs45538440	114634182	0.31	0.35
rs45577334	114634631	1.40	1.34
rs6589490	114637110	37.5	37.2
rs11215455	114639795	20.3	20.8
rs2154690	114640754	38.0	37.7
rs11215456	114640983	17.6	16.9
rs4938193	114641217	20.0	20.7
rs4597099	114641818	37.4	37.2
rs10891818	114642013	35.7	35.5
rs10891819	114642457	18.9	18.0
rs11215458	114645061	3.71	3.65
rs7950069	114645763	15.6	14.8

Table S1. Continued

		Minor allele frequency, %	
***CADM1* variant**	**Position**	**Patients**	**Controls**
rs11215459	114646718	1.62	1.26
rs45539832	114648118	5.51	5.41
rs4938194	114648551	38.1	37.9
rs10891820	114649664	12.6	13.0
rs12577839	114649744	0.01	0.00
rs45451094	114649852	0.01	0.03
rs17118172	114650309	5.27	5.02
rs12788053	114652701	20.3	21.0
rs10502203	114655447	1.40	1.16
rs17519855	114656695	0.23	0.24
rs7944529	114657017	11.6	11.2
rs7944955	114657247	31.8	31.8
rs7931895	114657509	31.7	31.8
rs11215462	114658528	0.03	0.00
rs17118198	114660163	0.09	0.09
rs45595941	114662359	0.79	0.70
rs11215466	114663198	18.2	17.5
rs10891823	114663444	6.74	6.48
rs2014270	114664443	12.6	13.0
rs17441594	114664964	11.6	11.0
rs7936399	114665469	38.6	38.3
rs17441610	114667144	11.6	11.0
rs4938195	114668875	12.7	13.0
rs7104872	114670321	19.5	19.1
rs7928044	114670523	6.53	6.41
rs11215470	114671854	0.03	0.00
rs45581535	114674341	3.69	3.70
rs45488901	114674457	11.6	11.0
rs11215474	114674839	19.0	18.3
rs7104113	114675467	38.8	38.6
rs45505692	114676989	0.01	0.02
rs10891825	114678381	34.8	33.8
rs2040456	114683727	0.03	0.00
rs2157612	114684281	10.9	10.4
rs7949084	114685949	46.5	47.0
rs12290790	114688338	10.8	10.3
rs45616036	114688640	0.82	0.84
rs17442145	114688855	1.26	1.15
rs17442179	114689108	3.28	3.43
rs45626034	114693674	0.01	0.03

Table S1. Continued

		Minor allele frequency, %	
***CADM1* variant**	**Position**	**Patients**	**Controls**
rs988873	114694438	21.0	21.3
rs11607436	114694623	2.18	2.42
rs2366904	114695046	46.7	46.1
rs12577709	114695169	14.0	13.6
rs17118264	114695758	13.6	13.3
rs4396320	114696475	46.8	46.3
rs12284145	114696918	0.03	0.00
rs45467696	114699386	0.07	0.03
rs45474291	114701226	11.0	10.6
rs45508698	114701763	0.69	0.96
rs10891829	114703906	46.6	46.0
rs45543336	114707093	10.9	10.4
rs45469396	114707350	0.53	0.55
rs17118279	114707907	0.10	0.12
rs10891832	114710833	43.7	42.6
rs10488710	114712386	33.1	33.3
rs10891833	114712918	38.5	38.5
rs7952231	114713208	38.5	38.6
rs9888216	114714603	44.3	43.4
rs2105976	114715710	44.2	43.2
rs7105871	114717935	20.7	21.3
rs45465296	114718461	12.2	11.8
rs11215504	114718584	4.94	4.35
rs4938201	114723923	40.4	40.7
rs12575143	114726812	2.51	2.39
rs45599536	114727833	0.39	0.49
rs10891836	114728167	44.3	43.1
rs2105982	114729014	44.3	43.1
rs7120311	114729924	22.7	23.4
rs11215512	114732381	44.2	43.1
rs10891839	114733207	33.0	33.3
rs10891840	114734721	44.3	43.1
rs17521934	114735633	12.0	11.7
rs11215515	114738087	4.41	4.26
rs45559239	114738583	0.04	0.03
rs45455497	114740709	0.40	0.49
rs11215517	114742555	10.4	10.1
rs10891842	114744233	39.1	39.1
rs10160742	114744607	7.06	7.06
rs45545346	114745259	3.13	3.42

Table S1. Continued

		Minor allele frequency, %	
***CADM1* variant**	**Position**	**Patients**	**Controls**
rs45580634	114746210	1.66	1.73
rs17118309	114746787	0.97	1.10
rs220850	114753565	49.4	48.8
rs4938202	114754917	39.2	39.1
rs11608105	114756400	4.57	4.72
rs45585234	114761471	0.00	0.03
rs45608938	114761668	5.33	5.42
rs17443832	114762977	5.31	5.38
rs220869	114767246	0.04	0.09
rs45578937	114769761	6.58	7.07
rs45514899	114771373	0.00	0.03
rs220872	114771575	50.5	49.9
rs7114341	114774371	44.5	43.5
rs45555732	114775788	1.02	1.14
rs11215532	114776409	44.7	43.8
rs4938203	114780571	44.6	43.7
rs220828	114782015	42.9	41.6
rs2366914	114784746	36.6	36.8
rs45559131	114786337	0.00	0.03
rs220842	114787382	0.76	0.51
rs17118328	114787872	1.95	1.82
rs220843	114788745	16.0	16.8
rs220847	114791327	49.3	48.7
rs11215545	114791960	42.9	41.9
rs12273801	114795200	0.01	0.01
rs7106275	114797011	0.26	0.21
rs220860	114799274	16.1	16.7
rs220861	114799402	6.26	6.55
rs45455306	114799791	1.37	1.54
rs220862	114801129	14.1	15.0
rs45458294	114801307	6.51	6.92
rs220864	114801841	14.1	14.6
rs220865	114802160	22.2	23.3
rs10891854	114804638	38.8	39.0
rs220836	114807081	20.5	21.3
rs45522132	114807342	1.37	1.61
rs7122693	114809573	43.4	42.7
rs45587938	114810013	10.8	10.1
rs17444623	114812143	18.8	19.8
rs17451032	114813684	1.02	1.16

Table S1. Continued

		Minor allele frequency, %	
***CADM1* variant**	**Position**	**Patients**	**Controls**
rs45509898	114813930	2.09	2.26
rs45473492	114816541	1.29	1.44
rs45520832	114818047	0.14	0.11
rs45625135	114818415	1.55	1.71
rs220838	114819312	18.9	19.7
rs12801130	114820321	36.7	35.8
rs17118342	114820680	0.07	0.25
rs160604	114823801	0.01	0.00
rs544083	114825691	17.3	18.0
rs220840	114826173	17.3	18.1
rs314474	114826343	17.2	17.9
rs314476	114827516	18.8	19.8
rs10502202	114829700	21.5	22.1
rs10891856	114830116	6.17	5.75
rs1155756	114830467	37.4	36.6
rs7927390	114831701	18.8	19.8
rs10047420	114834362	38.0	37.5
rs45490692	114835734	0.62	0.54
rs314491	114840421	20.2	20.8
rs10891859	114840831	35.8	35.4
rs314494	114841812	20.3	20.9
rs314495	114842583	20.3	20.8
rs314496	114842787	20.2	20.9
rs45474398	114844445	3.52	3.88
rs17451771	114845558	6.55	7.03
rs314497	114847142	6.27	5.86
rs11827474	114848809	0.01	0.00
rs17118360	114849006	0.07	0.16
rs1460909	114851977	0.43	0.36
rs314503	114852071	6.52	7.01
rs314507	114854460	0.01	0.00
rs314512	114858104	6.46	7.03
rs314513	114858508	6.49	7.02
rs314514	114861898	1.74	1.62
rs7924765	114862746	0.00	0.01
rs12281277	114866132	0.00	0.01
rs11215574	114868653	25.9	25.8
rs17524208	114871498	3.36	3.89
rs973550	114872351	0.27	0.28
rs17524278	114875616	6.49	7.13

Table S1. Continued

		Minor allele frequency, %	
***CADM1* variant**	**Position**	**Patients**	**Controls**
rs314464	114878567	0.01	0.01
rs45583332	114880825	0.80	0.70
rs11215581	114884622	0.49	0.36
rs314469	114885900	7.07	7.54
rs314468	114887234	6.58	7.20
rs7101558	114892659	6.94	7.44

Table S2. Associations of positive variants identified in different protein C pathway subgroups with venous thrombosis.

CADM1	Overall MEGA		Low protein C activity		Low protein S levels		High factor VIII activity		FV Leiden carriers	
variants	cases	controls	cases	OR (95% CI)	cases	OR (95% CI)	cases	OR (95% CI)	cases	OR (95% CI)
rs11608105	4.57	4.72	7.22	1.57 (1.05-2.34)	13.0	2.98 (1.27-7.02)	2.73	0.56 (0.29-1.11)	3.36	0.93 (0.69-1.26)
rs45520832	0.14	0.11	0	NA	2.17	20.1 (2.45-166)	0	NA	0.09	0.77 (0.10-6.05)
rs11215504	4.94	4.35	5.67	1.32 (0.85-2.06)	2.17	0.49 (0.07-3.56)	7.58	1.79 (1.18-2.73)	5.44	1.26 (0.96-1.67)
rs220842	0.76	0.51	1.03	2.03 (0.72-5.73)	0	NA	1.52	3.02 (1.18-7.74)	0.69	1.35 (0.63-2.90)
rs45608938	5.33	5.42	5.93	1.10 (0.72-1.69)	6.52	1.21 (0.38-3.87)	4.55	0.84 (0.50-1.41)	3.89	0.71 (0.52-0.97)
rs45545346	3.13	3.42	2.84	0.83 (0.45-1.51)	4.35	1.27 (0.32-5.13)	2.44	0.71 (0.35-1.44)	1.73	0.50 (0.32-0.79)

The risk allele frequencies were calculated (shown in percentages) in the overall MEGA study population and in subgroups of patients with protein C pathway abnormalities.

Table S3. The *CADM1* variant found in the Vermont family assessed for associations with venous thrombosis in overall MEGA population and subgroups.

rs6589488	Risk allele frequency, %	OR (95% CI)
Overall controls	14.3	REF
Overall patients	15.2	1.07 (0.98-1.17)
Low protein C	14.2	0.99 (0.74-1.32)
Low protein S	17.4	1.26 (0.59-2.69)
High factor VIII	15.2	1.07 (0.79-1.45)
FVL carriers	14.5	1.02 (0.85-1.21)

CHAPTER 6

Multiple SNP testing improves risk prediction of first venous thrombosis

de Haan HG, Bezemer ID, Doggen CJ, Le Cessie S, Reitsma PH, Arellano AR, Tong CH, Devlin JJ, Bare LA, Rosendaal FR, Vossen CY.

Blood. 2012;120(3):656-63.

ABSTRACT

There are no risk models available yet that accurately predict an individual's risk for developing venous thrombosis. Our aim was therefore to explore whether inclusion of established thrombosis-associated SNPs in a venous thrombosis risk model improves the risk prediction. We calculated genetic risk scores by counting risk-increasing alleles from 31 venous thrombosis-associated SNPs for subjects of a large case-control study including 2712 patients and 4634 controls (MEGA). Genetic risk scores based on all 31 SNPs or on the 5 most strongly associated SNPs performed similarly (areas under receiver-operating characteristic curves (AUCs) of 0.70 and 0.69 respectively). For the 5-SNP risk score, the odds ratios for venous thrombosis ranged from 0.37 (95% CI 0.25-0.53) for individuals with 0 risk alleles to 7.48 (95% CI 4.49-12.46) for individuals with ≥6 risk alleles. The AUC of a risk model based on known non-genetic risk factors was 0.77 (95% CI 0.76-0.78). Combining the non-genetic and genetic risk models improved the AUC to 0.82 (95% CI 0.81-0.83), indicating good diagnostic accuracy. In order to become clinically useful, subgroups of high-risk individuals must be identified in whom genetic profiling will also be cost-effective.

INTRODUCTION

Venous thrombosis is the result of innate thrombotic tendency and non-genetic triggers. Many common genetic variants, mainly single-nucleotide polymorphisms (SNPs), with modest effects on risk of venous thrombosis have been reported.[1] Individual SNPs have little predictive value due to their modest effect on risk, but combinations of gene variants may improve the predictive ability and could be used to model susceptibility to venous thrombosis.

Simulation studies have shown that so-called genetic profiling may be useful to discriminate between individuals with high risk of disease and those with low risk. The discriminative accuracy of genetic profiling depends on the heritability and incidence of the disease and on the frequencies of risk alleles.[2,3]

Genetic profiling has become a popular aim in epidemiologic studies of many common diseases since a large amount of data from genome-wide association studies (GWAS) has become available.[2-8] For recurrent venous thrombosis, we previously investigated the potential clinical utility of multiple SNP testing for recurrent events.[9] In that study, individual SNPs were not significantly associated with recurrent venous thrombosis. However, when the risk alleles of the individual SNPs were combined, the risk estimates as well as the significance of the association increased. The predictive ability of multiple SNP analysis has not been studied for first events of venous thrombosis. Genetic profiling may guide decisions on prophylactic measures in high-risk groups such as cancer patients, individuals undergoing surgery, requiring a plaster cast or those subject to prolonged immobilization.

In order to explore to what extent venous-thrombosis associated SNPs can be used as predictors for a first venous thrombosis in the general population and in high-risk groups, we investigated 31 SNPs in two large population-based case-control studies, of which one was used as a validation set. We created genetic risk scores based on these SNPs and a risk score based on non-genetic risk factors. We also compared and combined our genetic risk score with the non-genetic risk score to determine whether genetic profiling with the currently known SNPs will improve the assessment of venous thrombosis risk.

METHODS

Study populations

The Multiple Environmental and Genetic Assessment of risk factors for venous thrombosis (MEGA study) is a population-based case-control study of venous thrombosis. Collection and ascertainment of events have been described in detail previously.[10,11] The MEGA analysis included 2712 consecutive patients with a diagnosis of a first deep vein thrombosis of the leg or arm (with or without pulmonary embolism) and 4634 control subjects (partners of patients and random population controls).

The Leiden Thrombophilia Study (LETS), another population-based case-control study of venous thrombosis, was used to validate the risk scores and included 443 consecutive patients with a diagnosis of a first deep vein thrombosis of the leg (with or without pulmonary embolism) and 453 control subjects (acquaintances or partners of patients), all without a known malignancy. Collection and ascertainment of events have been described in detail previously.[12] Both studies were approved by the Medical Ethics Committee of the Leiden University Medical Center, Leiden, the Netherlands.

SNP selection

Initially we selected 40 SNPs for the genetic risk score, based on the literature and our previous work. Eighteen SNPs had been reported and repeatedly confirmed to be associated with venous thrombosis.[1,13] Twelve SNPs were added from the Group Health study,[13,14] these SNPs were associated with venous thrombosis in the original study and replicated in the MEGA study. Nine SNPs were added from a large SNP association analysis including subsequent fine mapping that we performed recently in LETS and MEGA.[15,16] Another added SNP was recently identified in a follow-up study of a GWAS and replicated in the FARIVE study and the MEGA study.[17] Among the 40 SNPs in the initial selection, we studied linkage disequilibrium and mutually adjusted SNPs within genes. Four SNPs in *PROC* (rs1799808, rs1799810, rs2069915 and rs5937) were explained by rs1799809 in *PROC*; 4 SNPs in the fibrinogen genes (rs6050 and rs2070006 in *FGA*, rs1800788 in *FGB* and rs2066854 in *FGG*) were explained by rs2066865 in *FGG*; and rs3753305 in *F5* was explained by rs6025 (factor V Leiden). Consequently, we excluded 9 SNP associations that were explained by other SNPs. The remaining 31 SNPs (Table 1) were included in the genetic risk score.

Genetic risk score

We defined a genetic risk score that counts the total number of risk-increasing alleles in individuals. To take into account the stronger association of some SNPs with venous thrombosis, we also constructed a weighted risk score assigning weights to the risk alleles of each SNP corresponding to the logarithm of the average risk estimates found in literature. In addition to the full genetic model including 31 SNPs, we constructed a parsimonious model with fewer SNPs. To determine which SNPs should be included in this model, we added SNPs one-by-one to create the genetic risk score. We started with the SNP with the highest odds ratio in literature and assessed whether adding SNPs to the risk score improved the AUC after each SNP addition. The addition of SNPs was stopped when the AUC of the risk score including the newly added SNP did not differ from the AUC of the full genetic model.

Non-genetic risk factors

We constructed a non-genetic risk score, which included the following risk factors: recent (within three months prior to the index date) leg injury, surgery, pregnancy or postpartum, immobilization (i.e. plaster cast, bedridden at home, hospitalization), travel for more than four hours in two months prior to the index date, oral contraceptives (OC) use or hormone replacement therapy (HRT) at the index date, obesity (body mass index >30kg/m^2) and a cancer diagnosis between five years before and six months after the index date. The index date was defined as date of diagnosis for patients and their partner controls, and the date of completing the questionnaire for random controls. We also included family history in the non-genetic risk score. Family history was defined as positive when a parent or sibling had experienced venous thrombosis and negative when none of these relatives had experienced venous thrombosis, or when the participant was not aware of venous thrombosis in the family. We assigned weights to each non-genetic risk factor corresponding to the logarithm of the risk estimates in MEGA (Supplemental Table 1) and constructed a simple risk scoring system counting the weighted risk factors. We also constructed a combined risk score including both the genetic risk score and the non-genetic risk score using a logistic regression model.

Application of genetic profiling may be most useful in high-risk groups, i.e. individuals exposed to known non-genetic risk factors. We therefore studied the discriminative accuracy of our genetic risk score as well as the combined scores in high-risk situations of surgery, plaster cast, hospitalization, young women (under 50 years) using oral

contraceptives, women using HRT, pregnancy or postpartum, middle-aged individuals (above 50 years) and travel. We also studied individuals with a positive family history and individuals with malignant disorders.

Statistical analyses

Crude and sex-adjusted (in case SNPs were located on the X chromosome) odds ratios and 95% confidence intervals were calculated by logistic regression for individual SNPs and the genetic, non-genetic and combined risk scores. When assessing the magnitude of risk associated with number of risk alleles, we used the median number of risk alleles among control subjects as the reference group.

To assess how well a score classifies venous thrombosis patients and control subjects, we calculated the area under the receiver-operating characteristic (ROC) curve (AUC). The AUC ranges from 0.5 (no discrimination between patients and control subjects) to 1.0 (perfect discrimination). We compared the AUCs of the different genetic and non-genetic risk models according to the method of Hanley *et al.*[18] Nagelkerke's pseudo-r^2 statistic was used to approximate the proportion of variability explained by the different risk models. All analyses, including ROC curves and AUC calculation were performed in SPSS for Windows, 17.0.2 (SPSS Inc, Chicago, Ill).

RESULTS

SNPs associated with venous thrombosis

Table 1 lists all associations between SNPs and venous thrombosis in the MEGA population and the average estimated effect-size in literature.[13-17,19-26] Not all SNPs were associated with venous thrombosis in our study populations; nevertheless, we included all 31 SNPs in the genetic risk score because these SNPs had been associated with venous thrombosis in other studies.

Table 1. 31 SNP associations with venous thrombosis in MEGA and literature.[13-17; 19-26]

				MEGA				Literature
				Risk allele frequency, %				Average
Gene	**SNP**	**Chr**	**Position**	**Cases**	**Controls**	**OR**	**95% CI**	**OR**
F5	rs6025	1	167.785.673	10	3	4.30	(3.70 -4.99)	3.79
F2	rs1799963	11	46.717.631	6	2	3.01	(2.36 -3.85)	2.78
ABO	rs8176719	9	136.132.908	47	34	1.74	(1.63 -1.87)	1.85
FGG	rs2066865	4	155.744.726	34	27	1.41	(1.32 -1.51)	1.56
F11	rs2036914	4	187.429.475	59	52	1.35	(1.26 -1.44)	1.32
PROCR	rs2069951	20	33.227.425	7	5	1.32	(1.16 -1.51)	1.30
F11	rs2289252	4	187.444.375	48	41	1.36	(1.28 -1.45)	1.26
F9	rs4149755	X	138.451.778	7	6	1.11	(0.99 -1.24)	1.24
PROCR	rs2069952	20	33.227.612	64	60	1.21	(1.13 -1.29)	1.21
SERPINC1	rs2227589	1	172.152.839	11	9	1.27	(1.15 -1.41)	1.20
HIVEP1	rs169713	6	11.920.517	22	20	1.10	(1.01-1.19)	1.20
F2	rs3136516	11	46.717.332	52	49	1.12	(1.06 -1.20)	1.19
F5	rs1800595	1	167.776.972	6	5	1.18	(1.03 -1.36)	1.18
PROC	rs1799809	2	127.892.345	47	43	1.17	(1.10 -1.25)	1.17
PROCR	rs867186	20	33.228.215	14	12	1.18	(1.07 -1.29)	1.17
VWF	rs1063856	12	6.153.534	37	33	1.18	(1.10-1.26)	1.16
GP6	rs1613662	19	60.228.407	84	82	1.18	(1.09 -1.29)	1.15
F2	rs3136520	11	46.699.808	3	2	1.09	(0.89 -1.32)	1.13
F8	rs1800291	X	153.811.479	85	83	1.12	(1.05 -1.20)	1.13
STXBP5	rs1039084	6	147.635.413	42	45	0.90	(0.84-0.96)	0.90
NAT8B	rs2001490	2	73.781.606	40	37	1.13	(1.06 -1.20)	1.10
F13B	rs6003	1	195.297.644	9	10	1.11	(1.00 -1.24)	1.09
RGS7	rs670659	1	239.228.398	67	64	1.14	(1.06 -1.22)	1.09
F9	rs6048	X	138.460.946	72	70	1.09	(1.03 -1.16)	1.08
F5	rs4524	1	167.778.379	79	74	1.31	(1.22 -1.42)	0.92
F13A1	rs5985	6	6.263.794	76	76	1.03	(0.95 -1.10)	0.93
F3	1208 indel	1	94.780.000	46	46	1.02	(0.96 -1.09)	1.06
TFPI	rs8176592	2	188.040.937	69	68	1.04	(0.97 -1.11)	1.06
F11	rs3822057	4	187.425.146	55	49	1.31	(1.23 -1.39)	1.06
NR1I2	rs1523127	3	120.983.729	41	38	1.15	(1.08 -1.23)	1.05
CPB2	rs3742264	13	45.546.095	69	68	1.04	(0.97 -1.11)	1.01

Abbreviations: Chr=chromosome; OR=odds ratio; CI=confidence interval

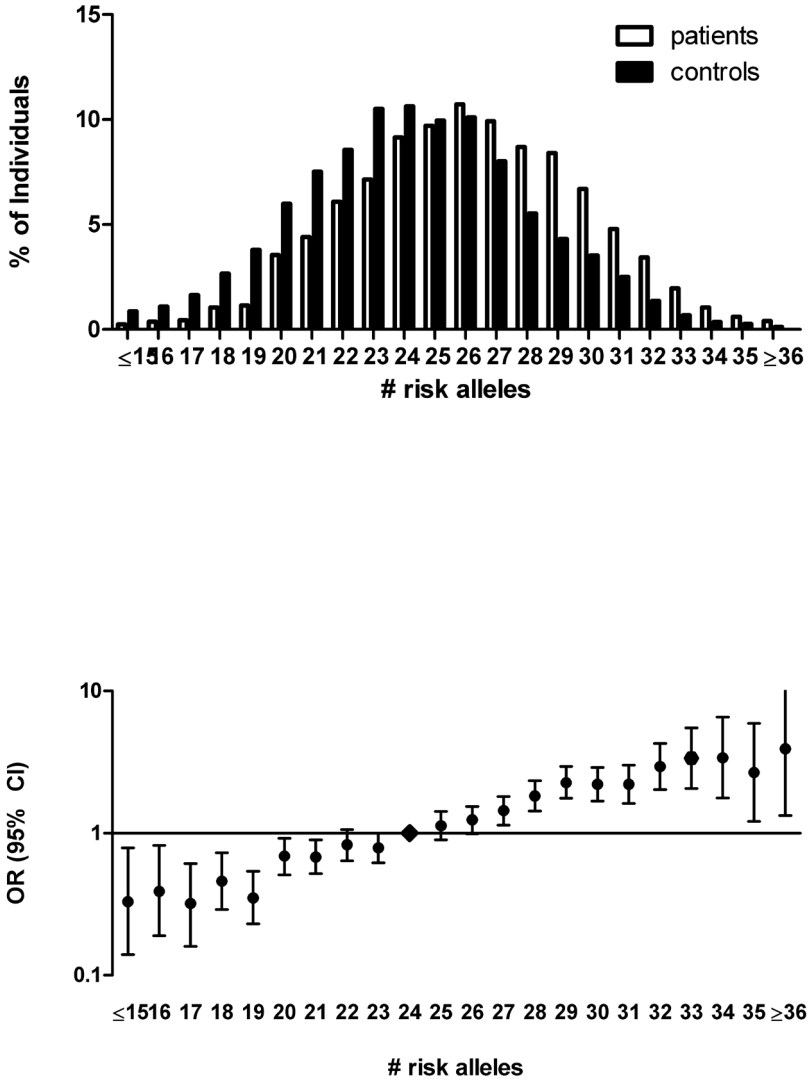


Figure 1. 31-SNP risk allele distribution in patients with venous thrombosis and control subjects (upper panel of figure) and corresponding odds ratios (lower panel).

Odds ratios (95% confidence interval) for venous thrombosis were calculated relative to the median number of risk alleles among control subjects (24 risk alleles). Individuals with 15 or less and 36 or more risk alleles were combined for the calculation of the odds ratio because of the low numbers of individuals with that few or many risk alleles.

Genetic risk score

We first included all 31 SNPs in the genetic risk score. For each individual we counted the number of risk-increasing alleles. The number of risk alleles ranged from 13 to 38 with a median of 24 among control subjects and 26 among cases (Figure 1). The risk for venous thrombosis was estimated for each number of risk alleles, relative to the median number of risk alleles of 24, and ranged from an odds ratio of 0.27 (95% CI 0.13-0.56) for 16 risk alleles to an odds ratio of 3.23 (95% CI 1.96-5.30) for 33 risk alleles. At the more extreme ends of the risk distribution, confidence intervals around risk estimates became very wide due to small numbers. The average relative risk increase per risk allele, when treated as an ordinal variable, however, could be estimated with a high level of precision, and was 1.14 (95% CI 1.12-1.16). This corresponds to an about 100-fold difference in risk between the lowest and the highest number of risk alleles in our population.

We also constructed a weighted risk score thereby assigning weight to the risk alleles according to their risk estimates found in literature (Table 1). A few SNPs have only been studied in the MEGA population; in that case we used the risk estimate in MEGA as weight. The ROC curve for the weighted 31-SNP risk score had an AUC of 0.71 (Table 2: 95% CI 0.69-0.72); i.e., there is a 71% probability that a randomly chosen patient will have a higher score than a randomly chosen control subject. The weighted 31-SNP risk score was a better predictor than the non-weighted 31-SNP risk score (AUC 0.64, 95% CI 0.63-0.65). The average relative risk increase per unit in the risk score, when treated as an ordinal variable, was 7.89 (95% CI 6.76-9.21). The proportion of variability explained by the 31-SNP risk score was 16.1% (Nagelkerke's pseudo-r^2; Table 2).

Table 2. Venous thrombosis prediction using genetic, non-genetic and combined risk scores. The LETS study was used as a validation set.

	MEGA (N=7092)		LETS (N=881)	
	AUC (95% CI)	Nagelkerke pseudo r^2	AUC (95% CI)	Nagelkerke pseudo r^2
31-SNP risk score	0.71 (0.69-0.72)	0.161	0.69 (0.65-0.72)	0.149
5-SNP risk score	0.69 (0.67-0.70)	0.135	0.67 (0.64-0.71)	0.138
Non-genetic risk score	0.77 (0.76-0.78)	0.288	0.71 (0.68-0.74)	0.200
Combined risk score	0.82 (0.81-0.83)	0.378	0.77 (0.74-0.80)	0.292

Abbreviations: AUC=area under the receiver-operating characteristic (ROC) curve; CI=confidence interval

In order to construct a genetic risk score using the most parsimonious model, we added SNPs one-by one to the genetic risk score, starting with the SNP with the highest OR in literature (Factor V Leiden, rs6025), and calculated the AUC after the addition of each SNP (Figure 2). The AUC for each single SNP ranged from 0.50 (95% CI 0.49-0.52) for rs3136520 in *F2* to 0.60 (95% CI 0.59-0.61) for rs8176719 in *ABO*. The discriminative accuracy of the model improved rapidly with the addition of each SNP, until 5 SNPs were included in the model (Figure 2). These SNPs were rs6025 (*F5*, factor V Leiden), rs1799963 (*F2*, 20210 G>A), rs8176719 (*ABO)*, rs2066865 (*FGG* 10034 C>T) and rs2036914 (*F11).* The AUC for this 5-SNP risk score was 0.69 (Table 2, 95% CI 0.67-0.70). Moreover, a model based on the three most well-known prothrombotic polymorphisms (i.e. rs6025, rs1799963 and rs8176719; AUC 0.65, 95% CI 0.64-0.66) performed significantly worse than the 5-SNP risk score. The average relative risk increase per unit in the risk score, when treated as an ordinal, was 9.50 (95% CI 7.92-11.39). The 5-SNP risk score explained 13.5% of the total variability (Nagelkerke's pseudo-r^2; Table 2).

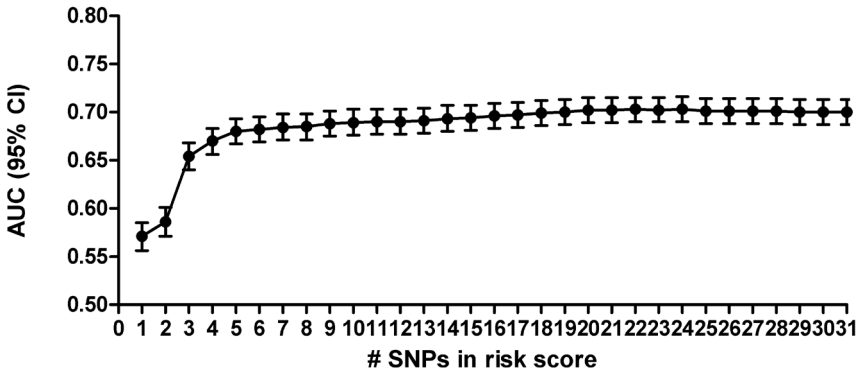


Figure 2. Area under the ROC of genetic risk scores based on increasing numbers of SNPs.

SNPs were added in order of the odds ratio as found in the literature, starting with rs6025 in the score based on 1 SNP, and ending with *CPB2* included in the score of 31 SNPs (Table 1).

The number of risk alleles in the 5-SNP risk score ranged from 0 (OR 0.37, 95% CI 0.26-0.53) to 8 (OR 7.48, 95% CI 4.49-12.46 for ≥6 risk alleles), with a median number of risk alleles of 2 among control subjects (Figure 3). The relative increase in risk per increase in number of risk alleles was 1.61 (95% CI 1.54-1.68), again corresponding to an over 100–fold difference in risk between the lowest and the highest number of risk alleles. The weighted 5-SNP risk score was a better predictor than a non-weighted model based on number of risk alleles (AUC 0.66, 95% CI 0.64-0.67).

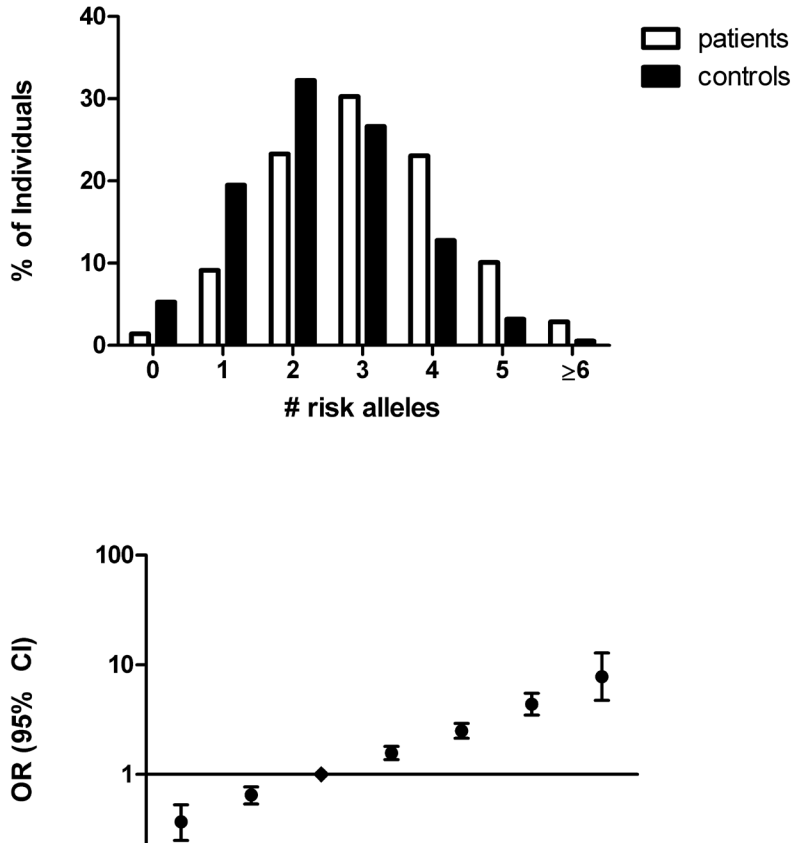


Figure 3. 5-SNP risk allele distribution in patients with venous thrombosis and control subjects (upper panel of figure) and corresponding odds ratios (lower panel).

Odds ratios (95% confidence interval) for venous thrombosis were calculated relative to the median number of risk alleles among control subjects (score 2). Individuals with 6 or more risk alleles were combined for the calculation of the odds ratio because of the low numbers of individuals with that few or many risk alleles.

No difference between the discriminative accuracy of the 5-SNP risk score in men (AUC 0.69, 95% CI 0.67-0.71) and women (AUC 0.67, 95% CI 0.65-0.69) was found. However, differences were found when we constructed and compared the 5-SNP genetic risk score in patients with DVT in the arm, patients with DVT in the leg and patients with DVT in the leg combined with PE. The AUC of the 5-SNP risk score in patients with DVT in the arm (AUC 0.62, 95% CI 0.57-0.67) was significantly lower than in patients with DVT in the leg (AUC 0.68, 95% CI 0.67-0.70) or for DVT combined with PE (AUC 0.68, 95% CI 0.67-0.70).

High-risk groups and SNP testing

To explore clinical applications of genetic profiling, we studied groups exposed to known non-genetic factors in more detail. The discriminative accuracy of the genetic risk scores in these subgroups was similar to the discriminative accuracy in the overall study population, except among cancer patients (Table 3). Sub-analysis in cancer patients according to therapy (chemotherapy, surgery, radiation) or tumor class (solid versus other) did not improve the discriminative accuracy of the weighted 5-SNP risk score (data not shown).

To assess whether the genetic risk score performs better than the current clinical practice of assessing family history, we compared the discriminative accuracy of the genetic risk score with a risk score with family history alone. The AUC of the 5-SNP risk score (0.68, 95% CI 0.67-0.70) was significantly higher than the AUC of family history (0.58, 95% CI 0.57-0.60), with a similar trend observed among all subgroups of high-risk individuals (Table 3).

Table 3. Risk score prediction in sub-groups of individuals exposed to known non-genetic risk factors.

High Risk group	Patients, N	Control subjects, N	Family history risk score, AUC (95%CI)	5-SNP risk score, AUC (95%CI)	Non-genetic risk score, AUC (95%CI)	Combined risk score, AUC (95%CI)
Surgery	292	111	0.60 (0.55-0.66)	0.66 (0.60-0.72)	0.67 (0.61-0.72)	0.73 (0.67-0.78)
Plaster cast	111	18	0.61 (0.48-0.73)	0.73 (0.59-0.87)	0.70 (0.56-0.84)	0.78 (0.64-0.91)
Hospitalization	278	93	0.57 (0.50-0.63)	0.66 (0.59-0.72)	0.60 (0.53-0.66)	0.66 (0.59-0.72)
Oral contraceptives*	513	327	0.58 (0.54-0.62)	0.71 (0.68-0.75)	0.73 (0.69-0.76)	0.81 (0.78-0.84)
HRT	58	90	0.59 (0.49-0.68)	0.71 (0.63-0.80)	0.74 (0.66-0.82)	0.80 (0.72-0.87)
Pregnancy/Postpartum*	67	46	0.54 (0.44-0.65)	0.70 (0.60-0.79)	0.68 (0.57-0.79)	0.76 (0.66-0.85)
Age>50 years	944	1534	0.57 (0.55-0.60)	0.68 (0.66-0.70)	0.73 (0.71-0.75)	0.79 (0.77-0.81)
Travel	379	610	0.58 (0.54-0.62)	0.70 (0.67-0.73)	0.77 (0.73-0.80)	0.82 (0.80-0.85)
Family history	659	551	-	0.68 (0.65-0.71)	0.74 (0.71-0.76)	0.81 (0.78-0.83)
Malignancies	156	65	0.57 (0.49-0.65)	0.60 (0.52-0.68)	0.71 (0.64-0.78)	0.72 (0.65-0.80)

Abbreviations: AUC=area under the receiver-operating characteristic (ROC) curve; CI=confidence interval; HRT=hormone replacement therapy
* Women under 50 years.

Combining non-genetic and genetic risk scores

We assessed the discriminative accuracy of a non-genetic risk score based on known non-genetic risk factors for venous thrombosis (leg injury, surgery, pregnancy, plaster cast, bedridden at home, hospitalization, travel, OC use, HRT, obesity and malignancy) and family history. For the individual components the AUC ranged from 0.50 (95% CI 0.48-0.51) for recent travel to 0.67 (95% CI 0.65-0.69) for OC use by women. The AUC for the non-genetic risk score including family history was 0.77 (95% CI 0.76-0.78). When we added the genetic risk score to the non-genetic score, the AUC significantly increased to 0.82 (Figure 4: 95% CI 0.81-0.83) compared with the non-genetic risk score alone (p-value <0.0001) using either the 31-SNP or the 5-SNP risk score. In addition, 28.8% of the total variability in venous disease risk was explained by the non-genetic risk score, which significantly improved to 37.8% (Nagelkerke pseudo r^2; Table 2) when combining the non-genetic and genetic risk scores. Both the non-genetic and the combined risk score models performed better in women than in men (non-genetic risk score: AUC 0.81, 95% CI 0.80-0.83 for women and AUC 0.74, 95% CI 0.72-0.75 for men; combined risk score: AUC 0.85, 95% CI 0.83-0.86 for women and AUC 0.80, 95% CI 0.78-0.81 for men).

We also studied the discriminative accuracy of the combined risk score model in the high-risk groups. For all subgroups the AUC improved when using the combined risk score compared with the non-genetic risk score, which was significant for individuals using oral contraceptives, individuals with a positive family history of venous thrombosis and individuals over 50 years old (Table 3).

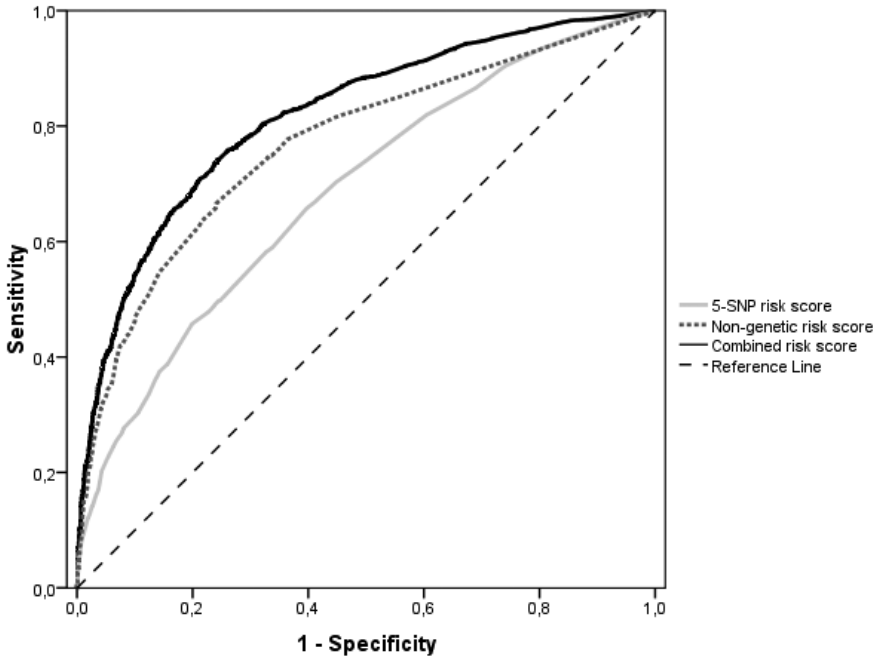


Figure 4. ROC (AUC) curves of the weighted 5-SNP risk score (light grey line), the non-genetic risk score (dotted grey line) and the combined risk score (black line).
The striped black line represents the reference line (no discrimination).

Validation of the risk scores

In order to validate the genetic, non-genetic and combined risk scores, we studied their discriminative accuracy in subjects from another population, the LETS population. As described in the Methods, LETS and MEGA are both population-based case control studies and are similar with respect to mean age at index of patients (45 years in LETS, 47 years in MEGA) or control subjects (45 years in LETS, 48 years in MEGA) and sex distribution (43% men in LETS, 47% men in MEGA). Associations between the 31 SNPs and venous thrombosis in LETS can be found in Supplemental Table 2. The discriminative accuracy of the weighted 31-SNP and 5-SNP risk scores in LETS were 0.69 (95% CI 0.65-0.72) and 0.67 (95% CI 0.64-0.71) respectively, which are similar to those found in MEGA (Table 2).

We also constructed the non-genetic risk score weighted according to the risk estimates of each risk factor from MEGA, except for malignancies as having cancer was an exclusion criterion in LETS. In addition, information of some non-genetic risk factors, i.e. HRT, recent travel, leg injury and plaster cast was not assessed in LETS or not in such detail as in MEGA. Therefore, these risk factors were excluded from the non-genetic risk score. The discriminative accuracy of the non-genetic risk score in LETS was 0.71 (95% CI 0.68-0.74) and improved to 0.77 (95% CI 0.74-0.80) when combed with the genetic risk score. Both risk scores performed slightly better in MEGA than in LETS (Table 2).

DISCUSSION

We calculated a genetic risk score based on SNPs consistently associated with venous thrombosis and observed a 'dose-response' relationship between this score and the risk of venous thrombosis. The more risk alleles or genotypes present, the higher the risk of venous thrombosis. A score constructed of the 5 most strongly associated SNPs appeared to differentiate between patients and control subjects equally as well as the initial genetic risk score based on 31 SNPs. The discriminative accuracy of both the 5-SNP and 31-SNP risk score was replicated in another study (LETS) suggesting robustness of the genetic models.

When preventive measures following a positive test are invasive or can have harmful side- effects, strict discrimination is required between those at high risk and low risk of developing a specific disease. In the case of venous thrombosis, indiscrimination may lead to an increased risk of thrombosis in high-risk individuals receiving insufficient prophylactic anticoagulant treatment, whereas individuals at low risk receiving treatment are at an increased risk of major bleeding. We investigated the extent to which genetic risk scores can improve the accuracy of thrombosis risk assessment by means of ROC curves. The 5-SNP genetic score performed better than family history assessment, which is the current clinical practice of risk assessment in individuals exposed to known non-genetic risk factors. However, the 5-SNP genetic risk score performed worse than a risk score of non-genetic risk factors. A recent study by Hippisley-Cox and Coupland[27] showed that an algorithm of non-genetic risk factors is able to discriminate between patients and control subjects with an AUC of 0.75. This is similar to the AUC observed

with our non-genetic risk score (0.77). However, the AUC may be an overestimation since we used (the logarithm of) the risk estimates from MEGA as weights.

Here, we showed that addition of the 5-SNP genetic risk score to the non-genetic risk score model significantly improved the AUC to 0.82, indicating good diagnostic accuracy. In our validation study, information on the non-genetic risk factors was less complete, which explains the lower discriminative accuracy of both the non-genetic risk score (0.71) and the combined risk score (0.77).

Identification of individuals at risk of developing venous thrombosis is most useful in high-risk populations. This is because the incidence of venous thrombosis in the general population is too low (1 per 1000 individuals a year[28]) to justify genotyping of all individuals. In all subgroups of high-risk individuals the combined risk score performed better than the non-genetic score alone, which may indicate the potential clinical value of genetic profiling in these high-risk individuals.

We defined a basic genetic risk score that counts the total number of risk-increasing alleles in individuals. To take into account the stronger association of some SNPs with venous thrombosis, we assigned literature-based weights to each SNP, which discriminated patients better from controls than a non-weighted genetic risk score. Although the proportion of variability explained by the 5-SNP risk score is smaller than by the 31-SNP risk score, we showed that the discriminative accuracy of the 5-SNP and 31-SNP risk scores was similar. The genetic risk score is still limited though by its assumption that all SNPs act independently and in an additive manner in venous thrombosis susceptibility. An additive effect was assumed for the different genotypes, whereas we cannot exclude a multiplicative effect. Gene-gene interaction and gene-environment interaction is not taken into account, while in reality many interactions exist. Examples for venous thrombosis are the synergistic effects between factor V Leiden (rs6025) and oral contraceptive use[29] and between the *F13A1* Val34Leu variant (rs5985) and fibrinogen levels.[30] We chose to include SNPs on their contribution to risk (effect size) and gave weights corresponding to the logarithm of this effect size. This is the most relevant for an individual who has a certain genotype. One could argue that on a population level, the prevalence of risk alleles is of relevance. However, this would not be expected to improve the performance of the risk prediction model, and indeed a genetic risk model based on the 5 SNPs with the highest risk allele frequency

in MEGA performed worse than the non-weighted 5-SNP risk score which is based on the 5 SNPs with the highest effect-size (AUC 0.54, 95% CI 0.53-0.56 and AUC 0.66, 95% CI 0.64-0.67, respectively).

In the future, adding newly discovered predictive SNPs to the model may further improve discrimination. In a simulation study, Janssens et al[2] showed that the AUC depends on the number of SNPs included, and their OR and risk allele frequency. The heritability of a disease determines the maximum obtainable AUC. For venous thrombosis the heritability is estimated to be about 60%.[31,32] The simulation study indicated that at this level high AUCs (>0.90) can be obtained, given that all genetic contributors are in the prediction model. Identification of new genetic predictors and validation of the genetic risk score in other study populations will reveal whether genetic profiling is useful in venous thrombosis.

In summary, we demonstrated that addition of a 5-SNP risk score to a risk scoring system based on non-genetic risk factors significantly improved the risk prediction of venous thrombosis. Although additional predictive markers may be required for a risk score to be clinically useful in the general population, the 5-SNP risk score may aid the management of subgroups of high-risk individuals.

REFERENCES

1. Bezemer ID, Rosendaal FR. Predictive genetic variants for venous thrombosis: what's new? *Semin Hematol*. 2007; 44(2):85-92.
2. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med*. 2006; 8(7):395-400.
3. Janssens AC, Moonesinghe R, Yang Q, Steyerberg EW, van Duijn CM, Khoury MJ. The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genet Med.* 2007; 9(8):528-535.
4. Brautbar A, Ballantyne CM, Lawson K, et al. Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the Atherosclerosis Risk In Communities Study. *Circ Cardiovasc Genet.* 2009; 2(3):279-285.
5. Paynter NP, Chasman DI, Pare G, et al. Association between a Literature-Based Genetic Risk Score and Cardiovascular Events in 19,313 Women. *JAMA*. 2010; 303(7): 631-637.
6. Davies RW, Dandona S, Stewart AFR, et al. Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. *Circ Cardiovasc Genet*. 2010; 3(5): 468-474.
7. Chen H, Poon A, Yeung C, et al. A genetic risk score combining ten psoriasis risk loci improves disease prediction. *PLoS One*. 2011; 6(4): e19454.
8. Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet*. 2010; 376(9750):1393-1400.
9. van Hylckama Vlieg A, Baglin CA, Bare LA, Rosendaal FR, Baglin TP. Proof of principle of potential clinical utility of multiple SNP analysis for prediction of recurrent venous thrombosis. *J Thromb Haemost*. 2008; 6(5):751-754.
10. Blom JW, Doggen CJM, Osanto S, Rosendaal FR. Malignancies, prothrombotic mutations, and the risk of venous thrombosis. *JAMA*. 2005; 293(6):715-722.
11. van Stralen KJ, Rosendaal FR, Doggen CJM. Minor injuries as a risk factor for venous thrombosis. *Arch Intern Med*. 2008; 168(1):21-26.
12. Koster T, Rosendaal FR, De Ronde H, Briët E, Vandenbroucke JP, Bertina RM. Venous thrombosis due to poor anticoagulant response to activated protein C: Leiden Thrombophilia Study. *Lancet*. 1993; 342(8886-8887):1503-1506.
13. Smith NL, Hindorff LA, Heckbert SR, et al. Association of genetic variations with nonfatal venous thrombosis in postmenopausal women. *JAMA*. 2007; 297(5):489-498.

14. Smith NL, Rice KM, Bovill EG, et al. Genetic variation associated with plasma von Willebrand factor levels and the risk of incident venous thrombosis. *Blood*. 2011;117(22):6007-6011.
15. Bezemer ID, Bare LA, Doggen CJM, et al. Gene variants associated with deep vein thrombosis. *JAMA*. 2008; 299(11):1306-1314.
16. Li Y, Bezemer I, Rowland CM, et al. Genetic variants associated with deep vein thrombosis: the F11 locus. *J Thromb Haemost*. 2009; 7(11):1802-1808.
17. Morange PE, Bezemer I, Saut N, et al. A follow-up study of a genome-wide association scan identifies a susceptibility locus for venous thrombosis on chromosome 6p24.1. *Am J Hum Genet.* 2010;86(4):592-595.
18. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983; 148(3):839-843.
19. Delluc A, Gourhant L, Lacut K, et al. Association of common genetic variations and idiopathic venous thromboembolism. Results from EDITh, a hospital-based case-control study. *Thromb Haemost*. 2010; 103(6):1161-1169.
20. Emmerich J, Rosendaal FR, Cattaneo M, et al. Combined effect of factor V Leiden and prothrombin 20210A on the risk of venous thromboembolism--pooled analysis of 8 case-control studies including 2310 cases and 3204 controls. Study Group for Pooled-Analysis in Venous Thromboembolism. *Thromb Haemost*. 2001; 86(3):809-816.
21. Jick H, Slone D, Westerholm B, et al. Venous thromboembolic disease and ABO blood type. A cooperative study. *Lancet.* 1969; 15;1(7594):539-542.
22. Grünbacher G, Weger W, Marx-Neuhold E, et al. The fibrinogen gamma (FGG) 10034C>T polymorphism is associated with venous thrombosis. *Thromb Res*. 2007; 121(1):33-36.
23. Uitte de Willige S, Pyle ME, Vos HL, et al. Fibrinogen gamma gene 3'-end polymorphisms and risk of venous thromboembolism in the African-American and Caucasian population. *Thromb Haemost*. 2009; 101(6):1078-1084.
24. Smith NL, Wiggins KL, Reiner AP, et al. Replication of findings on the association of genetic variation in 24 hemostasis genes and risk of incident venous thrombosis. *J Thromb Haemost*. 2009; 7(10):1743-1746.
25. Austin H, De Staercke C, Lally C, et al. New gene variants associated with venous thrombosis: a replication study in white and black Americans. *J Thromb Haemost*. 2011; 9(3):489-495.
26. Wells PS, Anderson JL, Scarvelis DK, et al. Factor XIII Val34Leu variant is protective against venous thromboembolism: a HuGe review and meta-analysis. Am J Epidemiol. 2006; 164:101-109.
27. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithm (QThrombosis) to estimate future risk of venous thromboembolism: prospective cohort study. *BMJ*. 2011; 343: d4656.

28. Naess IA, Christiansen SC, Romundstad P, Cannegieter SC, Rosendaal FR, Hammerström J. Incidence and mortality of venous thrombosis: a population-based study. *J Thromb Haemost*. 2007; 5(4):692-699.
29. Vandenbroucke JP, Koster T, Briët E, Reitsma PH, Bertina RM, Rosendaal FR. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet*. 1994; 344(8935):1453-1457.
30. Vossen CY, Rosendaal FR. The protective effect of the factor XIII Val34Leu mutation on the risk of deep venous thrombosis is dependent on the fibrinogen level. *J Thromb Haemost*. 2005; 3(5):1102-1103.
31. Souto JC, Almasy L, Borrell M, et al. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. *Am J Hum Genet*. 2000; 67(6):1452-1459.
32. Larsen TB, Sorensen HT, Skytthe A, Johnsen SP, Vaupel JW, Christensen K. Major genetic susceptibility for venous thromboembolism in men: a study of Danish twins. *Epidemiology*. 2003; 14(3):328-332.

SUPPLEMENTAL TABLES

Table S1. Associations between non-genetic risk factors and venous thrombosis risk in MEGA.

	MEGA			
	Risk factor frequency, %			
Non-genetic risk factors	**cases**	**Controls**	**OR**	**95% CI**
Plaster cast	5	1	5.35	(3.21-8.92)
Leg injury	14	3	5.11	(4.01-6.51)
Malignancy	8	2	4.91	(3.64-6.62)
OC use	28	11	3.98	(3.44-4.62)
Surgery	16	3	3.48	(2.66-4.55)
Hospitalization	15	3	2.69	(2.01-3.60)
Family history	32	17	2.68	(2.34-3.06)
Bedridden at home	12	4	2.29	(1.81-2.90)
Pregnancy or postpartum	3	1	2.23	(1.50-3.32)
Obesity	21	14	1.83	(1.57-2.13)
HRT	3	3	1.11	(0.77-1.60)
Travel	18	17	1.05	(0.90-1.22)

OR odds ratio; CI confidence interval; OC oral contraceptive; HRT hormone replacement therapy

Table S2. 31 SNP associations with venous thrombosis in LETS

				LETS			
				Risk allele frequency, %			
Gene	**SNP**	**Chr**	**Position**	**cases**	**Controls**	**OR**	**95% CI**
F5	rs6025	1	167.785.673	11	2	7.19	(4.05-12.76)
F2	rs1799963	11	46.717.631	3	1	2.99	(1.43-6.23)
ABO	rs8176719	9	136.132.908	44	36	1.43	(1.18-1.75)
FGG	rs2066865	4	155.744.726	34	26	1.45	(1.18-1.78)
F11	rs2036914	4	187.429.475	60	54	1.27	(1.05-1.53)
PROCR	rs2069951	20	33.227.425	7	5	1.40	(0.96-2.04)
F11	rs2289252	4	187.444.375	47	43	1.19	(0.99-1.42)
F9	rs4149755	X	138.451.778	7	7	0.96	(0.70-1.31)
PROCR	rs2069952	20	33.227.612	64	61	1.13	(0.94-1.37)
SERPINC1	rs2227589	1	172.152.839	12	9	1.42	(1.04-1.94)
HIVEP1	rs196713	6	11.920.517	23	20	1.02	(0.88-1.18)
F2	rs3136516	11	46.717332	51	50	1.03	(0.86-1.25)
F5	rs1800595	1	167.776.972	5	4	1.38	(0.89-2.16
PROC	rs1799809	2	127.892.345	47	43	1.19	(0.99-1.43)
PROCR	rs867186	20	33.228.215	15	13	1.22	(0.94-1.60)
VWF	rs1063856	12	6.153.534	35	36	0.96	(0.79-1.17)
GP6	rs1613662	19	60.228.407	85	80	1.36	(1.07-1.74)
F2	rs3136520	11	46.699.808	3	3	1.06	(0.62-1.79)
F8	rs1800291	X	153.811.479	85	82	1.15	(0.93-1.42)
STXBP5	rs1039084	6	147.635.413	44	40	1.19	(0.98-1.43)
NAT8B	rs2001490	2	73.781.606	43	38	1.22	(1.01-1.49)
F13B	rs6003	1	195.297.644	10	8	1.32	(0.96-1.83)
RGS7	rs670659	1	239.228.398	70	64	1.27	(1.04-1.54)
F9	rs6048	X	138.460.946	73	67	1.21	(1.02-1.44)
F5	rs4524	1	167.778.379	80	74	1.36	(1.09-1.69)
F13A1	rs5985	6	6.263.794	79	76	1.19	(0.95-1.49)
F3	1208 indel	1	94.780.000	43	49	0.78	(0.65-0.94)
TFPI	rs8176592	2	188.040.937	67	69	0.89	(0.73-1.10)
F11	rs3822057	4	187.425.146	55	51	1.19	(0.99-1.42)
NR1I2	rs1523127	3	120.983.729	42	33	1.43	(1.19-1.73)
CPB2	rs3742264	13	45.546.095	71	67	1.22	(1.00-1.50)

SNP single nucleotide polymorphism; Chr chromosome; OR odds ratio; CI confidence interval

CHAPTER 7

Mendeliaanse randomisatie

de Haan HG, Siegerink B, van Hylckama Vlieg A.

Ned Tijdschr Geneeskd. 2014;158:A7547.

ABSTRACT

- Classical observational studies into the causal relationship between a risk factor and a disease sometimes result in contradictory and spurious findings. This is due to confounding factors.

- It is not possible to conclude from the results of classical observational studies whether a specific risk factor may be a suitable target for future treatments.

- A solution is to conduct a Mendelian randomization analysis, which uses genetic variation as a surrogate marker for the risk factor.

- Mendelian randomisation is based on the idea that characteristics and environmental factors are proportionately divided into carriers and non-carriers of various genetic variants.

- Mendelian randomisation can be used only if there is a robust relationship between the genetic variant and the risk factor, if the genetic variant is not associated with other factors that confound the relationship between the risk factor and the disease, and if the genetic variant has an effect on the disease only via the risk factor, i.e. not via other biological mechanisms.

Als je een nieuwe behandelstrategie wil ontwikkelen, dan zou je idealiter van tevoren al willen vaststellen of de risicofactor waar je je strategie op gaat richten, daadwerkelijk een oorzaak is van het ziekteproces. Als die risicofactor geen oorzaak is, dan zullen eventuele behandelstrategieën immers weinig effect hebben op het ziekteproces.

Neem bijvoorbeeld het C-reactief proteïne (CRP). Enkele jaren geleden vond men in verschillende observationele onderzoeken een verband tussen de CRP-concentratie en het risico op cardiovasculaire ziekten: bij een hogere CRP-concentratie was het risico hoger.[1,2] Maar een verband betekent nog geen causaal verband. Uitsluitsel over het oorzakelijk verband tussen de CRP-concentratie en de verhoogde kans op cardiovasculaire ziekten was daarom wenselijk.

Om duidelijkheid te krijgen over een oorzaak-gevolgrelatie kan men een methode genaamd 'Mendeliaanse randomisatie' toepassen.[3,4] Deze methode gebruikt genetische variatie als 'proxy' (surrogaatmarker) voor de risicofactor; in termen van ons voorbeeld: we gebruiken variatie in het CRP-gen als proxy voor CRP-concentraties in bloed. Hierdoor kan men aan de hand van observationeel onderzoek toch uitspraken doen over de causaliteit van het verband.

Dit artikel beschrijft wat Mendeliaanse Randomisatie inhoudt, geeft een historisch voorbeeld en bespreekt wanneer deze methode wel of juist niet te gebruiken is. Voor alle duidelijkheid: we gaan niet in op de identificatie van genetische risicofactoren als voorspellers van ziekte.

Problemen in klassieke observationele studies

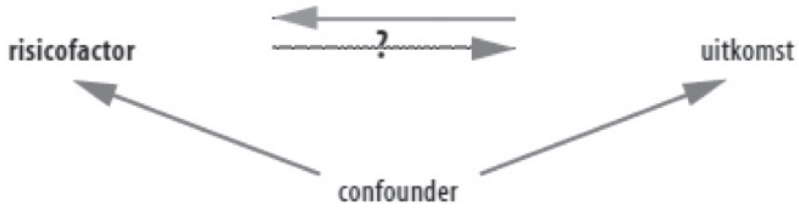
Het aantonen van een causaal verband is niet altijd mogelijk in klassiek observationeel onderzoek. 'Confounding' en 'reverse causation' kunnen namelijk een statistisch verband tussen de risicofactor en de uitkomst verklaren (Figuur 1).[5,6]

Bij reverse causation zijn oorzaak en gevolg omgedraaid; de ziekte of een subklinische vorm daarvan veroorzaakt een verandering in de parameter die beschouwd wordt als risicofactor, in plaats van andersom. In het CRP-voorbeeld dacht men dat verhoogde CRP-concentraties een oorzaak konden zijn van cardiovasculaire ziekten, bijvoorbeeld doordat het CRP de hechting van monocyten aan de vaatwand en hun migratie door de vaatwand zou bevorderen; op die manier zou het CRP indirect plaquevorming zou

stimuleren.[7] Aan de andere kant zouden CRP-concentraties ook verhoogd kunnen zijn door ontsteking van de vaatwand als gevolg van beginnende atherosclerotische processen. In dat geval is de ziekte de oorzaak en de verhoogde CRP-concentratie het gevolg.

Bij confounding verstoren andere factoren de onderzochte associatie.[5] Dat zijn dan factoren die geassocieerd zijn met zowel de risicofactor als het ziekteproces. Hierdoor lijkt er -ten onrechte- een verband tussen de risicofactor en het ziekteproces te zijn. Mogelijke bronnen van confounding in ons CRP-voorbeeld zijn bijvoorbeeld roken en veroudering. Traditioneel kan men confounding tegengaan door aangepaste analyses uit te voeren.[5] Als je echter niet alle bronnen van confounding volledig meeneemt in de analyses, blijft de associatie tussen risicofactor en uitkomst verstoord.

Waar klassieke observationele studies een vertekend resultaat kunnen opleveren door reverse causation en confounding, zijn deze problemen geminimaliseerd bij een Mendeliaanse randomisatie, zoals geïllustreerd met het historisch voorbeeld in de tabel.



Figuur 1. Relatie tussen risicofactor en uitkomst bij het gebruik van klassieke observationele analyses.

In klassieke observationele analyses wordt de risicofactor (bijvoorbeeld 'CRP-concentratie') gerelateerd aan de uitkomst ('risico op cardiovasculaire ziekten'). Vaak is het niet mogelijk om vast te stellen of het gevonden verband ook een oorzakelijke relatie weergeeft. Dit komt doordat er verstorende factoren aanwezig zijn, zoals 'confounders' (bijvoorbeeld 'leeftijd en BMI', rode pijlen) en 'reverse causation' (blauwe pijl), waarbij de uitkomst (bijvoorbeeld atherosclerose) de veronderstelde risicofactor beïnvloedt (bijvoorbeeld hogere CRP-concentraties).

Tabel. Historisch voorbeeld van Mendeliaanse randomisatie

Mendeliaanse randomisatie werd voor het eerst beschreven in de jaren 80, al werd die methode toen nog niet zo genoemd. In een brief aan *The Lancet* beschreef Katan hoe met behulp van variatie in het gen voor apolipoproteïne E (*apoE*) de causale relatie tussen cholesterolconcentraties en kanker onderzocht kon worden.[13] Resultaten van diverse klassieke observationele studies hadden geleid tot de gedachte dat lagere serumcholesterolconcentraties mogelijk kanker konden veroorzaken. Hiermee was cholesterolverlagende medicatie in een kwaad daglicht komen te staan.

Katan motiveerde hoe de resultaten uit observationele studies mogelijk verklaard konden worden door 'reverse causation': de lage serumcholesterolconcentraties bij kankerpatiënten zijn mogelijk een gevolg van de aanwezigheid van de tumor, zelfs als deze nog in een subklinisch stadium verkeert. Maar ook confounding speelt een rol: factoren zoals leeftijd en leefstijl zijn gerelateerd aan zowel de hoogte van de cholesterolconcentratie als aan het risico op kanker.

Om tot een definitieve uitspraak te komen suggereerde Katan om gebruik te maken van natuurlijke variaties in het *apoE*-gen, waarvan bekend was dat dragers hun leven lang gemiddeld lagere serumcholesterolconcentraties hebben dan niet-dragers, onafhankelijk van hun leefstijl of andere factoren die cholesterolconcentraties beïnvloeden. Katan redeneerde als volgt: als lage serumcholesterolconcentraties oorzakelijk zijn voor tumorgroei, dan zullen mensen met deze *apoE* gen-varianten ook een verhoogd risico op kanker hebben.[13]

De voorgestelde Mendeliaanse-randomisatie-analyse werd jaren later uitgevoerd in een studie waarin DNA-materiaal, cholesterolbepalingen en gegevens over kankerincidentie van bijna 3000 individuen beschikbaar waren. Uit deze analyse bleek dat lagere cholesterol- en LDL-concentraties weliswaar geassocieerd waren met een hoger risico op kanker, maar dat variatie in het *apoE2*-gen niet geassocieerd was met kankerincidentie en kankermortaliteit.[14] Hieruit kan geconcludeerd worden dat lage LDL- en cholesterolconcentraties geen oorzaak zijn van kanker. De eerder gevonden associaties in de klassieke epidemiologische analyses waren dus het gevolg van confounding of reverse causation.

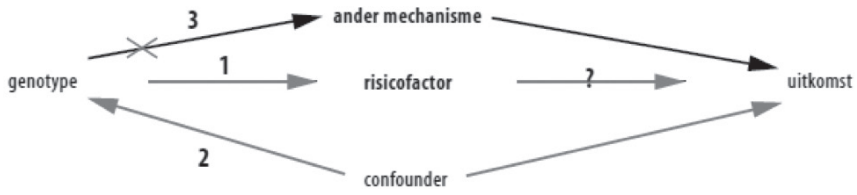
Principes van Mendeliaanse randomisatie

Zoals gezegd ligt het principe van Mendeliaanse randomisatie in het gebruik van genetische variatie als proxy (of 'instrument') voor de variatie in de risicofactor waarvan men graag wil vaststellen of deze daadwerkelijk een oorzaak is van de ziekte. En daarmee is Mendeliaanse randomisatie een bijzondere vorm van een zogenoemde instrumentele-variabele-analyse (zie uitlegkader).[8] In theorie kan men genetische variatie gebruiken als proxy voor allerlei risicofactoren, van eiwitten in het bloed tot leefstijl en psychologische factoren.[9] Terug naar ons voorbeeld: als er een causale relatie bestaat tussen CRP-concentraties en cardiovasculaire ziekten, dan is het aannemelijk dat variaties in het *CRP*-gen -die samenhangen met de CRP concentraties in het bloed- ook een oorzaak zijn van cardiovasculaire ziekten. Anders geformuleerd: doordat dragers van een bepaalde variant in het *CRP*-gen levenslang licht verhoogde CRP-concentraties hebben, zouden deze dragers een verhoogd risico op cardiovasculaire ziekten moeten hebben als het CRP daadwerkelijk een oorzaak is.

Het minimaliseren van confounding en reverse causation bij analyses met Mendeliaanse randomisatie voert terug op de tweede wet van overerving van Gregor Mendel. Deze wet stelt dat genen, maar ook genetische variatie, onafhankelijk van elkaar van ouders op kind overerven. Analoog hieraan zijn niet alleen alle genetische eigenschappen, maar ook omgevingsfactoren evenredig verdeeld over de dragers van de verschillende genetische varianten van het CRP-gen. Zo zal het percentage rokers even groot zijn onder dragers en niet-dragers van een variant in het *CRP*-gen. Hierdoor wordt confounding geminimaliseerd. Maar ook reverse causation wordt geëlimineerd, aangezien de genetische variatie van een individu wordt vastgelegd bij de conceptie en dus niet wordt beïnvloed door ziekte.

Voorwaarden voor Mendeliaanse randomisatie

Mendeliaanse-randomisatie-analyses zijn gebaseerd op een aantal aannames (Figuur 2). Men kan nagaan of deze geldig zijn, maar de geldigheid is vaak niet te garanderen.[8,10] De juiste interpretatie van analyse met Mendeliaanse randomisatie hangt hier echter wel van af. Hier bespreken we kort deze voorwaarden en manieren om de validiteit van een Mendeliaanse-randomisatie-analyse te controleren.



Figuur 2. Voorwaarden waaronder Mendeliaanse randomisatie toegepast mag worden.
Allereerst dient er een sterke associatie te zijn tussen het genotype (bijvoorbeeld een bepaalde variant in *CRP* gen) en de risicofactor (in dit voorbeeld: CRP-concentratie) (pijl 1). De tweede voorwaarde houdt in dat het genotype niet geassocieerd mag zijn met een andere risicofactor (bijvoorbeeld etniciteit) die gerelateerd is aan de uitkomst (in dit geval: cardiovasculaire ziekte); anders zou confounding ontstaan (pijl 2). Tot slot mag het genotype niet geassocieerd zijn met de uitkomst via een ander mechanisme (pleiotropie; pijl 3).

Robuust verband met de risicofactor

Allereerst dient er een robuust verband te zijn tussen de genetische variant en de risicofactor. In ons voorbeeld houdt dit in dat variaties in het *CRP*-gen een voldoende groot effect moet hebben op de CR-concentraties. Dat is niet altijd het geval, aangezien vaak vele genetische varianten, in één of zelfs meerdere genen, bijdragen aan de variatie in de risicofactor. Zo worden de CRP-concentraties ook beïnvloed door variaties in andere genen, bijvoorbeeld in *HNF1a* en *LEPR*[11].

Het gebruik van een genetische variant die een zeer zwakke associatie heeft met de risicofactor kan leiden tot bias.[12] Dit kun je voorkomen door de verschillende genetische varianten in één risicoscore te combineren. Verder is het mogelijk dat de veronderstelde associatie tussen genotype en risicofactor 'fout-positief' is. Bij het gebruik van zo'n genetische variant in de Mendeliaanse-randomisatie-analyse zal, wellicht ten onterechte, geconcludeerd worden dat de risicofactor géén oorzaak is van de ziekte. Dit kun je voorkomen door alleen genetische varianten te gebruiken die in meerdere, onafhankelijke studies geassocieerd zijn met de risicofactor.

Geen associatie met andere confounders

De tweede voorwaarde is dat de genetische variant niet geassocieerd mag zijn met andere factoren die het verband tussen de risicofactor en de ziekte vertekenen. Dit zou namelijk weer leiden tot confounding. Variaties in het *CRP*-gen mogen dus niet vaker gepaard gaan met andere factoren die het risico op cardiovasculaire ziekten

beïnvloeden dan men op basis van toeval zou verwachten. De frequentste voorbeelden hiervan zijn genetische varianten die tegelijk overerven met de genetische variant die in de Mendeliaanse randomisatie onderzocht wordt; er is dan sprake van 'linkage disequilibrium'. Dit gebeurt wanneer de varianten in dezelfde regio op hetzelfde chromosoom liggen. Dat is met behulp van publieke databases en biologische kennis redelijk goed te bestuderen.

Een andere bron van confounding kan optreden wanneer men een Mendeliaanse-randomisatie-analyse uitvoert bij een onderzoekspopulatie met mensen van verschillende etnische achtergronden, elk met hun eigen genetische variaties en eigen basisrisico op ziekte. Deze vorm van confounding -ook wel 'populatiestratificatie' genoemd- kan men voorkomen door aanpassingen in de onderzoeksopzet of door correcties in de data-analyse. Dit is natuurlijk alleen mogelijk als de verschillende etnische achtergronden binnen de onderzoekspopulatie voldoende nauwkeurig vastgesteld kunnen worden.

Geen andere biologische mechanismen

Als derde en laatste voorwaarde dient de genetische variant alleen via de tussenliggende risicofactor geassocieerd te zijn met de ziekte, en dus niet via andere biologische mechanismen. Dit betekent dat variaties in het *CRP*-gen niet mogen leiden tot veranderingen in het serumcholesterolconcentraties of andere factoren die geassocieerd zijn met cardiovasculaire ziekten. Als dat wél het geval is, dan is er sprake van 'pleiotropie' en is het niet langer duidelijk voor welke risicofactor de genetische variatie nu als proxy functioneert. De aanwezigheid van pleiotropie is soms bekend uit de literatuur en kan, soms worden vastgesteld aan de hand van eigen data, maar men kan dit fenomeen nooit uitsluiten.

Een ander fenomeen dat het verband tussen het genotype en de uitkomstmaat kan vertekenen is 'kanalisatie'. Dit houdt in dat er biologische, epigenetische aanpassingsmechanismen in werking komen, bijvoorbeeld verhoging van de concentratie van IL-10 -een cytokine met anti-atherosclerotisch effect-, die de effecten van de genetische variatie in *CRP* te compenseren. Het is moeilijk in te schatten in hoeverre kanalisatie plaatsvindt en in hoeverre dit ook daadwerkelijk het verband tussen genotype en uitkomstmaat vertekent.

Voor- en nadelen van Mendeliaanse randomisatie

Het grootste voordeel van Mendeliaanse randomisatie is het minimaliseren van bekende en onbekende confounding en reverse causation. Daarnaast kan men het effect van een levenslange blootstelling bepalen, iets wat in klassiek observationeel onderzoek vaak niet mogelijk is.

Om meer duidelijkheid te krijgen over de causale relatie tussen risicofactor en ziekteproces kan men ook kiezen voor een zogenaamde gerandomiseerde, gecontroleerde trial (RCT), waarbij de onderzoeker de blootstelling aan de risicofactor direct beïnvloedt. Door de randomisatie worden alle mogelijke verstorende factoren in principe gelijk verdeeld over de behandelgroepen en daardoor is confounding geminimaliseerd. Deze onderzoeksopzet is echter niet altijd mogelijk vanwege ethische en praktische overwegingen. Bovendien is de generaliseerbaarheid van een RCT doorgaans beperkt door strenge in- en exclusiecriteria en relatief gezonde deelnemers. In observationele studies daarentegen is het mogelijk om een representatieve steekproef van de algemene bevolking te includeren om daarmee de klinische toepasbaarheid van de resultaten te vergroten.

De methode van Mendeliaanse randomisatie kent ook beperkingen.[3,4,10] Zo moet de onderzoekspopulatie vaak erg groot zijn om met enige zekerheid de associaties - zowel tussen proxy en risicofactor als tussen proxy en ziekte- te kunnen bepalen. En zoals we al hebben aangegeven, steunt Mendeliaanse randomisatie op een aantal voorwaarden. Als aan deze voorwaarden niet voldaan wordt, kan de methode vertekende resultaten opleveren. Het is aan de onderzoeker om de lezer ervan te overtuigen dat aan alle voorwaarden zo goed als mogelijk is voldaan.

Conclusie

Klassieke observationele studies naar de causale relatie tussen een risicofactor en een ziekte resulteren soms in tegenstrijdige en foutieve bevindingen door de aanwezigheid van verstorende factoren (bias en confounding) of reverse causation. Hierdoor blijft het onduidelijk of de risicofactor een geschikt aangrijpingspunt kan zijn voor toekomstige behandelingen. In die situaties worden steeds vaker analyses op basis van Mendeliaanse randomisatie toegepast. Bij Mendeliaanse-randomisatie-analyse worden de associaties tussen genotype en risicofactor en tussen genotype en ziekte gebruikt om het causale verband tussen de risicofactor en de ziekte te herleiden. Doordat deze methode

uitgaat van onafhankelijke overerving van genen, zijn problemen als confounding en reverse causation geminimaliseerd. Mendeliaanse randomisatie vereist echter grote studiepopulaties, steunt op enkele cruciale aannames en kan vertekende resultaten opleveren als niet aan de voorwaarden is voldaan. Het is niet mogelijk te garanderen dat aan alle voorwaarden is voldaan, maar als er voldoende biologische kennis is over de risicofactor en het ziektemechanisme kan men op basis van analyses van eigen onderzoeksgegevens een goed beoordelen in hoeverre aan de voorwaarden voor Mendeliaanse randomisatie is voldaan.

Uitlegkader.

'Reverse causation'
Omdraaiing van oorzaak en gevolg: de ziekte beïnvloedt de risicofactor en niet andersom. Dit fenomeen kan optreden bij een klassieke observationele studie. Hierdoor kunnen onderzoeksresultaten vertekend zijn.

Confounding
Verstoring van het verband tussen risicofactor en ziekte door andere factoren die zowel gerelateerd zijn aan de risicofactor als aan de ziekte. Dit komt vaak voor bij klassieke observationele studies en kan vertekende onderzoeksresultaten opleveren.

Instrumentele-variabele-analyse
Instrumentele-variabele-analyse is een onderzoeksmethode in observationeel onderzoek waarbij een proxy ('instrument') voor een risicofactor gebruikt wordt om te bestuderen in hoeverre er een causale relatie tussen risicofactor en ziekte is. Mendeliaanse randomisatie is hier een bijzondere vorm van.

Mendeliaanse randomisatie
Observationele onderzoeksmethode waarbij genetische variatie wordt gebruikt als proxy voor een risicofactor. Omdat reverse causation en confounding minder snel de resultaten van deze methode verstoren is het mogelijk om de causale relatie tussen risicofactor en ziekte te bestuderen.

Pleiotropie
Situatie waarin een genetische variant via meerdere en verschillende mechanismen een effect heeft op het lichaam. Als er sprake is van pleiotropie, dan is het niet mogelijk om een uitspraak te doen over de precieze causale mechanismen die leiden tot de ziekte.

Kanalisatie
Biologische aanpassingsmechanismen die in werking treden bij bepaalde genetische varianten en die de effecten van genetische variate compenseren.

REFERENTIES

1. Ridker PM, Buring JE, Shih J, et al. Prospective Study of C-Reactive Protein and the Risk of Future Cardiovascular Events Among Apparently Healthy Women. *Circulation*. 1998;98:731-3
2. Koenig W, Sund M, Frohlich M, et al. C-Reactive Protein, a Sensitive Marker of Inflammation, Predicts Future Risk of Coronary Heart Disease in Initially Healthy Middle-Aged Men. *Circulation*. 1999;99:237-42
3. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1-22
4. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials and limitations. *Int J Epidemiol*. 2004;33:30-42
5. Groenwold RH. Verstoring in observationeel onderzoek: 'confounding'. *Ned Tijdschr Geneeskd*. 2012;156:A4221
6. Groenwold RHH. 3 vormen van bias. *Ned Tijdschr Geneeskd*. 2013;157:A6497
7. Jialal I, Devaraj S, Venugopal SK. C-Reactive Protein: Risk Marker or Mediator in Atherothrombosis? *Hypertension*. 2004;44:6-11
8. Boef AGC, le Cessie S, Dekkers OM. Instrumentele-variabele-analyse. *Ned Tijdschr Geneeskd*. 2013;157:A5481
9. Davey Smith G, Ebrahim S. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ*. 2005;330:1076-79
10. Lawlor DA, Harbord RM, Sterne JA, et al. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statist. Med*. 2008;27:1133-63
11. Wu Y, McDade TW, Kuzawa CW, et al. Genome-wide association with C-reactive protein levels in CLHNS: evidence for the CRP and HNF1A loci and their interaction with exposure to a pathogenic environment. *Inflammation.* 2012;35:574-83
12. Burgess S, Thompson SG, CRP HD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization. *Int J Epidemiol*. 2011;40:755-64
13. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*. 1986;1:507-8
14. Trompet S, Jukema JW, Katan MB, et al. Apolipoprotein e genotype, plasma cholesterol, and cancer: a Mendelian randomization study. *Am J Epidemiol*. 2009;170:1415-21

CHAPTER 8

Discussion

DISCUSSION

The aim of the research conducted for this thesis was to identify novel genetic risk factors for a first and recurrent venous thrombosis. In addition, we investigated whether previously identified genetic risk variants can be used to improve risk stratification for venous thrombosis and we discussed the potential value of using genetic variation to aid causal inferences in observational research. In this chapter, we discuss the main findings and some methodological considerations, and we provide directions for biological and clinical interpretations.

Main findings

So far, variation in seventeen genes, almost all encoding proteins related to hemostasis, have consistently been identified as genetic risk factors for a first VT.[1,2] Evidence from previous GWAS and family studies suggests that additional genetic risk variants are yet to be discovered.[3-6] In addition, the extent to which the identified risk variants contribute to recurrence risk is not clear, nor whether different genetic risk factors play a role in recurrence pathophysiology than those involved in a first event.[7-10] In chapters 2 to 5, we used various strategies to identify variants across the allele frequency spectrum that are associated with the risk of a first or recurrent VT.

In **chapter 2**, we studied the association between a first DVT and genetic variation in the coding regions of 734 genes related to hemostasis. More than 3,500 common variants, identified by next-generation DNA sequencing, were assessed in approximately 900 DVT patients and 600 controls. We confirmed, as expected, the association between DVT and variation in the *F11* region, *FGA-FGG*, *ABO*, and *F5*, which are all established risk loci for VT. At *F5* and the *F11* region we also found evidence for secondary association signals, suggesting that these risk loci contain multiple conditionally independent risk factors for DVT. Remarkably, we found only two suggestive association signals mapping to genes not previously implicated in VT pathophysiology, although these were not replicated in data from the INVENT consortium. In addition, an assessment of over 16,000 rare variants mapping to 647 genes did not reveal a burden of rare variants in DVT patients compared with controls. However, it is possible that associations of both common and rare variants conferring small effects on DVT risk were missed, as our study did not include sufficient patients and controls to identify such variants.

Instead of focusing on variation in candidate genes, we followed an agnostic approach in **chapter 3,** as for recurrence it is unknown whether the same or different genetic risk factors than those identified for a first VT play a role. We conducted a GWAS in which we studied the association between about 8 million common autosomal variants and recurrent VT, followed by a replication study. In addition to confirming the association between FV Leiden and recurrence risk, we identified a novel risk locus at 18q22.1, which was associated with recurrent VT with an odds ratio of 1.7 per minor allele copy in the replication analysis. This intergenic locus may affect recurrence risk by influencing the expression of nearby or distant genes, though further research is needed to unravel the underlying molecular mechanism. We found limited support for previously identified variant associations with recurrence, emphasizing the importance of replication in genetic association analyses.

A first investigation of variation in the Y chromosome and its effect on first and recurrent VT risk was reported in **chapter 4**. As men have an intrinsically higher risk of VT than women[11-16], we postulated that variation in the Y chromosome may increase the risk of VT in subgroups of men. We therefore explored the association between 13 common European Y chromosome haplogroups and the risk of a first and recurrent VT in over 3,700 men. Compared with the most common haplogroup R1b, none of the haplogroups were associated with the risk of a first VT. Specifically, no evidence for an association between haplogroup I, which was previously identified as a risk factor for coronary artery disease[17], and VT risk was observed, even though the analysis was powered to detect a similar association. In addition, we observed some suggestive evidence that carriers of haplogroup R1a had a decreased risk of recurrence compared with R1b-carriers. However, this cannot explain the difference in risk between men and women, as we observed a higher recurrence rate for R1a-carriers than for women.

We used a candidate gene approach in **chapter 5** to study common variation in *CADM1* and the association with the risk of a first VT. An earlier study in a protein C deficient family identified *CADM1*, encoding a cell adhesion molecule involved in endothelial cell migration, as a risk gene for VT.[18,19] To assess whether a joint effect of *CADM1* variation and protein C on VT risk also exists in the general population, we studied the association between over 300 variants in *CADM1* and VT risk in 962 individuals with an abnormality in the protein C pathway and 4004 controls. For six variants we observed a large joint effect on VT risk, of which one variant also showed evidence of an association with

VT in the overall study population of 3496 VT patients and 4004 controls. Due to the high number of statistical tests and low number of individuals with protein C pathway abnormalities, caution is needed when interpreting these results.

In the two remaining chapters, we discussed two of the main applications of genetic risk factors in research, that is risk stratification and Mendelian randomization. Using a panel of 31 previously reported VT risk variants, we constructed genetic risk scores and compared the discriminative values with a model based on clinical risk factors and a combined model (**chapter 6**). We showed that a score containing five risk variants (FV Leiden, PT G20210A, *ABO* non-O, *FGG*-rs2066865, and *F11*-rs2036914) added significant discriminative power to a clinical risk model for venous thrombosis in the general population. As genetic risk profiling is not (yet) cost-effective in the general population, we also explored risk discrimination in clinically relevant subgroups. Except among cancer patients, the genetic risk score performed similarly in the subgroups as in the general population. Replication of our findings in an independent study showed the robustness of our genetic risk score, although the genetic risk score may perform less well in populations with a different ethnic background.

In **chapter 7**, we discussed the possibilities of using genetic variation as an instrument for an exposure of interest to aid causal inference in observational studies. In this educational chapter, we explained that, if none of the Mendelian randomization (MR) assumptions are violated, a genetic instrument can be used to estimate the causal effect of the exposure on the outcome of interest, while minimizing confounding and reverse causation. Although not all assumptions are falsifiable, and a large study population is required, MR studies are increasingly successful applied in observational research, especially when randomized trials are not possible. Outside the scope of this chapter, where we merely described the concepts of MR in general, are the different analytical methods that have recently been developed, including those dealing with pleiotropy.[20]

Methodological considerations

Venous thrombosis is a common complex trait, driven by a multitude of genetic and environmental factors. The first genetic risk factor for VT was suspected over 60 years ago[21], and ever since, studies have aimed to unravel the genetic architecture underlying VT. At first, studies used linkage analyses in families and candidate gene approaches to identify risk genes as the genetic component of common complex traits was thought

to be based on a single gene or few genes each following Mendel's law of inheritance. Technological advances and large collaboratives such as the Human Genome Project[22] paved the way for systematic analysis of millions of (common) variants across the genome. These GWASs fitted the then popular 'common disease – common variant' hypothesis, which claimed that common traits such as VT would be the result of common variants each having a low penetrance.[23,24] Although GWASs identified many risk loci for common complex traits, including several for VT[4,25,26], these loci only explained part of the heritability of each trait.[27,28] For venous thrombosis, Germain *et al*.[3], estimated that common variants could explain around 35% of the genetic variance, of which only 3% could be attributed to the four most well-known risk variants (in *F5, ABO, FGG,* and *F11*). These observations fueled the 'common disease – rare variant' hypothesis, which argued that rare variants with high penetrance contribute substantially to complex trait genetics.[28,29] The advent of high-throughput exome and whole genome sequencing now allows large-scale investigations of rare and even 'private' variants using single-variant and aggregate association tests, though the effect sizes conferred by rare variants seem to be smaller than initially thought.[30-32] For VT risk, most studies have so far focused on rare variants associated with thrombophilia. Lotta *et al*.[33], observed a burden of rare coding variants in *ADAMTS13* associated with a 4.8-fold increased risk of DVT, but we did not replicate this finding in our sequencing data (**chapter 2**). Based on recent genetic studies on other common complex traits, the genetic architecture of VT is most likely characterized by a polygenic signature of common and rare variants conferring modest-to-small effects on disease risk.[28,30,34,35] The causal variants map most likely to both coding and noncoding sequence across the genome.[36-38] This has several important methodological consequences for studies aiming to identify novel risk factors for (recurrent) VT, which are discussed below.

First, sample size is of utmost importance when conducting large genetic association studies due to the small effect sizes that need to be detected with precision and the large number of statistical tests performed thereby requiring a stringent threshold to attain statistical significance. The number of tests conducted depends on the approach taken: a few to 500 (tagging) variants in a candidate gene study compared with several millions in a GWAS study imputed to a dense reference panel. As evidenced from the two largest GWAS studies on VT so far, the effects conveyed by low-frequency and common variants (MAF ≥ 1%) on VT risk are generally small, with odds ratios ranging between 1.1 and 1.8.[25,26] Exceptions are FV Leiden (MAF 3.0% in Europeans) and PT

G20210A (MAF 1.0% in Europeans) which are associated with a 3.5-fold and 2-fold increased risk of a first VT per copy of the minor allele, respectively.[25,39,40] As part of the INVENT consortium, we have previously meta-analyzed GWAS data from 7,507 VT patients and 52,632 controls, resulting in sufficient statistical power to detect odds ratios of >1.2 for common, but not low-frequency, variants.[25] We expect that with increasing sample sizes more genetic risk factors for VT will be identified, as has been the case for other common complex traits such as height and obesity.[41,42] Recent estimates suggest that for common complex traits sample sizes ranging from a few hundred thousand to multiple millions are required to identify variants that explain most heritability found in GWASs.[35,41] Sequencing studies focusing on rare variants across the exome or the entire genome require an even larger sample size to discover novel risk variants. Achieving these large sample sizes is a major bottleneck, as venous thrombosis occurs in only 1-2 per 1000 persons per year.[43,44] As such, several analyses conducted for this thesis were underpowered, and we may have missed relevant associations with venous thrombosis. To maximize statistical power, alternative strategies can be employed, such as we did in the sequencing study (**chapter 2**), where we specifically focused on DVT risk instead of DVT or PE in order to study a homogenous phenotype. In addition, we excluded individuals with major clinical risk factors for VT in order to study a population which is more likely to carry genetic risk variants. Further strategies to maximize power include studying population isolates, conducting transethnic analyses, or by using advanced statistical models such as Bayesian models that do not require Bonferroni correction for multiple testing.[45-47] Of note, sample size is not just critical for discovery analyses, but also for replication analyses in which the top candidates per locus, usually the variants with the lowest P-values, are tested in an independent sample. This P-value driven selection can lead to the so-called 'winner's curse', which is a bias away from the null similar to regression-to-the-mean.[48,49] Genetic variants passing the threshold for statistical significance are more likely to have overestimated effect sizes in the discovery sample due to chance. Therefore, if possible, replication analyses should be powered to detect effect sizes smaller than those reported in the initial discovery analysis.

Second, the genetic ancestry of the study population should be considered before and during genetic analyses. Genetic association studies in admixed populations may be hampered by confounding due to population structure.[50] As both allele frequencies and the incidence of VT vary according to genetic ancestry, the independence assumption is violated in studies of admixed populations resulting in potentially spurious associations.

To avoid this, studies should appropriately account for population structure. Therefore, most of our analyses, were limited to individuals of self-reported European origin. In the GWAS discussed in **chapter 3**, we used principal component analysis[51] to control for population structure and calculated the genomic inflation factor[52] to assess the presence of any remaining population substructure. Recent studies suggest that confounding by population structure may be more of a concern when studying rare variants, as these may show different stratification patterns compared with common variants due to selection pressure, founder effects, and as these are more likely to have arisen recently.[53,54] Of note, the downside of studying genetic risk variants in an ethnically homogenous population is that the results are only generalizable to that population. For example, the genetic risk score in **chapter 6** was constructed and validated in individuals of European origin and, therefore, performs less well in individuals of non-European ancestry as the included variants are less informative in non-European populations. For example, FV Leiden reaches a MAF of 3% in Europeans but is virtually absent in Africans and East Asians, thereby limiting its discriminative power in those populations.[55] While it has been shown that our genetic risk score has limited predictive value in African Americans,[56] another study reported some generalizability of VT risk variants identified in Europeans to other ancestries in a study on chronic venous disease.[57] As few and only small studies on genetic risk factors for VT have been performed in populations of non-European ancestry,[58-62] it is currently difficult to assess the generalizability of our findings.

Last but not least, linkage disequilibrium (LD), the non-random association of alleles at closely linked loci in a population, requires attention when conducting and interpreting genetic analyses. Specifically, LD may affect genetic association studies and Mendelian randomization studies, as associated variants may not be causal variants, but rather be in linkage with these. LD, amongst others determined by recombination rate and demographic aspects of a population, may extend for several megabases along a chromosome while sometimes interspersed with blocks of no or little LD.[63-65] As a result, causal variants may even map to different genes than the associated variants, complicating the interpretation of an association signal. Of the VT risk loci, LD blocks spanning multiple genes are, for example, observed at the *F11* locus and *FGA*-*FGG* locus.[66-68] We were therefore unable to disentangle the association between DVT and genetic variants *FGA*-rs6050 and *FGG*-rs2066865 (**chapter 2**), which have both previously been associated with VT risk[68,69] and are almost in complete LD (r^2 0.90 in

Europeans). In addition, a GWAS association signal at 11p11.2 has previously almost been misinterpreted as a novel risk locus for VT before it was tracked down to PT G20210A using LD and haplotype analyses.[70] Since LD patterns differ between genetic ancestries,[64,71] transethnic analyses could aid fine-mapping at regions with strong LD in Europeans.[72] Of note, even in regions with considerable LD, it is possible that multiple conditionally independent associations exist, either because there are multiple causal variants or the associated variants are all in moderate LD with the unmeasured causal variant(s). We and others have reported evidence for secondary associations at several of the known VT risk loci, including *ABO*, *CADM1*, *F2*, *F5,* and the *F11* locus (**chapters 2 and 5**).[4,25,58,66,67,73,74] Enlarging the sample size and extension to non-European populations will help to unravel the genetic structure at these loci.

Biological interpretation

Most of the established genetic risk factors for VT can be linked to the hemostatic system.[1,2] For some risk loci, the causal variant and the underlying biological mechanism have largely been elucidated. For example, a missense variant FV Leiden leads to loss of a cleavage site for activated protein C (APC), resulting in both APC resistance and decreased degradation of activated FVIII by APC and protein S.[39,75] PT G20210A results in increased PT plasma levels due to differential post-transcriptional regulation of PT mRNA,[40,76] whereas the *FGG*-haplotype containing rs2066865 yields lower levels of the γ'-fibrinogen and reduction of the γ'/γ ratio.[68] In addition, clearance of vWF is affected by the presence of A and B antigens of ABO on the surface of vWF.[77] The biological interpretation of other VT risk loci is more complex. VT risk variants in *F11* and *KNG1* are associated with increased FXI plasma and/or activity levels and with prolonged activated partial thromboplastin time.[25,66,67,78,79] However, it is suggested that their association with venous thrombosis cannot be completely explained by their effect on FXI levels.[79,80] Near *F11,* and part of the same LD block, lie *KLKB1* and *CYP4V2*, encoding prekallikrein and a cytochrome P450 family member, respectively. Several studies (including our sequencing study in **chapter 2**) have reported multiple conditionally independent associations between VT and variants in *KLKB1*, *CYP4V2*, and *F11*[25,66,67], but the exact causal mechanism has not been elucidated due to the extensive LD at this locus. Data from the Genotype-Tissue Expression Project[81] are also inconclusive: *F11*-rs2036914 is, for example, an expression quantitative trait locus (eQTL) for *F11* in lung tissue, whereas *F11*-rs1593 is an eQTL for *KLKB1* and *CYP4V2,* but not *F11*, in multiple tissues.

Furthermore, the link to venous thrombosis is unclear for the recently identified GWAS loci near *TSPAN15* and *SLC44A2*, which showed no evidence of an association with any of 25 hemostasis-related biomarkers.[25] It should be noted that the causal variant at these loci may also target a different gene, as many GWAS loci associated with common complex traits have shown not to impact the most nearby gene.[82,83] GWASs typically identify associations in noncoding sequence, which cannot be explained by linkage to coding variants, and are thought to impact a complex trait by affecting gene regulation, both transcriptionally and post-transcriptionally.[27,36-38,83] In order to elucidate the functional impact of such variants, integration with multiple genomics data, such as generated by ENCODE[84] and GTEx[81], is necessary. For example, colocalization analyses of GWAS hits with overlapping eQTL associations in relevant tissues can be used to pinpoint plausible causal variants and genes.[85,86] Further integration with methylation and epigenomic annotation data can help to dissect potential regulatory mechanisms, whereas chromatin interaction methods can detect long-range chromosomal interactions between variants in potential enhancers and their target genes.[87-89] These methods should also be applied to identify the causal variant and gene for the intergenic locus at 18q22.1, which was associated with recurrent VT (**chapter 3**). In addition, leveraging from data on endophenotypes, such as plasma coagulation factor levels, or metabolomics can help to dissect the biological link between the identified variants and the pathophysiology of VT.

Our lack of understanding of the biological underpinnings of GWAS loci also hampers the clinical translation of these genetic risk factors. Much effort is currently spent to increase our understanding of the role of regulatory variation in the genome. As this research field is evolving fast, with new methods and data becoming available on a regular basis, we expect that the biological mechanism underlying GWAS variants and other VT risk variants can be unraveled in the near future.

Clinical relevance

The ultimate goal of genetic association studies is to bring the genetic discoveries to the clinic, assuming that a better understanding of the biology underlying a disease leads to better treatments and preventive strategies. Specifically, elucidating risk genes and pathways may provide novel drug targets, for example, those that reduce thrombosis risk without (substantially) increasing the bleeding risk. Although the effect sizes of individual risk variants are small, their effect on molecular phenotypes

and the resulting drug effects can be large. A well-known example is the field of pharmacogenetics, which investigates genetic variation in metabolic pathways affecting individual responses to drugs. Variation in the vitamin K epoxide reductase (*VKORC1*) and hepatic drug-metabolizing enzyme cytochrome P450 2C9 (*CYP2C9*) genes largely determine the dose variability of coumarin anticoagulants.[90,91] As a result, patients taking these anticoagulants to prevent or treat thrombotic events have, depending on their genotypes, an increased risk of major bleeding due to over-anticoagulation. So far, several trials have investigated the use of genotype-guided dosing to reduce the number of adverse events during anticoagulant treatment, albeit with inconsistent results.[92-95] Besides guiding therapy, genetic variation may be informative in personalized risk prediction, i.e. identifying those who are at increased risk of developing VT and those who are not. In **chapter 6**, we showed that a genetic risk score of five well-known VT risk variants improved risk stratification in the general population and in clinically relevant subgroups. Our genetic risk score has been validated and extended in other studies of individuals of European ancestry, but showed limited discriminative power in African Americans.[56, 96-99] Identification of additional genetic variants, especially variants that increase VT risk in individuals of non-European ancestry, may further improve the discriminative power of such genetic risk scores. As the costs of genotyping continue to drop, the implementation of genetic risk factors into clinical prediction models may also become cost-effective. This may be most relevant for recurrence risk, as patients with a recurrent VT currently receive lifelong treatment with anticoagulants, which are associated with an increased risk of bleeding.

Two other clinically relevant applications of genetic findings are Mendelian randomization studies and studies focusing on the genetic correlation between traits. Specifically, GWAS results have shown that the same genetic variants can be associated with multiple traits, suggesting that some of the underlying causal mechanisms are shared.[100,101] This pleiotropic nature can also be exploited to quantify the genetic overlap between traits and diseases using methods such as cross-trait LD score regression.[100] As large-scale GWAS summary statistics for VT are not publicly available, a systematic analysis of genetic correlation between VT and other traits has not (yet) been published. A first study by Klarin *et al.*, based on a genetic risk score consisting of 10 VT risk variants, showed a statistically significant genetic overlap between VT and coronary artery disease risk, but not with 37 other disorders tested in data from the UK Biobank.[4] MR studies, on the other hand, can aid in unravelling the causal relationship between

clinical factors and VT risk (as explained in **chapter 7**). So far, MR studies on VT have shown that obesity and height, but not lipoprotein(a) and YKL-40, are causal risk factors for VT.[4,102-105] As more genetic variants are being identified and the analytical methods are being improved, we expect that both MR and genetic correlation analyses will become standard tools in genetic studies on VT and other common complex traits, ultimately advancing personalized medicine.

REFERENCES

1. Rosendaal FR, Reitsma PH. Genetics of venous thrombosis. *J Thromb Haemost*. 2009; 7 Suppl 1:301–304.
2. Trégouët DA, Morange PE. What is currently known about the genetics of venous thromboembolism at the dawn of next generation sequencing technologies. *Br J Haematol*. 2018;180(3):335-45.
3. Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, et al. Genetics of venous thrombosis: insights from a new genome wide association study. *PLoS One*. 2011;6(9):e25581.
4. Klarin D, Emdin CA, Natarajan P, Conrad MF; INVENT Consortium, Kathiresan S. Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor. *Circ Cardiovasc Genet*. 2017;10(2):e001643.
5. Souto J, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria J, Coll I, Felices R, Stone W, Fontcuberta J, Blangero J. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. *Am J Hum Genet*. 2000;67:1452–9.
6. Heit J, Phelps M, Ward S, Slusser J, Petterson T, De Andrade M. Familial segregation of venous thromboembolism. *J Thromb Haemost*. 2004;2:731–6.
7. Ho WK, Hankey GJ, Quinlan DJ, Eikelboom JW. Risk of recurrent venous thromboembolism in patients with common thrombophilia: a systematic review. *Arch Intern Med*. 2006;166(7):729-36.
8. Marchiori A, Mosena L, Prins MH, Prandoni P. The risk of recurrent venous thromboembolism among heterozygous carriers of factor V Leiden or prothrombin G20210A mutation. A systematic review of prospective studies. *Haematologica*. 2007;92(8):1107-14.
9. van Hylckama Vlieg A, Flinterman LE, Bare LA, Cannegieter SC, Reitsma PH, Arellano AR, et al. Genetic variations associated with recurrent venous thrombosis. *Circ Cardiovasc Genet*. 2014;7(6):806-13.
10. Bruzelius M, Ljungqvist M, Bottai M, Bergendal A, Strawbridge RJ, Holmström M, et al. F11 is associated with recurrent VTE in women. A prospective cohort study. *Thromb Haemost*. 2016;115(2):406-14.
11. Kyrle PA, Minar E, Bialonczyk C, Hirschl M, Weltermann A, Eichinger S. The risk of recurrent venous thromboembolism in men and women. *N Engl J Med.* 2004;350:2558–63.
12. Christiansen SC, Cannegieter SC, Koster T, Vandenbroucke JP, Rosendaal FR. Thrombophilia, clinical factors, and recurrent venous thrombotic events. *JAMA*. 2005;293:2352-61.
13. McRae S, Tran H, Schulman S, Ginsberg J, Kearon C. Effect of patient's sex on risk of recurrent venous thromboembolism: a meta-analysis. *Lancet.* 2006;368:371-8.

14. Rodger MA, Kahn SR, Wells PS, Anderson DA, Chagnon I, Le Gal G, Solymoss S, Crowther M, Perrier A, White R, Vickars L, Ramsay T, Betancourt MT, Kovacs MJ. Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy. *CMAJ.* 2008;179:417-426.
15. Douketis J, Tosetto A, Marcucci M, Baglin T, Cosmi B, Cushman M, Kyrle P, Poli D, Tait RC, Iorio A. Risk of recurrence after venous thromboembolism in men and women: patient level meta-analysis. *BMJ.* 2011;342:d813.
16. Roach RE, Lijfering WM, Rosendaal FR, Cannegieter SC, le Cessie S. Sex difference in risk of second but not of first venous thrombosis: paradox explained. *Circulation.* 2014;129:51-6.
17. Charchar FJ, Bloomer LD, Barnes TA, Cowley MJ, Nelson CP, Wang Y, Denniff M, Debiec R, Christofidou P, Nankervis S, Dominiczak AF, Bani-Mustafa A, Balmforth AJ, Hall AS, Erdmann J, Cambien F, Deloukas P, Hengstenberg C, Packard C, Schunkert H, et al. Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. *Lancet.* 2012;379:915-22.
18. Hasstedt SJ, Bezemer ID, Callas PW, Vossen CY, Trotman W, Hebbel RP, et al. Cell adhesion molecule 1: a novel risk factor for venous thrombosis. *Blood*. 2009;114:3084–91.
19. Tatsumi K, Taatjes DJ, Wadsworth MP, Bouchard BA, Bovill EG. Cell adhesion molecule 1 (CADM1) is ubiquitously present in the endothelium and smooth muscle cells of the human macro- and micro-vasculature. *Histochem Cell Biol*. 2012;138:815–20.
20. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet*. 2018;27(R2):R195-R208.
21. Jordan FL, Nandorff A. The familial tendency in thrombo-embolic disease. *Acta Med Scand.* 1956;156(4):267-75.
22. Collins FS, Morgan M, Patrinos A. The Human Genome Project: Lessons from Large-Scale Biology. *Science*. 2003;300(5617):286-90.
23. Lander ES. The new genomics: global views of biology. *Science*. 1996;274(5287):536–9.
24. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet*. 2001;17(9):502–10.
25. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng LC, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet*. 2015;96(4):532-42.
26. Hinds DA, Buil A, Ziemek D, Martinez-Perez A, Malik R, Folkersen L, Germain M, Mälarstig A, Brown A, Soria JM, Dichgans M, Bing N, Franco-Cereceda A, Souto JC, Dermitzakis ET, Hamsten A, Worrall BB, Tung JY; METASTROKE Consortium, INVENT Consortium, Sabater-Lleal M. Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. *Hum Mol Genet*. 2016;25(9):1867-74.

27. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci*. 2009;106:9362–9367.
28. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-53.
29. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*. 2001;69(1):124-37.
30. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A*. 2011;108:18026–31.
31. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*. 2014;506(7487):185-90.
32. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526:82–90.
33. Lotta LA, Tuana G, Yu J, Martinelli I, Wang M, Yu F, Passamonti SM, Pappalardo E, Valsecchi C, Scherer SE, Hale W 4th, Muzny DM, Randi G, Rosendaal FR, Gibbs RA, Peyvandi F. Next-generation sequencing study finds an excess of rare, coding single-nucleotide variants of ADAMTS13 in patients with deep vein thrombosis. *J Thromb Haemost*. 2013;11:1228-39.
34. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. *Nature*. 2017;542(7640):186-190.
35. Zhang Y, Qi G, Park JH, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat Genet.* 2018;50(9):1318-1326.
36. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet*. 2014;94(4):559-73.
37. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 2014;42(D1):D1001–D1006.
38. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, ReproGen Consortium. Schizophrenia Working Group of the Psychiatric Genomics Consortium. RACI Consortium Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 2015;47(11):1228–35.
39. Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, et al. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature*. 1994;369(6475):64-7.

40. Poort SR, Rosendaal FR, Reitsma PH, Bertina RM. A common genetic variation in the 3'-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis. *Blood*. 1996;88(10):3698-703.
41. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42:565–569.
42. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, Frayling TM, Hirschhorn J, Yang J, Visscher PM1; GIANT Consortium. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet*. 2018;27(20):3641-3649.
43. Anderson FA Jr, Wheeler HB, Goldberg RJ, Hosmer DW, Patwardhan NA, Jovanovic B, et al. A population-based perspective of the hospital incidence and case-fatality rates of deep vein thrombosis and pulmonary embolism. The Worcester DVT Study. *Arch Intern Med*. 1991;151(5):933-8.
44. Naess IA, Christiansen SC, Romundstad P, Cannegieter SC, Rosendaal FR, Hammerstrøm J. Incidence and mortality of venous thrombosis: a population-based study. *J Thromb Haemost*. 2007;5(4):692-9.
45. Mägi R, Horikoshi M, Sofer T, Mahajan A, Kitajima H, Franceschini N, McCarthy MI; COGENT-Kidney Consortium, T2D-GENES Consortium, Morris AP. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum Mol Genet*. 2017; 26(18): 3639–3650.
46. Kenny EE, Kim M, Gusev A, Lowe JK, Salit J, Smith JG, Kovvali S, Kang HM, Newton-Cheh C, Daly MJ, Stoffel M, Altshuler DM, Friedman JM, Eskin E, Breslow JL, Pe'er I. Increased power of mixed models facilitates association mapping of 10 loci for metabolic traits in an isolated population. *Hum Mol Genet*. 2011;20(4):827-39.
47. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N, Price AL. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*. 2015;47(3):284-90.
48. Capen EC, Clapp RV, Campbell WM. Competitive bidding in high-risk situations. *J Petrol Technol*. 1971;23:641–653.
49. Göring HHH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet*. 2001;69:1357–1369.
50. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet*. 2004;36(5):512-7.
51. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904-9.

52. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55(4):997-1004.
53. Babron MC, de Tayrac M, Rutledge DN, Zeggini E, Génin E. Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PLoS One*. 2012;7(10):e46519.
54. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012;44(3):243-6.
55. Rees DC, Cox M, Clegg JB. World distribution of factor V Leiden. *Lancet*. 1995;346(8983):1133-4.
56. Folsom AR, Tang W, Weng LC, Roetker NS, Cushman M, Basu S, Pankow JS. Replication of a genetic risk score for venous thromboembolism in whites but not in African Americans. *J Thromb Haemost*. 2016;14(1):83-8.
57. Wassel CL, Rasmussen-Torvik LJ, Callas PW, Denenberg JO, Durda JP, Reiner AP, Smith NL, Allison MA, Rosendaal FR, Criqui MH, Cushman M. A genetic risk score comprising known venous thromboembolism loci is associated with chronic venous disease in a multi-ethnic cohort. *Thromb Res*. 2015;136(5):966-73.
58. Uitte de Willige S, Pyle ME, Vos HL, de Visser MC, Lally C, Dowling NF, Hooper WC, Bertina RM, Austin H. Fibrinogen gamma gene 3'-end polymorphisms and risk of venous thromboembolism in the African-American and Caucasian population. *Thromb Haemost*. 2009;101(6):1078-84.
59. Tang L, Lu X, Yu JM, Wang QY, Yang R, Guo T, Mei H, Hu Y. PROC c.574_576del polymorphism: A common genetic risk factor for venous thrombosis in the Chinese population. *J Thromb Haemost*. 2012;10(10):2019–2026.
60. Tang L, Wang HF, Lu X, Jian XR, Jin B, Zheng H, Li YQ, Wang QY, Wu TC, Guo H, Liu H, Guo T, Yu JM, Yang R, Yang Y, Hu Y. Common genetic risk factors for venous thrombosis in the Chinese population. *Am J Hum Genet*. 2013;92(2):177-87.
61. Hernandez W, Gamazon ER, Smithberger E, O'Brien TJ, Harralson AF, Tuck M, Barbour A, Kittles RA, Cavallari LH, Perera MA. Novel genetic predictors of venous thromboembolism risk in African Americans. *Blood*. 2016;127(15):1923-9.
62. Heit JA, Armasu SM, McCauley BM, Kullo IJ, Sicotte H, Pathak J, Chute CG, Gottesman O, Bottinger EP, Denny JC, Roden DM, Li R, Ritchie MD, de Andrade M. Identification of unique venous thromboembolism-susceptibility variants in African-Americans. *Thromb Haemost*. 2017;117(4):758-768.
63. Peterson AC, Rienzo AD, Lehesjokl A-E, Chapelle Adl, Slatkin M, et al. The distribution of linkage disequilibrium over anonymous genome regions. *Hum Mol Genet*. 1995;4:887–894.
64. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*. 2002;3(4):299-309.

65. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. Linkage disequilibrium in the human genome. *Nature*. 2001;411(6834):199-204.
66. Bezemer ID, Bare LA, Doggen CJ, Arellano AR, Tong C, Rowland CM, Catanese J, Young BA, Reitsma PH, Devlin JJ, Rosendaal FR. Gene variants associated with deep vein thrombosis. *JAMA*. 2008;299(11):1306-14.
67. Li Y, Bezemer ID, Rowland CM, Tong CH, Arellano AR, Catanese JJ, Devlin JJ, Reitsma PH, Bare LA, Rosendaal FR. Genetic variants associated with deep vein thrombosis: the F11 locus. *J Thromb Haemost*. 2009;7(11):1802-8.
68. Uitte de Willige S, de Visser MC, Houwing-Duistermaat JJ, Rosendaal FR, Vos HL, Bertina RM. Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen gamma' levels. *Blood*. 2005;106(13):4176-83.
69. Lotta LA, Wang M, Yu J, Martinelli I, Yu F, Passamonti SM, Consonni D, Pappalardo E, Menegatti M, Scherer SE, Lewis LL, Akbar H, Wu Y, Bainbridge MN, Muzny DM, Mannucci PM, Gibbs RA, Peyvandi F. Identification of genetic risk variants for deep vein thrombosis by multiplexed next-generation sequencing of 186 hemostatic/pro-inflammatory genes. *BMC Med Genomics*. 2012;5:7.
70. Germain M, Saut N, Oudot-Mellakh T, Letenneur L, Dupuy AM, Bertrand M, Alessi MC, Lambert JC, Zelenika D, Emmerich J, Tiret L, Cambien F, Lathrop M, Amouyel P, Morange PE, Trégouët DA. Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. *PLoS One*. 2012;7(6):e38538.
71. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A. Linkage disequilibrium patterns of the human genome across populations. *Hum Mol Genet*. 2003;12(7):771-6.
72. Zaitlen N, Paşaniuc B, Gur T, Ziv E, Halperin E. Leveraging genetic variability across populations for the identification of causal variants. *Am J Hum Genet*. 2010;86(1):23-33.
73. Smith NL, Hindorff LA, Heckbert SR, Lemaitre RN, Marciante KD, Rice K, Lumley T, Bis JC, Wiggins KL, Rosendaal FR, Psaty BM. Association of genetic variations with nonfatal venous thrombosis in postmenopausal women. *JAMA*. 2007;297(5):489-98.
74. Heit JA, Armasu SM, Asmann YW, Cunningham JM, Matsumoto ME, Petterson TM, De Andrade M. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *J Thromb Haemost*. 2012;10(8):1521-31.
75. Nicolaes GA, Dahlbäck B. Factor V and thrombotic disease: description of a janus-faced protein. *Arterioscler Thromb Vasc Biol*. 2002;22(4):530-8.
76. Pollak ES, Lam HS, Russell JE. The G20210A mutation does not affect the stability of prothrombin mRNA in vivo. *Blood*. 2002;100(1):359-62.

77. Gallinaro L, Cattini MG, Sztukowska M, Padrini R, Sartorello F, Pontara E, Bertomoro A, Daidone V, Pagnan A, Casonato A. A shorter von Willebrand factor survival in O blood group subjects explains how ABO determinants influence plasma von Willebrand factor. *Blood*. 2008;111(7):3540-5.
78. Sabater-Lleal M, Martinez-Perez A, Buil A, Folkersen L, Souto JC, Bruzelius M, Borrell M, Odeberg J, Silveira A, Eriksson P, Almasy L, Hamsten A and Soria JM. A genome-wide association study identifies KNG1 as a genetic determinant of plasma factor XI Level and activated partial thromboplastin time. *Arterioscler Thromb Vasc Biol*. 2012;32:2008-16.
79. Sennblad B, Basu S, Mazur J, Suchon P, Martinez-Perez A, van Hylckama Vlieg A, Truong V, Li Y, Gadin JR, Tang W, Grossman V, de Haan HG, Handin N, Silveira A, Souto JC, Franco-Cereceda A, Morange PE, Gagnon F, Soria JM, Eriksson P, Hamsten A, Maegdefessel L, Rosendaal FR, Wild P, Folsom AR, Tregouet DA and Sabater-Lleal M. Genome-wide association study with additional genetic and post-transcriptional analyses reveals novel regulators of plasma factor XI levels. *Hum Mol Genet*. 2017.
80. Rohmann JL, de Haan HG, Algra A, Vossen CY, Rosendaal FR, Siegerink B. Genetic determinants of activity and antigen levels of contact system factors. *J Thromb Haemost*. 2019;17(1):157-168.
81. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648-60.
82. Brodie A, Azaria JR, Ofran Y: How far from the SNP may the causative genes be? *Nucleic Acids Res*. 2016;44:6046–54.
83. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337:1190–1195.
84. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
85. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet*. 2016;99(6):1245-1260.
86. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48:245–52.
87. Hannon E, Weedon M, Bray N, O'Donovan M, Mill J. Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci. *Am J Hum Genet*. 2017;100(6):954-959.

88. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*. 2010;28(10):1045-8.
89. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, Ren B. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* 2016;17:2042–59.
90. Bodin L, Verstuyft C, Tregouet DA, Robert A, Dubert L, Funck-Brentano C, Jaillon P, Beaune P, Laurent-Puig P, Becquemont L, Loriot MA. Cytochrome P450 2C9 (CYP2C9) and vitamin K epoxide reductase (VKORC1) genotypes as determinants of acenocoumarol sensitivity. *Blood* 2005;106:135-140.
91. Wadelius M, Chen LY, Lindh JD, Eriksson N, Ghori MJ, Bumpstead S, Holm L, McGinnis R, Rane A, Deloukas P. The largest prospective warfarin-treated cohort supports genetic forecasting. *Blood*. 2009;113:784-792.
92. Verhoef TI, Ragia G, de Boer A, Barallon R, Kolovou G, Kolovou V, Konstantinides S, Le Cessie S, Maltezos E, van der Meer FJ, Redekop WK, Remkes M, Rosendaal FR, van Schie RM, Tavridou A, Tziakas D, Wadelius M, Manolopoulos VG, Maitland-van der Zee AH; EU-PACT Group. A randomized trial of genotype-guided dosing of acenocoumarol and phenprocoumon. *N Engl J Med*. 2013;369(24):2304-12.
93. Pirmohamed M, Burnside G, Eriksson N, Jorgensen AL, Toh CH, Nicholson T, Kesteven P, Christersson C, Wahlstrom B, Stafberg C, Zhang JE, Leathart JB, Kohnke H, Maitland-van der Zee AH, Williamson PR, Daly AK, Avery P, Kamali F, Wadelius M; EU-PACT Group. A randomized trial of genotype-guided dosing of warfarin. *N Engl J Med.* 2013;369:2294–303.
94. Kimmel SE, French B, Kasner SE, Johnson JA, Anderson JL, Gage BF, Rosenberg YD, Eby CS, Madigan RA, McBane RB, Abdel-Rahman SZ, Stevens SM, Yale S, Mohler ER 3rd, Fang MC, Shah V, Horenstein RB, Limdi NA, Muldowney JA 3rd, Gujral J, *et al.* A pharmacogenetic versus a clinical algorithm for warfarin dosing. *N Engl J Med.* 2013;369:2283–93.
95. Gage BF, Bass AR, Lin H, Woller SC, Stevens SM, Al-Hammadi N, Li J, Rodríguez T Jr, Miller JP, McMillin GA, Pendleton RC, Jaffer AK, King CR, Whipple BD, Porche-Sorbet R, Napoli L, Merritt K, Thompson AM, Hyun G, Anderson JL, Hollomon W, Barrack RL, Nunley RM, Moskowitz G, Dávila-Román V, Eby CS. Effect of Genotype-Guided Warfarin Dosing on Clinical Events and Anticoagulation Control Among Patients Undergoing Hip or Knee Arthroplasty: The GIFT Randomized Clinical Trial. *JAMA*. 2017;318(12):1115-24.
96. van Hylckama Vlieg A, Flinterman LE, Bare LA, Cannegieter SC, Reitsma PH, Arellano AR, Tong CH, Devlin JJ, Rosendaal FR. Genetic variations associated with recurrent venous thrombosis. *Circ Cardiovasc Genet*. 2014;7(6):806-13.

97. Soria JM, Morange PE, Vila J, Souto JC, Moyano M, Trégouët DA, Mateo J, Saut N, Salas E, Elosua R. Multilocus genetic risk scores for venous thromboembolism risk assessment. *J Am Heart Assoc*. 2014;3(5):e001060.
98. Bruzelius M, Bottai M, Sabater-Lleal M, Strawbridge RJ, Bergendal A, Silveira A, Sundström A, Kieler H, Hamsten A, Odeberg J. Predicting venous thrombosis in women using a combination of genetic markers and clinical risk factors. *J Thromb Haemost*. 2015;13(2):219-27.
99. McDaid A, Logette E, Buchillier V, Muriset M, Suchon P, Pache TD, Tanackovic G, Kutalik Z, Michaud J. Risk prediction of developing venous thrombosis in combined oral contraceptive users. *PLoS One*. 2017;12(7):e0182041.
100. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, Duncan L, Perry JR, Patterson N, Robinson EB, ReproGen Consortium. Psychiatric Genomics Consortium. Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 An atlas of genetic correlations across human diseases and traits. *Nat Genet*. 2015;47:1236–41.
101. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*. 2016;48:709–717.
102. Kamstrup PR, Tybjærg-Hansen A, Nordestgaard BG. Genetic evidence that lipoprotein(a) associates with atherosclerotic stenosis rather than venous thrombosis. *Arterioscler Thromb Vasc Biol*. 2012;32(7):1732-41.
103. Kjaergaard AD, Johansen JS, Bojesen SE, Nordestgaard BG. Observationally and Genetically High YKL-40 and Risk of Venous Thromboembolism in the General Population: Cohort and Mendelian Randomization Studies. *Arterioscler Thromb Vasc Biol*. 2016;36(5):1030-6.
104. Lindström S, Germain M, Crous-Bou M, Smith EN, Morange PE, van Hylckama Vlieg A, de Haan HG, Chasman D, Ridker P, Brody J, de Andrade M, Heit JA, Tang W, DeVivo I, Grodstein F, Smith NL, Tregouet D, Kabrhel C; INVENT Consortium. Assessing the causal relationship between obesity and venous thromboembolism through a Mendelian Randomization study. *Hum Genet*. 2017;136(7):897-902.
105. Roetker NS, Armasu SM, Pankow JS, Lutsey PL, Tang W, Rosenberg MA, Palmer TM, MacLehose RF, Heckbert SR, Cushman M, de Andrade M, Folsom AR. Taller height as a risk factor for venous thromboembolism: a Mendelian randomization meta-analysis. *J Thromb Haemost*. 2017;15(7):1334-43.

CHAPTER 9

Nederlandse samenvatting

Dankwoord

Curriculum Vitae

Publicatielijst

NEDERLANDSE SAMENVATTING

Bij veneuze trombose wordt de doorstroming in een veneus bloedvat belemmerd door een bloedstolsel. De aandoening presenteert zich voornamelijk als een diep veneuze trombose in het been of als een longembolie. Per jaar komt het bij 1 tot 2 per 1000 personen voor. Ongeveer 25% van de patiënten met een eerste veneuze trombose krijgt binnen vijf jaar een recidief. Het risico op veneuze trombose is niet voor iedereen hetzelfde. Verschillende klinische en leefstijlfactoren spelen een rol, zoals leeftijd, immobilisatie, hormonale factoren en kanker. Ook genetische factoren dragen bij aan het ontstaan van veneuze trombose. De genetische component van veneuze trombose wordt op basis van onderzoek in families en tweelingen tussen de 50 en 60% geschat. We weten tot nu toe van 17 genen dat bepaalde varianten in deze genen het risico op veneuze trombose beïnvloeden, hiervan is de variant Factor V Leiden in het gen voor stollingsfactor V het meest bekend. De bekende genetische risicofactoren verklaren maar een klein deel van de genetische component van veneuze trombose; de overige genetische factoren zijn nog niet goed in kaart gebracht. Ook weten we nog weinig van de genetische risicofactoren voor een recidief veneuze trombose.

Het doel van het in dit proefschrift beschreven onderzoek was om nieuwe genetische risicofactoren te identificeren voor een eerste en een recidief veneuze trombose. Daarnaast hebben we onderzocht of de bekende genetische risicofactoren kunnen bijdragen aan de risicostratificatie voor veneuze trombose. Ook hebben we beschreven hoe genetische variatie gebruikt kan worden voor het doen van causale uitspraken in observationeel onderzoek.

Identificatie van genetische risicofactoren voor een eerste veneuze trombose

We hebben in hoofdstuk 2 tot en met 5 diverse strategieën gebruikt om genetische varianten te identificeren die geassocieerd zijn met het risico op een eerste veneuze trombose of een recidief. In hoofdstuk 2 bestudeerden we met DNA sequencing de variatie in voornamelijk de coderende delen van 734 genen die betrokken zijn bij hemostase. Meer dan 3500 veelvoorkomende varianten werden onderzocht in ongeveer 900 patiënten met een eerste diep veneuze trombose en 600 controlepersonen. We bevestigden eerder gerapporteerde associaties tussen diep veneuze trombose en variatie in de *F11* locus, *FGA-FGG*, *ABO*, en *F5.* Dit was geen verrassing, want in deze gengebieden liggen bekende risicofactoren voor veneuze trombose. In *F5* en de *F11*

locus vonden we aanwijzingen voor secundaire associatiesignalen, wat suggereert dat deze genen meerdere onafhankelijke risicofactoren voor veneuze trombose dragen. Opmerkelijk genoeg vonden we slechts twee suggestieve assocatiesignalen in genen die nog niet eerder in verband gebracht waren met de pathofysiologie van veneuze trombose. We konden deze associaties echter niet repliceren in data van het INVENT-consortium. In een analyse van meer dan 16000 zeldzamen varianten in 647 genen werd daarnaast geen opeenstapeling van zeldzame varianten gevonden in patiënten met een eerste diep veneuze trombose in vergelijking tot controlepersonen. Het is echter mogelijk dat we associaties van zowel veelvoorkomende als zeldzame varianten met diep veneuze trombose gemist hebben omdat onze studie onvoldoende groot was om varianten met een gering effect op trombose te identificeren.

In plaats van een focus op variatie in kandidaatgenen hadden we in hoofdstuk 3 voor een agnostische aanpak gekozen. Het is voor een recidief trombose namelijk niet duidelijk in welke mate de bekende genetische risicofactoren een rol spelen en of andere genetische varianten ook van belang zijn. We hebben een genoombrede associatiestudie, ook wel een 'GWAS' genoemd, uitgevoerd waarbij we de associatie tussen ongeveer 8 miljoen veelvoorkomende varianten en recidief veneuze trombose hebben bestudeerd. Onze resultaten bevestigden de associatie tussen FV Leiden en het risico op een recidief. Daarnaast hebben we een nieuwe risicolocus gevonden op 18q22.1, welke in de replicatie analyse geassocieerd was met het risico op recidief veneuze trombose met een odds ratio van 1.7 per kopie van het minor allel. Mogelijk beïnvloedt deze intergene locus het risico op recidief door het moduleren van de expressie van genen die dichtbij of juist verder weg op het chromosoom liggen. Er is echter meer onderzoek nodig om het onderliggende moleculaire mechanisme te ontrafelen. We vonden beperkt bewijs voor een aantal eerder gerapporteerde associaties tussen varianten en recidiefrisico, wat het belang van replicatie in genetische associatiestudies nogmaals benadrukt.

Een eerste onderzoek naar het effect van variatie in het Y-chromosoom op het risico op een eerste en recidief veneuze trombose beschreven we in hoofdstuk 4. Eerdere studies hebben aangetoond dat mannen een intrinsiek hoger risico op veneuze trombose hebben dan vrouwen. Daarom hadden we de hypothese dat variatie in het Y-chromosoom het risico op veneuze trombose in bepaalde subgroepen van mannen zou kunnen verhogen. We onderzochten hiervoor de associatie tussen 13 veelvoorkomende haplogroepen in het Y-chromosoom en het risico op een eerste en recidief veneuze

trombose in ruim 3700 mannen van Europese afkomst. Geen van de haplogroepen was geassocieerd met het risico op een eerste veneuze trombose ten opzichte van de meest voorkomende haplogroep R1b. We vonden met name geen aanwijzingen voor een associatie tussen haplogroep I en het risico op veneuze trombose, terwijl onze analyse wel voldoende statistische power had om een vergelijkbaar effect te vinden zoals dat eerder gerapporteerd is voor haplogroep I en het risico op coronaire hartziekte. Daarnaast vonden we aanwijzingen dat dragers van haplogroep R1a een verlaagd risico op recidief veneuze trombose hebben in vergelijking tot dragers van R1b. Deze bevinding kan echter niet het verschil in tromboserisico tussen mannen en vrouwen verklaren aangezien we voor dragers van haplogroep R1a een hoger recidiefrisico vonden dan voor vrouwen.

In hoofdstuk 5 hadden we een kandidaatgen-aanpak: we bestudeerden de associatie tussen veelvoorkomende variatie in het *CADM1* gen en het risico op een eerste veneuze trombose. Een eerdere studie in een familie met proteïne C deficiëntie had laten zien dat *CADM1*, wat codeert voor een celadhesiemolecuul betrokken bij celmigratie in het endotheel, een risicogen voor veneuze trombose is. In onze studie wilden we bekijken of er ook in de algemene bevolking een gecombineerd effect van variatie in *CADM1* en proteïne C op het risico op veneuze trombose bestaat. We bestudeerden hiervoor de associatie tussen ruim 300 varianten in *CADM1* en het risico op veneuze trombose in 962 patiënten met een afwijking in het proteïne C systeem en 4004 controlepersonen. We zagen voor zes varianten een groot gecombineerd effect op het risico op veneuze trombose. Voor een van deze zes varianten vonden we ook aanwijzingen voor een associatie met het risico op veneuze trombose in de gehele studiepopulatie van 3496 trombosepatiënten en 4004 controlepersonen. Voorzichtigheid is geboden bij de interpretatie van onze resultaten vanwege het grote aantal statistische toetsen en het kleine aantal patiënten met een afwijking in het proteïne C systeem.

Toepassingen van genetische risicofactoren voor veneuze trombose

In hoofdstuk 6 en 7 hebben we twee van de hoofdtoepassingen van genetische risicofactoren in wetenschappelijk onderzoek beschreven, namelijk risicostratificatie en Mendeliaanse randomisatie. Op basis van een set van 31 eerder gerapporteerde risicovarianten voor veneuze trombose hebben we genetische risicoscores gebouwd. Zoals beschreven in hoofdstuk 6 hebben we het onderscheidend vermogen van deze scores vergeleken met een predictiemodel gebaseerd op klinische risicofactoren en met

een gecombineerd model. We lieten zien dat het onderscheidend vermogen van een klinisch predictiemodel voor veneuze trombose in de algemene bevolking significant verbeterde door het toevoegen van een genetische risicoscore bestaande uit vijf varianten (FV Leiden, PT G20210A, *ABO* non-O, *FGG*-rs2066865, en *F11*-rs2036914). Om de klinische toepasbaarheid van de risicoscores nader te onderzoeken, hebben we ook het onderscheidend vermogen van de verschillende risicoscores bekeken in relevante risicogroepen. Met uitzondering van kankerpatiënten was de voorspellende waarde van de genetische score hetzelfde in de risicogroepen als in de gehele studiepopulatie. Replicatie van onze bevindingen in een onafhankelijk studie laten de robuustheid van de onze genetische risicoscore zien, al kan het zo zijn dat de score minder goed functioneert in populaties met een andere etnische achtergrond.

In hoofdstuk 7 hebben we beschreven hoe genetische variatie gebruikt kan worden voor het doen van causale uitspraken in observationeel onderzoek. Bij Mendeliaanse randomisatie wordt genetische variatie gebruikt als een instrument voor een blootstelling van interesse. In dit educatieve hoofdstuk hebben we uitgelegd dat indien geen van de assumpties voor Mendeliaans randomisatie geschonden wordt, dat dan een genetisch instrument gebruikt kan worden om het causale effect van de blootstelling op de uitkomst te schatten. Door het gebruik van de genetische variatie zijn confounding en reverse causation geminimaliseerd. Alhoewel niet alle assumpties te testen zijn en er ook een grote studiepopulatie nodig is, worden Mendeliaanse randomisatie analyses steeds vaker succesvol toegepast in observationeel onderzoek, met name als gerandomiseerd onderzoek niet mogelijk is. In dit hoofdstuk werden voornamelijk de concepten van Mendeliaanse randomisatie besproken en gingen we niet in op de verschillende analytische modellen die recent ontwikkeld zijn, zoals bijvoorbeeld methodes waarbij rekening gehouden wordt met pleiotropie.

DANKWOORD

Dit proefschrift zou er niet zijn geweest zonder de inzet, steun en hulp van anderen. Ik kijk dan ook dankbaar terug op de diverse samenwerkingen!

Een aantal mensen wil ik nog extra in het zonnetje zetten.

Allereerst wil ik alle deelnemers aan de verschillende studies roemen voor hun bereidheid om hun medische gegevens en lichaamsmateriaal beschikbaar te stellen voor wetenschappelijk onderzoek.

Mijn (oud-) collega's wil ik bedanken voor de leuke en leerzame tijd. De afdeling Klinische Epidemiologie is een hele goede leerschool geweest. En gelachen hebben we ook: met jullie hulp wist ik altijd mijn credo 'elke dag een tegenslag' te pareren.

Daarnaast wil ik graag Annelies Hoenderdos, Lejla Mahic en Petra Noordijk bedanken voor het uitzoeken, meten en verdunnen van alle DNA-monsters en voor het runnen van de diverse Taqmanassays. Ook Ingeborg de Jonge, die database na database met onderzoeksgegevens voor mij heeft gemaakt, mag in dit rijtje niet ontbreken. Bedankt!

Tot slot, lieve vrienden en familie: jullie waren niet zozeer inhoudelijk betrokken, maar toch minstens zo belangrijk bij de totstandkoming van dit proefschrift. Bedankt dat ik altijd en eindeloos tegen jullie mocht doorratelen over genen, varianten en veneuze trombose. En minstens zoveel dank voor alle broodnodige afleiding.

CURRICULUM VITAE

Hugoline Georgette de Haan werd geboren op 8 juli 1987 in De Bilt. In 2005 behaalde ze haar gymnasiumdiploma aan het Utrechts Stedelijk Gymnasium. Hierna vervolgde ze haar opleiding aan de Universiteit Leiden: eerst de propedeuse Geneeskunde, vervolgens de bachelor en master Biomedische Wetenschappen. Tijdens haar opleiding studeerde ze een semester aan het Karolinska Institutet in Stockholm en deed ze een onderzoeksstage aan Trinity College in Dublin. Ook liep ze stage bij LURIS, het technology transfer office van Universiteit Leiden en het Leids Universitair Medisch Centrum. Na het behalen van haar masterdiploma begon ze haar promotieonderzoek onder begeleiding van Prof. dr F.R. Rosendaal en dr C.Y. Vossen aan de afdeling Klinische Epidemiologie van het Leids Universitair Medisch Centrum. Na een jaar werd dr C.Y. Vossen opgevolgd door dr A. van Hylckama Vlieg. Tijdens het promotietraject volgde ze verschillende (genetisch) epidemiologische cursussen voor de opleiding tot Epidemioloog B. Verder ontving ze een Virchow scholarship om drie maanden onderzoek te doen aan het Center for Stroke Research van het Charité in Berlijn. De resultaten van het promotieonderzoek staan beschreven in dit proefschrift. Tevens zijn de resultaten gepresenteerd op verschillende nationale en internationale congressen.

PUBLICATIELIJST

Publicaties in dit proefschrift

de Haan HG, Bezemer ID, Doggen CJ, Le Cessie S, Reitsma PH, Arellano AR, Tong CH, Devlin JJ, Bare LA, Rosendaal FR, Vossen CY. Multiple SNP testing improves risk prediction of first venous thrombosis. *Blood*. 2012;120(3):656-63.

de Haan HG, Bezemer ID, Vossen CY, van Hylckama Vlieg A, Böehringer S, Hasstedt SJ, Levy S, Rosendaal FR, Bovill EG. Genetic variants in Cell Adhesion Molecule 1 (CADM1): a validation study of a novel endothelial cell venous thrombosis risk factor. *Thromb Res*. 2014;134(6):1186-92.

de Haan HG, Siegerink B, van Hylckama Vlieg A. Mendeliaanse randomisatie. *Ned Tijdschr Geneeskd*. 2014;158:A7547.

de Haan HG, van Hylckama Vlieg A, van der Gaag KJ, de Knijff P, Rosendaal FR. Male-specific risk of first and recurrent venous thrombosis: a phylogenetic analysis of the Y chromosome. *J Thromb Haemost*. 2016;14(10):1971-7.

de Haan HG, van Hylckama Vlieg A, Germain M, Baglin TP, Deleuze JF, Trégouët DA, Rosendaal FR. Genome-wide association study identifies a novel genetic risk factor for recurrent venous thrombosis. *Circ Genom Precis Med*. 2018;11:e001827.

de Haan HG, van Hylckama Vlieg A, Lotta LA, Gorski MM, Bucciarelli P, Martinelli I, Baglin TP, Peyvandi F, Rosendaal FR; INVENT consortium. Targeted sequencing to identify novel genetic risk factors for deep vein thrombosis: a study of 734 genes. *J Thromb Haemost*. 2018;16(12):2432-41.

Gerelateerde publicaties

Germain M, Chasman DI, de Haan HG, Tang W, Lindström S, Weng LC, de Andrade M, de Visser MC, Wiggins KL, Suchon P, Saut N, Smadja DM, Le Gal G, van Hylckama Vlieg A, Di Narzo A, Hao K, Nelson CP, Rocanin-Arjo A, Folkersen L, Monajemi R, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet*. 2015;96(4):532-42.

Gorski MM, Lotta LA, Pappalardo E, de Haan HG, Passamonti SM, van Hylckama Vlieg A, Martinelli I, Peyvandi F. Single nucleotide variant rs2232710 in the protein Z-dependent protease inhibitor (ZPI, SERPINA10) gene is not associated with deep vein thrombosis. *PLoS One*. 2016;11(3):e0151347.

Pagliari MT, Lotta LA, de Haan HG, Valsecchi C, Casoli G, Pontiggia S, Martinelli I, Passamonti SM, Rosendaal FR, Peyvandi F. Next-generation sequencing and in vitro expression study of ADAMTS13 single nucleotide variants in deep vein thrombosis. *PLoS One*. 2016;11(11):e0165665.

Sennblad B, Basu S, Mazur J, Suchon P, Martinez-Perez A, van Hylckama Vlieg A, Truong V, Li Y, Gådin JR, Tang W, Grossman V, de Haan HG, Handin N, Silveira A, Souto JC, Franco-Cereceda A, Morange PE, Gagnon F, Soria JM, Eriksson P, et al. Genome-wide association study with additional genetic and post-transcriptional analyses reveals novel regulators of plasma factor XI levels. *Hum Mol Genet*. 2017;26(3):637-49.

Vossen CY, van Hylckama Vlieg A, Teruel-Montoya R, Salloum-Asfar S, de Haan HG, Corral J, Reitsma P, Koeleman BPC, Mart ínez C. Identification of coagulation gene 3'UTR variants that are potentially regulated by microRNAs. *Br J Haematol*. 2017;177(5):782-90.

Lindström S, Germain M, Crous-Bou M, Smith EN, Morange PE, van Hylckama Vlieg A, de Haan HG, Chasman D, Ridker P, Brody J, de Andrade M, Heit JA, Tang W, DeVivo I, Grodstein F, Smith NL, Trégouët DA, Kabrhel C; INVENT Consortium. Assessing the causal relationship between obesity and venous thromboembolism through a Mendelian Randomization study. *Hum Genet*. 2017;136(7):897-902.

Gorski MM, de Haan HG, Mancini I, Lotta LA, Bucciarelli P, Passamonti SM, Cairo A, Pappalardo E, van Hylckama Vlieg A, Martinelli I, Rosendaal FR, Peyvandi F. Next-generation DNA sequencing to identify novel genetic risk factors for cerebral vein thrombosis. *Thromb Res*. 2018;169:76-81.

Sabater-Lleal M, Huffman JE, de Vries PS, Marten J, Mastrangelo MA, Song C, Pankratz N, Ward-Caviness CK, Yanek LR, Trompet S, Delgado GE, Guo X, Bartz TM, Martinez-Perez A, Germain M, de Haan HG, Ozel AB, Polasek O, Smith AV, Eicher JD, et al. Genome-wide association transethnic meta-analyses identifies novel associations

regulating coagulation factor VIII and von willebrand factor plasma levels. *Circulation*. 2019;139(5):620-35.

Rohmann JL, de Haan HG, Algra A, Vossen CY, Rosendaal FR, Siegerink B. Genetic determinants of activity and antigen levels of contact system factors. *J Thromb Haemost*. 2019;17(1):157-68.

de Vries PS, Sabater-Lleal M, Huffman JE, Marten J, Song C, Pankratz N, Bartz TM, de Haan HG, Delgado GE, Eicher JD, Martinez-Perez A, Ward-Caviness CK, Brody JA, Chen MH, de Maat MPM, Frånberg M, Gill D, Kleber ME, Rivadeneira F, Soria JM, et al. A genome-wide association study identifies new loci for factor VII and implicates factor VII in ischemic stroke etiology. *Blood*. 2019;133(9):967-77.

Overige publicaties

Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, Nutile T, Jackson AU, Schurmann C, Smith AV, Zhang W, Okada Y, Stančáková A, Faul JD, Zhao W, Bartz TM, Concas MP, Franceschini N, Enroth S, Vitart V, Trompet S, et al. Directional dominance on stature and cognition in diverse human populations. *Nature*. 2015;523(7561):459-62.

Kola V, de Haan HG, Siegerink B. Clinical and laboratory predictors of deep vein thrombosis after acute stroke; does D-dimer really improve predictive power? *Thromb Res*. 2016;146:131-2.

Horikoshi M, Beaumont RN, Day FR, Warrington NM, Kooijman MN, Fernandez-Tajes J, Feenstra B, van Zuydam NR, Gaulton KJ, Grarup N, Bradfield JP, Strachan DP, Li-Gao R, Ahluwalia TS, Kreiner E, Rueedi R, Lyytikäinen LP, Cousminer DL, Wu Y, Thiering E, et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature*. 2016;538(7624):248-52.

Barban N, Jansen R, de Vlaming R, Vaez A, Mandemakers JJ, Tropf FC, Shen X, Wilson JF, Chasman DI, Nolte IM, Tragante V, van der Laan SW, Perry JR, Kong A; BIOS Consortium, Ahluwalia TS, Albrecht E, Yerges-Armstrong L, Atzmon G, Auro K, et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet*. 2016;48(12):1462-72.

van den Berg ME, Warren HR, Cabrera CP, Verweij N, Mifsud B, Haessler J, Bihlmeyer NA, Fu YP, Weiss S, Lin HJ, Grarup N, Li-Gao R, Pistis G, Shah N, Brody JA, Müller-Nurasyid M, Lin H, Mei H, Smith AV, Lyytikäinen LP, et al. Discovery of novel heart rate-associated loci using the Exome Chip. *Hum Mol Genet*. 2017;26(12):2346-63.

Mahajan A, Wessel J, Willems SM, Zhao W, Robertson NR, Chu AY, Gan W, Kitajima H, Taliun D, Rayner NW, Guo X, Lu Y, Li M, Jensen RA, Hu Y, Huo S, Lohman KK, Zhang W, Cook JP, Prins BP, et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet*. 2018;50(4):559-71.

Opstal-van Winden AWJ, de Haan HG, Hauptmann M, Schmidt MK, Broeks A, Russell NS, Janus CPM, Krol ADG, van der Baan FH, De Bruin ML, van Eggermond AM, Dennis J, Anton-Culver H, Haiman CA, Sawyer EJ, Cox A, Devilee P, Hooning MJ, Peto J, Couch FJ, et al. Genetic susceptibility to radiation-induced breast cancer after Hodgkin lymphoma. *Blood*. 2019;133(10):1130-9.

Kraja AT, Liu C, Fetterman JL, Graff M, Have CT, Gu C, Yanek LR, Feitosa MF, Arking DE, Chasman DI, Young K, Ligthart S, Hill WD, Weiss S, Luan J, Giulianini F, Li-Gao R, Hartwig FP, Lin SJ, Wang L, et al. Associations of mitochondrial and nuclear mitochondrial variants and genes with seven metabolic traits. *Am J Hum Genet*. 2019;104(1):112-38.

Warrington NM, Beaumont RN, Horikoshi M, Day FR, Helgeland Ø, Laurin C, Bacelis J, Peng S, Hao K, Feenstra B, Wood AR, Mahajan A, Tyrrell J, Robertson NR, Rayner NW, Qiao Z, Moen GH, Vaudel M, Marsit CJ, Chen J, et al. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat Genet*. 2019;51(5):804-14.

Clark DW, Okada Y, Moore KHS, Mason D, Pirastu N, Gandin I, Mattsson H, Barnes CLK, Lin K, Zhao JH, Deelen P, Rohde R, Schurmann C, Guo X, Giulianini F, Zhang W, Medina-Gomez C, Karlsson R, Bao Y, Bartz TM, et al. Associations of autozygosity with a broad range of human phenotypes. *Nat Commun*. 2019;10(1):4957.