OXFORD

Whole-exome sequencing of 14 389 individuals from the ESP and CHARGE consortia identifies novel rare variation associated with hemostatic factors

Nathan Pankratz [ID][1,*], Peng Wei[2], Jennifer A. Brody[3], Ming-Huei Chen[4,5,6], Paul S. de Vries[2,7], Jennifer E. Huffman[5,6,8],

Mary Rachel Stimson[1], Paul L. Auer [ID][9], Eric Boerwinkle[2,10], Mary Cushman[11], Moniek P.M. de Maat[12], Aaron R. Folsom[13],

Oscar H. Franco[7], Richard A. Gibbs[10], Kelly K. Haagenson[1], Albert Hofman[7,14], Jill M. Johnsen[15,16], Christie L. Kovar[10], Robert Kraaij[17],

Barbara McKnight[18], Ginger A. Metcalf[10], Donna Muzny[10], Bruce M. Psaty[3,19,20], Weihong Tang[13], André G. Uitterlinden[7,17],

Jeroen G.J. Rooij[17], Abbas Dehghan[7,21], Christopher J. O'Donnell[5,22,23], Alex P. Reiner[19,23],

Alanna C. Morrison[2] and Nicholas L Smith[19,24,25]

[1]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA
[2]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA
[3]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA
[4]Department of Neurology, Boston University School of Medicine, Boston, MA, USA
[5]Framingham Heart Study, National Heart, Lung and Blood Institute, Framingham, MA, USA
[6]Population Sciences Branch, National Heart, Lung and Blood Institute, Framingham, MA, USA
[7]Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[8]Center for Population Genomics, MAVERIC, VA Boston Healthcare System, Boston, MA, USA
[9]Division of Biostatistics, Institute for Health and Equity, and Cancer Center, Medical College of Wisconsin, Milwaukee, WI, USA
[10]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
[11]Departments of Medicine and Pathology, University of Vermont, Colchester, VT, USA
[12]Department of Hematology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[13]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN, USA
[14]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[15]Research Institute Bloodworks, Seattle, WA, USA
[16]Department of Medicine, University of Washington, Seattle, WA, USA
[17]Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands
[18]Department of Biostatistics, University of Washington, Seattle, WA, USA
[19]Department of Epidemiology, University of Washington, Seattle, WA, USA
[20]Department of Health Services, University of Washington, Seattle, WA, USA
[21]Department of Biostatistics and Epidemiology, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK
[22]Cardiology Section, Department of Medicine, Boston Veterans Administration Healthcare, Harvard Medical School, Boston, MA, USA
[23]Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[24]Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle WA, USA
[25]Seattle Epidemiologic Research and Information Center, Veterans Administration Office of Research and Development, Seattle, WA, USA

*To whom correspondence should be addressed at: Department of Laboratory Medicine and Pathology, University of Minnesota, Mayo Building Room 780, 420 Delaware St SE, Minneapolis, MN 55455, USA. Tel: +1-612-624-2456; Email: nathan.pankratz.rps@gmail.com

Abstract

Plasma levels of fibrinogen, coagulation factors VII and VIII and von Willebrand factor (vWF) are four intermediate phenotypes that are heritable and have been associated with the risk of clinical thrombotic events. To identify rare and low-frequency variants associated with these hemostatic factors, we conducted whole-exome sequencing in 10 860 individuals of European ancestry (EA) and 3529 African Americans (AAs) from the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium and the National Heart, Lung and Blood Institute's Exome Sequencing Project. Gene-based tests demonstrated significant associations with rare variation (minor allele frequency < 5%) in *fibrinogen gamma chain* (*FGG*) (with fibrinogen, $P = 9.1 \times 10^{-13}$), *coagulation factor VII* (*F7*) (with factor VII, $P = 1.3 \times 10^{-72}$; seven novel variants) and *VWF* (with factor VIII and vWF; $P = 3.2 \times 10^{-14}$; one novel variant). These eight novel rare variant associations were independent of the known common variants at these loci and tended to have much larger effect sizes. In addition, one of the rare novel variants in *F7* was significantly associated with an increased risk of venous thromboembolism in AAs (Ile200Ser; rs141219108; $P = 4.2 \times 10^{-5}$). After restricting gene-based analyses to only loss-of-function variants, a novel significant association was detected and replicated between factor VIII levels and a stop-gain mutation exclusive to AAs (rs3211938) in *CD36 molecule* (*CD36*). This variant has previously been linked to dyslipidemia but not with the levels of a hemostatic factor. These efforts represent the largest integration of whole-exome sequence data from two national projects to identify genetic variation associated with plasma hemostatic factors.

Introduction

Fibrinogen, factor VII (FVII), factor VIII (FVIII) and von Willebrand factor (vWF) are circulating plasma hemostatic factors that have been associated with the development of arterial and venous thrombosis or atherothrombotic cardiovascular disease in human populations (1,2). Estimates of heritability range from 0.28 to 0.44 for fibrinogen (3–5), 0.33 to 0.63 for FVII (3–5), 0.29 to 0.61 for FVIII (3–5) and 0.32 to 0.75 for vWF (4,5). The characterization of common and low-frequency variation influencing interindividual and interpopulation differences in circulating fibrinogen, FVII, FVIII and vWF may lead to improved understanding of the role of hemostasis in inflammation and athero-thrombotic risk and potentially reveal novel biologic pathways influencing these hemostatic factors.

Recent genome-wide association studies (GWAS) have demonstrated that common polymorphisms with minor allele frequencies (MAFs) greater than 0.05 contribute to the heritability of all of these traits (6–14) and that the variants underlying variation in FVIII and vWF heavily overlap (7). Using these variants, Sabater-Lleal *et al.* (15) performed a Mendelian randomization experiment demonstrating causal effects of plasma FVIII activity levels on venous thrombosis and coronary artery disease risk and plasma VWF levels on ischemic stroke risk, while de Vries *et al.* (16) did the same for FVII levels and also saw a significant causative effect for ischemic stroke as well as positive trends for venous thromboembolism (VTE) and coronary artery disease. The common polymorphisms identified to date, however, explain only a small proportion of the heritability (7,17), and the amount of variation that they explain is modest: 12.8% for vWF, 7.7% for FVII, 10.0% for FVIII and 3.0% for fibrinogen (6,7,13). This suggests that additional loci or variation within known genes may account for interindividual variability in these hemostatic factors.

An analysis of exome chip data identified rare and low-frequency coding variants at known loci (*fibrinogen beta chain (FGB), fibrinogen gamma chain (FGG), coagulation factor VII (F7), von Willebrand factor (VWF)* and *stabilin 2 (STAB2)* genes), as well as at two novel loci (*potassium sodium-activated channel subfamily T member 1 (KCNT1)* and *HID1 domain containing (HID1)*) in African Americans (AAs), which were associated with levels of hemostatic factors (12). The exome chip content was designed to capture variation as rare as 0.03%; however, the samples used to design the chip were mostly of European descent and so would not capture those variants derived in Africa or other continents. Moreover, the chip is missing variation present only in a single family or in a single individual (i.e. de novo mutations) and also is not well suited to assaying indels, other structural variation and variants close to other variants. For example, a recent population-based analysis identified rare coding variants not present on the exome chip that were significantly associated with VTE, especially in *protein C, inactivator of coagulation factors Va And VIIIa (PROC)* (18). Therefore, the aim of this study

was to characterize additional rare and low-frequency variants associated with plasma levels of fibrinogen, FVII, FVIII and vWF by analyzing whole-exome sequence data in individuals of European and African ancestries from two large, coordinated exome sequencing projects (ESPs).

Results

Participant characteristics

Final sample sizes for each of the four traits are summarized in Table 1, and characteristics of the participating cohorts are summarized in Supplementary Material, Table S1. Taking into account sample size, the mean age was 58.2 ± 6.2 years and 56.8% were female. The means [standard deviations (SDs)] for the hemostatic factors were 3.1 (0.7) g/L for fibrinogen, 115.9 (27.2) IU/dL for FVII, 128.7 (38.4) IU/dL for FVIII and 117.7 (46.2) IU/dL for vWF.

Single-variant test results

Single-variant analyses were performed on the 269 877 single-variant sites with a minor allele count (MAC) ≥ 40; therefore, an association was considered to be study-wide significant at a Bonferroni-corrected threshold of $P < 1.9 \times 10^{-7}$. In single-variant analyses, we did not observe significant inflation or deflation of the meta-analysis P-values ($0.96 \leq$ lambdas ≤ 1.06; see Supplementary Material, Table S2 and Supplementary Material, Fig. S1), indicating that there was no serious overcorrection or confounding by any covariate, population substructure or lab effects. Table 2 summarizes the significant loci in the trans-ethnic meta-analysis and lists the index variant (the variant with the smallest P-value). For each hemostatic factor, an exhaustive list of all variants with a minimum P-value < 0.001 in AAs alone, EAs alone or the meta-analysis across AAs and EAs is available in Supplementary Material, Table S3a–d. After performing a *post-hoc* sensitivity analysis that included all variants with a MAC ≥ 10 (449 733 variants; Bonferroni threshold $P < 1.1 \times 10{-}7$), lambdas were similar (see Supplementary Material, Table S2), no new loci were identified, and all loci identified at MAC ≥ 40 would have also been significant at the more conservative threshold.

Fibrinogen

There were two loci significantly associated with fibrinogen levels: one within the fibrinogen gene cluster (*fibrinogen alpha chain (FGA), FGB* and *FGG*; codon numbers in these genes reflect the cleaved form, reflected in Supplementary Material, Table S4) and the other at *interferon regulatory factor 1 (IRF1)* (Table 2). There were five study-wide significant variants within the fibrinogen gene cluster [Ala82Gly (rs148685782) in *FGG*; Tyr345Tyr (rs4681), Arg448Lys (rs4220) and Ser159Ser (rs6056) in *FGB* and Thr312Ala (rs6050) in *FGA*; variants in the gene cluster were annotated using the sequence of the mature, circulating protein]. The three *FGB* variants were all in high linkage disequilibrium (LD) with one another ($r^2 = 0.99$), but all other combinations of variants across loci were not in strong LD ($r^2 \leq 0.01$). The Ala82Gly variant in *FGG*

Table 1. Number of samples for each study with whole-exome sequencing data listed by hemostatic factor

Study	Race	Fibrinogen	FVII	FVIII	vWF	Maximum
ARIC	EA	5652	5527	5658	5682	5682
CHS	EA	737	742	734	-	742
FHS	EA	741	667	-	667	741
RS	EA	987	263	906[a]	788[a]	906+788[a]
NHLBI ESP[b]	EA	2001	1204	1282	1181	2001
EA subtotal		10 118	8403	8580	8318	10 860
ARIC	AA	2655	2602	2659	2664	2664
NHLBI ESP[b]	AA	865	644	598	439	865
AA subtotal		3520	3246	3257	3103	3529
Total EA + AA		13 638	11 649	11 837	11 421	14 389

[a]These subsets of the individuals are mutually exclusive. [b]ESP consists of nonoverlapping samples from ARIC, CHS, FHS, CARDIA, MESA and WHI.

Table 2. Index variants for the loci with single nucleotide variant associations exceeding study-wide significance ($P < 1.9 \times 10^{-7}$)

Trait index variant[a]	Gene	# sig[b]	Function	Beta EA+AA[c]	Est. effect[d]	Effect in SD[e]	P-value EA + AA	EA	AA	EA + AA MAF[f]	EA	AA
Fibrinogen												
rs148685782	FGG	5	Ala82Gly	−0.256	−0.70	−1.03	**2.4E−28**	**4.5E−27**	0.02	0.003	0.004	0.0006
rs2706379	IRF1	1	intronic	−0.020	−0.06	−0.09	**4.2E−09**	**1.6E−07**	0.01	0.23	0.21	0.28
FVII												
rs1260326	GCKR	1	Leu446Pro	−0.018	−2.09	−0.08	**6.3E−09**	**4.3E−09**	0.38	0.33	0.41	0.14
rs1126670	ADH4	4	Pro255Pro	0.019	2.16	0.08	**5.2E−09**	**4.0E−07**	0.003	0.27	0.31	0.17
rs12453	MS4A6A	2	Leu137Leu	0.017	1.95	0.07	**2.5E−08**	**3.5E−08**	0.20	0.34	0.40	0.20
rs6046	F7	8	Arg413Gln	−0.157	−17.47	−0.64	**1.8E−261**	**2.3E−230**	**7.3E−38**	0.11	0.11	0.12
rs867186	PROCR	4	Ser219Gly	0.054	6.36	0.23	**4.9E−29**	**1.1E−26**	0.0002	0.10	0.10	0.09
FVIII												
rs8176749	ABO	28	Leu310Leu	0.132	17.76	0.46	**4.3E−94**	**5.7E−46**	**1.8E−51**	0.10	0.07	0.16
rs57950734	VWF	7	His817Gln	−0.097	−11.98	−0.31	**9.1E−15**	0.39	**6.0E−15**	0.03	0.0001	0.10
rs7296626	STAB2	2	intronic	0.057	7.51	0.20	**2.5E−11**	**1.3E−10**	0.07	0.05	0.06	0.01
vWF												
rs1039084	STXBP5	2	Asn436Ser	0.030	3.49	0.08	**1.2E−09**	**1.0E−07**	0.003	0.48	0.46	0.44
rs8176741	ABO	36	His219His	0.198	24.81	0.54	**8.6E−115**	**2.3E−60**	**1.7E−57**	0.09	0.07	0.16
rs1063856	VWF	8	Thr789Ala	0.059	6.81	0.15	**1.8E−30**	**3.5E−20**	**1.7E−12**	0.42	0.36	0.41
rs35102665	STAB2	1	Ala1996Ala	0.089	10.83	0.23	**1.1E−09**	**8.4E−10**	0.74	0.03	0.04	0.004
rs17564	STX2	1	Ser42Thr	−0.037	−4.44	−0.10	**1.7E−12**	**9.5E−11**	0.004	0.45	0.35	0.27

[a]Additional properties of each variant, including genomic position and presence on the exome chip, are available in Supplementary Material, Table S3a–d.
[b]sig = number of significant markers ($P < 1.9 \times 10^{-7}$) in the same region. [c]EAs, Europeans/European Americans; AAs, African Americans; EA + AA, the combined multi-ethnic meta-analysis; MAF, minor allele frequency. [d]All traits were natural log (ln) transformed, so the effect on the trait in the original units can be estimated using algebra; see Supplementary Methods for the formulas. [e]Dividing the estimated effect[d] by the pooled SD in Table 1 gives the effect in SD units.

was rare (EA MAF = 0.0036; AA MAF = 0.0006; EA + AA $P = 2.4 \times 10^{-28}$) and was associated with 0.70 g/L lower levels, on average, for each copy of the minor allele, which is more than one SD from the trait mean. A known common variant within *IRF1* (EA + AA $P = 4.2 \times 10^{-9}$) also reached genome-wide significance (−0.06 g/L per allele). The index variants (the marker with the smallest *P*-value) for *IRF1* and for each of the genes in the fibrinogen cluster demonstrated at least a trend ($P < 0.05$ in the same direction of effect) in both races (see Supplementary Material, Table S3a).

Fibrinogen conditional analyses

Sensitivity analyses were conditioned on the five variants identified at the fibrinogen gene cluster that met our study-wide significance threshold [Ala82Gly (rs148685782), Tyr345Tyr (rs4681), Arg448Lys (rs4220), Ser159Ser (rs6056) and Thr312Ala (rs6050)]. In addition, a

subthreshold rare variant in *FGB*, Pro176Leu (rs6054; EA MAF = 0.004; AA MAF = 0.0009; EA + AA $P = 4.6 \times 10^{-6}$) was included because it was shown to be significantly associated with fibrinogen in previous studies (12,17). When the two rare variants (Ala82Gly and Pro235Leu) were included as covariates in the analysis model, the common variants in *FGB* and *FGA* maintained their level of significance (Tyr345Tyr $P < 2.1 \times 10^{-10}$; Thr312Ala $P < 2.4 \times 10^{-9}$; see Supplementary Material, Table S5a). Similarly, when the common variants in *FGB* were included as covariates, the rare variants maintained a similar level of significance (Ala82Gly $P = 6.9 \times 10^{-24}$; Pro235Leu $P = 1.2 \times 10^{-5}$), indicating that the effects are independent. The common Thr312Ala variant in *FGA* represents a third independent effect, as it remained significant when conditioning on either the common *FGB* variants or the rare variants. No additional variation was significantly associated at the *IRF1* locus.

Factor VII

There were five loci significantly associated with FVII levels, encompassing known genes *glucokinase regulator* (*GCKR*), *alcohol dehydrogenase 4* (*class II*), *ADH4*), *Pi polypeptide* (*ADH4*), *membrane spanning 4-domains A6A* (*MS4A6A*) *F7* and *protein C receptor* (*PROC*) (Table 2). All of the index variants in these regions were common. The variant at the *F7* locus had the largest effect, where the index variant [Arg413Gln (rs6046)] decreased FVII values by an average of 17 IU/dL (0.68 SD units) for each copy of the minor allele (EA MAF = 0.11; AA MAF = 0.12; EA + AA $P = 1.8 \times 10^{-261}$). All loci demonstrated significant association within EAs. In AAs, suggestive association in the same direction was observed for variants in *ADH4*, *PROCR* and *F7* ($P < 0.003$) but not for variants in *GCKR* and *MS4A6A* ($P > 0.20$). See Supplementary Material, Table S3b for full results.

FVII conditional analyses

Conditional analyses were conducted at the *F7* locus to better understand the contribution of common and rare variation. When the Arg413Gln (rs6046) index variant in *F7* was included as a covariate in the model, signals for two additional common variants in strong LD with the index single nucleotide polymorphism (SNP) ($r^2 = 0.78$–0.90) were severely attenuated but not abolished (rs6039 went from $P = 1.2 \times 10^{-250}$ to $P = 5.1 \times 10^{-32}$ and rs6042 went from $P = 1.7 \times 10^{-260}$ to $P = 1.5 \times 10^{-10}$; see Supplementary Material, Table S5b). Since they are in high LD and losing 90–95% of their signal, because rs6039 is intronic and rs6042 is a synonymous variant, they are not likely independent signals. The conditional analyses also identified six rare (MAC > 10 but MAF < 1%) missense variants in F7 significantly associated with FVII levels that maintained their level of significance after conditioning on Arg413Gln ($1.8 \times 10^{-25} < P < 4.7 \times 10^{-6}$; see Supplementary Material, Table S5b) and were associated with lower FVII levels between 17 and 50 IU/dL. MACs for these variants ranged from 11 to 49. Four of these variants were present almost exclusively in AAs (AA MAC ≥ 20; EA MAC ≤ 2), while one was present only in EAs (EA MAC = 11) and one was present in both (EA MAC = 9; AA MAC = 19). See Supplementary Material, Table S5b for more detail.

Factor VIII

Three loci were significantly associated with FVIII: multiple variants in *alpha 1-3-N acetylgalactosaminyl-transferase and alpha 1-3-galactosyltransferase* (*ABO*) (index variant = rs8176749; EA MAF = 0.07; AA MAF = 0.16; EA + AA $P = 4.3 \times 10^{-94}$), a variant in *VWF* common only in AAs (His817Gln; rs57950734; EA MAF = 0.0001, EA $P = 0.39$; AA MAF = 0.10, AA $P = 6.0 \times 10^{-15}$) and a variant in *STAB2* that is more common in EAs (rs7296626; EA MAF = 0.06, EA $P = 1.3 \times 10^{-10}$; AA MAF = 0.01, AA $P = 0.072$). The index variant at the *ABO* locus tags the O deletion. See Supplementary Material, Table S3c for more detail.

FVIII conditional analyses

Supplementary Material, Table S5c describes the conditional analyses performed at the *ABO* locus. When the O deletion at the *ABO* locus was included as a covariate in the model, the *P*-value for a variant that tags the A2 blood type [Pro156Leu (rs1053878)] became more extreme (EA conditional model 1 $P = 1.4 \times 10^{-13}$). Pro156Leu was not significant in AAs before ($P = 0.13$) or after ($P = 0.55$) despite being more common in AAs (EA MAF = 0.07; AA MAF = 0.22). When both variants (Type O and Type A2) were included in the model, then a variant that tags the B blood type that was significant in the unconditional results (Leu266Met; EA MAF = 0.07, EA unconditional model $P = 1.0 \times 10^{-44}$; AA MAF = 0.16, AA unconditional model $P = 4.0 \times 10^{-52}$) regained significance in AAs (EA conditional model 2 $P = 0.10$; AA conditional model 2 $P = 4.8 \times 10^{-7}$). When all three blood types were included, an uncommon missense variant that tags the O² blood group haplotype gained significance in EAs [Gly268Arg (rs41302905); EA MAF = 0.02, EA $P = 2.9 \times 10^{-25}$; AA MAF = 0.0003, AA $P = 0.03$].

When the His817Gln (rs57950734) variant in *VWF* was included as a covariate in the model, a variant common in both AAs and EAs remained significant [Thr789Ala (rs1063856); EA MAF = 0.36, EA $P = 1.1 \times 10^{-8}$; AA MAF = 0.41, AA $P = 3.9 \times 10^{-6}$]. When conditioning on Thr789Ala, a third independent signal that was common only in AAs remained [Arg2185Gln (rs2229446); AA MAF = 0.19, AA $P = 2.6 \times 10^{-13}$]. In addition, three rare missense variants (MAC > 10 but MAF < 0.01) also remained significant in the trans-ethnic analyses after conditioning on each of the three common variants [Tyr1584Cys (rs1800386) $P = 3.2 \times 10^{-13}$; Arg854Gln (rs41276738) $P = 8.6 \times 10^{-8}$; and Arg2287Trp (rs61750625) $P = 8.7 \times 10^{-6}$]. See Supplementary Material, Table S5d for more detail. When conditioning on rs7296626 near *STAB2*, several synonymous variants remained nominally significant ($P < 0.001$) after serial conditional analysis but did not achieve study-wide significance (Asn1113Asn, Ala1996Ala, Leu80Leu).

von Willebrand factor

Five loci were significantly associated with vWF. Three of these loci—the *ABO* O deletion tag (rs8176749), the common *VWF* variant [Thr789Ala (rs1063856)] and a synonymous variant in *STAB2* [Ala1996Ala (rs35102665)]—were also significantly associated with FVIII (see Table 2 for a comparison). Two variants associated with vWF but not FVIII were in *STXBP5* [Asn436Ser (rs1039084)] and *STX2* [Ser42Thr (rs17564)]. See Supplementary Material, Table S3d for more detail.

vWF conditional analyses

The same pattern of associations with ABO blood types found to be significant for FVIII was also significant for vWF (Supplementary Material, Table S5e), with the O deletion having the strongest effect, the A2 group tagged by Pro156Leu (rs1053878) (EA $P = 6.1 \times 10^{-12}$; AA

Table 3. Results for gene-based tests of association

Trait	Gene	# Variants	Burden P-values			SKAT P-values		
			EA + AA	EA	AA	EA + AA	EA	AA
Fibrinogen	FGG	78	0.0001	**4.7E−10**	0.001	9.1E−13	**3.0E−18**	3.0E−06
FVII	F7	115	**1.3E−72**	**1.1E−19**	**1.2E−55**	**2.3E−46**	**6.0E−19**	**3.9E−39**
FVIII	REXO4 (ABO locus)	58	**9.7E−07**	0.03	5.3E−06	**9.9E−12**	0.24	**5.7E−12**
FVIII	VWF	640	0.0009	1.9E−06	0.04	**3.2E−14**	4.3E−06	1.1E−05
vWF	REXO4 (ABO locus)	58	0.0002	0.08	0.0004	**8.8E−11**	0.15	**1.9E−09**
vWF	VWF	640	3.7E−05	**4.9E−07**	2.1E−05	**1.0E−07**	0.0002	**5.7E−10**

EA + AA, the combine multi-ethnic meta-analysis; The burden test included all functional variants with an MAC less than 5%.

$P = 0.31$), the B group tagged by Leu266Met (rs8176746) (EA $P = 5.0 \times 10^{-4}$; AA $P = 1.8 \times 10^{-6}$) and the O^2 group tagged by Gly268Arg (rs41302905) (EA $P = 1.2 \times 10^{-26}$; AA $P = 0.11$). Conditional analyses in *STX2* revealed no secondary signal (all 26 variants with MAC > 10 had $P \geq 0.08$).

Despite the high correlation between FVIII and vWF, the index variant for FVIII that was common in AAs but not EAs (His817Gln; rs57950734; EA MAF = 0.00007; AA MAF = 0.10) was not associated with vWF ($P = 0.089$). The index variant in VWF for vWF was Thr789Ala (rs1063856); after conditioning on that, the variant with the next smallest P-value was Arg2185Gln (rs2229446) (EA MAF = 0.002, EA $P = 0.13$; AA MAF = 0.19, AA $P = 2.7 \times 10^{-17}$). The P-value for His817Gln was nominally significant when conditioning on Thr789Ala ($P = 0.000084$), but not when conditioning on the other variant (Arg2185Gln) common to AAs but not EAs ($P = 0.72$). Similar to FVIII, additional rare missense variants [Tyr1584Cys (rs1800386), Arg2287Trp (rs61750625) and Ser1486Leu (rs149424724)] remained significant after conditioning on the common variants. See Supplementary Material, Table S5f for more detail.

Gene-based test results

Results for the gene-based tests [burden and sequence kernel association test (SKAT)] are summarized in Table 3. Two different rare variant (MAF < 5%) gene-based tests were performed on the 17 150 genes with a total MAC for the whole gene ≥40; therefore, an association was considered to be significant at a Bonferroni-corrected threshold of $P < 1.5 \times 10^{-6}$. Aggregate testing revealed significant gene-level associations between fibrinogen levels and *FGG* as well as FVII level and *F7* gene. FVIII and vWF levels were both significantly associated with *VWF* and several genes at the *ABO* locus (*ABO, REX4 homolog, 3'-5' exonuclease, (REXO4), ADAM metallopeptidase with thrombospondin type 1 motif 13 (ADAMTS13), surfeit 2 (SURF2), serine/threonine kinase like domain containing 1 (C9orf96) SURF2, C9orf96*). However, the burden tests for these other genes surrounding *ABO* were no longer significant when conditioning on the variants tagging the common ABO blood types, indicating no evidence of rare functional variants in the region with an independent association with either FVIII or vWF.

Further analyses were conducted to assess the variation contributing to the gene-based tests. There were a large number of polymorphic but rare missense or nonsense variants (labeled as 'n var') contributing to each gene-based test, of which a subset were showing a trend toward significance ($P < 0.05$; 'n trend') and/or were gene-wide significant ($P < 0.05/\#$ rare missense or nonsense polymorphic variants in the gene; 'n sig'): *FGG* (n var = 50, n trend = 7, n sig = 2), *F7* (n var = 63, n trend = 27, n sig = 10), *VWF* associated with FVIII (n var = 353, n trend = 42, n sig = 4) and *VWF* associated with vWF (n var = 349, n trend = 42, n sig = 3). See Supplementary Material, Table S6a–d for more detail. While variants in each of these genes have been previously reported, many of the rare variants underlying these associations are novel. *FGG* had two gene-wide significant results: the rare Ala82Gly (rs148685782) variant identified in the single-variant analyses and Ser219Phe (rs145051028), which is polymorphic only in AAs (AA MAF = 0.001, AA $P = 7.4 \times 10^{-6}$). There were 10 variants in *F7* that had a P-value less than the gene-wide threshold of $P < 8.0 \times 10^{-4}$ (see Supplementary Material, Table S6b for the full list of markers and results). Within *VWF*, five rare variants were gene-wide significant (see Supplementary Material, Table S6c and d). The rare variants significantly associated with FVIII [Tyr1584Cys (rs1800386), Arg2287Trp (rs61750625), Arg854Gln (rs41276738), Gly2705Arg (rs7962217)] were slightly different than those that were significantly associated with vWF [Tyr1584Cys, Arg2287Trp, Ser1486Leu (rs149424724)]. However, if a variant was associated with FVIII, then the variant tended to be associated with vWF as well, and vice versa. The correlation of $-\log_{10}(P$-values) for FVIII and vWF was 0.652, and the correlation of the betas was 0.613 in EAs and 0.744 in AAs. These values are very similar to the correlation between the actual trait values [0.689 in the Atherosclerosis Risk in Communities Study (ARIC)]. Of these 17 rare variants significantly associated with a hemostatic factor, eight of them were novel (the novel variants are listed in Table 4 and flagged in the full list of significant variants in Supplementary Material, Table S7a). These variants tended to have higher CADD (CADD) scores on average (median: 23.7; interquartile range: 13.1–29.7), and additional functional predictions and annotation indicate that nearly all of these variants are expected to affect protein function in a deleterious fashion (see Supplementary Material, Table S7a). Gene-level annotations suggest that FGG, F7 and VWF are tolerant to loss-of-function (LOF) variants (probability of being

Table 4. Summary of the nine novel variants associated with hemostatic factors

| Trait | chr:position:REF:ALT[a] (rsID) | Gene | Function | MAF | | Beta | SE | P-values[b] | | |
				EA	AA			EA + AA	EA	AA
FVII	chr13:113773044:C:T (rs137919286)	*F7*	Arg375Trp	5.5E−04	2.9E−03	−0.412	0.044	**5.9E-21**	**2.1E-07**	**3.8E-15**
FVII	chr13:113772982:C:T (rs36209567)	*F7*	Ala354Val	6.8E−04	.	−0.473	0.065	**4.8E-13**	**4.8E-13**	NA
FVII	chr13:113771107:T:G (rs141219108)	*F7*	Ile200Ser	6.9E−05	7.9E−03	−0.176	0.034	**1.8E-07**	3.5E−01	**2.8E-07**
FVII	chr13:113771790:CTGT:C (rs767092036)	*F7*	.	.	9.7E−04	−0.487	0.107	5.8E−06	NA	5.8E−06
FVII	chr13:113760157:T:C (rs766294997)	*F7*	Met1Thr	.	5.8E−04	−0.571	0.139	3.8E−05	NA	3.8E−05
FVII	chr13:113770022:A:G (rs200016360)	*F7*	Gln160Arg	1.4E−04	.	−0.569	0.152	1.7E−04	1.7E−04	NA
FVII	chr13:113773216:G:T (rs1450120320)	*F7*	Gly432Val	.	1.9E−04	−0.831	0.240	5.4E−04	NA	5.4E−04
FVIII	chr7:80300449:T:G (rs3211938)	*CD36*	Tyr325Stop	.	8.3E−02	−0.189	0.014	2.5E−05	NA	2.5E−05
FVIII	chr12:6127833:T:C (rs1800386)	*VWF*	Tyr1584Cys	2.8E−03	3.1E−04	−0.281	0.037	**2.4E-14**	**4.8E-14**	0.24
vWF	chr12:6127833:T:C (rs1800386)	*VWF*	Tyr1584Cys	3.0E−03	1.6E−04	−0.371	0.051	**5.3E-13**	**6.3E-13**	0.52

[a]Additional properties of each variant are available in Supplementary Material, Table S9a. [b]**Bold** indicates study-wide significance (Bonferroni correction for the number of variants tested in the genome); *italics* indicates locus-wide significance (Bonferroni correction for the number of variants tested at that locus; detail on the thresholds is present in Supplementary Material, Table S9a).

Loss-of-function Intolerant (pLI) score < 0.1; however, the data in gnomAD (19) suggest that the general population tends to have fewer missense variants than expected based on other genes with similar properties (see Supplementary Material, Table S7b).

Loss-of-function variants

There were only 2247 genes that contained one or more LOF variants (stop-gain, stop-loss, splice site and indels, but not missense variants) and had a cumulative MAC > 40, leading to a Bonferroni-corrected threshold for two gene-based tests of 1.1×10^{-5}. Only one gene exceeded this threshold and was between FVIII levels and LOF variants in *CD36 molecule* (*CD36*) ($P = 7.9 \times 10^{-06}$). While the gene contained 26 LOF variants that were polymorphic in this study, a single stop gain variant (Tyr325Stop; rs3211938; MAF = 0.08 in AA; $P = 2.5 \times 10^{-5}$) dominated the signal, with carriers having an average decrease in plasma FVIII levels of 9.6 IU/dL. This variant was also nominally associated with reduced vWF levels (-8.57 IU/dL; $P = 0.00045$) and increased FVII levels (3.31 IU/dL; $P = 0.01$), but not with fibrinogen levels (-0.049 g/L; $P = 0.12$). See Supplementary Material, Table S8a–c for more detail. We were able to replicate this association between Tyr325Stop and FVIII levels in an independent set of 1087 AA samples from ARIC (decrease of 13.2 IU/dL; $P = 0.0005$) using genotype data from the exome chip (see Supplementary Methods). The intensity plot for this exome chip variant revealed three distinct clusters with no ambiguous genotype calls indicating that genotype quality, and the accuracy was excellent (see Supplementary Material, Fig. S2).

While the CD36 signal was overwhelmingly dominated by the Tyr325Stop, other LOF variants in the CD36 gene also led to a nominally significant decrease in F8 levels ($P = 0.027$), as did very rare (MAF < 0.001) missense variants ($P = 0.032$). The effect size for those missense variants that were predicted to be deleterious by a majority of bioinformatic algorithms tested (beta = -0.026; $P = 0.30$) was in the same direction of those for the LOF model (beta = -0.040; $P = 7.9 \times 10^{-06}$) but was not significant.

Those missense variants that were predicted to be benign by a majority of algorithms did not show this trend (beta = 0.003; $P = 0.88$). Bioinformatic predictions of all CD36 variants are available in Supplementary Material, Table S8d. For more comparisons of CD36 burden and SKAT tests across a range of MAF thresholds, see Supplementary Methods and Supplementary Material, Table S8e.

Association with clinical outcomes

The 18 variants found to be associated with one of the hemostatic factors (summarized in Supplementary Material, Table S9a) were then tested for association with VTE. The Longitudinal Investigation of Thromboembolism Etiology study (19) has identified and validated hospitalized VTE events (leg deep vein thromboses or pulmonary emboli) in ARIC and the Cardiovascular Health Study (CHS) through 2011 (383 EA cases and 214 AA cases with whole exome sequencing (WES) data). An association was considered to be study-wide significant at a Bonferroni-corrected threshold of $P < 0.003$. One of the rare variants in F7 that is exclusive to AAs (Ile200Ser; rs141219108; AA P-value = 4.2×10^{-5}) was significantly associated with an increased risk of VTE, while another AA variant was nominally associated with an increased risk of VTE (Arg139Gln; rs150525536; AA P-value = 0.02) (see Supplementary Material, Table S9b).

We then attempted to see if these rare variants have an aggregate effect on VTE. We took an unweighted approach, because harmonizing weights across different hemostatic factors, each with their own range and unit of measurement, could be problematic. We did, however, take the direction of thrombotic effect into account. If a variant led to a decrease in a prothrombotic coagulation factor, then that allele was subtracted from the score instead of being added to it. This rare variant polygenic risk score was then tested for an effect on VTE using a Cox proportional hazards model regressing out age, sex, study center and population-specific principal components (PCs). There was a trend in the EA samples [HR: 1.17 (0.97, 1.40)], but not in the AA samples [HR: 1.00 (0.81,

1.23)]. However, when limiting the analyses to only those with idiopathic VTE (i.e. not provoked by a hospitalization or cancer diagnosis), then we saw a consistent effect across both AAs (29% increase in risk per thrombotic allele) and EAs (33% increase in risk per thrombotic allele) which was statistically significant when meta-analyzed together [HR: 1.31 (1.04, 1.65); $P = 0.02$].

Association between CD36 and other phenotypes

When ARIC AAs were analyzed for the same traits noted to be different in previous studies (20) of the *CD36* stop-gain variant (rs3211938), carriers were found to have higher high-density lipoprotein (HDL) cholesterol ($P = 2.2 \times 10^{-6}$) and lower triglycerides ($P = 0.001$). The association with FVIII levels remained similar after correcting for age, sex, center, body mass index (BMI), HDL and triglycerides (see Supplementary Material, Table S8a–c for full detail).

Discussion

Analysis of exome sequence data allows for the opportunity to identify and analyze coding variation across the full allele-frequency spectrum (from common to rare) and to distinguish independent signals from common and rare coding variation within significant loci. Analysis of these four hemostatic factors in a large ($n = 10\,860$ EA individuals and 3529 AA individuals) meta-analysis of multiple multiethnic human population studies identified eight new independent signals in known loci using exome sequence data as well as an LOF variant at a novel locus (*CD36*) (see Table 4, for a summary of these nine novel variants). The rare novel variants tended to have large effect sizes and in total explained 1–2% of the variance of their respective traits (see Supplementary Material, Table S9a). The variance explained tended to be larger for AAs, especially for FVII (see additional detail below). Of note, none of the rare variants that we identified were correlated with each other (all $r^2 < 0.02$) and are therefore considered independent signals (see Supplementary Material, Table S9c). When we combined these into a rare variant polygenic risk score, we saw a significant association with idiopathic VTE in both AAs (29% increase in risk per thrombotic allele) and EAs (33% increase in risk per thrombotic allele).

Fibrinogen

Single-variant analyses identified associations at the *FGA*, *FGB* and *FGG* gene cluster on chromosome 4, including the rare *FGG* Ala82Gly (rs148685782) mutation that has been reported in a pair of cases with hypofibrinogenemia as well as in previous GWAS (12,13,21,22). In addition, assessment of the significant gene-based test for *FGG* revealed that there were a large number of rare missense or nonsense variants ($n = 51$), including two rare variants that were gene-wide significant: Ala82Gly, which was identified in the single-variant analyses and

Ser219Phe (rs145051028), which is polymorphic only in AAs (AA MAF = 0.001, AA MAC = 8, AA $P = 7.4 \times 10^{-6}$). Both of these variants were also identified in a recent exome chip analysis (12). The current study also showed that there are independent signals from common [*FGB* Tyr345Tyr (rs4681) and *FGA* Thr312Ala (rs6050)] and rare variation [*FGG* Ala82Gly and *FGB* Pro235Leu (rs6054)] at this locus. Importantly, common and rare variation at this locus may either increase (Tyr345Tyr) or decrease (Ala82Gly, Pro235Leu, Thr312Ala) fibrinogen levels, substantiating the fact that these variants each have important independent contributions to the trait. The contribution from a known variant in *FGA* (Thr312Ala) is independent of the common and rare variation in *FGB* and *FGG*. The common signals at the *IRF1* gene and the fibrinogen locus have been observed in previous studies (6,11). The most recent meta-analysis of fibrinogen levels found that the 47 independent common variants explained 3.0% of the variance of the trait (13). The independent rare variants in our study explained $\sim 1.1\%$ of the variance.

Factor VII

This study detected associations between FVII levels and common variation in genes (*GCKR*, *ADH4*, *MS4A6A*, *PROCR* and *F7*) identified in a prior GWAS (7). We also identified six rare variants in *F7* (all MAF < 0.0025) that are independent ($1.8 \times 10^{-25} < P < 5 \times 10^{-6}$) from the common index variant (Arg413Gln (rs6046)). These variants are not in LD with each other ($r^2 < 0.001$) and were not detected via prior GWAS; however, three of them [Arg139Gln (rs150525536), Arg364Gln (rs121964926) and Glu445Lys (rs3093248)] were identified by a study focusing on the rare variants present on the exome chip (12). These rare variants have ancestry-specific contributions: four in AAs [Arg139Gln, Arg364Gln, Ile200Ser (rs141219108), Glu445Lys], one in EAs [Ala354Val (rs36209567)] and one in both [Arg375Trp (rs137919286)]. Assessment of the significant gene-based test for *F7* revealed that there were a large number of rare missense or nonsense variants ($n = 63$), including 10 that were gene-wide significant. These include very rare variation not tested in single-variant analyses due to their low MAC (MAC < 40), including an in-frame deletion of three bases present in only five AAs (chr13:113771790:CTGT:C; AA $P = 5.8 \times 10^{-6}$; see Supplementary Material, Table S6b). Seven of the significant rare variants (delineated in Supplementary Material, Table S9a) and 18 of the rare variants showing a trend ($P < 0.05$) have not been previously associated with FVII levels in a population-based study (7,8,12); however, some of these variants are present in the F7 mutation database and a subset of those have been noted in individuals with FVII deficiency. The most recent meta-analysis of FVII levels found that the eight independent common variants explained 19.0% of the variance of the trait (16). The independent rare variants in our study explained $\sim 3.9\%$ of the variance. Of note, the variance explained by these rare variants

was dramatically higher for AAs (10.6%) compared with EAs (1.3%), which was due to 8 of the 10 rare variants having dramatically higher allele frequencies in AAs.

Factor VIII

The signal at the *ABO* locus can be fully explained after taking into account variants tagging the major ABO blood types (A2, B, O and O²); however, novel variation was identified at the *VWF* and *STAB2* loci. A previous study (23) reported an association with Pro2039Thr in *STAB2* (chr12:104139034), and here, we report an unlinked ($r^2 = 0.003$ within ARIC EAs) novel intronic variant near a splice site of *STAB2* (rs7296626; EA MAF = 0.06, EA $P = 1.3 \times 10^{-10}$; AA MAF = 0.01, AA $P = 0.072$). Conditional analyses revealed several independent signals at *VWF*, including His817Gln (rs57950734) and known common variants Thr789Ala (rs1063856) and Arg2185Gln (rs2229446) (24). Importantly, this study identified three rare independent missense variants in VWF, one of which is novel [Tyr1584Cys (rs1800386)]. The Try1584Cys signal is driven by the EAs and led to an average decrease in FVIII levels of 32 IU/dL. Assessment of the significant gene-based test for *VWF* revealed that there were a large number of rare functional variants ($n = 353$). These included the four gene-wide significant rare variants driving the gene-based finding: the novel Tyr1584Cys variant identified by this study, the Arg854Gln variant identified by Huffman *et al.* (12), the Arg2287Trp (rs61750625) variant identified by Johnsen *et al.* (24) that is exclusive to AAs and the Gly2705Arg (rs7962217) variant identified by Tang *et al.* (25) that is more common in EAs. The most recent meta-analysis of FVIII levels found that independent common variants explained 19.0% of the variance of the trait (15). The independent rare variants in our study explained 2.7% of the variance.

The *CD36* finding was the only novel locus identified in this study and was identified only when restricting our search to LOF variants. The stop-gain variant driving the gene-based association (rs3211938) is found at an allele frequency of 0.08 in AAs and is nonexistent in EAs. We were able to replicate the association with this variant in an independent set of 1087 AAs ($P = 0.0005$). The few ARIC individuals homozygous for the variant had FVIII levels that were similar to those of heterozygotes, suggesting a dominant effect. This variant was found at a high frequency in the Yoruban population, but not other HapMap populations, and was on a background with extended haplotype homozygosity that spanned hundreds of kilobases, suggesting that there was recent selective pressure on this variant (26). There is conflicting evidence of whether it protects against malarial infection in a way similar to sickle cell trait (26). The variant has been shown to cause CD36 deficiency in the homozygous state (20) and has previously been associated with increased HDL ($P = 0.00018$) and decreased triglycerides ($P = 0.0059$) (20). Analyses performed in ARIC AAs corroborate the association with higher HDL ($P = 2.2 \times 10^{-6}$)

and lower triglycerides ($P = 0.001$) and revealed that the association with FVIII levels remains virtually unchanged when correcting for age, sex, center, BMI, HDL and triglycerides. Another study found that this variant led to reduced CD36 protein expression and confirmed that it may promote a protective metabolic profile (27). CD36, also known as glycoprotein IV, is a cell adhesion molecule expressed by several cell types and in particular has been linked to platelet activation in response to oxidized low-density lipoprotein (28). Moreover, studies in mice have shown that CD36 may be involved in recruiting microparticles into thrombi (29) and that knocking out CD36 protects against a prothrombotic phenotype (28). This is corroborated by how the stop-gain variant identified in this study leads to lower FVIII levels. While this variant was less frequent in VTE cases, the association was not significant [odds ratio: 0.863 (95% CI: 0.566, 1.316)], perhaps due to lack of power.

von Willebrand factor

Among the five loci significantly associated with vWF, there were two common variants that were not associated with FVIII. The Ser42Thr (rs17564) variant in *STX2* is near an intronic *STX2* variant (rs7978987) reported to be associated with vWF in (7). Similarly, Asn436Ser (rs1039084) in *STXBP5* is located near synonymous rs9390459 in *STXBP5* in (7). As with FVIII, the association at the *ABO* locus for vWF can be explained after taking into account variants tagging the major ABO blood types. Conditional analyses at VWF revealed independent significant variants, all of which are known (24) except for the novel rare missense variant Tyr1584Cys (rs1800386) that was also associated with FVIII. Assessment of the significant gene-based test for *VWF* revealed that there were a large number of rare functional variants ($n = 349$), including three that were gene-wide significant: Tyr1584Cys, Arg2287Trp (rs61750625) and Ser1486Leu (rs149424724), which is seen only in AAs. Associations with Arg2287Trp and Ser1486Leu have been reported previously (24), while Tyr1584Cys is novel. The most recent meta-analysis of vWF levels found that the independent common variants explained 21.3% of the variance of the trait (15). The independent rare variants in our study explained 1.0% of the variance.

Comparison to findings from the exome chip

As expected, we were able to identify new associations with variants that were not on the exome chip ($n = 6$; see Supplementary Material, Table S9a), all of which were exclusive or almost exclusive to AAs. This confirms the expected bias of the exome chip content toward variants from those with EA. Of the 12 000 sequenced individuals that were used to design the exome chip, only ∼ 15% had African, Asian or Hispanic ancestry. Of the 18 rare variants found to be significantly associated with a hemostatic factor in this study, 12 were on the exome chip, and five of them were reported here for the first time. One of these five novel variants [*VWF* Tyr1584Cys

(rs1800386); exm976674] was dropped from the exome chip project during the quality assessment due to diffuse genotype clusters (see Supplementary Material, Fig. S3). Another was called as monomorphic by the Illumina software because the small heterozygote cluster was overlapping the reference homozygote cluster [F7 Ala354Val (rs36209567); exm1081312; see Supplementary Material, Fig. S4]. The other three variants had similar *P*-values and effect sizes as in the exome chip study (see Supplementary Material, Table S9a) but were slightly below the significance level for that study.

Strengths and limitations

To maximize power to identify new associations, we used all available samples at the time with whole-exome sequencing and these phenotypes in the primary analyses. As the cost of sequencing continues to drop, it is likely that other studies with these hemostasis phenotypes will generate sequence data, which will allow for replication and future meta-analyses. The inclusion of a large number of AAs has allowed us to identify ancestry-specific variants that would not have been possible with European samples alone. This is one particular advantage over analyses of the exome chip, which was designed primarily using data from White, non-Hispanic samples. In the future, sequencing Asian and Hispanic samples will likely identify additional rare variants associated with hemostatic factors. Whole-genome sequencing will do the same for regulatory regions outside of the exome.

Conclusion

Using a large biracial sample of individuals with whole-exome sequencing data (total *n* > 14 000), we have extended prior studies and identified eight novel associations between hemostatic factors and rare variants in *F7* and *VWF*. While these variants are in genes where common variation is known to be associated with the same trait, the discoveries herein from whole-exome sequencing identify novel independent signals with generally much larger effect sizes than previously reported and explain an appreciable proportion of the variance in these traits. In addition, we have also identified and replicated a novel association between a stop-gain mutation in the *CD36* gene and lower FVIII levels. This study validates the use of exome sequencing to identify novel variation associated with disease endophenotypes.

Materials and Methods
Study subjects and hemostatic factor measurements

Exome sequencing data came from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (30) and from the National Heart, Lung and Blood Institute's ESP. Descriptions and ethnic composition of participating cohorts are found in the Supplementary Methods. Individuals from CHARGE came from four population-based cohorts: the ARIC, CHS, Framingham Heart Study (FHS) and Rotterdam Study (RS). Participants in ESP were sampled from six population-based cohorts—ARIC, CHS, FHS, Coronary Artery Risk Development in Young Adults (CARDIA) study, Multi-Ethnic Study of Atherosclerosis (MESA) and Women's Health Initiative (WHI)—and do not overlap the CHARGE participants.

Detailed descriptions of each of the seven cohorts and the techniques used to measure hemostatic factor levels are provided in previous publications (31–43) and are summarized in the Supplementary Methods. Fibrinogen was available in all seven cohorts, FVII activity in six and FVIII activity or vWF antigen in five. Plasma levels of fibrinogen were measured in g/L, and FVII, FVIII and vWF were measured in international units (IU/dL, which are sometimes denoted as a percentage). All participants provided written informed consent as approved by local human subjects committees.

Exome sequencing and variant calling

Deoxyribonucleic acid (DNA) samples from ARIC, CHS and FHS participants from CHARGE were prepared using the same Baylor human genome sequencing center (HGSC) VCRome 2.1 design (44) (42 Mb, NimbleGen), sequenced, and called together. DNA samples from RS participants were prepared using Roche NimbleGen SeqCap v2 (44 Mb), and DNA samples from ESP participants were prepared using either Roche Nimblegen SeqCap EZ or Agilent SureSelect Human All Exon 50 Mb. All samples were paired-end sequenced (2 × 100 bp for the CHARGE cohorts and 2 × 76 bp for ESP) using Illumina GAII or HiSeq instruments. For more details on sequencing, variant calling and variant quality control, see Supplementary Methods. After taking into account available hemostatic factor measures, there were 8859 EA and 2664 AA samples from CHARGE and 2001 EA and 865 AA samples from ESP. Analyses were conducted using a total of 14 389 unique individuals (10 860 EA and 3529 AA) across CHARGE and ESP.

Annotation of whole-exome sequence variants

To facilitate meta-analysis between CHARGE and ESP, we created a combined variant annotation file that included all quality-controlled variant sites observed in CHARGE or ESP. Variants were annotated using ANNOtate VARiation (ANNOVAR) (45) and dbNSFP v2.0 (46) according to the reference genome GRCh37 and the National Center for Biotechnology Information RefSeq. The combined variant information file contained 6 605 975 unique sites, including the 2 706 509 sites that were polymorphic in the samples with hemostatic factors. A more detailed description of the annotation procedure is found in the Supplementary Methods.

Association analyses

Samples with extreme values for hemostatic factors (>3 SDs from the mean) were excluded from analyses to prevent spurious associations with rare variants. Final sample sizes for each of the four traits are summarized in Table 1. All four traits were natural log (ln) transformed, and the distributions for all studies were approximately normal following this transformation. Cohort-level analyses were carried out using the seqMeta R package. Data from the CHARGE cohorts (ARIC, CHS, FHS and RS) were each analyzed separately, while the five cohorts that make up ESP were included in a single pooled analysis. Fixed-effect inverse-variance weighted meta-analyses of single-variant and gene-based tests were conducted using seqMeta for ancestry-specific results as well as trans-ethnic analyses.

Single-variant testing

The first test of association was for individual variants for which the MAC was at least 40 after combining across cohorts (MAC = 40 translates to MAF > 0.0014 for this meta-analysis). There was no MAC threshold for individual studies before they were meta-analyzed. This was an a priori-determined threshold that was designed to reduce the chance for false positive associations caused by extreme phenotypic values as well as to reduce the number of tests performed, thereby increasing power to detect true associations. Within each meta-analysis group (trans-ethnic, EA, AA), we tested for single-variant association with hemostatic factor levels by linear regression with an additive genetic model adjusting for age, sex and ancestry-specific PCs. The trans-ethnic analysis was the primary approach, and an association was considered to be study-wide significant at $P < 1.9 \times 10^{-7}$ for single variants given a Bonferroni correction for testing as many as 269 877 single-variant sites with MAC $\geq$ 40. Conditional analyses were conducted to establish statistical independence among identified variants using a Bonferroni correction for the number of variants in each gene region ($P < 0.05/\#$ of polymorphic variants tested). In this conditional analysis, the dosage values for one or more index variants were included in the model as covariates and the variant being tested was considered independent of the index variants if the conditional P-value exceeded this Bonferroni threshold. Since the goal of conditional analyses is to establish allelic heterogeneity in genes known to be associated with the trait, the MAC requirement was dropped to > 5 for all conditional analyses. The estimated proportion of total variance explained was computed using the formula from Shim et al. (47) (see specific formula here: https://doi.org/10.1371/journal.pone.0120758.s001).

Gene-based testing

We performed gene-based tests that included only rare and low-frequency variants (MAF < 0.05) annotated as stop-gain, stop-loss, splicing, missense or small insertion or deletion sites (indels). Using the seqMeta package, two gene-level tests were performed. The first was a burden test where all variants passing the previously mentioned filters were summed to generate a gene burden score (48,49). The second test was the SKAT (50), which analyzes the same variants as the burden test but has greater power when effects are in both directions and up-weights the contribution of rarer variants. All gene-level tests were adjusted for the same covariates as the single-variant test and required the gene to have a cumulative MAC of at least 40 after combining across all cohorts, similar to the single-variant tests. The trans-ethnic analysis was the primary approach, and an association was considered to be significant at $P < 1.5 \times 10^{-6}$, which is the Bonferroni-corrected significance threshold for two gene-based tests and the 17 150 qualifying genes (i.e. MAC > 40). A secondary analysis was performed which was limited to LOF variants (only stop-gain, stop-loss, splicing and indels).

Association with VTE

Each of the 17 rare variants identified by our analysis was tested for association with VTE using a Cox proportional hazards model in ARIC and CHS. There were 214 incident VTE events among 3159 AAs and 383 incident VTE events among 9523 EA individuals. Analyses were run in R using the coxph function in the Survival package with age, sex, center, ancestry-specific PCs and the variant to be examined as the predictors.

Association with other phenotypes

The stop-gain variant driving the association in *CD36* for FVIII levels was previously reported to be associated with HDL and lower triglycerides (20). We attempted to replicate these associations using the ARIC AAs. We used linear regression with an additive genetic model that controlled for age, sex, center and PCs. To test for confounding, we reran the regression model for the CD36 stop-gain variant with FVIII levels after adding three additional covariates: HDL, triglycerides and BMI.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

Authors' contribution

N.P., P.S.dV., J.E.H., A.P.R., N.L.S. and A.C.M. designed the research. N.P., E.B., M.C., A.R.F., B.M.P., W.T., A.P.R. and A.C.M. performed the research. E.B., M.C., M.P.M.dM., A.R.F., O.H.F., A.H., R.K., D.M., B.M.P., A.G.U., J.G.J.vR. and C.J.O. collected data. N.P., P.W., J.A.B., M.C., P.S.dV., M.R.S., P.L.A., M.P.M.dM., O.H.F., R.A.G., K.K.H., A.H., J.M.J., C.L.K., R.K., B.M., W.T., A.G.U., J.G.J.vR., A.D., C.J.O., A.P.R., N.L.S. and A.C.M. analyzed and interpreted data. N.P., P.W., J.A.B., M.C., P.S.dV., M.R.S., A.D. and A.C.M. performed statistical analysis. N.P., J.E.H., M.R.S., K.K.H., N.L.S. and A.C.M. wrote the manuscript. All authors were given the opportunity to comment and provide revisions to the manuscript.

References

1. Collaboration, F.S., Kaptoge, S., White, I.R., Thompson, S.G., Wood, A.M., Lewington, S., Lowe, G.D.O. and Danesh, J. (2007) Associations of plasma fibrinogen levels with established cardiovascular disease risk factors, inflammatory markers, and other characteristics: individual participant meta-analysis of 154,211 adults in 31 prospective studies: the fibrinogen studies collaboration. *Am. J. Epidemiol.*, **166**, 867–879.

2. Collaboration, F.S., Danesh, J., Lewington, S., Thompson, S.G., Lowe, G.D.O., Collins, R., Kostis, J.B., Wilson, A.C., Folsom, A.R., Wu, K. *et al.* (2005) Plasma fibrinogen level and the risk of major cardiovascular diseases and nonvascular mortality: an individual participant meta-analysis. *JAMA*, **294**, 1799–1809.

3. Freeman, M.S., Mansfield, M.W., Barrett, J.H. and Grant, P.J. (2002) Genetic contribution to circulating levels of hemostatic factors in healthy families with effects of known genetic polymorphisms on heritability. *Arterioscler. Thromb. Vasc. Biol.*, **22**, 506–510.

4. Souto, J.C., Almasy, L., Borrell, M., Garí, M., Martínez, E., Mateo, J., Stone, W.H., Blangero, J. and Fontcuberta, J. (2000) Genetic determinants of hemostasis phenotypes in Spanish families. *Circulation*, **101**, 1546–1551.

5. de Lange, M., Snieder, H., Ariëns, R.A., Spector, T.D. and Grant, P.J. (2001) The genetics of haemostasis: a twin study. *Lancet*, **357**, 101–105.

6. Dehghan, A., Yang, Q., Peters, A., Basu, S., Bis, J.C., Rudnicka, A.R., Kavousi, M., Chen, M.-H., Baumert, J., Lowe, G.D.O. *et al.* (2009) Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. *Circ. Cardiovasc. Genet.*, **2**, 125–133.

7. Smith, N.L., Chen, M.-H., Dehghan, A., Strachan, D.P., Basu, S., Soranzo, N., Hayward, C., Rudan, I., Sabater-Lleal, M., Bis, J.C. *et al.* (2010) Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation*, **121**, 1382–1392.

8. Taylor, K.C., Lange, L.A., Zabaneh, D., Lange, E., Keating, B.J., Tang, W., Smith, N.L., Delaney, J.A., Kumari, M., Hingorani, A. *et al.* (2011) A gene-centric association scan for Coagulation Factor VII levels in European and African Americans: the Candidate Gene Association Resource (CARe) Consortium. *Hum. Mol. Genet.*, **20**, 3525–3534.

9. De Visser, M.C.H., Sandkuijl, L.A., Lensen, R.P.M., Vos, H.L., Rosendaal, F.R. and Bertina, R.M. (2003) Linkage analysis of factor VIII and von Willebrand factor loci as quantitative trait loci. *J. Thromb. Haemost.*, **1**, 1771–1776.

10. Keightley, A.M., Lam, Y.M., Brady, J.N., Cameron, C.L. and Lillicrap, D. (1999) Variation at the von Willebrand factor (vWF) gene locus is associated with plasma vWF:Ag levels: identification of three novel single nucleotide polymorphisms in the vWF gene promoter. *Blood*, **93**, 4277–4283.

11. Sabater-Lleal, M., Huang, J., Chasman, D., Naitza, S., Dehghan, A., Johnson, A.D., Teumer, A., Reiner, A.P., Folkersen, L., Basu, S. *et al.* (2013) Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. *Circulation*, **128**, 1310–1324.

12. Huffman, J.E., de Vries, P.S., Morrison, A.C., Sabater-Lleal, M., Kacprowski, T., Auer, P.L., Brody, J.A., Chasman, D.I., Chen, M.-H., Guo, X. *et al.* (2015) Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF. *Blood*, **126**, e19–e29.

13. de Vries, P.S., Chasman, D.I., Sabater-Lleal, M., Chen, M.-H., Huffman, J.E., Steri, M., Tang, W., Teumer, A., Marioni, R.E., Grossmann, V. *et al.* (2016) A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. *Hum. Mol. Genet.*, **25**, 358–370.

14. de Vries, P.S., Sabater-Lleal, M., Chasman, D.I., Trompet, S., Ahluwalia, T.S., Teumer, A., Kleber, M.E., Chen, M.-H., Wang, J.J., Attia, J.R. *et al.* (2017) Comparison of HapMap and 1000 genomes reference panels in a large-scale Genome-Wide Association Study. *PLoS One*, **12**, e0167742.

15. Sabater-Lleal, M., Huffman, J.E., de Vries, P.S., Marten, J., Mastrangelo, M.A., Song, C., Pankratz, N., Ward-Caviness, C.K., Yanek, L.R., Trompet, S. *et al.* (2019) Genome-wide association transethnic meta-analyses identifies novel associations regulating coagulation factor VIII and von Willebrand factor plasma levels. *Circulation*, **139**, 620–635.

16. de Vries, P.S., Sabater-Lleal, M., Huffman, J.E., Marten, J., Song, C., Pankratz, N., Bartz, T.M., de Haan, H.G., Delgado, G.E., Eicher, J.D. *et al.* (2019) A genome-wide association study identifies new loci for factor VII and implicates factor VII in ischemic stroke etiology. *Blood*, **133**, 967–977.

17. Wassel, C.L., Lange, L.A., Keating, B.J., Taylor, K.C., Johnson, A.D., Palmer, C., Ho, L.A., Smith, N.L., Lange, E.M., Li, Y. *et al.* (2011) Association of genomic loci from a cardiovascular gene SNP array with fibrinogen levels in European Americans and African-Americans from six cohort studies: the Candidate Gene Association Resource (CARe). *Blood*, **117**, 268–275.

18. Tang, W., Stimson, M.R., Basu, S., Heckbert, S.R., Cushman, M., Pankow, J.S., Folsom, A.R. and Pankratz, N. (2020) Burden of rare exome sequence variants in PROC gene is associated with venous thromboembolism: a population-based study. *J. Thromb. Haemost.*, **18**, 445–453.

19. Cushman, M., Tsai, A.W., White, R.H., Heckbert, S.R., Rosamond, W.D., Enright, P. and Folsom, A.R. (2004) Deep vein thrombosis and pulmonary embolism in two cohorts: the longitudinal investigation of thromboembolism etiology. *Am. J. Med.*, **117**, 19–25.

20. Love-Gregory, L., Sherva, R., Sun, L., Wasson, J., Schappe, T., Doria, A., Rao, D.C., Hunt, S.C., Klein, S., Neuman, R.J. *et al.* (2008) Variants in the CD36 gene associate with the metabolic syndrome and high-density lipoprotein cholesterol. *Hum. Mol. Genet.*, **17**, 1695–1704.

21. Brennan, S.O., Fellowes, A.P., Faed, J.M. and George, P.M. (2000) Hypofibrinogenemia in an individual with 2 coding (gamma82 A–>G and Bbeta235 P–>L) and 2 noncoding mutations. *Blood*, **95**, 1709–1713.

22. Wyatt, J., Brennan, S.O., May, S. and George, P.M. (2000) Hypofibrinogenaemia with compound heterozygosity for two gamma

chain mutations - gamma 82 Ala–>Gly and an intron two GT–>AT splice site mutation. *Thromb. Haemost.*, **84**, 449–452.

23. Antoni, G., Oudot-Mellakh, T., Dimitromanolakis, A., Germain, M., Cohen, W., Wells, P., Lathrop, M., Gagnon, F., Morange, P.-E. and Tregouet, D.-A. (2011) Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. *BMC Med. Genet.*, **12**, 102.

24. Johnsen, J.M., Auer, P.L., Morrison, A.C., Jiao, S., Wei, P., Haessler, J., Fox, K., McGee, S.R., Smith, J.D., Carlson, C.S. *et al.* (2013) Common and rare von Willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in African Americans: the NHLBI Exome Sequencing Project. *Blood*, **122**, 590–597.

25. Tang, W., Cushman, M., Green, D., Rich, S.S., Lange, L.A., Yang, Q., Tracy, R.P., Tofler, G.H., Basu, S., Wilson, J.G. *et al.* (2015) Gene-centric approach identifies new and known loci for FVIII activity and VWF antigen levels in European Americans and African Americans. *Am. J. Hematol.*, **90**, 534–540.

26. Fry, A.E., Ghansa, A., Small, K.S., Palma, A., Auburn, S., Diakite, M., Green, A., Campino, S., Teo, Y.Y., Clark, T.G. *et al.* (2009) Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. *Hum. Mol. Genet.*, **18**, 2683–2692.

27. Love-Gregory, L., Sherva, R., Schappe, T., Qi, J.-S., McCrea, J., Klein, S., Connelly, M.A. and Abumrad, N.A. (2011) Common CD36 SNPs reduce protein expression and may contribute to a protective atherogenic profile. *Hum. Mol. Genet.*, **20**, 193–201.

28. Podrez, E.A., Byzova, T.V., Febbraio, M., Salomon, R.G., Ma, Y., Valiyaveettil, M., Poliakov, E., Sun, M., Finton, P.J., Curtis, B.R. *et al.* (2007) Platelet CD36 links hyperlipidemia, oxidant stress and a prothrombotic phenotype. *Nat. Med.*, **13**, 1086–1095.

29. Ghosh, A., Li, W., Febbraio, M., Espinola, R.G., McCrae, K.R., Cockrell, E. and Silverstein, R.L. (2008) Platelet CD36 mediates interactions with endothelial cell-derived microparticles and contributes to thrombosis in mice. *J. Clin. Invest.*, **118**, 1934–1943.

30. Psaty, B.M., O'Donnell, C.J., Gudnason, V., Lunetta, K.L., Folsom, A.R., Rotter, J.I., Uitterlinden, A.G., Harris, T.B., Witteman, J.C.M., Boerwinkle, E. *et al.* (2009) Cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ. Cardiovasc. Genet.*, **2**, 73–80.

31. (1989) The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. The ARIC investigators. *Am. J. Epidemiol.*, **129**, 687–702.

32. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B. and Newman, A. (1991) The Cardiovascular Health Study: design and rationale. *Ann. Epidemiol.*, **1**, 263–276.

33. Feinleib, M., Kannel, W.B., Garrison, R.J., McNamara, P.M. and Castelli, W.P. (1975) The Framingham Offspring Study. Design and preliminary data. *Prev. Med.*, **4**, 518–525.

34. Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N. and Stokes, J., 3rd (1961) Factors of risk in the development of coronary heart disease–six year follow-up experience. The Framingham Study. *Ann. Intern. Med.*, **55**, 33–50.

35. Hofman, A., Grobbee, D.E., de Jong, P.T. and van den Ouweland, F.A. (1991) Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. *Eur. J. Epidemiol.*, **7**, 403–422.

36. Hofman, A., Breteler, M.M.B., van Duijn, C.M., Janssen, H.L.A., Krestin, G.P., Kuipers, E.J., Stricker, B.H.C., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R. *et al.* (2009) The Rotterdam Study: 2010 objectives and design update. *Eur. J. Epidemiol.*, **24**, 553–572.

37. Hofman, A., Darwish Murad, S., van Duijn, C.M., Franco, O.H., Goedegebure, A., Ikram, M.A., Klaver, C.C.W., Nijsten, T.E.C., Peeters, R.P., Stricker, B.H.C. *et al.* (2013) The Rotterdam Study: 2014 objectives and design update. *Eur. J. Epidemiol.*, **28**, 889–926.

38. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.

39. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J. *et al.* (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.

40. Friedman, G.D., Cutter, G.R., Donahue, R.P., Hughes, G.H., Hulley, S.B., Jacobs, D.R., Jr., Liu, K. and Savage, P.J. (1988) CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J. Clin. Epidemiol.*, **41**, 1105–1116.

41. (1998) Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control. Clin. Trials*, **19**, 61–109.

42. Hays, J., Hunt, J.R., Hubbell, F.A., Anderson, G.L., Limacher, M., Allen, C. and Rossouw, J.E. (2003) The Women's Health Initiative recruitment methods and results. *Ann. Epidemiol.*, **13**, S18–S77.

43. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Jr., Kronmal, R., Liu, K. *et al.* (2002) Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol.*, **156**, 871–881.

44. Bainbridge, M.N., Wang, M., Wu, Y., Newsham, I., Muzny, D.M., Jefferies, J.L., Albert, T.J., Burgess, D.L. and Gibbs, R.A. (2011) Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol.*, **12**, R68.

45. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

46. Liu, X., Jian, X. and Boerwinkle, E. (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, **34**, E2393–E2402.

47. Shim, H., Chasman, D.I., Smith, J.D., Mora, S., Ridker, P.M., Nickerson, D.A., Krauss, R.M. and Stephens, M. (2015) A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS One*, **10**, e0120758.

48. Li, B. and Leal, S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

49. Morris, A.P. and Zeggini, E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, **34**, 188–193.

50. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.