Published in final edited form as:
Hum Mutat. 2019 September ; 40(9): 1321–1329. doi:10.1002/humu.23824.

Identifying mutation-driven changes in gene functionality that lead to venous thromboembolism

Yanran Wang,
Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey, USA

Yana Bromberg
Department of Genetics, Rutgers University, New Jersey, USA

Technical University of Munich Institute for Advanced Study, (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany

Abstract

Venous thromboembolism (VTE) is a common hematological disorder. VTE affects millions of people around the world each year and can be fatal. Earlier studies have revealed the possible VTE genetic risk factors in Europeans. The 2018 Critical Assessment of Genome Interpretation (CAGI) challenge had asked participants to distinguish between 66 VTE and 37 non-VTE African American (AA) individuals based on their exome sequencing data. We used variants from AA VTE association studies and VTE genes from DisGeNET database to evaluate VTE risk via four different approaches; two of these methods were most successful at the task. Our best performing method represented each exome as a vector of predicted functional effect scores of variants within the known genes. These exome vectors were then clustered with k-means. This approach achieved 70.8% precision and 69.7% recall in identifying VTE patients. Our second-best ranked method had collapsed the variant effect scores into gene-level function changes, using the same vector clustering approach for patient/control identification. These results show predictability of VTE risk in AA population and highlight the importance of variant-driven gene functional changes in judging disease status. Of course, more in-depth understanding of AA VTE pathogenicity is still needed for more precise predictions.

Keywords

Venous thromboembolism; CAGI; warfarin; function; prediction

Background

Venous thromboembolism (VTE) is a disorder that causes formation blood clots, primarily affecting veins deep in the body. VTE that affects legs, groins, or arms, specifically, is

* Corresponding author yanab@rci.rutgers.edu, http://bromberglab.org, Tel: +1-732-932-9763 Ext. 218; Fax +1-732-932-8965, Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey, USA .

The authors declare that they have no conflict of interests.

designated deep vein thrombosis (DVT). Clots traveling to the lungs and blocking arteries result in pulmonary embolisms (PE) (Bartholomew, 2017). VTE affects 300,000 – 600,000 individuals in the U.S. each year (Beckman, Hooper, Critchley, & Ortel, 2010). In an estimated 100,000 – 180,000 of these, the disease is fatal (Goldhaber, 2012). VTE incidence is much higher in older people (≥ 80 years of age) compared with that in the younger ones. European and African descents are reported to have the highest VTE incidence in multiple studies (Beckman et al., 2010; Heit, Spencer, & White, 2016; Zakai et al., 2014).

The pathogenesis of VTE is not fully understood. Multiple factors contributes to development of VTE, including genetics, but also factors such as advanced age, pregnancy, obesity, trauma, surgery, hospitalization, etc. (Beckman et al., 2010). Previous studies have identified several potential genetic causes, including factor V Leiden variation (Ridker et al., 1995) and prothrombin mutations (Rosendaal et al., 1998). Despite the recent approval of new direct-acting anticoagulants (Tellor et al., 2018), warfarin is commonly used for treatment and prevention of VTE (Sterne et al., 2017), due to its low renal function impairment, good drug adherence record, and accessible cost (Burn & Pirmohamed, 2018). Warfarin is also used for clot prevention in other diseases, *e.g.* Atrial Fibrillation (AF) (Shariff, Aleem, Singh, Y, & S, 2012), where an abnormal heart rhythm affects blood flow and predisposes to clot formation (Zimetbaum, 2017).

CAGI (Critical Assessment of Genome Interpretation, http://genomeinterpretation.org) clotting disease challenge (https://genomeinterpretation.org/content/clotting-disease-dvt-or-pe-exomes) provided a dataset of exomes of VTE (unprovoked VTE, i.e. not caused by external factors such as trauma, surgery, etc.) and non-VTE (mostly AF) patients. All 103 patients in this cohort are African-American (AA) and all were on warfarin for treatment or prevention of thrombosis.

Some VTE risk prediction methods using genetic biomarkers have been published (Ahmad et al., 2018; Folsom et al., 2016), focusing primarily on populations of European descent. However, none of these methods specifically address VTE in AA patients. We implemented two different methods for the prediction of VTE risk in this challenge. First, we applied the commonly used genetic risk score (GRS) (Cooke Bailey & Igo, 2016), imputing missing variants and also making use of the warfarin dose information. Second, we clustered exomes using different features, *e.g.* functional deficiency of the known disease variant/genes (Pinero et al., 2017). Note that warfarin dosage is a confounding factor for identifying VTE signal. For example, patients with VTE often have underlying genetic defects, which may cause hypercoagulation and, thus, require a higher warfarin dose compared to patients with AF, who take warfarin as a prophylactic measure (James, Britt, Raskino, & Thompson, 1992); *i.e.* it is likely simpler to predict VTE status when the warfarin dose is known. All predictions submitted in this challenge were evaluated by the assessors according to the data provider labels.

Among all methods that relied solely on genetic information (i.e. not including warfarin doses), our three clustering-based methods were ranked 1st, 3rd, and 4th. Our best performing method represented each exome in the cohort as a vector of predicted effect scores of variants within the known VTE genes (extracted from the DisGeNET database). These exome vectors were then clustered into two clusters with k-means. Our second-best method (ranked 3rd overall) had collapsed the variant effect scores into gene-level function changes, using the same vector clustering approach for patient/control identification. Our last clustering approach (ranked 4th) represented each exome as a vector of genotypes of all variants within the known VTE genes, without any functional annotation. The genetic risk score method, based on known GWAS variants, performed worse than the clustering methods even with warfarin dosage information included but, as expected, had very high precision, albeit for a smaller number of patients identified.

Transforming our current knowledge of AA VTE variants into computationally predicted variant effects, we achieved the highest (62%) overall accuracy of prediction. However, integrating other elements into our method, including, for example, patient clinical features (Zhai et al., 2019), is likely to improve performance. Since VTE is preventable (Beckman et al., 2010), it is critical to have an accurate prediction of VTE risk for clinical use. Notably, unprovoked VTE is also a sign of other diseases, e.g. some forms of cancer (van Es et al., 2017). An early diagnosis of VTE could thus potentially lower the disease prevalence and help elevate the patients' quality of life.

Materials and Methods

Challenge Data.

The study cohort contained 103 African-American (AA) individuals who were taking warfarin for either VTE (unprovoked VTE, 66 individuals) or non-VTE (most are AF patients, 37 individuals) treatment and/or prevention of blood clots. Specifically, CAGI participants were provided with the whole exome sequencing variant call files (VCFs) and the clinical covariates files for each individual. The clinical covariates included gender, age, height, weight, whether the individual was also taking aspirin and/or amiodarone, and his/her warfarin dose. There were 58 individuals taking a high dose (> 49 mg/week) of warfarin and 45 individuals taking a low dose (< 35 mg/week). This dataset and its detailed clinical covariate statistics were also reported in Daneshjou *et al.* (Daneshjou et al., 2014).

Data cleaning.

We retained only the PASS variants according to VQSR (Variant Quality Score Recalibration) standard (McKenna et al., 2010). Principle Component Analysis (PCA) of SNPs indicated three different clusters of the subjects (Supp. Figure S1A). Other analysis of the quality metrics, including individual number of variants, number of heterozygous variants, number of singletons, individual call rate, individual Ti-Tv ratio, and individual PASS (VQSR standard) rate, indicated that the separation of subjects might be due to the difference in quality of the variant calls that may result from differences in sequencing batches or other systematic errors. We removed all variants in the VCF files that failed the GTAK VQSR PASS qualification. The cleaning resulted in higher quality of data (Ts-Tv ratio rising from 2.20 to 2.48) and loss of obvious clustering of subjects (Supp. Figure S1B). While for the purposes of this challenge we did not apply further cleaning, the cleaned data still had low quality calls. Thus, we suggest that a more comprehensive clean-up could benefit all further analyses.

Genetic risk scoring (GRS) methodology (Method 1).

A variant in Protein S (*PROS1*), Valine in position 510 to Methionine (V510M, rs138925964) was shown to be associated of VTE (Daneshjou et al., 2016). To the best of our knowledge, there are two AA VTE genome-wide association studies (GWAS) (Heit et al., 2017; Hernandez et al., 2016). Hernandez, *et al.* have found three SNPs on chromosome 20 which increased the risk of VTE by 2.3-fold. These SNPs are in the eQTLs (expression quantitative trait loci) for the *THBD* gene; here, the VTE patients had lower *THBD* expression than the healthy controls. Heit, *et al.* found three intragenic SNPs of genome-wide significance in the *LEMD3*, *LY86*, *LOC100130298* genes, respectively. Note that there were VTE GWAS done in the European population, but the identified SNPs are generally not observed in the AA population (Hotoleanu, 2017).

The genetic risk scoring (GRS) using SNPs identified in the Europeans does not work in African-Americans (Folsom et al., 2016). Thus, to apply the GRS strategy, we had to develop a strategy using AA relevant SNPs. We used variants from the most recent GWAS study from Heit *et al.* and one study from data provider (Daneshjou et al., 2016) to construct our GRS. None of the three significant GWAS variants (Heit et al., 2017) were covered by the challenge exome data. Therefore, we imputed the variants using IMPUTE2 (Howie, Donnelly, & Marchini, 2009) with reference to the 1000 Genomes Phase 3 data (NCBI build b37) (Genomes Project et al., 2015). Unfortunately, imputation accuracies (R^2, *i.e.* info value in IMPUTE2) were relatively low: 0.336, 0.206, 0.288 for the GWAS-significant variants rs138916004 (*LEMD3*), rs3804476 (*LY86*), rs142143628 (*LOC100130298*), respectively. We calculated three versions of scores using the GRS method (Method 1.1, Method 1.2, and Method 1.3) as described below:

$$r_i = \sum_{j=1}^{m} w_j x_{ij} \quad 1$$

***Method 1.1* (not submitted to CAGI):** Only the three significant GWAS loci from Heit *et al.* (imputed for our data) were included in the GRS equation (Eqn. 1), where w_j was the j^{th} SNP log odds ratio from the GWAS study (for a total of m SNPs) and the total risk of the i^{th} individual (r_i) was the sum of the weighted genotypes (x_{ij}) in his/her exome.

***Method 1.2* (not submitted to CAGI):** We included (i) the three significant Heit *et al.* GWAS loci (imputed), (ii) loci reported in Heit *el al.* that were below GWAS significance but were covered by the challenge data, (iii) and also variants in the *PROS1* gene reported in Daneshjou *et al.* that were covered by the challenge data, into Eqn. 1. Note here that w_j log odds ratios for (i and ii) came from the Heit et al. study, while the *PROS1* variants had ratios assigned by the Daneshjou *et al.* study. While combining multiple studies into a single score is not ideal, we felt that additional strongly-associated variants could contribute to the resolution of the method.

***Method 1.3* (submitted):** Based on the warfarin dose we adjusted the predictions of Method 1.2, such that individuals with high and low warfarin dosage (heuristically) scored 1.5-fold and 0.8-fold of the Method 1.2 predictions, respectively.

Note that since imputed variants had low imputation quality, the probabilistic values of the imputed genotypes were used in the equation instead of the hard call values (Li, Willer, Sanna, & Abecasis, 2009).

Clustering methodology (Methods 2–4).

From the DisGeNET (Pinero et al., 2017) database we extracted a full list of VTE genes. DisGeNET contains standardizes annotations of gene-disease relationships extracted from various sources. Each relationship is assigned a Gene-Disease Association Score (GDA Score) according to the level of source evidence, *i.e.* the higher the GDA Score, the more reliable the gene-disease relationship is. We searched (April 12th, 2018) for keywords "Venous thromboembolisms" ("VTE", C1861172; 111 results), "Deep vein thrombosis" ("DVT", C0149871; 96 results), "Pulmonary embolism" ("PE", C0034065; 71 results), obtaining three gene lists and retaining only the genes with a GDA Score ≥ 0.2 (8, 8, and 21 genes from each keyword search, respectively).

We further categorized the genes into *Level 1*, *Level 2*, and *Level 3* genes (Supp. Table S1; 8, 3, 14 genes in *Level 1, 2, 3* gene list, respectively), where *Level 1* genes were in the VTE gene list, *Level 2* genes were in both DVT and PE lists, but not in VTE list, and *Level 3* genes were unique to the DVT or PE list.

Further, VCF variants were annotated with ANNOVAR (K. Wang, Li, & Hakonarson, 2010) to retain those that affected the genes of interest. We applied three different clustering strategies to cluster exomes in our cohort: (i) each individual was represented as a vector of genotypes (0/0, 0/1, or 1/1) of all variants within *Level 1* and *2* genes (125 variants within 11 genes, Method 2) and then clustered using k-modes (Huang, 1997) clustering; (ii) k-means (Hartigan & Wong, 1979) clustering by SNAP (Bromberg & Rost, 2007) predicted functional effects of all non-neutral variants within all three gene lists (67 non-neutral variants in total, Method 3); (iii) k-means clustering by gene function deficiency scores (27 genes total) with the 67 non-neutral variants in (ii) (Method 4). The gene function deficiency score for each gene was calculated via Eqn. 2, where $score_i$ was the SNAP prediction, normalized to range 0 to 1, for the i^{th} variant (for a total of N variants) within the gene. We heuristically set a factor of 0.35 (het_i) for heterozygous genotype to account for the fact that heterozygous variants are generally less functionally effective than homozygous variants. The *gene score* approximated the amount of gene function left after the mutation(s).

$$gene\ score = \prod_{i=1}^{N}\left(1 - het_i \times score_i\right) \quad 2$$

Note that k-modes was used in (i) for clustering nominal features (genotypes) and k-means was used in (ii) and (iii) for clustering numerical features (function deficiency scores for variant or gene). For all three clustering methods, we chose the bigger cluster as the VTE group and smaller one as non-VTE group for the submissions, as we knew that there were more VTE than non-VTE individuals in the dataset (Daneshjou et al., 2014).

Prediction evaluation.

Method 1 produced numeric predictions (ranging from 0 to 1) while Methods 2–4 results were binary (0 is non-VTE and 1 is VTE). We used the ROC (receiver operating characteristic) curve AUC (area under the curve) to evaluate the prediction performance of Method 1. For all methods, we calculated the overall accuracy (Eqn. 3), precision and recall (Eqn. 4), and Matthews correlation coefficient (MCC) (Eqn. 5), where TP (true positives) were the correctly classified VTE patients, FP (false positives) were the incorrectly classified non-VTE controls, TN (true negatives) were the correctly classified non-VTEs, and FN (false negatives) were the incorrectly classified VTEs.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad 3$$

$$Precision = \frac{TP}{TP + FP}\; Recall = \frac{TP}{TP + FN} \quad 4$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP\ + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad 5$$

Results

In post-CAGI evaluation, the performance of the 3-locus GRS method (Method 1.1) was worse than random guessing (ROC/PR AUC = 0.432/0.610). The performance of Method 1.2 (not submitted to CAGI) was also poor (ROC/PR AUC = 0.546/0.688). This is not unexpected as the score included low significance variants of low imputation confidence and also heuristically combined variant log-odds scores across different studies. The submitted GRS+warfarin (Method 1.3) predictions were significantly better (Figure 1, Table 1, ROC/PR AUC = 0.646/0.788). However, since Method 1.3 used warfarin dosage information in prediction, it was not in the final ranking of evaluated submissions.

For Method 2, we considered all possible variants within the known *Level 1* and *2* disease genes as features (Materials and Methods). By choosing the larger cluster to be VTE, 69 and 34 individuals were predicted to be VTEs and non-VTEs, respectively. There was also an obvious separation between the two clusters by Multiple Correspondence Analysis (Husson, Lê, & Pagès, 2017) (Supp. Figure S2).

For Method 3, we chose only the functional significant (as per SNAP predictions variants as clustering features. Only the 67 non-neutral variants from SNAP predictions were kept for k-means clustering. Here, 65 and 38 individuals were predicted to be VTEs and non-VTEs, respectively. Since this CAGI challenge assessors chose the overall accuracy (Eqn. 3) as the primary metric for ranking methods, we report that this method (accuracy = 62%) was ranked the first among all methods (ours and others') that did not use warfarin dosage.

For Method 4, we converted the functional annotation from the variant-level to the protein-level, taking the product of all variant function scores within one gene as the protein function deficiency score (Eqn. 2). By k-means clustering of the 27 gene function deficiency scores, 73 and 30 individuals were predicted to be VTEs and non-VTEs, respectively. This method was our second-best and ranked 3rd overall (accuracy = 58%).

The GRS method (Method 1.1) had the highest precision (100%) among all predictions, but it recognized only very few VTE patients (recall=3%; 2 of 66 patients). On the other hand, Method 3 had more balanced precision and recall values, identifying 70% of the VTE patients with moderately high precision (71%). Method 4 identified more VTE patients (73%) but less accurately (precision of 66%). Note that the two methods that included variant functional annotations had higher accuracies than the GRS methods and Method 2, the clustering method without functional annotation.

Discussion

Similarity between VTE and AF complicates prediction.

In the challenge, non-VTE samples came mostly from the AF (Atrial Fibrillation patients) population. The reason for this choice is clear – it is uncommon to find a study where individuals would be taking a drug for no reason. Warfarin in VTE is mainly used for treatment of blood clots in veins (for DVT) or in the lungs (for PE), while in non-VTE cases it is often prophylactic. However, separating two types of patients is arguably a more difficult task than separating patients from healthy controls.

VTE and AF are both blood-related polygenic conditions and are not fully understood genetically (Bapat, Anderson, Ellinor, & Lubitz, 2018; Hotoleanu, 2017). They share genetic risk factors and patho-physiological bases for clot formation (Shariff et al., 2012). In fact, DisGeNET (Pinero et al., 2017) suggests 39 overlapping disease risk genes between the two diseases, *e.g.* coagulation factor *F5* (Tang et al., 2013; Zateyshchikov, Brovkin, Chistiakov, & Nosikov, 2010), *CEPT* (cholesteryl ester transfer protein) (Asselbergs et al., 2006; Deguchi, Banerjee, Elias, & Griffin, 2016), and *TFPI* (tissue factor pathway inhibitor) (Efthymiou et al., 2018; Xie et al., 2017) are involved in thromboembolic risk of both VTE and AF. Additionally note that VTE often results from the activation of the coagulation system (Shariff et al., 2012), where tissue factors (TF) play an important role in initiating the clot formation. Similarly TF are over-expressed in AF patients with thromboembolism (Watson, Shantsila, & Lip, 2009). This observation may explain why an anticoagulative drug warfarin is effective in preventing blood clots in both VTE and AF, as well as why studies found that the two diseases often co-occur and that the presence of one increases the risk for another (Enga et al., 2015; Lutsey et al., 2018; Sundboll et al., 2017).

Thus, while AF is indeed different from VTE, including a separate cohort of healthy controls may help better explore the disease pathogenesis of either VTE or AF in future studies.

VTE genetics vary between European and African-American populations.

VTE-associated genetic risk loci were reported in several studies. The first VTE GWAS (Tregouet et al., 2009) done in the European population. Later, more VTE GWA studies were done (Germain et al., 2015; Germain et al., 2011; Heit et al., 2012; Hinds et al., 2016; Tang et al., 2013), but no African-American VTE GWAS was published until 2016 (Hernandez et al., 2016); another AA GWAS closely followed (Heit et al., 2017).

In all European population GWAS, variants in genes *F2*, *F5*, *F11*, *FGG*, *FGA*, *ABO*, *ZFPM2*, *LCN1P2*, *NME7*, *etc*. were found to be significantly VTE-associated. However, the allele frequencies (AFs) of the risk alleles in the European population are different in people of African ancestry (AFR in 1000 Genomes). For example, the T allele in variant rs6025 (gene *F5*) is significantly associated with VTE in European and other ancestry cases (no AA included, log-odds ratio of up to 3.57 (Heit et al., 2012)), but this allele is not observed at all in the AFR population. On the other hand, the two AA-specific GWAS (Heit et al., 2017; Hernandez et al., 2016) have found significantly VTE-associated SNPs in *LEMD3*, *LY86*, *LOC100130298* or *CLVS1*, *LOC102723446*, *CD93* genes (Table 2). These variants were not discovered in any other GWA studies in European populations.

Thus, the differences in genetic underpinnings of this disease in different populations require population-specific model building for further insight. That is, we suspect that our predictors built for this challenge will not work as well for Europeans as for African-Americans. However, we can likely use the same approaches with European-specific data.

Warfarin dose predicts VTE risk.

One consideration for this challenge was whether to take the warfarin dose and/or warfarin dose-related genes into consideration during prediction. At the time of the challenge, all 103 individuals were taking warfarin for treatment or prevention of thrombosis. For the VTE status prediction in the current CAGI challenge, warfarin dose alone (without any genetic data) performed the best. That is, high-dose individuals were more likely to be VTE and low-dose warfarin individuals were more likely to be non-VTE. Note that as the purpose of the CAGI challenge was to interpret the genetic data, the assessor had excluded from assessment all methods that used the provided warfarin dose information.

This result, however, highlights two questions: (1) Does warfarin dose vary from disease to disease? and (2) Do the warfarin dose-related genes affect VTE risk? In response to the first question, James *et al.* have found that AF patients required lower and VTE patients require higher doses of warfarin (James et al., 1992). This trend remained significant even after adjusting for age and other factors. Thus, if used in hind-sight, once the warfarin dosage is established it is arguably easier to identify the patient's VTE status. This approach, however, has limited, if any, clinical utility.

As for the second question, aside from clinical factors such as age, weight, sex, *etc*., two major categories of genes affect warfarin dose: the pharmacodynamic genes (*e.g. VKORC1*, *EPHX1*, *GGCX*, *CALU*), which warfarin works on to block the vitamin K dependent clotting pathway, and the pharmacokinetic genes (*CYP3A4* for R-warfarin, *CYP2C9* for S-warfarin, and other Cytochrome P450 enzymes), which metabolize warfarin (Whirl-Carrillo et al., 2012). One variant in *VKORC1* (1173 C > T) was shown to be VTE-associated (Lacut et al., 2007). Another in *VKORC1* (−1639 G > A) was shown to be DVT-associated (Vesa, Trifa, Crisan, & Buzoianu, 2016). To make a generalizable conclusion about relationship between warfarin dose-relevant genes and VTE risk across populations, further specifically designed experiments are needed.

Use of prediction method in clinical practice needs more work.

Although the GRS result (Method 1.1, Supp. Figure S3) had the lowest overall accuracy, it is likely more useful in the clinical diagnostic settings, where the number of healthy individuals is much larger than the number of VTE-affected people. Here, precision in "diagnosing" someone with VTE is of utmost importance as it guides further treatments and interventions. While identifying only two patients with VTE (3% of all patients), GRS VTE diagnosis precision was 100%--that is no healthy people were misdiagnosed. Our clustering-based Method 3, on the other hand, identified significantly more VTE patients (46 people; 70%), but had a precision (71%) likely useless in the clinic. Thus, our methods are not yet ready for prime-time in real world applications.

Note further that clustering-based approaches require two pieces of information to be practical: the presumed number of clusters (i.e. phenotypes to split people by) and knowledge of which cluster represents which phenotypic group. Of course, this would not be a problem if a classification (e.g. clustering) method was developed using an unrelated "training set", for which both of these pieces of information were available. However, this approach would then be further complicated by the need to ensure biological, and methodological similarity between training and testing samples (Y. Wang et al., 2017).

Variant/gene functional changes are important for prediction.

Our two function-annotated methods (Methods 3 and 4) performed better than the one without function annotation (Method 2). In terms of predicting disease risk, using variant/gene function has two important merits over simply using variant genotypes: (1) decreasing the number of features and (2) giving biologically meaningful weight to individual features.

High-dimensionality of feature space and small numbers of samples consistently plague high-throughput experimentation. For complex diseases, such as VTE, hundreds or even thousands of genetic loci are within the known disease genes, but we generally have much fewer individuals to analyze. For example, in this challenge, we had only 103 individuals evaluate as compared to the 561 variants in the known disease-related genes. Keeping only the 67 variants that had a predicted effect on protein function significantly changed the prediction space. Moreover, as most clustering algorithms assume roughly equal importance for all features (Dash & Liu, 2000), the variants of low or no likely contribution to disease should be ignored as noise. These observations could explain the reduced Method 2 performance where all variants were treated equally.

When comparing the performance of top two methods (Methods 3 and 4), Method 4 had a slightly lower performance. One way to explain this is that the heuristic gene function deficiency score used here possibly did not represent function change accurately; *e.g.* due to the semi-probabilistic combination of the multiple variants per gene. Nevertheless, the fact that Method 4 still worked better than Method 2 suggests that weighing genes in a biologically meaningful fashion is a more reasonable approach than simply using genotypes. Current work in the lab is promising a more meaningful gene score formula in the near future.

Future prospects.

Since African-American and European populations have very different genetic architecture (Park, Cheng, & Haiman, 2018) precise and validated biomarkers of VTE useful in the AA population need more targeted GWAS. Equipped with this new information, GRS calculations can be modified to be more accurate. However, it is worth repeating that the GWAS significant variants are often only biomarkers of the disease instead of disease-causing mutations. Methods that consider variants in the marked regions from a functional level could contribute to our understanding of disease and, potentially, outperform allele-count based approaches. Moreover, as the next-generation sequencing techniques develop and drop in price, we will be able to access more variants, including the rare and individual ones, within an individual and across larger cohorts. Information contained in these new data sets cannot, by definition, be assessed with strictly statistically driven methods. On the other hand, new and specifically targeted tools, could help reveal previously unseen disease-causative variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We'd like to express our gratitude to Dr. John Moult (University of Maryland), Dr. Steven Brenner (University of California), and Dr. Gaia Andreoletti (University of California), and all other CAGI organizers for giving us the opportunity to participate in the challenges and for creating the environment for discussions. We are thankful to Dr. Russ Altman and Dr. Roxana Daneshjou (both Stanford University School of Medicine) for providing the data and Gregory McInnes (Stanford University) for the assessment. Also a special thanks to Dr. Roxana Daneshjou for the discussion. We'd also like to thank all members of the Bromberg Lab for all their support.

YB and YW were supported by the NIH U01 GM115486 grant.

The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650.

References

Ahmad A, Sundquist K, Palmer K, Svensson PJ, Sundquist J, & Memon AA (2018). Risk prediction of recurrent venous thromboembolism: a multiple genetic risk model. J Thromb Thrombolysis. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30368761 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6394443/pdf/11239_2018_Article_1762.pdf. doi:10.1007/s11239-018-1762-7

Asselbergs FW, Moore JH, van den Berg MP, Rimm EB, de Boer RA, Dullaart RP, . . . van Gilst WH (2006). A role for CETP TaqIB polymorphism in determining susceptibility to atrial fibrillation: a nested case control study. BMC Med Genet, 7, 39 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16623947. doi:10.1186/1471-2350-7-39 [PubMed: 16623947]

Bapat A, Anderson CD, Ellinor PT, & Lubitz SA (2018). Genomic basis of atrial fibrillation. Heart, 104(3), 201–206. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28893835. doi:10.1136/heartjnl-2016-311027 [PubMed: 28893835]

Bartholomew JR (2017). Update on the management of venous thromboembolism. Cleve Clin J Med, 84(12 Suppl 3), 39–46. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29257737. doi: 10.3949/ccjm.84.s3.04 [PubMed: 29257737]

Beckman MG, Hooper WC, Critchley SE, & Ortel TL (2010). Venous thromboembolism: a public health concern. Am J Prev Med, 38(4 Suppl), S495–501. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20331949 https://ac.els-cdn.com/S0749379709009465/1-s2.0-S0749379709009465-main.pdf?_tid=c1af9235-9bf9-487d-89eb-8e9ede02776a&acdnat=1549496252_8d9b51eb4a2315d62036819b0bab3b5f. doi:10.1016/j.amepre.2009.12.017 [PubMed: 20331949]

Bromberg Y, & Rost B (2007). SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res, 35(11), 3823–3835. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17526529. doi:10.1093/nar/gkm238 [PubMed: 17526529]

Burn J, & Pirmohamed M (2018). Direct oral anticoagulants versus warfarin: is new always better than the old? Open Heart, 5(1), e000712. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29531758. doi:10.1136/openhrt-2017-000712

Cooke Bailey JN, & Igo RP Jr. (2016). Genetic Risk Scores. Curr Protoc Hum Genet, 91, 1 29 21-21 29 29. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27727440. doi:10.1002/cphg.20

Daneshjou R, Cavallari LH, Weeke PE, Karczewski KJ, Drozda K, Perera MA, . . . Altman RB (2016). Population-specific single-nucleotide polymorphism confers increased risk of venous thromboembolism in African Americans. Mol Genet Genomic Med, 4(5), 513–520. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27652279 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5023936/pdf/MGG3-4-513.pdf. doi:10.1002/mgg3.226 [PubMed: 27652279]

Daneshjou R, Gamazon ER, Burkley B, Cavallari LH, Johnson JA, Klein TE, . . . Perera MA (2014). Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. Blood, 124(14), 2298–2305. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25079360 http://www.bloodjournal.org/content/bloodjournal/124/14/2298.full.pdf. doi:10.1182/blood-2014-04-568436 [PubMed: 25079360]

Dash M, & Liu H (2000). Feature selection for clustering. Knowledge Discovery and Data Mining, Proceedings, 1805, 110–121. Retrieved from <Go to ISI>://WOS:000170556400010.

Deguchi H, Banerjee Y, Elias DJ, & Griffin JH (2016). Elevated CETP Lipid Transfer Activity is Associated with the Risk of Venous Thromboembolism. J Atheroscler Thromb, 23(10), 1159–1167. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27169917. doi:10.5551/jat.32201 [PubMed: 27169917]

Efthymiou M, Arachchillage DRJ, Lane PJ, O'Keeffe AG, McDonnell T, Cohen H, & Mackie IJ (2018). Antibodies against TFPI and protein C are associated with a severe thrombotic phenotype in patients with and without antiphospholipid syndrome. Thromb Res, 170, 60–68. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30121005. doi:10.1016/j.thromres.2018.08.003 [PubMed: 30121005]

Enga KF, Rye-Holmboe I, Hald EM, Lochen ML, Mathiesen EB, Njolstad I, . . . Hansen JB (2015). Atrial fibrillation and future risk of venous thromboembolism:the Tromso study. J Thromb Haemost, 13(1), 10–16. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25330989. doi: 10.1111/jth.12762 [PubMed: 25330989]

Folsom AR, Tang W, Weng LC, Roetker NS, Cushman M, Basu S, & Pankow JS (2016). Replication of a genetic risk score for venous thromboembolism in whites but not in African Americans. J Thromb Haemost, 14(1), 83–88. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26565658 https://onlinelibrary.wiley.com/doi/pdf/10.1111/jth.13193. doi:10.1111/jth.13193 [PubMed: 26565658]

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, . . . Abecasis GR (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26432245. doi:10.1038/nature15393 [PubMed: 26432245]

Germain M, Chasman DI, de Haan H, Tang W, Lindstrom S, Weng LC, . . . Morange PE (2015). Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. Am J Hum Genet, 96(4), 532–542. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25772935. doi:10.1016/j.ajhg.2015.01.019 [PubMed: 25772935]

Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, . . . Morange PE (2011). Genetics of venous thrombosis: insights from a new genome wide association study. PLoS One, 6(9), e25581. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21980494. doi:10.1371/journal.pone.0025581

Goldhaber SZ (2012). Venous thromboembolism: epidemiology and magnitude of the problem. Best Pract Res Clin Haematol, 25(3), 235–242. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/22959540. doi:10.1016/j.beha.2012.06.007 [PubMed: 22959540]

Hartigan JA, & Wong MA (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100–108.

Heit JA, Armasu SM, Asmann YW, Cunningham JM, Matsumoto ME, Petterson TM, & De Andrade M (2012). A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. J Thromb Haemost, 10(8), 1521–1531. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/22672568 https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-7836.2012.04810.x. doi:10.1111/j.1538-7836.2012.04810.x [PubMed: 22672568]

Heit JA, Armasu SM, McCauley BM, Kullo IJ, Sicotte H, Pathak J, . . . de Andrade M (2017). Identification of unique venous thromboembolism-susceptibility variants in African-Americans. Thromb Haemost, 117(4), 758–768. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28203683 https://www.thieme-connect.com/products/ejournals/pdf/10.1160/TH16-08-0652.pdf. doi:10.1160/TH16-08-0652 [PubMed: 28203683]

Heit JA, Spencer FA, & White RH (2016). The epidemiology of venous thromboembolism. J Thromb Thrombolysis, 41(1), 3–14. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26780736. doi:10.1007/s11239-015-1311-6 [PubMed: 26780736]

Hernandez W, Gamazon ER, Smithberger E, O'Brien TJ, Harralson AF, Tuck M, . . . Perera MA (2016). Novel genetic predictors of venous thromboembolism risk in African Americans. Blood, 127(15), 1923–1929. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26888256 http://www.bloodjournal.org/content/bloodjournal/127/15/1923.full.pdf. doi:10.1182/blood-2015-09-668525 [PubMed: 26888256]

Hinds DA, Buil A, Ziemek D, Martinez-Perez A, Malik R, Folkersen L, . . . Sabater-Lleal M (2016). Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. Hum Mol Genet, 25(9), 1867–1874. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26908601. doi:10.1093/hmg/ddw037 [PubMed: 26908601]

Hotoleanu C (2017). Genetic Risk Factors in Venous Thromboembolism. Adv Exp Med Biol, 906, 253–272. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27638626. doi:10.1007/5584_2016_120 [PubMed: 27638626]

Howie BN, Donnelly P, & Marchini J (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet, 5(6), e1000529. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19543373. doi:10.1371/journal.pgen.1000529

Huang Z (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. DMKD, 3(8), 34–39.

Husson F, Lê S, & Pagès J (2017). Exploratory multivariate analysis by example using R: Chapman and Hall/CRC.

James AH, Britt RP, Raskino CL, & Thompson SG (1992). Factors Affecting the Maintenance Dose of Warfarin. Journal of Clinical Pathology, 45(8), 704–706. Retrieved from <Go to ISI>://WOS:A1992JG75400014 https://jcp.bmj.com/content/jclinpath/45/8/704.full.pdf. doi:DOI 10.1136/jcp.45.8.704 [PubMed: 1401182]

Lacut K, Larramendy-Gozalo C, Le Gal G, Duchemin J, Mercier B, Gourhant L, . . . Verstuyft C (2007). Vitamin K epoxide reductase genetic polymorphism is associated with venous thromboembolism: results from the EDITH study. Journal of Thrombosis and Haemostasis, 5(10), 2020–2024. Retrieved from <Go to ISI>://WOS:000249827700006. doi:DOI 10.1111/j.1538-7836.2007.02706.x [PubMed: 17883698]

Li Y, Willer C, Sanna S, & Abecasis G (2009). Genotype imputation. Annu Rev Genomics Hum Genet, 10, 387–406. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19715440. doi:10.1146/annurev.genom.9.081307.164242 [PubMed: 19715440]

Lutsey PL, Norby FL, Alonso A, Cushman M, Chen LY, Michos ED, & Folsom AR (2018). Atrial fibrillation and venous thromboembolism: evidence of bidirectionality in the Atherosclerosis Risk in Communities Study. J Thromb Haemost, 16(4), 670–679. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29431904. doi:10.1111/jth.13974 [PubMed: 29431904]

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, . . . DePristo MA (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20(9), 1297–1303. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20644199. doi:10.1101/gr.107524.110 [PubMed: 20644199]

Park SL, Cheng I, & Haiman CA (2018). Genome-Wide Association Studies of Cancer in Diverse Populations. Cancer Epidemiol Biomarkers Prev, 27(4), 405–417. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28637795. doi:10.1158/1055-9965.EPI-17-0169 [PubMed: 28637795]

Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, . . . Furlong LI (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res, 45(D1), D833–D839. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27924018. doi:10.1093/nar/gkw943 [PubMed: 27924018]

Ridker PM, Miletich JP, Stampfer MJ, Goldhaber SZ, Lindpaintner K, & Hennekens CH (1995). Factor V Leiden and risks of recurrent idiopathic venous thromboembolism. Circulation, 92(10), 2800–2802. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7586244. [PubMed: 7586244]

Rosendaal FR, Doggen CJ, Zivelin A, Arruda VR, Aiach M, Siscovick DS, . . . Reitsma PH (1998). Geographic distribution of the 20210 G to A prothrombin variant. Thromb Haemost, 79(4), 706–708. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9569177. [PubMed: 9569177]

Shariff N, Aleem A, Singh M, Y ZL, & S JS (2012). AF and Venous Thromboembolism-Pathophysiology, Risk Assessment and CHADS-VASc score. J Atr Fibrillation, 5(3), 649 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28496776. doi:10.4022/jafib.649 [PubMed: 28496776]

Sterne JA, Bodalia PN, Bryden PA, Davies PA, Lopez-Lopez JA, Okoli GN, . . . Hingorani AD (2017). Oral anticoagulants for primary prevention, treatment and secondary prevention of venous thromboembolic disease, and for prevention of stroke in atrial fibrillation: systematic review, network meta-analysis and cost-effectiveness analysis. Health Technol Assess, 21(9), 1–386. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28279251. doi:10.3310/hta21090

Sundboll J, Hovath-Puho E, Adelborg K, Ording A, Schmidt M, Botker HE, & Sorensen HT (2017). Risk of arterial and venous thromboembolism in patients with atrial fibrillation or flutter: A nationwide population-based cohort study. Int J Cardiol, 241, 182–187. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28473169. doi:10.1016/j.ijcard.2017.04.081 [PubMed: 28473169]

Tang W, Teichert M, Chasman DI, Heit JA, Morange PE, Li G, . . . Smith NL (2013). A genome-wide association study for venous thromboembolism: the extended cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium. Genet Epidemiol, 37(5), 512–521. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23650146. doi:10.1002/gepi.21731 [PubMed: 23650146]

Tellor KB, Nguyen SN, Bultas AC, Armbruster AL, Greenwald NA, & Yancey AM (2018). Evaluation of the impact of body mass index on warfarin requirements in hospitalized patients. Ther Adv Cardiovasc Dis, 12(8), 207–216. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29914293. doi:10.1177/1753944718781295 [PubMed: 29914293]

Tregouet DA, Heath S, Saut N, Biron-Andreani C, Schved JF, Pernod G, . . . Morange PE (2009). Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. Blood, 113(21), 5298–5303. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19278955 http://www.bloodjournal.org/content/bloodjournal/113/21/5298.full.pdf. doi:10.1182/blood-2008-11-190389 [PubMed: 19278955]

van Es N, Le Gal G, Otten HM, Robin P, Piccioli A, Lecumberri R, . . . Carrier M (2017). Screening for cancer in patients with unprovoked venous thromboembolism: protocol for a systematic review and individual patient data meta-analysis. BMJ Open, 7(6), e015562. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28601834. doi:10.1136/bmjopen-2016-015562

Vesa SC, Trifa AP, Crisan S, & Buzoianu AD (2016). VKORC1–1639 G > A Polymorphism in Romanian Patients With Deep Vein Thrombosis. Clinical and Applied Thrombosis-Hemostasis, 22(8), 760–764. Retrieved from <Go to ISI>://WOS:000386019900008. doi:10.1177/1076029615585993 [PubMed: 25976278]

Wang K, Li M, & Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 38(16), e164. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20601685. doi:10.1093/nar/gkq603

Wang Y, Astrakhan Y, Petersen B-S, Schreiber S, Franke A, & Bromberg Y (2017). Identifying Crohns disease signal from variome analysis. BioRxiv, 216432.

Watson T, Shantsila E, & Lip GY (2009). Mechanisms of thrombogenesis in atrial fibrillation: Virchow's triad revisited. Lancet, 373(9658), 155–166. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19135613. doi:10.1016/S0140-6736(09)60040-4 [PubMed: 19135613]

Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, . . . Klein TE (2012). Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther, 92(4), 414–417. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/22992668. doi:10.1038/clpt.2012.96 [PubMed: 22992668]

Xie J, Zhu S, Dai Q, Lu J, Chen J, Li G, . . . Xu W (2017). Oncostatin M was associated with thrombosis in patients with atrial fibrillation. Medicine (Baltimore), 96(18), e6806. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28471981. doi:10.1097/MD.0000000000006806

Zakai NA, McClure LA, Judd SE, Safford MM, Folsom AR, Lutsey PL, & Cushman M (2014). Racial and regional differences in venous thromboembolism in the United States in 3 cohorts. Circulation, 129(14), 1502–1509. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/24508826. doi:10.1161/CIRCULATIONAHA.113.006472 [PubMed: 24508826]

Zateyshchikov DA, Brovkin AN, Chistiakov DA, & Nosikov VV (2010). Advanced age, low left atrial appendage velocity, and factor V promoter sequence variation as predictors of left atrial thrombosis in patients with nonvalvular atrial fibrillation. J Thromb Thrombolysis, 30(2), 192–199. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20082208. doi:10.1007/s11239-010-0440-1 [PubMed: 20082208]

Zhai Z, Kan Q, Li W, Qin X, Qu J, Shi Y, . . . Dissol V. E. i. (2019). VTE Risk Profiles and Prophylaxis in Medical and Surgical Inpatients: The Identification of Chinese Hospitalized Patients' Risk Profile for Venous Thromboembolism (DissolVE-2)-A Cross-sectional Study. Chest, 155(1), 114–122. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30300652. doi:10.1016/j.chest.2018.09.020 [PubMed: 30300652]

Zimetbaum P (2017). Atrial Fibrillation. Ann Intern Med, 166(5), ITC33-ITC48. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28265666. doi:10.7326/AITC201703070

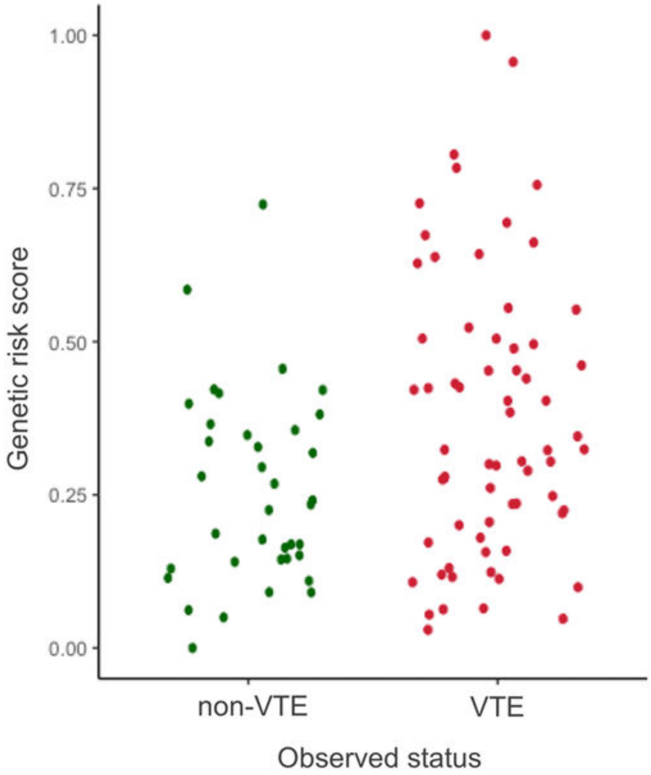

Figure 1. GRS (Method 1.3) predictions separate VTE and non-VTE individuals.

GRS (genetic risk scores) were normalized to a 0 to 1 range (Y axis). The samples are scattered within a status along the X-axis for better visibility. Note that non-VTE individuals, on average, score lower than VTE ones.

Table 1.

Method performance summary

Method#	**Method**	**Known variants/genes used**	**Function effect**	**Accuracy**[†]	**Precision**[†]	**Recall**[†]	**MCC**[†]
1.1	GRS	Heit *et al.* (3 loci)	No	37.9%	**100.0%**	3.0%	0.105
1.2	GRS	Heit *et al.* and *PROS1* variants	No	47.6%	68.8%	33.3%	0.065
1.3	GRS+warfarin dose	Heit *et al.* and *PROS1* variants	No	50.5%	89.5%	25.8%	**0.252**
2	Kmodes clustering	*Level 1, 2* genes	No	52.4%	62.3%	65.2%	0.052
3	Kmeans clustering	*Level 1, 2, 3* genes	variant-level	**62.1%**	70.8%	69.7%	0.182
4	Kmeans clustering	*Level 1, 2, 3* genes	protein-level	58.3%	65.8%	**72.7%**	0.054

[†] Overall accuracy (Eqn. 3), precision and recall (Eqn. 4), and MCC (Eqn. 5). Default cutoff of 0.5 was used for calling an exome VTE or non-VTE. The best performance among four methods is indicated in bold.

Table 2.

Significantly VTE-associated variants in AA population

Variant	**Mapped gene**	**Risk allele**	**AF[†] in AFR**	**AF[†] in EUR**
rs138916004	*LEMD3*	G	2%	0%
rs3804476	*LY86*	G	8%	44%
rs142143628	*LOC100130298, CLVS1*	T	1%	0%
rs73692310	*LOC102723446*	T	7%	0%
rs1998081	*CD93*	T	22%	8%
rs2144940	N.A.	C	26%	8%

[†] AFs (allele frequencies) of the risk allele from 1000 Genomes Project Phase 3.