Genome-wide association study identifies first locus associated with susceptibility to cerebral venous thrombosis

Running head: Genetic susceptibility to cerebral venous thrombosis.

Gie Ken-Dror, PhD[1], Ioana Cotlarciuc, PhD[1], Ida Martinelli, MD, PhD[2], Elvira Grandone, MD, PhD[3-4], Sini Hiltunen, MD[5], Erik Lindgren, MD[6-7], Maurizio Margaglione, MD[8], Veronique Le Cam Duchez, MD, PhD[9], Aude Bagan Triquenot, MD[10], Marialuisa Zedde, MD[11], Michelangelo Mancuso, MD, PhD[12], Ynte M Ruigrok, MD[13], Thomas Marjot, MRCP[14], Brad Worrall, MD, MSc[15], Jennifer J Majersik, MD, MS[16], Tiina M. Metso, MD[5], Jukka Putaala, MD, PhD[5], Elena Haapaniemi, MD, PhD[5], Susanna M. Zuurbier, MD, PhD[17], Matthijs C. Brouwer, MD, PhD[17], Serena M. Passamonti, BSc[2], Maria Abbattista, MSc[2], Paolo Bucciarelli, MD[2], Braxton D. Mitchell, PhD, MPH[18-19], Steven J. Kittner, MD, MPH[20-21], Robin Lemmens, MD, PhD[22], Christina Jern, MD, PhD[23-24], Emanuela Pappalardo, MS[2], Paolo Costa, MD[25], Marina Colombi, PhD[12], Diana Aguiar de Sousa, MD, MSc[26], Sofia Rodrigues, MD[26], Patrícia Canhão, MD, PhD[26], Aleksander Tkach, MD[16], Rosa Santacroce, MD[8], Giovanni Favuzzi, MD[3], Antonio Arauz, MD, MSc[27], Donatella Colaizzo, BSc[3], Kostas Spengos, MD, PhD[28], Amanda Hodge, MSc[29], Reina Ditta, MSc[29], Alessandro Pezzini, MD[12], Stephanie Debette, MD, PhD[30], Jonathan M. Coutinho, MD, PhD[17], Vincent Thijs, PhD[31], Katarina Jood, MD[6-7], Guillaume Pare, MD, MSc[29], Turgut Tatlisumak, MD, PhD[5-6-7], José M. Ferro, MD, PhD[26], and Pankaj Sharma, MD, PhD[1-32*]; on behalf of the International Stroke Genetics Consortium (ISGC) & Bio-Repository to Establish the Aetiology of Sinovenous Thrombosis (BEAST) collaborators.

[1]Institute of Cardiovascular Research Royal Holloway, University of London (ICR2UL), London, UK; [2]A. Bianchi Bonomi Hemophilia and Thrombosis Center, Fondazione IRCCS Ca'Granda – Ospedale Maggiore Policlinico, Milan, Italy; [3]Atherosclerosis and Thrombosis Unit, I.R.C.C.S. Casa Sollievo della Sofferenza, S. Giovanni Rotondo, Foggia, Italy; [4]Ob/Gyn Dept. The First I.M. Sechenov Moscow State Medical University; [5]Neurology, Helsinki University Hospital and University of Helsinki, Helsinki, Finland; [6]Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden; [7]Department of Neurology, Sahlgrenska University Hospital, Gothenburg, Sweden; [8]Medical Genetics, Dept. of Clinical and Experimental Medicine, University of Foggia, Italy; [9]Normandie University, UNIROUEN, INSERM U1096, Rouen University Hospital, Vascular Hemostasis Unit and Inserm CIC-CRB 1404, F 76000 Rouen, France; [10]Rouen University Hospital, Department of Neurology, F 76000 Rouen, France; [11]Neurology Unit, Stroke Unit, Azienda Unità Sanitaria Locale-IRCCS di Reggio Emilia, Italy; [12]Department of Molecular and Translational Medicine, Division of Biology and Genetics, University of Brescia, Italy; [13]UMC Utrecht Brain Center, Department of Neurology and Neurosurgery, University Medical Center Utrecht, the Netherlands; [14]Oxford Liver Unit, Translational Gastroenterology Unit, Oxford University Hospitals NHS Foundation Trust, Oxford, UK; [15]Department of Neurology, University of Virginia, Charlottesville, VA, USA. [16]Department of Neurology, University of Utah, Salt Lake City, UT, USA; [17]Department of Neurology, Amsterdam University Medical Centers, location AMC, Amsterdam Neuroscience, University of Amsterdam, the Netherlands; [18]Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA; [19]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD, USA; [20]Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA; [21]Department of Neurology, Veterans Affairs Medical Center,

Baltimore, MD, USA; [22]KU Leuven – University of Leuven, Department of Neurosciences, Experimental Neurology; VIB Center for Brain & Disease Research; University Hospitals Leuven, Department of Neurology, Leuven, Belgium; [23]Department of Laboratory Medicine, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Sweden; [24]Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Gothenburg, Sweden; [25]Department of Clinical and Experimental Sciences, Neurology Clinic, University of Brescia, Italy; [26]Department of Neurosciences, Hospital de Santa Maria, University of Lisbon, Lisbon, Portugal; [27]Stroke Clinic, National Institute of Neurology and Neurosurgery Manuel Velasco Suarez, Mexico City, Mexico; [28]Department of Neurology, University of Athens School of Medicine, Eginition Hospital, Athens, Greece; [29]McMaster University, Pathology and Molecular Medicine. Population Health Research Institute and Thrombosis and Atherosclerosis Research Institute, Hamilton Health Sciences, Ontario, Canada; [30]Department of Neurology, Bordeaux University Hospital, Bordeaux University, France; [31]Stroke Division, Florey Institute of Neuroscience and Mental Health, University of Melbourne, Heidelberg, Victoria, Australia; [32]Department of Clinical Neuroscience, Imperial College Healthcare NHS Trust, London.

Corresponding author: Pankaj Sharma MD PhD FRCP

Institute of Cardiovascular Research, Royal Holloway University of London (ICR2UL) TW20 0EX UK. Tel: 01784443807.

*Email: pankaj.sharma@rhul.ac.uk

Number of characters in the title: 101 characters; Running head 47 characters.

Number of words: Abstract 250 words; Introduction 253 words; Discussion 1050 words; Body of the manuscript 4129 words; 5 figures; 1 table; 1 supplementary material file.

Abbreviations: CVT, cerebral venous thrombosis; GWAS, genome-wide association study; SNP, Single Nucleotide Polymorphisms; LD, linkage disequilibrium; MAF, minor allele frequency; OR, odds ratios; CI, confidence interval; SE, standard error; PRS, polygenetic risk score; JAM, joint analysis of marginal; PC, principal component.

Summary for Social Media

What is the current knowledge on the topic?

Cerebral venous thrombosis (CVT) is an uncommon form of stroke affecting mostly young individuals and likely to be influenced by common genetic variants.

What question did this study address?

Is there a genetic aetiology to cerebral venous thrombosis (CVT)?

What does this study add to our knowledge?

The *ABO* gene is a strong candidate for CVT risk. We further demonstrate a 5.6 times increased risk of CVT in those with an AB blood group.

How might this potentially impact on the practice of neurology?

These findings provide robust evidence for a genetic basis for CVT and provide new insights into the pathophysiology of this important but previously largely neglected disease.

Abstract

Objective: Cerebral venous thrombosis (CVT) is an uncommon form of stroke affecting mostly young individuals. Although genetic factors are thought to play a role in this cerebrovascular condition, its genetic etiology is not well understood.

Methods: Genome-wide association study performed to identify genetic variants influencing susceptibility to CVT. A two-stage genome-wide study was undertaken in 882 Europeans diagnosed with CVT and 1205 ethnicity-matched control subjects divided into discovery and independent replication datasets.

Results: In the overall case-control cohort, we identified highly significant associations with 37 SNPs within 9q34.2 region. The strongest association was with rs8176645 (combined $P=9.15\times10^{-24}$; OR=2.01, 95%CI: 1.76-2.31). The discovery set findings were validated across an independent European cohort. Genetic risk score for this 9q34.2 region increases CVT risk by a pooled estimate OR=2.65 (95%CI: 2.21-3.20, $P=2.00\times10^{-16}$). SNPs within this region were in strong linkage disequilibrium (LD) with coding regions of the *ABO* gene. *ABO* blood group was determined using allele combination of SNPs rs8176746 and rs8176645. Blood groups A, B or AB, were at 2.85 times (95%CI: 2.32–3.52, $P=2.00\times10^{-16}$) increased risk of CVT compared with individuals with blood group-O.

Interpretation: We present the first chromosomal region to robustly associate with a genetic susceptibility to CVT. This region more than doubles the likelihood of CVT, a risk greater than any previously identified thrombophilia genetic risk marker. That the identified variant is in strong LD with the coding region of the *ABO* gene with differences in blood group prevalence provides important new insights into the pathophysiology of CVT.

Introduction

Cerebral venous thrombosis (CVT) is an uncommon cerebrovascular condition mainly affecting young people and accounting for less than 1% of all stroke cases[1,2]. Incidence rates range from 1.32-1.57 cases per 100,000 persons-years with 3.7–5.3 times greater rate in females[1,3] and is associated with a high mortality rate of 10%-35%[4,5]. The prevalence of CVT in high-income countries is 1.3 to 1.6 per 100,000 persons and is higher in low- and middle-income countries[6]. Numerous risk factors have been reported for adult CVT including oral contraceptive use, pregnancy, head and neck infections, obesity, anaemia, hematologic diseases, antiphospholipid syndrome, inflammatory bowel disease, high altitude, tamoxifen use, erythropoietin, phytoestrogens, and thalidomide use in multiple myeloma[3,7].

The genetic component of CVT has only been assessed using candidate gene studies with an *a priori* hypothesis[8]. CVT, being an unusual form of stroke, is likely to be influenced by rare genetic variants with large effects making their likelihood of identification higher compared to common types of sporadic ischemic stroke[9]. Further, as CVT is a clinically homogenous form of stroke, subtyping issues known to cause problems in sporadic stroke are less burdensome[9]. No single genome-wide association study (GWAS) has been performed to discover CVT genetic marker susceptibility, although about a dozen studies have been conducted for the most common forms of venous thromboembolism[10], deep vein thrombosis[11] and pulmonary embolism[12].

We present the first GWA study in CVT utilizing a two-stage design to identify susceptibility loci in CVT and provide evidence for involvement of a candidate gene within the identified chromosomal region.

Methods

Study Populations

An international collaboration (Bio-Repository to Establish the Aetiology of Sinovenous Thrombosis, BEAST) was established from 11 research centres (Belgium, Finland, Greece, Italy, Mexico, the Netherlands, Portugal, Sweden, France, UK, and USA). The genetic analysis was undertaken using data from 9 countries to maximise homogeneity (Mexican population substructure was significantly different while the Greek sample set was very small). Detailed description of the BEAST protocol has been published elsewhere[13].

A total of 405 European CVT patients and 434 control subjects were recruited as a discovery set and a second independent European set of 477 patients and 771 control subjects to replicate the findings (Table 1).

In all cases extensive clinical phenotyping was undertaken, as previously published[13]. Briefly, CVT diagnosis was confirmed by computed tomography (CT) or magnetic resonance (MR) brain imaging and dedicated venography CTA (computed tomography angiography), MRA (magnetic resonance angiography) or conventional angiogram. Extensive blood investigations were undertaken, as previously described[13]. All recruits were age≥18 years with patient or relative informed written consent. Age was documented at time of CVT diagnosis.

The inclusion criteria for the control population were age≥18 years at the time of enrolment, no previous history of CVT/stroke or any other thrombotic or chronic condition, not pregnant and no history of malignancy or any autoimmune disorder. Informed written consent was obtained in all recruits.

The study meets all ethical and consent standards set by local institutional review boards at each of the participating sites.

Genotyping and Quality Control Procedures

DNA samples for all CVT cases and controls were genotyped on the HumanCoreExome BeadChip v1.0 (Illumina, Inc., San Diego, CA) using standard protocols[14] at the Genetic and Molecular Epidemiology Laboratory, McMaster University, Canada.

The Illumina Infinium HumanCoreExome BeadChip contains functional exome markers include non-synonymous variants, stop altering variants, splice coding variants and variants located in promoter regions (>550,000 SNPs). In addition, it includes common tagSNP markers. The replication (Finnish and Sweden) data was genotyped using this chip. However, the later recruited replication samples included (Netherlands, USA, Portugal, French, Italy) data which was genotyped using the Infinium Global Screening Array BeadChip from Illumina containing 665,608 SNPs focused on genome-wide tag-SNP variant because of changes in commercial chip technology and availability. Genotype imputation conducts Michigan Imputation Server[15] to impute polymorphic SNPs that were not covered by the Illumina Infinium HumanCoreExome BeadChip. The reference panel used for imputation was the Haplotype Reference Consortium panel (HRC version-1.1) and run using Minimac4.

Genotype quality control was performed on samples and SNPs for the discovery and replication set separately. Samples were excluded for any of the following reasons: (i) a call rate of less than 98% (of total number of SNPs); (ii) evidence of non-European ancestry from principal components analysis (PCA) after combining with data from unrelated samples taken from four HapMap Phase-III populations: Utah residents with European ancestry (CEU), Yoruba from Ibadan, Nigeria with West African ancestry (YRI), Han Chinese (CHB), and Japanese with East Asian ancestry (JPT). Multidimensional scaling analysis[16] showed that the genotypes displayed by the European subjects intersect with those of CEU and are clearly unambiguous from JPT, CHB or YRI, according to International HapMap Project information; (iii) sex discrepancy checks (inbreeding coefficient-*F*) confirmed by genotyping that did not match the reported sex; (iv) evidence of a first-degree relationship or identity with another

sample and in this case the sample with a lower call rate of pair (identity by descent (IBD) pi-hat >0.1875) was excluded. Among the discovery set excluded 359 samples and among the replication set excluded 487 samples as they did not meet our above criteria. Regions of interest were pre-defined using UCSC genome browser GRCh37/hg19 (https://genome.ucsc.edu/cgi-bin/hgGateway). SNPs were excluded for any of the following reasons: (i) significant deviations from Hardy-Weinberg equilibrium (HWE) *P*-value<10^{-6}; (ii) call rate <98%; (iii) minor allele frequency (MAF)<5%. Among the discovery set excluded 9,869,083 SNPs and among the replication set excluded 6,244,867 SNPs.

A total of 6,041,026 autosomal SNPs in 405 cases and 434 controls were interrogated within the discovery set and, 3,981,455 autosomal SNPs in 477 cases and 771 controls were interrogated within the replication set. All genotyped and imputed SNPs were analysed.

Statistical Analysis

The association between SNP markers and disease susceptibility were calculated in the discovery and replication dataset using logistic regression to estimate per-allele odds ratios (ORs) and 95% confidence intervals (CIs) under an additive model (SNPs coded 0, 1 or 2 with respect to minor allele dosage) adjusted for age, sex, and the three first eigenvectors from principal components (PCs). To account for potential population stratification in the discovery set and the replication set, the first three PCs were included as covariates in the logistic regression model. To be more conservative, the analysis of the most significant SNPs was repeated including the first five PCs but had no effect on the results. The genomic inflation factor (λ) was calculated based on median χ^2 statistic. Manhattan charts were constructed using the negative logarithm at the base 10 of the *P*-values, abbreviated as -log10(*P*) via the physical map. SNPs with -log10(*P*)>8 were selected for further characterization. Linkage disequilibrium (LD) block constructed for significant SNP markers within a chromosome using

HAPLOVIEW v4.2[17] to assign markers to short blocks. The Haploview Tagger function (based on analysis of marker pairwise R^2 values) was used to select tag-SNPs with a tagger filter set at R^2=0.8 in HAPLOVIEW v4.2[17].

Primary evidence for secondary effects was assessed at each site using forward stepwise logistic regression. The highest order SNP in the region was included as a covariate and association statistics were recalculated for the remaining test SNPs. This process was iterated until no remaining SNPs reached a minimum level of significance. Independent effects were defined as a *P*-value $<5\times10^{-4}$, not closely correlated with the highest-ranking SNP, and the conditional *P*-value not substantially different from the unconditioned value. Next, it was tested whether the two-SNPs fit the risk in-situ significantly better than the single-SNP model using the likelihood ratio test. To confirm the stepwise results and identify secondary effects a second micro-mapping approach used JAM (Joint Analysis of Marginal Summary Statistics)[18]. JAM is a multivariate Bayesian variable selection framework that uses GWAS summary statistics to determine the potential number of independent associations within a site and to identify reliable sets of variables that drive these associations. For further verification of the independence of the selected loci, additional approximate conditional analyses were performed joint association method using the genome-wide complex trait analysis (GCTA-COJO)[19]. Linkage disequilibrium score regression (LDSC)[20] was used to estimate heritability by regressing summary statistics LD scores from our GWAS result. LD was characterized as a non-random association of alleles (alternative genetic variants at the same genomic locus) between differing genetic loci[21].

Haplotypes reconstructed for individual samples were estimated only for variants from a significant critical threshold (-log10(*P*)>8) by Bayesian estimation use of PHASE software (version 2.1.1; University of Washington, Seattle)[22] the algorithm that proved most accurate. A permutation test for assessing significant differences in haplotype frequencies between case

and control was performed by use of the PHASE program. Only haplotypes with an estimated frequency ≥0.5% were tested. The estimated heritability was calculated as the proportion of phenotype variance due to additive genotype that were measured in this study using GREML analysis in GCTA[23].

The heterogeneity of the ORs across the studies was estimated using I^2 and Cochran's Q tests[24,25]. The marginal risk estimates of odds ratios and standard errors were combined across the discovery set and the replication set using a fixed effects model for the meta-analysis except if there was evidence of heterogeneity between the study arrays (I^2>31%), a random-effects model was used in each case.

The four major blood groups are determined by two SNPs in the *ABO* gene: rs8176746 and rs8176719 the top-ranking SNP which is in strong LD (R^2=0.98) with rs8176719 located within the *ABO* gene[26]. The rs8176719 is a G deletion which generates a premature termination codon and is recessive for O blood group. The rs8176746 is non-synonymous polymorphisms determining the B blood group and the A allele, which changes a leucine to methionine amino acid.

Results

A total of 882 CVT patients and 1205 control subjects of European descent were enrolled in the study to identify common genetic variants associated with CVT. The demographic characteristics of individuals with CVT among the discovery dataset is presented in Table 1. CVT cases and controls were appropriately age and sex matched. No significant differences were observed in CVT patients and controls among thrombophilia genotype risk factors of lupus anticoagulant, antiphospholipid antibodies, protein C deficiency, protein S deficiency, antithrombin deficiency, homocysteine plasma levels and factor VIII.

In genome-wide association analysis within the European discovery set (n=882), a total of 37 SNPs (Figure 1) exceeded genome-wide significance with -log10(*P*)>8 using logistic regression model adjusted for age, sex and top three PCs as covariates (the SNPs are listed in S2 Table). The top-ranked SNP was rs8176645 (OR=2.35; 95%CI: 1.88-2.93, $P=5.02\times10^{-14}$) followed by 35 SNPs in strong linkage disequilibrium (LD, $r^2\geq0.80$). All these SNPs (S2 Table) locate within a 13-kilobase region of 9q34.2 encompassing all SNPs in the intronic region of the *ABO* gene. The Tagger function based on analysis of marker pairwise R^2 values was used to select tag-SNPs with a tagger filter set at $R^2=0.8$, 2 SNPs in 2 tests captured all 37 alleles at $r^2>0.8$ (mean max r^2 is 0.975).

The observed *P*-value distributions for association tests across all SNPs presents no evidence of a general systematic bias ($\lambda=1.04$) from the expected *P*-values, and the increase in lower *P*-values was consistent with true associations, indicating that the samples are genetically homogeneous and that any significant associations are attributable to genetic differences in CVT susceptibility.

Following a genome wide analysis, a replication set in an independent cohort from European (477 CVT cases and 771 controls) confirmed the 34 SNPs at 9q34.2, with three SNPs dropped following QC because of missing values during genotyping (Table S2). The

demographic characteristics of individuals with CVT among the replication dataset is presented in Table 1. The *P*-values of the SNPs ranged between 5.02×10^{-14} and 2.50×10^{-9} in the discovery set, 2.05×10^{-12} and 5.10×10^{-11} in the European replication set, 2.24×10^{-24} and 3.97×10^{-19} in the pooled set of European cohorts (Figure 2, 3 and S2 Table).

The directions of associations in the replication set were identical to those in the European discovery sets, and the associations were still significant after genome-wide Bonferroni correction, suggesting that the locus 9q34.2 is associated with susceptibility to CVT across multiple European populations. The comparison between the *P*-value of the discovery set and the replication set is presented (Figure 2). The 37 loci explained 1.63% of the variance in CVT, and heritability was estimated at 6.54% (SE=5.43) among discovery set, 1.12% of the variance and 4.79% (SE=4.73) among replication set. According to LD score regression there was no evidence of residual population stratification among the discovery and the replication set (intercept<1.019). The estimated h^2 to be 6.4% (SE=0.004, liability scale) among the discovery set, and 4.7% (SE=0.004, liability scale) among the replication set. There was no difference in the association between top ranking SNPs and disease among males and females in the discovery set (males: OR=2.53, 95%CI: 1.62-4.06, $P=7.38\times10^{-5}$; females: OR=2.35, 95%CI: 1.81-3.07, $P=1.82\times10^{-10}$) and the replication set (males: OR=1.71, 95%CI: 1.25-2.36, $P=9.00\times10^{-4}$; females: OR=1.92, 95%CI: 1.55-2.40, $P=4.06\times10^{-9}$).

Pooling the European discovery and replication data into one overall dataset resulted in an enhanced significance of the rs8176645 (*P*-values: 4.17×10^{-24}, Figure 2). Following this dataset pooling, a further signal at chromosome 4 (rs56810541) was seen with a *P*-value: 2.33×10^{-12}. However, this SNP did not reach genome wide significance when datasets were assessed separately. We present all SNPs with *P*-value ranging from -log10(*P*)>6 to -log10(*P*)<8 in Supplementary Table S4.

Regional association plots indicate the chromosome 9 locus is located near the 5′-end of the *ABO* gene (Figure 4). The $-\log_{10}(P)$ values abruptly dropped when they crossed a recombination hotspot located in the upstream region of *ABO* (Figure 4). The SNPs considered were not completely independent of each other with LD ranging from an r^2 value of 0.52–0.99 (Figure 4, bottom panel). Conditional analysis has been used as a tool to identify secondary association signals at a locus. We undertook a conditional analysis starting with our top associated SNP (rs8176645) found across the whole significant locus, followed by a stepwise procedure of selecting additional SNPs, one by one, according to their significant conditional *P*-values. Following this conditional logistic regression analysis, we find no secondary effects within the locus of interest in either discovery or replication datasets ($P>0.19$). A second micro-mapping approach, JAM (Joint Analysis of Marginal Summary Statistics), used GWAS summary statistics identifying reliable sets of variables that find independent association signals in regions of susceptibility. The 95% discovery cohort of JAM analysis confirmed the independent signal from the stepwise analysis except for rs8176645, where the evidence for an association was weak (Bayes' specific variable factor (BF)=21.66). In addition, no independent risk loci were identified using the approximate conditional and co-association method applied in GCTA (GCTA-COJO) among the discovery and replication cohort. To test the possibility that the number risk alleles tag an untyped SNP, haplotype analysis carried out of all significant SNPs but found no evidence for haplotype specific effects at any locus among discovery set and replication set. Haplotype analysis presents an additive effect of independent risk variants consistent with those expected in the single-variable test, and co-presence of 9q34.2 risk alleles on the same haplotype increases CVT risk (S1 Table).

The haplotype carrying the minor allele for rs8176645 (42.3% frequency in the discovery set and 44.6% in the replication set) is associated with an increasing risk of CVT (discovery: OR=2.10, 95%CI: 1.72-2.56, $P=2.01\times10^{-13}$; replication: OR=1.75, 95%CI: 1.49-2.07,

P=2.99×10^{-11}), suggesting that having the T allele confers a risk allele. The combinations of these SNPs together in a weighted genetic score increase the risk of an individual to CVT with OR=2.68 (95%CI: 2.10–3.46, P=1.15×10^{-14}) per increasing allele among discovery set, replication set (OR=2.79: 95%CI: 2.14–3.65, P=4.71×10^{-14}), and pooled estimates across the studies (OR=2.65: 95%CI: 2.21–3.20, P=2.00×10^{-16}).

Regional association plots of SNPs around the detected clusters of significant association peaks in the same regions are shown in Figure 4. Approximately 13 annotated genes, including potentially CVT related candidate genes of *ABO*, *OBP2B*, *SURF6*, *MED22*, *RPL7A*, and *SNORD24*, were located on or near these loci.

All the exceeded genome-wide significance SNPs locate in the intronic region of the *ABO* gene and are in strong LD with coding regions of the *ABO* gene. *ABO* blood group can be determined using allele combinations of SNP rs8176746 and rs8176719[27,28]. Only rs8176746 (MAF=10%) was directly genotyped on our array so rs8176719 was replaced by rs8176645 (MAF=42%) which is in strong LD[26] (R^2=0.98) with rs8176719 located within the *ABO* gene. Alleles were matched according to previous literature based on frequency. Figure 5 present the blood group distribution between CVT cases and controls.

The non-O blood groups (A, B or AB) of the cases was significantly higher than that of controls (82% vs. 57%, respectively, P=8.53×10^{-14}). A similar result was also observed in the replication set (82% vs. 63%, respectively, P=8.26×10^{-12}). The European discovery and replication sets were pooled, the non-O blood groups (A, B or AB) of the cases was significantly higher than that of controls (67%, 11%, 3% vs. 52%, 8%, 1%, respectively, P=2.20×10^{-16}). Blood group A associated with higher risk of CVT (OR=2.77, 95%CI: 2.25-3.44, P=2.20×10^{-16}), blood group B (OR=2.92, 95%CI: 2.08-4.09, P=4.64×10^{-10}) and blood group AB (OR=5.60, 95%CI: 2.96-11.01, P=2.18×10^{-7}), compared with blood group O adjusted for age, sex and 3 PCs. The frequency of the ABO blood groups in our control

population were within the expected range of previously published frequencies[28,29]. Further, we subsequently imputed rs8176719 and demonstrated very strong LD (r^2=0.9801 among discovery set, and r^2=0.9977 among replication set) with rs8176645 (i.e., the most significant SNP in our GWAS). The assessment of blood group results did not change following imputation.

The MAF of factor V Leiden mutation among CVT cases in the discovery and replication European pooled sets was higher than that of controls (1.2% vs. 0.4%, δ=0.8%, 95%CI: 0.26%-1.45%, P=2.90×10^{-3}) OR=2.94 (95% CI: 1.36–6.27, P=0.006), and a similar trend was also observed in Prothrombin G20210A mutation (1.9% vs. 0.2%, δ=1.7%, 95%CI: 0.76%-2.90%, P=5.00×10^{-4}) OR=7.65 (95% CI: 1.73–33.83, P=0.007), although the 9q34.2 locus was independently associated with CVT. The odds ratio for the most significant SNP (rs8176645) was enhanced (OR=2.38, 95%CI: 1.89-2.98, P=6.97×10^{-14}) in the Europeans discovery and replication sets after adjusted for these mutations.

No significant differences were observed between *ABO* blood groups and thrombophilia genotype risk factors of lupus anticoagulant, antiphospholipid antibodies, protein C deficiency, protein S deficiency, antithrombin deficiency, homocysteine plasma levels and factor VIII.

Discussion

We show that 9q34.2 is strongly associated with CVT with a pooled OR of 2.7, a result we validated in an additional independent cohort. There was evidence that the SNPs with the largest association were in strong LD with coding regions of the *ABO* gene. We go on to demonstrate that non-O blood group is more prevalent in CVT cases across multiple populations studied.

In European populations, 9q34.2 locus is associated with other comorbidities such as malaria[30], type 2 diabetes[31] and chronic obstructive pulmonary disease[32] and interstitial lung disease[33] and, importantly, with several diseases associated with clotting disorders such as pancreatic cancer[34], coronary artery disease[35], heart failure[36], ischemic stroke[37], thrombosis[11], venous thromboembolism[10]. However, with OR ranging from 1.06-1.81 the effect sizes are minor when compared with our new OR (2.68) finding for this locus, although direct comparisons of risks between differing populations remain challenging.

Several genes are located within a 1-Mb region of the locus 9q34.2, including Odorant Binding Protein 2B (*OBP2B)*, Surfeit locus protein 6 *(SURF6)*, Mediator of RNA polymerase-II transcription subunit-22 *(MED22)*, and 60S ribosomal protein-L7a (*RPL7A)* (Figure 2). These genes are included in the surfeit gene cluster, a group of very tightly linked genes that do not share sequence similarity. However, it is not known whether *OBP2B*, *SURF6*, *MED22*, and *RPL7A* are involved in thrombophilia. The chromosomal position of the SNP in 9q34.2 is found in strong LD with the coding sequence variant SNP rs8176719 which plays a key part in determining blood group status[27]. Since the present findings indicate that the SNPs of *ABO* may be involved in CVT development, it is assumed that *ABO* gene polymorphisms could also cause thrombophilia.

The association between venous thromboembolic disease and *ABO* blood type has been previously described[38] and recently reviewed[39] relating to CVT, and our results now provide a

potential genetic mechanism for that observation. As our results implicated the *ABO* gene with CVT we determined blood group characterization in our studied populations. Blood group phenotypes, based on inherited allelic combinations[27,28], and the *ABO* blood group system are thought to contribute to risk of developing thromboembolic diseases[28]. CVT cases had significant higher prevalence of non-O blood groups compared to controls. Compared with individuals with blood group O, Europeans with blood group A, B or AB, were at 2.9 times increased risk of CVT (OR=2.85, 95%CI: 2.32–3.52, $P=2.00\times10^{-16}$).

Notwithstanding our results, it is clear that a substantial fraction of the clinical response remains unmapped, raising questions as to the location of the missing heritability. However, even quite large GWA studies have only at best identified moderate proportions of the genetic variants contributing to disease heritability. Because of the relatively low incidence of CVT it is difficult to recruit a very large number of cases (>1000) for GWA. Despite this, we were able to perform a genome-wide analysis on the largest number of CVT patients ever assembled and validated the findings across independent European samples. We calculated a polygenetic risk score (PRS) of the known VTE SNPs[40] and found CVT was associated with an OR=1.35 (95% CI: 1.11–1.64, *P*=2.43e-03) among the discovery set and OR=1.61 (95% CI: 1.30–2.00, *P*=1.40e-05) among the replication set (Supplementary Table S5).

The analysis includes sample size of 882 patients in the discovery set and 1205 patients in the replication set resulting in limited statistical power to detect modest effects, such as those observed here, for the association between genetic variability and disease susceptibility. Our study can detect effects (OR) of 2.00 for any SNP with a MAF>20% but can only detect large effects of 3.10 from a SNP with <10% MAF.

Factor V Leiden (rs6025 or *F5* p.R506Q) is a variant which causes hypercoagulability. In addition, prothrombin G20210A (rs1799963) factor II mutation increases levels of the clotting factor prothrombin creating a greater tendency towards blood clotting[8,9]. There is variability in

the reports of the frequency of these variants for venous thrombosis depending on the location of that thrombosis. In VTE overall the reported frequency is small[41] but in CVT the frequency seems larger[42,43], and this is supported by our findings.

A total of 37 SNPs demonstrated genome-wide significance of between -log10(*P*)>6 to -log10(*P*)<8 in association with CVT susceptibility in the discovery and replication set (Supplementary Table S4). Thirty-three SNPs from chromosome 9 are within the region of the *ABO* gene. Four SNPs in chromosome 4 are in the region of Factor XI gene (plasma thromboplastin antecedent). These SNPs of lower significance may become important in future larger studies.

Regardless the fairly large sample size of members of European ancestry, there was only adequate statistical power to detect associations with common genetic variants. It is likely that other variants are also implicated in CVT, but we present the most common. However, it is likely we are underestimating the association of 9q34.2 as our conditional analysis was based on one, albeit highly significant, SNP within the LD region of interest. Further, although all study sites used well-established clinical protocols to diagnose CVT, the heterogeneity of the CVT phenotype may have limited the ability to discover some genetic associations by biasing the effect estimates towards the null hypothesis. However, unlike ischaemic stroke, CVT is relatively homogenous clinically which may serve to mitigate this issue. We determined ABO phenotype based on previously reported SNP allele combinations as blood group was not initially characterised. The distribution of the ABO blood groups was, however, consistent with previously reported distributions in European populations. However, any possible measurement errors or misclassifications are likely biased towards the null and would therefore, underestimate the presented risks associated with the ABO blood groups. Moreover, we were not able to take into account the possible role of Rhesus and other blood group systems which could provide other and more detailed insights. Finally, while the association observed

between *ABO* and CVT risk was significant, and we demonstrate blood group association with CVT, we are not in a position to confirm a causal mechanism.

We present the first GWA study in CVT. We report a chromosomal locus, 9q34.2 associated with CVT susceptibility in multiple populations of European descent with an effect size more than double (greater than any previously reported in other thrombophilia related diseases) and implicate the *ABO* gene along with blood group in its aetiology providing new insights into the pathophysiology of CVT.

Acknowledgements:

We are grateful to our patients for allowing us to enter them into this collaboration without whom this work would not have been possible.

Sources of research funding: This study was funded in part by grants awarded to Pankaj Sharma from The Stroke Association (UK), the Dowager Countess Eleanor Peel Trust (UK), and the Interreg 2 Seas programme 2014-2020 co-funded by the European Regional Development Fund under subsidy contract 2S01-059_IMODE. Pankaj Sharma was funded by a Dept of Health (UK) Senior Fellowship at Imperial College London for part of this study. The cohort of 231 French cases was constituted during a hospital protocol of clinical research approved by the French Ministry of Health; biological collection was kept and managed by the INSERM CIC-CRB 1404, F-76000 Rouen, France. The controls from Belgium were genotyped as part of the SIGN study. Robin Lemmens is a senior clinical investigator of FWO Flanders. Turgut Tatlisumak is the recipient of funding from the Sigrid Juselius Foundation (FIN), Helsinki University Central Hospital (FIN), Sahlgrenska University Hospital (SWE), and University of Gothenburg (SWE). Swedish Research Council (2018-02543) and the Swedish Heart and Lung Foundation (20190203). Jonathan M. Coutinho has received funding from the Dutch Thrombosis Foundation. This material is the result of work supported with resources and the use of facilities at the VA Maryland Health Care System, Baltimore, Maryland USA and was also supported in part by the National Institutes of Health (U01NS069208, R01NS105150, R01NS100178). The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

Author Contributions:

P.S. contributed to the conception and design of the study. G.K.D., P.S., J.M.F., I.C., V.T., R.L., S.H., T.M.M., J.P., E.H., T.T., K.S., I.M., E.G., M.M., M.Z., M.M., S.M.P., M.A., P.B., E.P., P.C., M.C., R.S., G.F., D.C., A.P., A.A., Y.M.R., S.M.Z., M.C.B., J.M.C., D.A.S., S.R., P.C., J.M.F., E.L., C.J., K.J., V.L.C.D., A.B.T., S.D., T.M., B.W., J.J.M., B.D.M., S.J.K., A.T., A.H., R.D., and G.P. contributed to the acquisition and analysis of data. P.S., and G.K.D. contributed to drafting a significant portion of the manuscript or figures.

Potential Conflicts of Interest:

Nothing to report.

References

1. Luo Y, Tian X, Wang X. Diagnosis and Treatment of Cerebral Venous Thrombosis: A Review. *Front Aging Neurosci.* 2018;10:2.
2. Ferro JM, Canhao P, Stam J, Bousser MG, Barinagarrementeria F, Investigators I. Prognosis of cerebral vein and dural sinus thrombosis: results of the International Study on Cerebral Vein and Dural Sinus Thrombosis (ISCVT). *Stroke.* 2004;35(3):664-670.
3. Stam J. Thrombosis of the cerebral veins and sinuses. *N Engl J Med.* 2005;352(17):1791-1798.
4. Caplan L, Caplan LR, ScienceDirect. *Primer on cerebrovascular diseases.* London, United Kingdom: Academic Press is an imprint of Elsevier; 2017.
5. Coutinho JM, Zuurbier SM, Stam J. Declining mortality in cerebral venous thrombosis: a systematic review. *Stroke.* 2014;45(5):1338-1341.
6. Ferro JM, Coutinho JM, Dentali F, et al. Safety and Efficacy of Dabigatran Etexilate vs Dose-Adjusted Warfarin in Patients With Cerebral Venous Thrombosis: A Randomized Clinical Trial. *JAMA Neurol.* 2019.
7. Silvis SM, Middeldorp S, Zuurbier SM, Cannegieter SC, Coutinho JM. Risk Factors for Cerebral Venous Thrombosis. *Semin Thromb Hemost.* 2016;42(6):622-631.
8. Marjot T, Yadav S, Hasan N, Bentley P, Sharma P. Genes associated with adult cerebral venous thrombosis. *Stroke.* 2011;42(4):913-918.
9. Sharma P, Meschia JF, SpringerLink. *Stroke genetics.* 2017.
10. Tregouet DA, Heath S, Saut N, et al. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. *Blood.* 2009;113(21):5298-5303.

11. Hinds DA, Buil A, Ziemek D, et al. Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. *Hum Mol Genet.* 2016;25(9):1867-1874.

12. de Haan HG, V.A. vH, Germain M, et al. Genome-Wide Association Study Identifies a Novel Genetic Risk Factor for Recurrent Venous Thrombosis. *Circ Genom Precis Med.* 2018;11(2):e002094.

13. Cotlarciuc I, Marjot T, Khan MS, et al. Towards the genetic basis of cerebral venous thrombosis-the BEAST Consortium: a study protocol. *BMJ Open.* 2016;6(11):e012351.

14. Guo Y, He J, Zhao S, et al. Illumina human exome genotyping array clustering and quality control. *Nat Protoc.* 2014;9(11):2643-2662.

15. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284-1287.

16. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet.* 2010;11(7):459-463.

17. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005;21(2):263-265.

18. Newcombe PJ, Conti DV, Richardson S. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genet Epidemiol.* 2016;40(3):188-201.

19. Zhu Z, Zheng Z, Zhang F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun.* 2018;9(1):224.

20. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015;47(3):291-295.

21. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet.* 2008;9(6):477-485.

22. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet.* 2005;76(3):449-462.

23. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82.

24. Higgins JPT, Thomas J, Chandler J, et al. *Cochrane handbook for systematic reviews of interventions.* Hoboken, NJ: Wiley-Blackwell; 2019.

25. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* 2002;21(11):1539-1558.

26. Li-Gao R, Carlotti F, de Mutsert R, et al. Genome-Wide Association Study on the Early-Phase Insulin Response to a Liquid Mixed Meal: Results From the NEO Study. *Diabetes.* 2019;68(12):2327-2336.

27. Melzer D, Perry JR, Hernandez D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* 2008;4(5):e1000072.

28. Groot HE, Villegas Sierra LE, Said MA, Lipsic E, Karper JC, van der Harst P. Genetically Determined ABO Blood Group and its Associations With Health and Disease. *Arterioscler Thromb Vasc Biol.* 2020;40(3):830-838.

29. Racial and ethnic distribution of ABO blood types. Bloodbook.com. Archived from the original on 4 March 2010. Retrieved 1 August 2010.

30. Timmann C, Thye T, Vens M, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature.* 2012;489(7416):443-446.

31. Scott RA, Scott LJ, Magi R, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes.* 2017;66(11):2888-2902.

32. Pillai SG, Ge D, Zhu G, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.* 2009;5(3):e1000421.

33. Figueroa JD, Ye Y, Siddiq A, et al. Genome-wide association study identifies multiple loci associated with bladder cancer risk. *Hum Mol Genet.* 2014;23(5):1387-1398.

34. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet.* 2009;41(9):986-990.

35. Dichgans M, Malik R, Konig IR, et al. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke.* 2014;45(1):24-36.

36. Shah S, Henry A, Roselli C, et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun.* 2020;11(1):163.

37. Cheng YC, Stanne TM, Giese AK, et al. Genome-Wide Association Analysis of Young-Onset Stroke Identifies a Locus on Chromosome 10q25 Near HABP2. *Stroke.* 2016;47(2):307-316.

38. Jick H, Slone D, Westerholm B, et al. Venous thromboembolic disease and ABO blood type. A cooperative study. *Lancet.* 1969;1(7594):539-542.

39. Tufano A, Guida A, Coppola A, et al. Risk factors and recurrent thrombotic episodes in patients with cerebral venous thrombosis. *Blood Transfus.* 2014;12 Suppl 1:s337-342.

40. Klarin D, Busenkell E, Judy R, et al. Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nat Genet.* 2019;51(11):1574-1579.

41. Lindstrom S, Wang L, Smith EN, et al. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood.* 2019;134(19):1645-1657.

42. Zuber M, Toulon P, Marnet L, Mas JL. Factor V Leiden mutation in cerebral venous thrombosis. *Stroke.* 1996;27(10):1721-1723.

43. Gonzalez JV, Barboza AG, Vazquez FJ, Gandara E. Prevalence and Geographical Variation of Prothrombin G20210A Mutation in Patients with Cerebral Vein Thrombosis: A Systematic Review and Meta-Analysis. *PLoS One.* 2016;11(3):e0151607.

Legends and Tables

Figure 1: The associations of SNPs by chromosome with CVT. The horizontal line indicates a critical threshold of -log10(*P*-value)>8, the selected genomic locus is shown. Below posterior probability of Bayesian variable selection framework (JAM) and *P*-value of condiation analysis.

Figure 2: Forest plot of the validated SNPs associated with CVT susceptibility.

Figure 3: Comparing the *P*-value of the discovery set to the replication set.

Figure 4: Regional association and linkage disequilibrium (LD) plot of the 9q34.2 locus. Known genes are shown in the relevant region. A LD map based on r^2 values is shown at the bottom panel.

Figure 5: Blood group distribution between CVT cases and control determined using allele combinations of SNP rs8176746 and rs8176645 located in the *ABO* gene.

Table 1: Demographic characteristics of individuals with CVT among discovery and replication datasets.