Journal of Thrombosis and Haemostasis, **16**: 2432–2441 DOI: 10.1111/jth.14279

ORIGINAL ARTICLE

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

H. G. DE HAAN,* A. VAN HYLCKAMA VLIEG,* L. A. LOTTA,† M. M. GORSKI,† P. BUCCIARELLI,† I. MARTINELLI,† T. P. BAGLIN,‡ F. PEYVANDI,† and F. R. ROSENDAAL,* FOR THE INVENT CONSORTIUM[1],

Department of Clinical Epidemiology, Leiden University Medical Center, Leiden,the Netherlands; †Angelo Bianchi Bonomi Hemophilia and Thrombosis Center, Fondazione IRCCS Ca' Granda – Ospedale Maggiore Policlinico, Università degli Studi di Milano and Fondazione Luigi Villa, Milan, Italy; and ‡Cambridge Haemophilia and Thrombophilia Centre, Addenbrooke's Hospital, Cambridge University Hospitals National Health Service Foundation Trust, Cambridge, UK

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

Essentials

- Deep vein thrombosis (DVT) has a large unknown genetic component.
- We sequenced coding areas of 734 hemostasis-related genes in 899 DVT patients and 599 controls.
- Variants in *F5*, *FGA-FGG*, *CYP4V2-KLKB1-F11*, and *ABO* were associated with DVT risk.
- Associations in *KLKB1* and *F5* suggest a more complex genetic architecture than previously thought.

Summary. *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:* To identify novel genetic determinants by using targeted DNA sequencing. *Patients/Methods:* We included 899 DVT patients and 599 controls from three 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) 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 use of the false discovery rate (FDR). *Results:* Sixty-two of 3617 common variants were associated with DVT risk (FDR < 0.10). Most of these mapped to *F5*, *ABO*, *FGA–FGG*, and *CYP4V2–KLKB1–F11*. The lead variant at *F5* was rs6672595 (odds ratio [OR] 1.58, 95% confidence interval [CI] 1.29–1.92), in moderate linkage with the 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 were not replicated in the meta-analysis data from the International Network against Thrombosis (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.

Keywords: deep vein thrombosis; DNA sequencing; genetics; risk factors; single-nucleotide polymorphisms.

Correspondence: Frits R. Rosendaal, Department of Clinical Epidemiology, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, the Netherlands
Tel.: +3171 526 4037
Email: f.r.rosendaal@lumc.nl

[1]See Appendix for full list of investigators.

Received: 25 March 2018
Manuscript handled by: J. Douketis
Final decision: J. Douketis, 9 August 2018

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 a 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 hundreds of (mainly rare) mutations have since been reported [10]. Two recent meta-analyses of genome-wide association studies (GWASs), each including > 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.

Although VT has a strong genetic basis, with heritability estimates of 50–60% [14–16], the established genetic risk factors explain only 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 hemostatis-related genes with 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. In this respect, 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 by using meta-analysis data from the International Network against Thrombosis (INVENT) consortium [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 2139 consecutive patients with a first DVT at the Angelo Bianchi Bonomi Hemophilia and Thrombosis Center in Milan (Italy) between 1995 and 2010 [19]. Controls were non-consanguineous relatives, partners, or friends who accompanied patients to center visits. In MEGA, 4956 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 control subjects. Additional controls were recruited from the general population by the use of random digit dialing. Patients and controls were invited to provide a blood sample up to 2002, after which we switched, for logistical reasons, to buccal swabs. The THE-VTE study 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 (UK) between 2003 and 2008 [21]. Again, partners of eligible patients were invited to participate as controls.

From each study, we included patients and controls on the basis of 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 THE-VTE study ($N = 241$), 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 shown in Table S1.

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 THE-VTE study were approved by the Medical Ethics Committee of the Leiden University Medical Center. The THE-VTE study was also approved by the National Health Service 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, and sequenced the coding regions plus 10 bp flanking the exons to cover the splice junctions. For a subset of 48 genes, we additionally sequenced the 3′-untranslated region (UTR) and the 5′-UTR. In addition, we performed whole gene sequencing including the 10-kbp promoter areas of three genes, i.e. *F5*, *VWF*, and *F8*, which are of particular interest for VT. *F5* harbors the strongest genetic risk factor for VT, i.e. FV Leiden, in the general population. Von Willebrand factor and FVIII, encoded by *VWF* and *F8*, are tightly interconnected proteins whose 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 bp of target region surrounding each variant. A list of the targeted genes and variants is shown in Table S2.

The target area was designed with the Reference Sequence (RefSeq) Database by the use of tools in the University of California Santa Cruz 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, TX, USA). A complete sequencing protocol can be accessed on the HGSC website (https://www.hgsc.bcm.edu/content/protocols-sequencing-library-construction). Briefly, Illumina paired-end pre-capture libraries were constructed from extract DNA 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 with the HiSeq 2000 platform (Illumina, San Diego, CA, USA).

Sequence analysis was performed with the MERCURY analysis pipeline [25]. In short, sequence reads and base-call confidence values were generated for demultiplexed pools with the vendor's primary analysis software (CASAVA). Next, reads and qualities were mapped to reference genome hg19 by use of the Burrows–Wheeler aligner [26], resulting in BAM files per sample [27]. Realignment around insertions and deletions (indels), and recalibration of quality scores, were performed with the GENOME ANALYSIS TOOLKIT [28]. Variant calling was conducted with the ATLAS2 suite [29], and this was 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 of < 0.95, < 3 variant reads, variant read ratio of < 0.1, variant reads in a single strand direction, and total coverage of < 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 with VCFTOOLS [30] to identify high-quality variants, requiring a sequencing depth of ≥ 10 reads, a call rate of $\geq 80\%$, a Phred score of ≥ 30, and Hardy–Weinberg equilibrium $P > 1.0 \times 10^{-4}$ in the controls separately per study. A total of 20 054 variants passed quality control.

Statistical analysis

We conducted single-variant association analyses for 3617 low-frequency and common variants, defined as a minor allele frequency (MAF) of $\geq 1\%$, by using logistic regression as implemented in PLINK [31]. We calculated effect estimates as odds ratios (ORs) with corresponding 95% confidence intervals (CIs) 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 1.38×10^{-5} (0.05 divided by 3617 variants) to account for multiple testing. We additionally calculated false discovery rates (FDRs), and variants with a FDR of < 0.10 were carried forward for replication.

Rare variants (MAF $\leq 1\%$) were collapsed per gene and analyzed with the T1 burden test and the Sequence Kernel Association Test (SKAT) [33], the latter allowing differential effect directions. In total, we analyzed 16 188 variants in 647 genes with a cumulative minor allele count (cMAC) of ≥ 5. Analyses were adjusted for sex, age, (study) origin, and carriership of FV Leiden and PT G20210A. In the burden test, we used adaptive permutations to calculate empirical *P*-values, which were stratified by northwestern versus southern 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 the INVENT consortium. Details of the meta-analysis and the included studies are provided

Table 1 Study population characteristics

	DVT patients	Controls
N	897	598
Age (years), mean (SD)	48.1 (13.7)	47.1 (13.3)
Male sex, *n* (%)	449 (50.1)	277 (46.3)
Northwestern European origin, *n* (%)	599 (67.8)	300 (50.2)
DVT only, *n* (%)	755 (84.2)	NA
PT G20210A carriers*, *n* (%)	15 (1.67)	NA
FVL carriers*, *n* (%)	75 (8.36)	NA

DVT, deep vein thrombosis; FVL, factor V Leiden; NA, not applicable; PT, prothrombin; SD, standard deviation. *These formed part of a subgroup of 241 DVT patients who had a recurrence during follow-up in MEGA and THE-VTE study (prevalences of FVL and PT G20210A in that subgroup of 31.1% and 6.2%, respectively).

elsewhere [11]. In short, GWAS data from 12 studies, including 7507 VT patients and 52 632 controls, were meta-analyzed with an inverse-variance weighting fixed-effects model. It is of note that 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 the INVENT consortium.

Results

Targeted DNA sequencing was successfully performed in 897 DVT patients and 598 controls. The study population characteristics are shown in Table 1. In total, 20 054 high-quality variants were identified, of which 11 268 were singletons (median of seven singletons per person; interquartile range of 4–10). An overview of the functional classes and the MAF distribution is shown in Fig. S1. The majority of the variants were rare and mapped to protein-coding sequences ($N = 10\ 131$), including several stop-loss and stop-gain variants. We also observed 168 indels and 530 splice variants. In addition, we identified a total of 5210 variants that had not been reported in any database.

Single-variant association analyses

We tested 3617 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 Fig. S2. 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 harbor established genetic risk factors for VT. Interestingly, only three of the 12 variants mapped to coding sequences. The exclusion of recurrent VT patients in a sensitivity analysis resulted in similar associations with DVT risk (Table S3). The lead variant in *ABO* was the well-known risk variant rs8176719 (frameshift variant; risk allele frequency [RAF] of 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), although 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 was associated with DVT risk conditional on rs8176719 (Table S4). In *CYP4V2–KLKB1–F11*, the 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, one of which remained associated upon conditioning on

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

							Discovery analysis		Conditional analysis*	
rsID	Chromosome	Position	Class	Gene	A_1/A_2	RAF	OR (95% CI)	*P*-value	OR (95% CI)*	*P*-value
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

A_1, reference allele; A_2, risk allele; CI, confidence interval; NA, not applicable; OR, odds ratio; RAF, risk allele frequency. Single-variant association analyses for 3617 low-frequency and common variants (minor allele frequency of > 1%) were conducted with logistic regression on the assumption of an additive mode of inheritance. Analyses were adjusted for sex, age, (study) origin, carriership of factor V Leiden per copy of the risk allele, and carriership of prothrombin 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).

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 (Table S5). We did not identify additional associations after conditioning on the lead variants (Table S4). Four intronic *F5* variants were associated with DVT risk at the Bonferroni threshold, were in almost complete LD (the 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 independently 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 was 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 CIs, with adjustment for rs4524 (Table S6). No secondary association signals were observed in the *F5* region (Fig. S3).

In addition, we observed 50 variants that did not exceed the Bonferroni threshold for statistical significance, but did have low FDRs (< 0.10). Almost all of these mapped to the four main loci and did not represent new association signals (Table S7). 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. *MASP1* 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, when activated by thrombin and activated platelets [42], MASP1 can cleave several coagulation factors, including PT, thrombin-activatable fibrinolysis inhibitor, and FXIII [41]. Of the other 16 *MASP1* variants, one was also associated with DVT risk (Table S8), and was in complete LD with rs72549167. The other novel variant mapped to a synonymous variant in *CBS*, encoding cystathionine β-synthase, and was

Table 3 Novel variant associations with deep vein thrombosis (false discovery rate [FDR] < 0.10) and replication effort

						Discovery analysis			Replication			
rsID	Chromosome	Position	Class	Gene	A_1/A_2	RAF	OR (95% CI)	*P*-value	FDR	RAF	OR (95% CI)	*P*-value	Mean imputation quality score (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)

A_1, reference allele; A_2, risk allele; CI, confidence interval; OR, odds ratio; RAF, risk allele frequency; SD, standard deviation; UTR, untranslated region. Discovery analysis was performed with logistic regression on the assumption of an additive mode of inheritance. Analyses were adjusted for sex, age, (study) origin, carriership of factor V Leiden per copy of the risk allele, and carriership of prothrombin G20210A. Replication analysis was performed on data from the INVENT consortium. Genome-wide association study results from 12 studies were meta-analyzed with a fixed-effect meta-analysis model based on inverse-variance weighting. Heterogeneity was assessed by use of Cochran's *Q* statistic and the *I2* index. For rs1801181, we observed a *Q*-value of 8.69, an *I2*-value of 0.00, and a *P*-value of 0.65. For rs72549167, we observed a *Q*-value of 9.06, an *I2*-value of 0.00, and a *P*-value of 0.62.

associated with DVT risk with an allelic OR of 1.31 (95% CI 1.11–1.55, FDR 9%). Cystathionine β-synthase catalyzes the conversion of homocysteine to cystathionine, and specific genetic defects in *CBS* lead to homocystinuria, a disorder that has been linked to increased VT risk [43]. We observed two additional common variants in *CBS*, not associated with rs1801181, and neither of these was associated with DVT risk (Table S9). We next aimed to replicate the two novel variant associations by using the meta-analysis data from the INVENT consortium, which included 7507 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) or with rs1801181 in *CBS* (OR 1.00, 95% CI 0.96–1.05).

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 (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 as compared with three alleles in controls. All 10 variants mapped to protein-coding sequences. None of the variants was solely driving the association signal (data not shown).

Discussion

To identify novel genetic risk factors for DVT that 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 and via proxy variants [11–13,19,34–36]. The effect sizes observed in our study were slightly higher than those 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 sequences, whereas 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 a causal risk factor, potentially influencing DVT risk by affecting gene regulation. Four co-inherited intronic variants in *F5*, which have not been implicated in VT risk, were associated with DVT risk at the Bonferroni threshold. Missense *F5* variant rs4524, which is an established risk variant independent of FV Leiden [38] and is 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 after adjustment for our lead *F5* variant (rs6672595). Both variants are part of a large, strongly linked cluster of variants that 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*, i.e. missense *KLKB1* variant rs3733402, which remained associated with DVT risk with an allelic OR of 1.4 after adjustment 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 harbor multiple causal variants. In addition, we were unable to disentangle the effects of *FGA* rs6050 and *FGG* rs2068865 on DVT risk, owing to their strong, although 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 that have not been linked to VT risk, with low FDRs but association tests that did not pass the Bonferroni threshold. These were a synonymous variant in *CBS* and a 3′-UTR variant in *MASP1*. Neither variant, however, was replicated in the meta-analysis data from the INVENT consortium. Imputation quality was sufficient and there was no evidence of statistical heterogeneity. We cannot rule out the possibility that differences in the discovery and the replication study populations, e.g. owing 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 being observed in DVT patients as compared with three alleles in controls, although 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 probably 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], because this would have further increased the multiple testing burden and lowered cMACs. 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 as compared with four 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 as compared with 42 alleles in controls). Although the majority of studied rare *ADAMTS13* variants mapped to coding sequences (75%), the inclusion of non-coding variants may explain the difference in the results of the burden analyses. However, when we focused only on rare coding variants in *ADAMTS13*, we observed a similar association with DVT risk ($P = 0.066$, 55 alleles in DVT patients as compared with 27 alleles in controls). Larger studies are needed to elucidate the role of rare coding and non-coding variants in *ADAMTS13* in DVT risk.

The major limitation of our study is its limited sample size, which prevented us from detecting associations across the entire allele frequency spectrum. Given the multicausal nature of DVT, estimates of genetic effects 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 could therefore 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 or variation in genes not previously linked to the hemostatic system. Therefore, we were unable to identify variants in untargeted regions of the candidate genes or novel DVT-associating genes outside the hemostatic system, or 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.

Addendum

H. G. de Haan designed the study, analyzed and interpreted the data, and wrote the manuscript. A. van Hylckama Vlieg designed the THE-VTE study, interpreted the data, and revised the manuscript. L. A. Lotta designed the study and revised the manuscript. M. M. Gorski interpreted the data and revised the manuscript. P. Bucciarelli recruited patients, performed data analysis, and revised the manuscript. I. Martinelli recruited patients and revised the manuscript. The INVENT Consortium contributed the meta-analysis data. T. P. Baglin designed the THE-VTE study, recruited patients, and revised the manuscript. F. Peyvandi supervised work and revised the manuscript. F. R. Rosendaal designed MEGA, supervised work, and revised the manuscript.

Acknowledgements

We would like to thank all individuals who participated in DVT-Milan, MEGA and the THE-VTE study. MEGA was supported by the Netherlands Heart Foundation (NHS98.113 and NHS208B086), the Dutch Cancer Foundation (RUL 99/1992), and the Netherlands Organization for Scientific Research (912-03-033|2003). The THE-VTE study was supported by the Netherlands Organization for Scientific Research (916.56.157). In addition, DVT-Milan received support from Fondazione Cariplo (grant N. 2011-0524) and the Italian Ministry of Health (N. RF-2009-1530493). The funding organizations played no role in the design and conduct of the study, the collection, management, analysis and interpretation of the data, the preparation, review and approval of the manuscript, or the decision to submit the manuscript for publication.

Disclosure of Conflict of Interests

F. Peyvandi has received consulting fees from Kedrion and Octopharma, all unrelated to the work presented in this manuscript. She is also a member of the scientific advisory boards of Ablynx, F. Hoffmann-La Roche, and

Shire. The other authors state that they have no conflict of interest.

Appendix

INVENT Consortium Investigators

P. Amouyel, Lille, France; M. de Andrade and J. Heit, Rochester, Minnesota, USA; S. Basu, A. R. Folsom, W. Tang and L.-C. Weng, Minneapolis, Minnesota, USA; C. Berr, Montpellier, France; J. A. Brody, B. M. Psaty, N. L. Smith and K. L. Wiggins, Seattle, Washington, USA; D. I Chasman, C. Kabrhel, P. Kraft, S. Lindström, L. M. Rose and P. M. Ridker, Boston, Massachusetts, USA; J.-F. Dartigues and C. Tzourio, Bordeaux, France; M. Germain, D. Smadja and D.-A. Trégouët, Paris, France; H. G. de Haan, J. Houwing-Duitermaat, R. Monajemi, P. H. Reitsma, F. R. Rosendaal, E. Slagboom, M. C. H. de Visser and A. van Hylckama Vlieg, Leiden, the Netherlands; G. Le Gal, Brest, France and Ottawa, Ontario, Canada; P.-E. Morange, N. Saut and P. Suchon, Marseille, France; K. D. Taylor, Torrance, California, USA.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Table S1. Included participants per study
Table S2. Targeted genes and variants
Table S3. Sensitivity analysis excluding recurrent VT patients
Table S4. Discovery and conditional association analyses of variants in *ABO*, *CYP4V2–KLKB1–F11*, and *FGA–FGG* ($P > 1.38 \times 10^{-5}$)
Table S5. Reciprocal conditional association analyses on rs6050 and rs2066865
Table S6. Association analysis at the *F5* association locus conditional on rs4524
Table S7. Suggestive single-variant associations (FDR < 0.10)
Table S8. Single-variant association analyses of common *MASP1* variants
Table S9. Single-variant association analyses of common *CBS* variants
Fig. S1. Minor allele frequency distribution of identified variants (left) and overview of functional classes (right)
Fig. S2. Quantile–quantile plot of single-variant association analyses
Fig. S3. Regional association plots for single-variant associations in the *F5* region

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): 301–4.

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* 2017; **136**: 665–77.

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, *et al*. 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): 610–15.
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**: 715–22.
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**: 1374–5.

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.