A TWAS-based investigation of gene expression mediated VTE risk

William Gordon

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2018

Committee:

Sara Lindström

Bruce Weir

Alexander Reiner

Program Authorized to Offer Degree:

Public Health Genetics

University of Washington

Abstract

A TWAS-based investigation of gene expression mediated VTE risk

William Gordon

Chair of the Supervisory Committee:

Dr. Sara Lindström

Department of Epidemiology

The corpus of GWAS has been successful in identifying many genetic loci associated with a large array of diseases, phenotypes, and other outcomes. However, the underlying molecular mechanisms behind such outcomes remain challenging to elucidate. The mediation of gene expression by genotypic variability is potentially one such molecular mechanism. Here, we utilize complementary TWAS methods and leverage four distinct transcriptomic studies (n = 153 - 1414) along with a recently completed meta-GWAS of VTE risk (n = 187,204) to investigate potential correlations between gene expression and genetic risk of VTE. We have imputed predicted expression into the much larger GWAS dataset and have identified six TWAS significant genes that do not overlap with previous VTE GWAS loci. The work here demonstrates the utility of combining the large sample sizes of summary-level GWAS and the denser, less accessible gene expression datasets.

Introduction

VTE: Biology, epidemiology, and economics

Venous thromboembolism, or VTE, is a disease term that describes two consecutive but distinct cardiovascular events. The first, a deep vein thrombosis (DVT), occurs when a blood clot--or thrombosis--occurs in a large interior vein, most often in the leg such as the femoral vein. These DVTs tend to form at the valves where blood is locally hypoxic and relatively slow moving, and result in swelling and redness of the leg along with severe discomfort. They are a major health concern and are to be considered a disease outcome of their own. However, a more severe manifestation of a DVT results in an embolism: if the clot, or some portion of it, is freed from the vein wall and travels up towards the heart, it is considered an embolus and the phenomenon as a whole is known as venous thromboembolism. Upon reaching the heart, the embolus may travel to the lung where it will become lodged in the narrowing arteries, restricting the supply of oxygenated blood to the lung tissue. This often-fatal condition is the second stage of VTE known as pulmonary embolism (PE).

In the United States, most individuals suffering from symptoms of VTE present only the first phase of the disease, DVT, with approximately one third of individuals presenting the more serious PEs (White, 2003). Estimates of this proportion tend to skew towards more PE when autopsy data is analyzed rather than clinical presentation, suggesting that undiagnosed PE is enriched for in end-of-life patients (White, 2003). Estimates of incidence rates in the US center around approximately 120 per 100,000 person-years, with modestly higher incidences observed

in men (Heit, 2015). Incidences have risen slightly over the past four decades alongside a rising proportion of PE versus DVT (Heit, 2015).

Financially, VTE is a major health care cost for the United States. The average acute case of VTE costs between $12,000 and $15,000 over the first year, with cumulative costs from later complications raising the figure by approximately $7,000 (Grosse, 2016). The most conservative estimates of national expenditures in response to acute VTE incidences sum to at least $7,000,000,000 annually, however, when lost economic output and other costs are considered, estimates grow to a staggering $60,000,000,000+ (Grosse, 2016). Between the substantial loss of life, downstream disability, and massive economic cost of VTE, it is clear that a robust understanding of the causes of VTE is a warranted and necessary public health undertaking towards the prevention and treatment of the disease.

VTE: Genetic risk factors

Though many environmental and behavioral risk factors for VTE are well known including stasis, advanced age, trauma, cancer, obesity, and pregnancy, the work presented here is concerned with the genetic bases of VTE risk. Before the genome-wide association (GWAS) era, the earliest clearly understood genetic risk factors for VTE were relatively rare high penetrance variants, observable by high VTE prevalence in “thrombophilic families” (Rosendaal, 2005). The first such set of variants to be identified were those that resulted in antithrombin (AT) deficiencies, which were first described in 1965 (Egeberg, 1965). AT deficiencies result from heterozygous dominant variants, which occur at an allele frequency of less than 0.2% in the general population but are highly predictive of early age VTE with up to 85% of carriers

suffering a VTE event by the age of 50 (Martinelli, 2001). Those born homozygous for these variants are severely thrombophilic from birth and are not expected to live past infancy (Martinelli, 2001).

Two decades later, analogous heterozygous variants resulting in deficiencies of protein C and its cofactor protein S were also found to greatly increase the risk of VTE (Broekmans, 1983) (Comp, 1984). Deficiencies in these proteins among heterozygous individuals lead to an approximately 10-fold increase in VTE incidence, and as with AT deficiency, homozygous status is generally lethal (Rosendaal, 2005). The commonality between these variants is that each affected protein is an inhibitor of naturally occurring coagulation, and for both, the lowered coagulation inhibition activities in the blood result in hypercoagulability and increased VTE risk (Simioni, 1999). It is worth noting that these studies, being based on pedigrees and observed protein activities, do not necessarily point to an individual genetic locus as the sole cause (or quantified portion of a cause) of elevated VTE incidence (Rosendaal, 2005). Subsequent studies have shown that genotypes in "thrombophilic families" harbor multiple variants that contribute to VTE risk, and that these deficiencies of anticoagulant inhibitors in individuals with "thrombophilic backgrounds" tend to have more pronounced and earlier acting VTE risks than similar deficiencies without the "thrombophilic background" (Rosendaal, 2005).

Blood type has also been known to be associated with VTE risk as far back as the late 1960s, but the mechanistic underpinnings were unclear (Jick, 1969). The initial observation which prompted investigation was that O blood types were underrepresented in VTE sufferers, while non-O blood types conferring approximately twice the risk versus O types, with a greater effect seen in pregnant women (Jick, 1969). Later studies would suggest that the association is

due to higher levels of von Willebrand factor and factor VIII in non-O blood types, themselves both VTE risk factors (Martinelli, 2001) (Rosendaal, 2005). The fundamental difference between blood type and protein deficiencies as risk factors is that the former is relatively common and of modest effect, while the latter are quite rare and produce pronounced risks. With this in mind, it is useful to consider the merits of relative risk versus population attributable risk when considering the impact of different risk factors.

In terms of population attributable risk, the factor V Leiden variant (factor V R506Q, G1691A) is a major cause of VTE, largely due to its high prevalence in some populations, especially individuals of European ancestry (5% are heterozygotes) (Rosendaal, 2005). The variant prevents proper inactivation of factor V, a clotting factor, leading to hypercoagulability, and confers a relative risk for VTE somewhere in the range of 3-7 with a US population attributable risk of at least 20% (De Stefano, 1995) (Rosendaal, 2005).

A variant related to factor V Leiden--prothrombin G20210A--also deserves mention. This mutation is private to European ancestry populations, and it is approximately twice as common among Southern Europeans (4%) versus Northern Europeans (2%), while the opposite is true for factor V Leiden (Martinelli, 2001). The relative VTE risk for G20210A carriers falls between 2 and 4, and its mechanism is conceptually opposite that of factor V Leiden: prothrombin G20210A is a gain-of-function mutation for the coagulation system, with carriers' blood prothrombin levels observed to be 30% higher than that of matched non-carriers (Martinelli, 2001). As neither variant is exceedingly rare in European ancestry populations, researchers have been able to identify and study individuals who carry both variants. Martinelli et al.'s work has

found that the relative risk of primary VTE for dual-carriers is approximately the product of the relative risks for single-carriers of both mutations (Martinelli, 2000).

The advent of GWAS has enabled researchers to investigate genetic risk factors in a manner agnostic to the causal mechanisms described in each of these previous examples. By scanning the genome for marker SNPs closely associated with outcome, it is possible to identify genomic loci that are likely harboring variation important to VTE risk. The corpus of such studies has produced a modest set of loci clearly associated with VTE (Trégouët, 2009) (Kong, 2014) (Desch, 2015), with the aforementioned ABO (blood type) and F5 (Factor V) genes dominating the GWAS signal space. Many additional variants have been identified that have suggestive association with VTE, with some loci colocalizing with coagulation family genes, and others appearing near genes with unknown relations to coagulation or VTE (Morange, 2015). Often, these "hits" cannot explain how (or if) a genetic variant causes VTE risk to change, and the fact that many hits occur outside of coding regions suggests that their association with VTE is not a result of alternative protein structure.

Expression-mediated phenotypes and TWAS

Of the pre-GWAS genetic risk factors identified, Protein C and Protein Cofactor S offer some insight into one possible alternative explanation for the causal pathway of genetic risk factors. Laboratory work has showed that abnormalities of these two proteins in thrombophilic families fell into two distinct types: Type 1 deficiencies belong to those in which the concentration of these proteins in the blood is lower than expected, while in Type 2 deficiencies the structures of these proteins are modified resulting in lower activities (though their

concentration may be normal) (Bertina, 1984) (Doray, 1986). The observation of Type 1 deficiencies, along with the fact that some GWAS signals fall outside of coding regions, suggest that heritable gene expression patterns may play a role in determining phenotypic outcomes such as VTE (Germain, 2015).

The obvious method for investigating a potential link between gene expression and phenotype would involve the comparison of tissue-specific gene expression data with outcomes of interest. However, the transcriptomic data--whether generated through RNA arrays or RNA sequencing, are costly, noisy, difficult to analyze, require prospective samples, and represent transient expression signals which must be integrated over time when scanning for associations with long-term outcomes such as VTE. In contrast, the cost of GWAS genotyping based on SNP arrays ("SNP-chips") has plummeted to less than $95 per sample for basic, "core" arrays with the most comprehensive ready-made arrays costing less than $500 per sample (CIDR Pricing, 2018). The substantially denser data produced by Whole Genome Sequencing (WGS) have also become within reach of many more researchers as prices have fallen to under $1,000 per sample for low pass (4X), and under $2,500 per sample for high pass (30X) services (CIDR Pricing, 2018). An additional attractive feature of GWAS data is its availability: thanks to databases of GWAS data from previously published work, such as dbGaP, large amounts of GWAS summary data (and sometimes individual-level data) can be relatively easily obtained without the cost and effort required for novel genotyping (Mailman, 2007). An investigative method that could leverage this readily available genotype-phenotype data towards understanding the relationship between a gene's expression and a phenotype of study could potentially be an efficient and practical alternative to costly transcriptomic techniques. In recent years, a variety of related methods have

attempted to accomplish this; these methods have been collectively referred to as Transcriptome Wide Association Studies (TWAS).

The broad goal of TWAS is to identify correlations between gene expression and phenotype. More specifically, most TWAS methods attempt to identify correlations through one of two strategies: the first calculates correlations between *predicted* expression of genes and a GWAS phenotype, while the second calculates posterior probabilities for colocalization of causal variants acting on both gene expression and GWAS phenotype. The former strategy ("predicted expression TWAS") is the basis for the FUSION TWAS analysis described here, while the latter ("colocalization TWAS") is also used in this analysis in a complementary manner.

Predicted expression TWAS

"Predicted Expression TWAS" is conducted in two distinct steps: gene expression prediction, followed by analysis of associations observed between these predicted gene expressions and phenotype as predicted by GWAS. The gene expression prediction step requires individual-level genotype data and tissue-specific gene expression data from the same individual. As mentioned, this data is costly to acquire, and not easily transferable between studies based on the temporal and tissue-specific nature of the data. Here, it is used to build gene expression prediction models based on the correlation between cis-genetic variability and the expression of a local gene.

In the simplest conceptual situation, a single expression quantitative locus (eQTL) near a gene explains the variance of the nearby gene's expression; thus, knowledge of the eQTL's genotypes in a sample from a second dataset can be used to predict the expression of the gene in

that sample. This second dataset is a simple summary-level GWAS matrix containing variant genotypes and some measure of association with outcome (e.g. a Z-score or *p* value). Previous work has used the simple single eQTL expression prediction framework under the name of "eQTL guided GWAS" or "eGWAS," however, these methods ignore many potentially valuable predictors in the form of secondary eQTLs (Zou, 2012). Thus, though the single eQTL example is simplest to conceptualize, the method is not the most useful when many cis-variants are associated with a given gene's expression, as it does not allow for the creation of a prediction model based on the genotype of multiple variants. For the purposes of TWAS, this prediction model can be conceptualized as a reweighting of each SNP's association with the outcome: if a given gene's expression is predicted to be strongly upregulated or downregulated by its best eQTL, the gene's constituent GWAS signals will be reweighted accordingly.

For more complex prediction models, the aggregate effect of weighted-SNPs within the neighborhood of a given gene (cis-eQTLs), after accounting for linkage disequilibrium (LD), determines the TWAS statistic for the gene-outcome association. The exact method for creating these multivariate prediction models must be selected by the researcher, with our selection of prediction models discussed later in this paper.

A number of analytical pipelines have been developed based on this general approach, including *PrediXcan/MetaXcan* (Barbeira, 2016) and *Summary-data-based Mendelian Randomization (SMR)* (Wu, 2018). Here, we use the Gusev Lab's *FUSION* pipeline (Gusev, 2016) which is primarily a Predicted Expression TWAS approach, though colocalization analysis is included for additional insight.

Colocalization TWAS

"Colocalization TWAS" is a distinct but complementary approach to TWAS. Rather than scanning the correlations between a cis-predicted gene expression signal and a GWAS signal across genes, colocalization TWAS methods estimate the distribution of estimated effects amongst variants in the queried gene's "neighborhood" for both expression and phenotype associations. Based on the overlap of these two distributions, posterior probabilities are calculated for five possible scenarios: (0) No signal, (1) GWAS signal only, (2) eQTL signal only, (3) Independent GWAS and eQTL signals, and (4) Colocalized GWAS and eQTL signals. For genes with significant TWAS statistics based on the Predicted Expression approaches, it is expected that the probabilities of the 3rd and 4th scenarios will dominate, since these genes will have correlated GWAS and eQTL signals. For these genes, a strong probability of colocalization is evidence of a shared causal variant or variants, which in turn is evidence of a cis-expression mediated outcome. A major difference between these approaches and the Predicted Expression approaches is that the colocalization approach is agnostic towards the strength of the genetic effect on phenotype and expression. In other words, Predicted Expression TWAS aims to identify genes that are associated with phenotype through cis-expression mediation, while Colocalization TWAS aims to identify genes that harbor cis-variants with effects on both expression and phenotype, regardless of their magnitudes. Recent examples of published Colocalization TWAS analysis pipelines include *coloc* (Giambartolomei, 2014) *, enloc* (Wen, 2017)*, Sherlock* (He, 2013)*,* and *eCAVIAR* (Hormozdiari, 2016).

Goals of analysis

The work presented here is the first TWAS of VTE conducted to identify genes whose expression are associated with VTE risk. Specifically, we aim to identify genes that have a strong correlation between their cis-genetic component of expression in whole blood and/or liver tissue and their genetic component of VTE risk. Such signals will be considered candidates for genes involved in the manifestation of VTE through genetic control of expression, and these candidacies will be strengthened if analysis suggests colocalization of variants responsible for correlation of predicted expression and VTE association. This technique can potentially identify candidate genes not previously identified by GWAS-only analyses, and may also offer clues towards the mechanism by which previously known VTE GWAS loci affect VTE risk.

Methods

FUSION

The FUSION TWAS pipeline relies upon three main data inputs: summary-level GWAS statistics for the phenotype of interest, individual-level and tissue-specific gene expression data alongside individual-matched genotype data, and a linkage disequilibrium reference panel representative of the studied population (Gusev, 2016).

VTE meta-GWAS dataset

In investigating the correlation between the genetic component of a phenotype and its predicted cis-expression component, a well-powered GWAS is crucial for estimating the genetic component of this relationship. Here, we have used an unpublished meta-analysis of 20 VTE

GWAS as part of the INVENT consortium (Table 1) (INVENT 2016). Together, these studies sample a total of 29,435 VTE cases and 157,769 controls; individuals sampled were all of European ancestry. All studies were imputed to either 1000 Genomes (1000 Genomes Project, 2015) or the Haplotype Reference Consortium (McCarthy, 2016). For each study, we excluded variants that had imputation quality score $r^2 < 0.3$ or a minor allele count in either cases or controls less than 20. We only included autosomal chromosomes and excluded variants that were seen in less than three of the studies. In total, we assessed 12.7M variants.

Previous GWAS have identified 14 unique loci associated with VTE risk at genome wide significance ($p < 5.0 \times 10^{-8}$) (Germain, 2015); the greatly increased power of this meta-analysis identified an additional ~10 novel loci (unpublished data). The sample size of the meta-analysis lends substantial power to the association analysis, resulting in extreme *p* values for the most significant "hits," for example, the *p* values for known risk factors Factor V and ABO (blood group system) are less than 10^{-171}.

Many of these loci, both known and novel, are found in intronic, intragenic, or otherwise non-coding regions. This suggests that if these variants are causal, their causal mechanism is not based on alternative coding of RNA transcripts (nor their resultant protein structures). One alternative causal mechanism is regulation of expression (GTEx Consortium, 2015). Here, we investigate the possibility that a set of these VTE-associated loci are acting on VTE risk through such mechanisms--specifically, through cis-regulation of the expression of nearby genes. Additionally, by "reweighting" GWAS variants based on their association with local gene expression, it is possible that some GWAS loci that did not previously meet genome-wide significance will be found to be TWAS-significant. The aggregative nature of studying

associations at the gene-level--rather than SNP-level--also makes it possible that a collection of SNPs which do not reach the significance threshold on their own can reach the gene-level significance threshold when considered as a gene-unit.

Expression datasets

In order to investigate a potentially expression-mediated mechanism, it is necessary to estimate the association between each gene's cis-genotype (i.e. a gene's "genetic neighborhood") and its expression. We obtained these estimates by leveraging available datasets containing both genotype and tissue-specific gene expression measurements in a sample of individuals. From this, a gene expression prediction model can be created based on genotype for each gene and tissue. For example, cis-genetic variants that are robustly associated with upregulation or downregulation of the gene of interest (in a given tissue) would be assigned relatively large coefficients in the prediction model.

Ideally, these expression datasets will pertain to gene expression in tissues believed to be relevant to the outcome of interest. Here, we have focused on datasets measuring expression in whole blood, an obvious choice due to the hematological nature and physiological occurrence of VTE. These datasets also have the additional advantage of having relatively large sample sizes resulting from the relevance of blood expression to a wide collection of outcomes as well as the non-invasive nature of sampling. Specifically, we have used RNA-array expression/genotype data from the Young Finns Study (YFS, n = 1414) and the Netherlands Twin Registry (NTR, n = 1247), as well as RNA-seq expression/genotype data from the GTEx Consortium (GTEx, n = 338 (version 6) / n = 369 (version 7)) (Raitakari, 2003) (Willemsen, 2013) (GTEx Consortium,

2015) (Table 2). The individuals sampled in these datasets were of European ancestry (Raitakari, 2003) (Willemsen, 2013) (GTEx Consortium, 2015).

In addition to the blood expression data, the GTEx Consortium's liver-derived expression data (n = 97 (version 6) / n = 153 (version 7)) were used in a parallel analysis. Our interest in the genetic predictors of gene expression in liver comes from the organ's involvement in coagulation and general hemostasis. The liver is the site of the synthesis and degradation of the majority of the body's coagulant, anticoagulant, and fibrinolytic proteins (Fasel, 2007). A previous RNA-seq based transcriptomic investigation has found that the KEGG pathway annotation "Complement and coagulation cascades" is the single most enriched pathway ($p = 10\text{x}10^{-39}$) in liver tissue (Kampf, 2014). As such, variation of gene expression in liver tissue and its correlation with the genetic component of VTE risk is of great interest.

From these genotype/expression datasets, prediction models can be constructed for each gene's expression based on the variants found within 500 kb in either direction of the gene. Because it is unreasonable to believe that a gene without heritable variance in expression could have a meaningful expression prediction model based on cis-genetic variants, only genes with a significant ($p < 0.01$) expression heritability were included in each analysis. As a consequence of this thresholding, the size of the gene set is dependent on the dataset's sample size.

Though we had initially used the GTEx Consortium's version 6 data, over the course of the project the Consortium released their 7th version containing additional samples. These updated datasets increased whole blood sample size from 338 to 368 individuals and liver sample size from 97 to 153 individuals. By recomputing our "weights" for predicted expression imputation with the expanded datasets, we were able to glean some understanding of the

dependence of the expression prediction modeling on sample size. For both liver and whole blood datasets, increased sample sizes lead to modest improvements in our gene expression prediction models, as estimated by the r^2 values generated through five-fold cross-validation (Figures 1, 2). Perhaps more importantly for our aims, the increase in sample size had a strong effect on the number of genes for which the heritability of expression was considered significant ($p < 0.01$), thus allowing for the TWAS analysis of a larger set of genes (Figures 1, 2). Based on these improvements, the version 6 GTEx dataset was dropped from further analysis in favor of GTEx version 7.

Computing expression "weights"

From these expression/genotype datasets, it is possible to build expression prediction models in a variety of ways. Here, we have chosen to build three models for each gene, with the most predictive model selected for TWAS analysis. Models are ranked based on their r^2 values, determined by 5-fold cross validation of the training data. The 1000 Genomes Project Reference Genome is used as an LD reference panel to impute expression "weights" for SNPs within the GWAS dataset.

The first model, known as "top1," is conceptually straightforward and essentially identical to "eGWAS" analysis: the 1Mb window surrounding the gene of interest is scanned for variants that are closely associated with the gene's observed expression, and the single best predictor is chosen as the sole independent variable in the prediction model. Thus, the prediction model for each gene is a simple linear regression with a single explanatory variable.

The second model is based on the LASSO (Least Absolute Shrinkage and Selection Operator) regression analysis, which penalizes regressions coefficients, potentially producing a more useful balance between bias and variance, at the cost of departure from a truly unbiased method (Tibshirani, 1996). Benefits are especially pronounced for the sort of highly dimensional modeling we are attempting here, partially due to LASSO's effect of dropping variables of negligible predictive power (i.e. setting regression coefficients to zero) (Tibshirani, 1996). The LASSO beta estimates are defined as:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$ (Tibshirani, 1996)

Lastly, the third model relies on Elastic Net, an extension of LASSO which combines LASSO's penalization with the penalization of the Ridge method in order to better cope with situations in which there are large numbers of predictors relative to the number of observations (Zou, 2005). Briefly, Elastic Net extends LASSO by using a penalization scheme that is a compromise between LASSO's penalties ($\boldsymbol{L_1}$) and Ridge Regression's penalties ($\boldsymbol{L_2}$) (Zou, 2005). Elastic Net tends to outperform LASSO when predictors are tightly correlated (Zou, 2005). Elastic Net beta estimates are defined as:

$$\underset{\beta}{\operatorname{argmin}} (\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1)$$ (Zou, 2005)

In each model, every variant included within the model is assigned a beta coefficient, which can conceptually be considered a variant-specific "weight" for the GWAS statistic. The result is a linear model that aggregates the product of these coefficients with a precomputed vector of GWAS statistics for a given gene. Thus, a TWAS statistic can be generated for any

GWAS dataset that contains GWAS statistics for each variant in the model. For the NTR and YFS datasets, precomputed weights were used, while we calculated our own weights for the GTEx datasets. Expression levels in the GTEx datasets were adjusted based on sex, platform, three genetic principal components and 30 PEER factors, as per the GTEx Consortium's recommendations (Stegle, 2012).

Testing correlation of associations

Once the best model for each gene has been selected, the gene's TWAS score can be calculated based on the vector of SNP Z-scores, their correlation (LD), and their corresponding gene prediction weights. The TWAS test statistic (TWAS Z-score, Z_{TWAS}) is calculated for each gene as:

$$Z_{TWAS} = w'Z/(w'Dw)^{½}$$

Where *w* represents weights and D represents the LD matrix (Gusev, 2016). We restricted our analysis to SNPs with a minor allele frequency greater than 0.01.

COLOC analysis

Finally, for genes that showed a significant TWAS Z-score, we reanalysed the relationship between expression-associated loci and phenotype-associated loci using COLOC. COLOC is a Bayesian statistical procedure that estimates the probability that two observed association signals are consistent with a shared causal variant (Giambartolomei, 2014). We are particularly interested in the posterior probability of "Hypothesis 4" from this method, which suggests that a single shared causal variant is behind both associations at a given locus.

Results

Overview

Each of our four separate TWAS analyses identified between 3 and 16 genes whose predicted expression is significantly associated with the genetic component of VTE risk (Table 3). Significance was conservatively determined at a Bonferroni corrected α of 0.05, wherein the number of multiple comparisons corrected for was the summation of the number of genes tested across all datasets (i.e. 0.05 / (5959 + 2875 + 2448 + 4690) = 3.13×10^{-6}) (Bonferroni, 1936). We have produced TWAS Manhattan plots of *p* values across the genome (figures 3, 4, 5, and 6); it is worth noting that these plots do not exhibit the "skyscrapper" behavior of GWAS plots due to the near-independence of neighboring genes (as opposed to the LD-induced dependence of neighboring GWAS SNPs).

After filtering these significant TWAS genes based on permutational analysis, 6 genes that did not overlap with previously identified GWAS significant SNPs remained significant. Under these stringent restrictions, no genes without previous GWAS significance were replicated across TWAS based on different expression prediction datasets, and no genes were found to be significant in the TWAS based on NTR blood data. Four of these genes--*SPSB1*, *RP11-747H7_3*, *SBNO1*, and *SH2B3*--demonstrated positive TWAS Z-scores, indicating that their predicted expression was positively correlated with VTE risk. The other two genes in this set--*ERAP1* and *SBNO1*--demonstrated negative TWAS Z-scores. All absolute TWAS Z-scores were in the range of 4.87 to 5.27 corresponding to *p* values ranging from 1.09×10^{-6} to 1.34×10^{-7}.

Permutation analysis

A substantial caveat to these results is the fact that while the TWAS association statistic is well-behaved when no major GWAS association is present, the statistic can be inflated if a QTL signal colocalizes with a strong GWAS signal in a region of extensive LD (Gusev, 2016). In other words, a modest by-chance association between phenotype and predicted gene expression can produce a significant TWAS statistic if the gene in question is located next to an especially strong GWAS signal. In practice, this results in GWAS "bleed over," wherein some genes in the immediate vicinity of the most significant GWAS peaks may be spuriously detected as TWAS significant.

A straightforward method for checking if this coincidental colocalization is driving the TWAS signal is to run a permutational analysis on each TWAS "hit" in which QTL weights are shuffled many times (here, adaptively up to 10,000 times) while keeping the GWAS signal's locus fixed. Doing so produces an empirical null distribution of *p* values that the original TWAS *p* value can be compared to. The empirical *p* value is then defined as the proportion of times the observed *p* value is as extreme or more extreme than each of the 10,000 permuted *p* values. If the observed *p* is not more extreme than the vast majority of empirical p values, the GWAS signal is likely driving the TWAS signal and the gene should be discarded as a false TWAS signal. By including only those signals with empirical *p* values less than 0.05 (i.e. at least 99% of permuted TWAS *p* values are less significant) we can filter out the vast majority of "hits" that arise from coincidental colocalization of QTL signals with strong GWAS signals. Here, this filtering reduced our total number of significant TWAS signals across all four datasets from 42 to 33 (Table 3).

Colocalization TWAS

As discussed in the methods section, a complementary approach towards better elucidation of TWAS signals involves SNP-level analysis of colocalization between predicted expression signal and GWAS signal. Because both signals are comprised of distributions of effects across a locus, the overlap of these distributions can be analyzed within a Bayesian framework to determine the probability that the causal variants between the signals are shared. Table 3 includes counts of TWAS significant genes with probabilities of colocalization greater than 0.5 (i.e. those more likely than not to be sharing a causal variant between signals). Figure 8 demonstrates the distribution of various colocalization scenarios in greater detail; in these plots of TWAS significant genes, dark green area represents the probability of colocalization, while light green area represents the probability of independent GWAS and eQTL signals. The less prevalent yellow, orange, and red areas represent the probabilities of a GWAS signal alone, an eQTL signal alone, and no signal at all, respectively. As expected for TWAS significant genes, the vast majority of candidate genes have posterior probabilities dominated by the hypotheses that include both expression and GWAS signals, with these two hypothesis probabilities approximately evenly split across the genes. A smaller set of genes have substantial GWAS-only posterior probabilities, and these genes tend to be those that do not pass permutation filtering, which is consistent with the assumption that they are the result of the GWAS "bleed over" effect.

GWAS overlap

In relation to traditional GWAS, TWAS can serve two distinct, complementary purposes. Should a TWAS signal replicate a GWAS signal, the replication serves as evidence that the relationship between genotype and phenotype may be mediated through differential expression of a cis-gene. Alternatively, a TWAS can "boost" the signal of marginal GWAS signals by considering the effect of predicted expression, resulting in a TWAS signal that does not overlap with a previous GWAS signal. Table 5 presents the collection of such signals identified here. A "replicated" signal, in this context, can be defined as one in which at least one GWAS variant within a TWAS gene's expression prediction model meets genome-wide significance. The majority of significant TWAS signals identified in this analysis overlap with SNPs identified as GWAS significant in the VTE meta-GWAS, as shown in Table 4.

Conditional analysis

In the search for genes that affect outcome via genetic control of their expression, an intuitive sanity check can be conducted via a counterfactual conditional: if a TWAS signal is the result of the mediation of GWAS outcome by predicted gene expression, conditioning on the predicted gene expression should mitigate the GWAS signal. To investigate if this behavior occurs in candidate genes we created GWAS plots conditional on predicted expression of TWAS candidate genes, with eQTL data included for interpretation of colocalization. *SLC44A2* has previously been identified and replicated as a gene local to a significant GWAS signal, and our TWAS work based on both GTEx Blood and YFS Blood datasets suggests that this may be due to genetic control of its expression (Figure 9). In the GTEx 7 Whole Blood dataset, this is

supported by substantial heritability of expression ($h^2 = 0.285$), a highly significant TWAS *p* value (3.5×10^{-15}), an extremely high probability of colocalization between GWAS and eQTL colocalization (0.987), and a significant *p* value from permutational analysis ($p_{perm} = 0.0001$). Indeed, the conditional GWAS plot supports this observation, as shown in Figure 9: the grey points show unconditional GWAS significance, which is largely mitigated after conditioning on the gene's predicted expression, shown as blue points. The lower panel shows that the set of SNPs acting as eQTLs on the gene are clustered around the gene locus, as would be expected for a colocalized signal.

Alternatively, TWAS analysis of some genes may produce significant scores but, upon closer inspection, suggest spurious correlation. Again working from the GTEx7 Whole Blood dataset, the conditional plot of *NME7* shows that it is in the neighborhood of the extremely GWAS-significant *F5* (Factor V) locus. Conditioning on the predicted expression of *NME7* has very little effect on the GWAS scores, suggesting that it is not a dominant factor with regards to VTE risk. Further, the bulk of eQTLs for the genes fall between *NME7* and *F5*, suggesting that these may be shared eQTLs, or that some *NME7* eQTLs may be in LD with *F5* GWAS signals, both potential causes of spurious correlation. These plots, along with premutational analyses and colocalization analyses, demonstrate that TWAS scores should not be interpreted alone as evidence of a causal mechanism; the confidence of a TWAS score describes only the probability of correlation between predicted gene expression and GWAS phenotype, and should be followed up with additional analyses.

Best candidate genes

By applying restrictions based on all of these measures (TWAS significance, permutation *p* value < 0.01, Pr(colocalization) > 0.9, support from visual inspection of conditional analysis), we've created a short list of best candidate genes (Table 6). The previously mentioned *SLC44A2* is included in this table twice as it is identified as a best candidate gene from both GTEx7 and YFS Whole Blood expression prediction models. A mutation in this gene has previously been identified through replicated GWAS and fine mapping to be associated with VTE. It should also be noted that the lncRNA *ILF3-AS1* shares a TWAS eQTL "neighborhood" with the neighboring *SLC44A2*, suggesting that there may be a "bleed over" effect between the two. However, as *SLC44A2* has previously been reported in GWAS as a VTE-related locus, as well as the fact that it is the only TWAS significant gene replicated across datasets under the most stringent restrictions here, there is some reason to believe that *ILF3-AS1* is a spurious result (Germain, 2015).

Discussion

Overview of findings and biology of candidate genes

Our analysis has identified six TWAS significant genes (after permutation filtering) that do not overlap with previous VTE GWAS loci. The known biology of two of these genes is worth noting. *ERAP1* has previously been associated with blood pressure in a meta-GWAS, and mutations of *ERAP1* have been linked to ankylosing spondylitis, a long term, mobility limiting inflammation of the spine (Tragante, 2014) (Brionez, 2008). Such long term mobility restrictions could possibly act as a risk factor for VTE through hemostasis. Perhaps more suggestively,

SH2B3 (also known as lymphocyte adapter protein) is a regulator of signaling pathways involved in hematopoiesis (Devallière, 2011) and is involved with blood diseases and diseases of the vasculature (Levy, 2009) (Perez-Garcia, 2013). Mutations in the gene have previously been found to be strongly associated with thrombophilia in antiphospholipid antibody carriers (Ochoa, 2013).

An alternative ranking of candidate genes placed restrictions on colocalization prior probabilities, while allowing for genes that overlapped with previous GWAS significant loci. The purpose of this ranking was to identify genes that are most likely to be affecting VTE risk through gene expression mediation, regardless of whether or not their signals had been identified by GWAS. Even with this relaxation, three of the seven candidate genes do not overlap with previous GWAS loci. These genes include the two lncRNAs *RP4-737E23_2* and *RP11-747H7_3*, as well as the Innate Immune System related *SPSB1*. As mentioned, *ILF3-AS1* may be a spurious result, arising from its proximity to *SLC44A2,* a gene shown to be associated with VTE risk by previous GWAS and whose mechanistic relationship with disease is an area of active study (Heestermans, 2016).

Little is known about the association between *EIF5A* and VTE, however, the protein product is noteworthy in that it is the only known protein to contain the amino acid hypusine (Kaiser, 2012). Posttranslational hypusination of EIF5A is essential for its function as a cell proliferation promoter, and the protein has been found to be involved in the pathogenicity of a wide range of diseases, including HIV, diabetes, malira, and some cancers (Kaiser, 2012).

Strengths of analysis

The main purpose of the TWAS technique described here is to extend a basic GWAS analysis by predicting the expression of genes potentially involved in the GWAS signal, and in doing so to consider the potential for expression-mediation as an explanation for genotype-phenotype association. TWAS results, in comparison to their constituent GWAS results, can prioritize loci for subsequent expression-based analysis. Additionally, some loci that were not observed to be GWAS-significant due to an underpowered GWAS design may be elevated to significance when predicted expression is considered. To this end, shifting from SNPs to genes as the basic unit of measure reduces the number of comparisons, allowing for a lower multiple comparison corrected significance threshold. Further, the complementary nature of prediction expression and colocalization approaches can provide better insight into the nature of the relationship between expression associated and phenotype associated variants around a gene.

Though TWAS, like GWAS, can only identify associations rather than causality, TWAS results can be used to rule out expression-mediation as a causal pathway for some GWAS signals. An extremely strong GWAS signal (such as, here, Factor V) with little evidence of correlation between predicted gene expression and genotype-predicted phenotype indicates that variation in cis-expression mediators are not responsible for the observed association. This, in turn, strengthens the argument that other causal mechanisms (such as a change in protein function) are at work.

As previously noted, the construction of expression prediction models based on many cis-variants allows for more informative and flexible predictions than the single eQTL guided

"eGWAS" technique. Beyond the obvious scenario in which multiple cis-variants may have a causal relationship with gene expression, the multivariate nature of the prediction models can also provide better predictions when multiple cis-variants measured act as markers for one or more unmeasured causal variants. Additionally, it is not necessary for any one variant to be significantly associated with a given gene's expression to be included in a prediction model (Gusev, 2016).

Assumptions

Because expression prediction models only consider the genotype of variants within 500kb of the gene of interest, it is assumed that all eQTLs are cis-acting, and any long range or trans-acting eQTLs will not be considered. This is a practical limitation of the computationally intensive nature of prediction model construction.

The TWAS framework also assumes that the causal pathway consists of cis-variants affecting gene expression, which in turn affects phenotype. This could be violated if, for example, genotype was affecting phenotype through some other pathway, and the phenotype was affecting the expression of the gene downstream (Gusev, 2016). A similar causality concern arises from the possibility that eQTLs shared between multiple nearby genes are associated with both the expression of one gene unrelated to phenotype as well as with SNPs in tight linkage to a strong GWAS signal that is unrelated to expression. In such a situation, the correlation between the predicted expression of the gene and the separate GWAS signal may produce a spurious TWAS signal, referred to as "bleed over" in the results section here.

Appropriate tissue choice is another assumption made by this analysis. We assume that if a gene is heavily mediating VTE risk through its expression levels, this expression variance will be observed in blood and/or liver tissue. However, if the meaningful variance is private to an unmeasured tissue, it cannot be discovered by our analysis. Another expression-based assumption is that, because of the linear nature of the expression prediction models, allele dosage acts linearly. In other words, we assume that expression-mediating allele dosages act additively.

Additionally, any SNP that cannot be imputed cannot contribute to either the prediction models nor the genotype-phenotype associations. Therefore, any rare variants that are not directly genotyped nor present in our LD reference panel will not be considered in the analysis. At the gene level, a similar restriction arises from heritability of expression: if a gene's predicted expression does not exhibit significant heritability, it's expression cannot be considered to be under genetic control and is therefor dropped from analysis. This is largely dependent on sample size and is further discussed in the methods section.

Consequences of expression dataset choice

It is clear that the choice of expression dataset is an essential consideration in the design of a TWAS experiment, as the results produced by our TWAS studies based on whole blood expression data vary considerably. We exclude the liver expression datasets from consideration here as they are not directly comparable to the blood datasets. Because expression data are inherently noisy, the quality of the dataset will be a crucial factor in the construction of expression prediction models for imputation of GWAS summary statistics. From our limited experience with four different whole blood expression datasets, it appears that the RNA-seq

datasets are substantially more useful than RNA array datasets. Though the sample sizes of the RNA array datasets (*n = 1247, n = 1414)* dwarf those of the RNA-seq dataset (*n = 369)*, our TWAS analysis identified more significant genes when using the smaller RNA-seq dataset for prediction modeling.

Comparing GTEx v6 and v7 for both liver and whole blood, we've found that sample size increases, at least in the range observed in our GTEx datasets, have a more prominent effect on the number of genes with significantly heritable expression rather than the mean model prediction accuracy (Figures 1 and 2). This is a crucial improvement, as it widens the scope of genes available for analysis, and it is expected that future expression datasets will allow TWAS analysis of a more nearly complete gene set. However, improvement in prediction models is another method for improving TWAS confidence, though one that can potentially be sought through Bayesian statistical methods such as BLUP and BSLMM without investing more resources into expression sampling.

Investigative TWAS vs. confirmatory TWAS

The analysis conducted here was investigative in nature, in that we scanned the entire genome for potential TWAS signals without any prior knowledge of likely TWAS significant genes. An alternative approach to TWAS could be confirmatory in nature: given a subset of genes that are believed to potentially be affecting phenotype through their expression, a TWAS analysis could be conducted on only these genes. This approach would save considerable computational costs by reducing the number of gene expression prediction models constructions, and could improve power by reducing the burden of multiple comparison correction.

In a similar vein, if such a subset of genes was known, but the researcher was agnostic towards the physiological localization of effect, a hybrid confirmatory/investigative TWAS could be conducted in which only the subset of genes was included, but with prediction models created from a large collection of expression datasets. One such collection could be the entire set of GTEx 7 expression datasets.

Two approaches to TWAS

The two approaches to TWAS presented here, Predicted Expression TWAS and Colocalization TWAS, are fundamentally distinct and complementary. We have focused primarily on the former as it is equipped to handle relative differences in effect size for both predicted expression and predicted outcome (i.e. GWAS estimated risk). Because the relative risk of VTE is of primary importance to this work, it is rational that this type of TWAS analysis should be considered the primary finding. Conversely, the Colocalization TWAS statistics are agnostic to effect size and are only concerned with probabilities of signal colocalization. This provides helpful complementary information, but is of little use on its own towards investigating the genes which have the greatest impact on VTE risk via their genetically controlled expression. It should also be understood that a gene's relationship with a phenotype or risk can still be of interest in terms of expression even without colocalization: for example, it is possible that multiple gene variants mediate phenotype through both change in gene expression as well as change in coding. However, a strong probability of colocalization between predicted expression and predicted phenotyped is evidence of a shared causal variant modifying phenotype through gene expression, which supports a hypothesis of outcome via gene expression mediation.

Implications and future directions

The analysis here has identified a set of genes worthy of follow-up. Such follow-up could take the form of fine-mapping, targeted RNA-seq, or knockout/knockdown analyses. Investigation of VTE risk via gene expression mediation in other tissues could also be of interest.. It is expected that this method will be refined in coming years by improvements in expression prediction modeling, as well as imputation methods and transcript calling. Finally, as with so many research methods in the omic era, increased sample sizes for both GWAS and expression datasets will undoubtedly increase the capacity for discovering true associations.

Figures and Tables

Study	Design	n Cases	n Controls
JUPITER	Case-control	77	8672
CHS	Cohort	95	3024
FHS	Cohort	222	7629
ARIC	Cohort	241	8646
EOVT	Case-control	411	1228
Tromso	Case-control	528	526
WGHS	Cohort	618	22032
WHI	Cohort	622	9139
HUNT	Cohort	811	4392
MAYO	Case-control	1238	1287
MEGA	Case-control	1289	1049
MARTHA	Case-control	1542	1110
eMERGE	Case-control	1558	10027
HVH	Case-control	1684	1641
HPFS/NHS/NHSII	Cohort	4636	33028
UKBB	Cohort	13863	44339
Total	Meta	29435	157769

Table 1. Constituent GWAS of meta-GWAS

Dataset	n	Platform
GTEx 7 (Blood)	369	RNA-seq
GTEx 7 (Liver)	153	RNA-seq
GTEx 6 (Blood)	338	RNA-seq
GTEx 6 (Liver)	97	RNA-seq
Netherland Twins Registry (Blood)	1247	RNA Array
Young Finns Study (Blood)	1414	RNA Array

Table 2. Expression training datasets

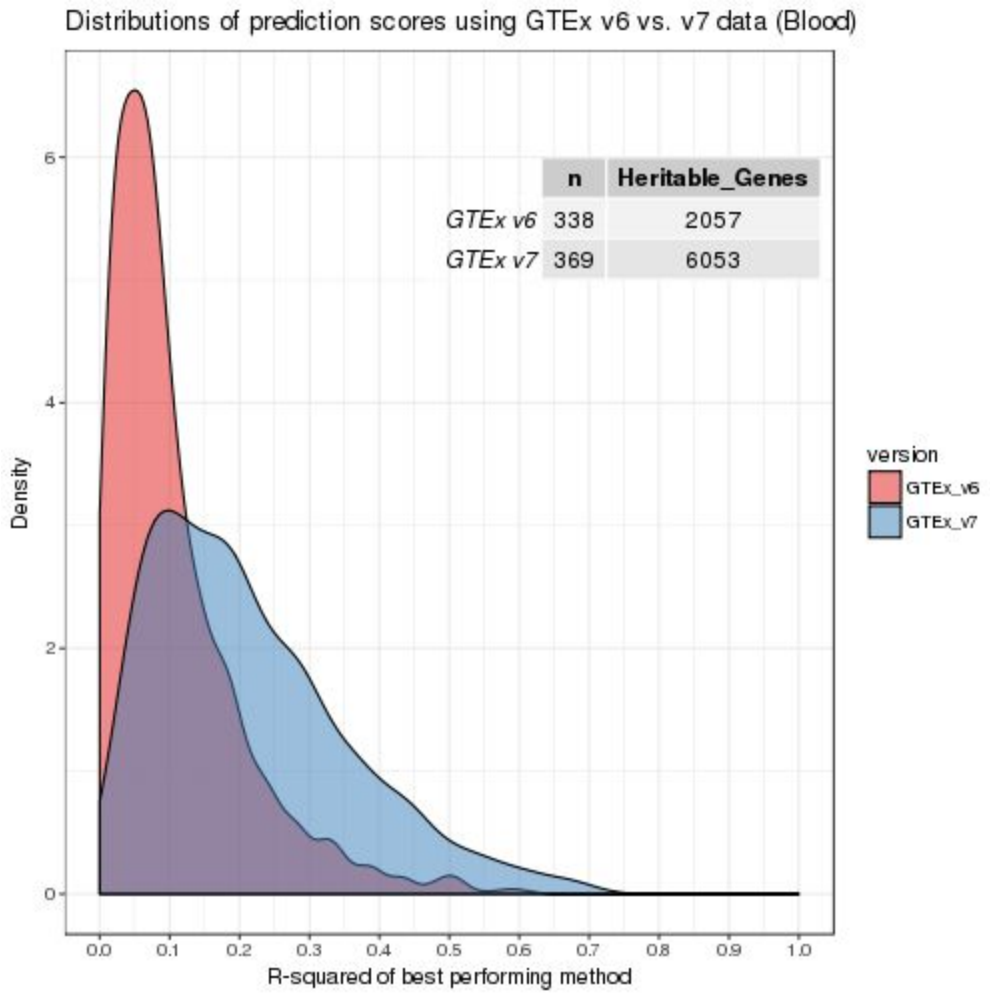

Figure 1. Performances of best expression prediction model (r^2), GTEx 6 vs. GTEx 7 (blood).

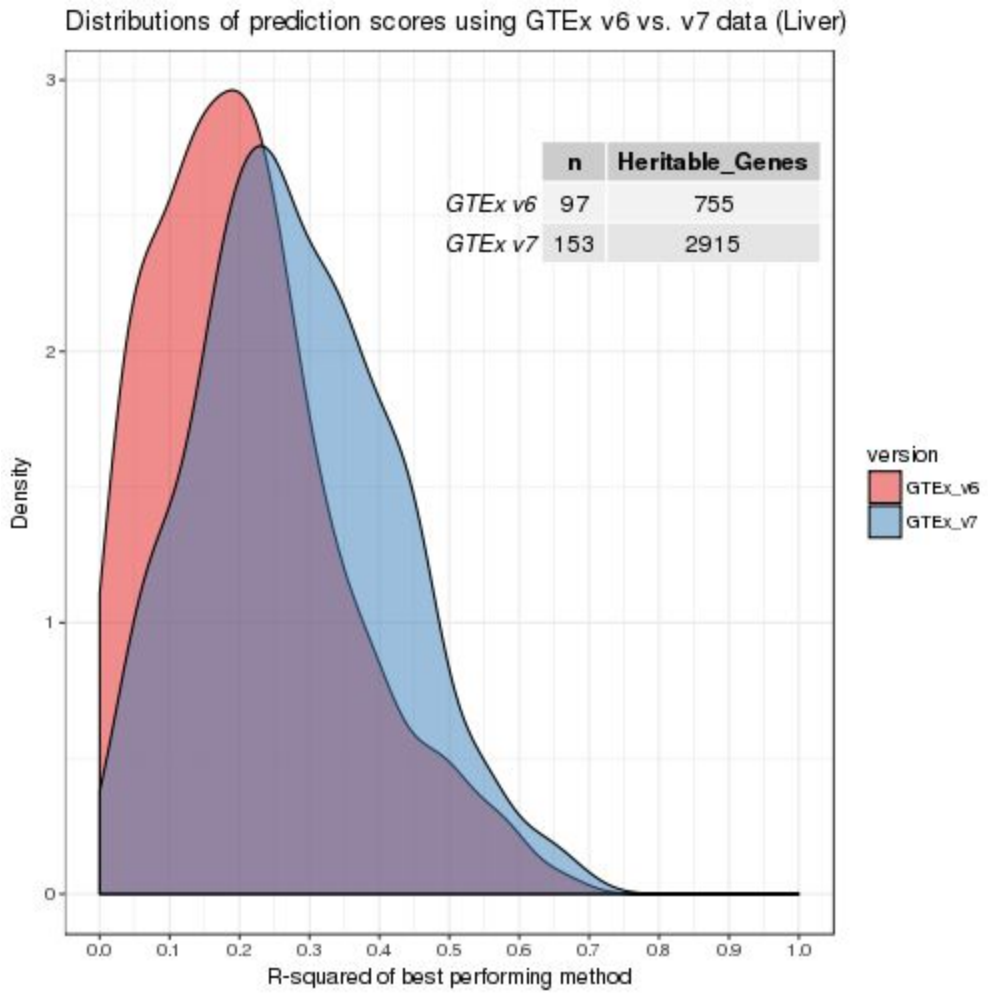

Figure 2. Performances of best expression prediction model (r^2), GTEx 6 vs. GTEx 7 (liver)

Dataset	TWAS Signif. Genes	With Perm. p < 0.01	With Pr(Coloc.) > 0.5
GTEx7 Whole Blood	16	11	10
GTEx7 Liver	12	4	3
NTR Whole Blood	3	0	0
YFS Whole Blood	11	7	5

Table 3. TWAS Results Overview: Permutation and Colocalization. TWAS Signif. Genes: count of genes in dataset with significant TWAS signal; With Perm. P< 0.01: size of subset of TWAS significant genes with a *p* value more extreme than at least 99% of permutations; With

Pr(Coloc.) > 0.5: size of subset of TWAS significant genes that are more likely than not to have colocalized eQTL and GWAS signals (as estimated by COLOC).

Dataset	Genes Tested	Signif. Genes	With GWAS Overlap	Without GWAS Overlap
GTEx7 Whole Blood	5969	16	12	4
GTEx7 Liver	2875	12	10	2
NTR Whole Blood	2448	3	3	0
YFS Whole Blood	4690	11	8	3

Table 4. TWAS overlap with GWAS. Genes Tested: total number of genes tested (with significant heritability of expression) in dataset; Signif. Genes: subset of these genes found to be TWAS significant; With GWAS Overlap: subset of significant TWAS genes that contain a significant GWAS variant in their prediction model; Without GWAS Overlap: subset of significant TWAS genes that do not contain a significant GWAS variant in their prediction model.

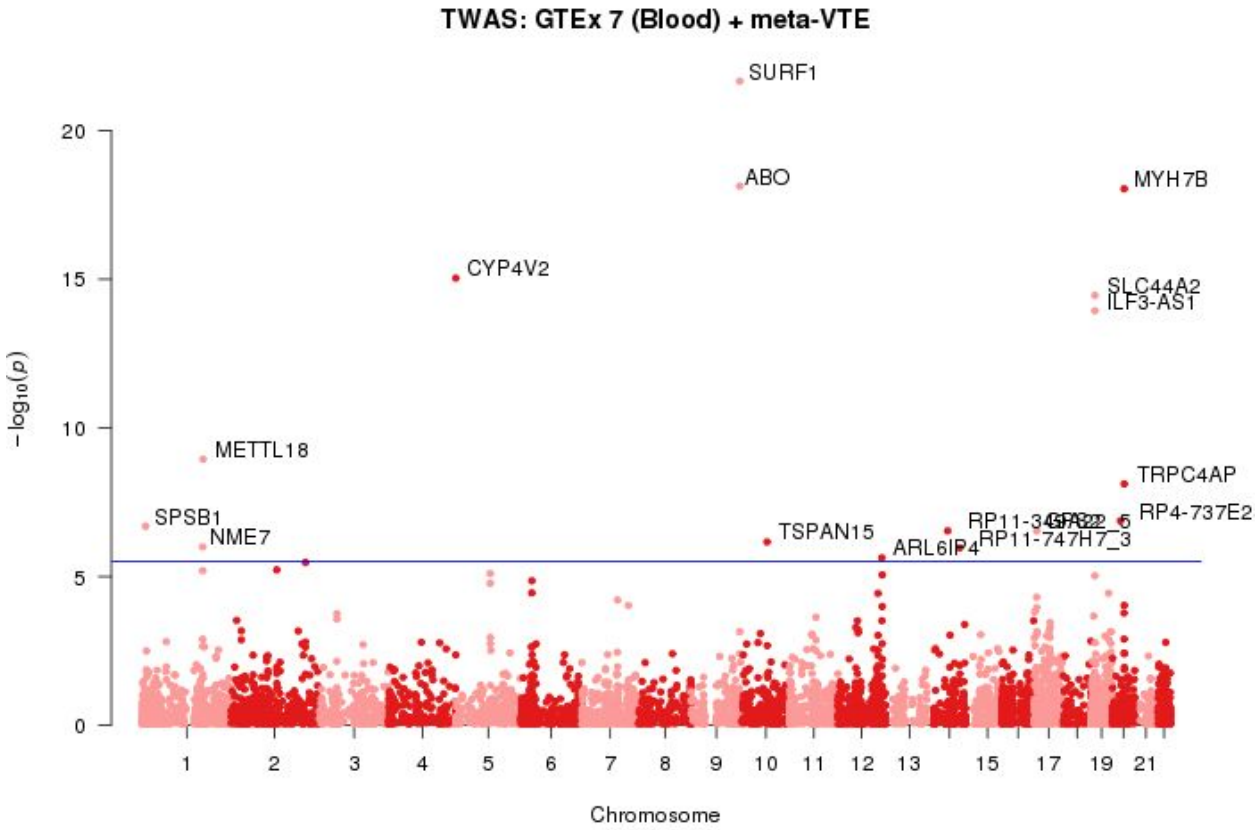

Figure 3. Manhattan plot of VTE TWAS generated from GTEx7 (blood) expression prediction dataset.

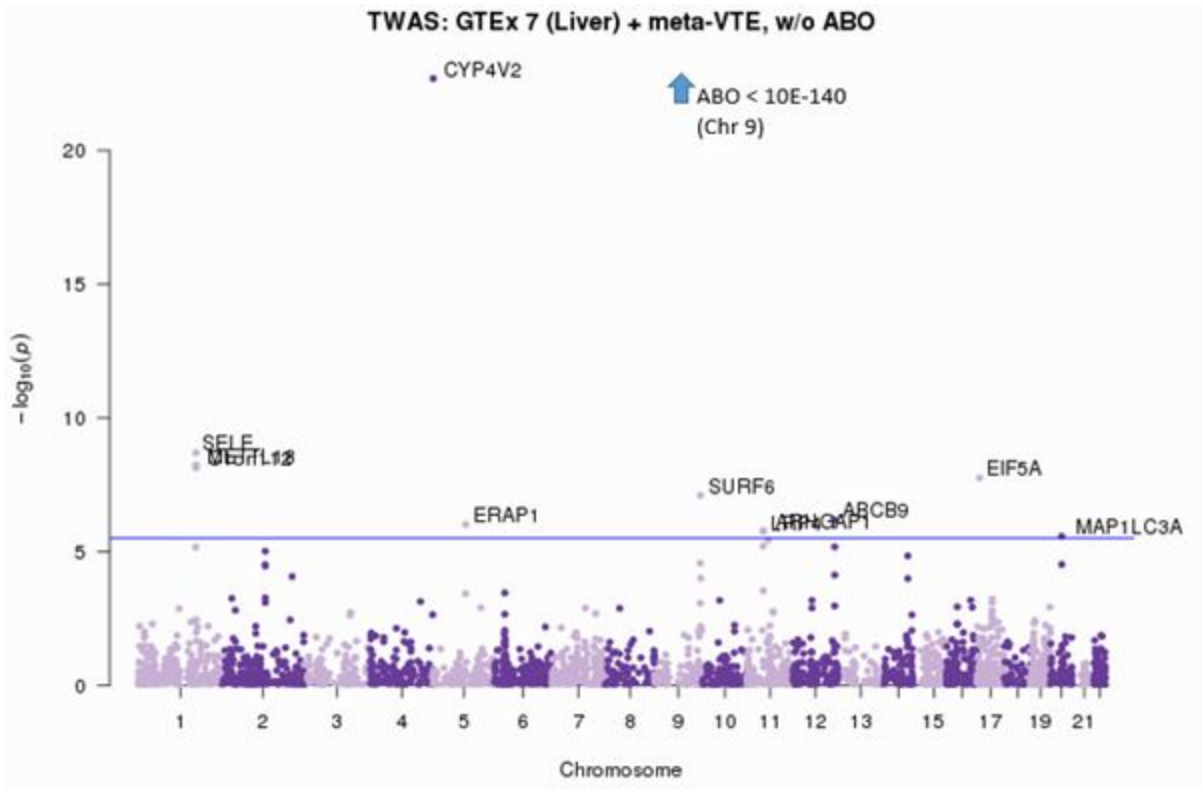

Figure 4. Manhattan plot of VTE TWAS generated from GTEx7 (liver) expression prediction dataset (y-axis limited due to dominant ABO signal).

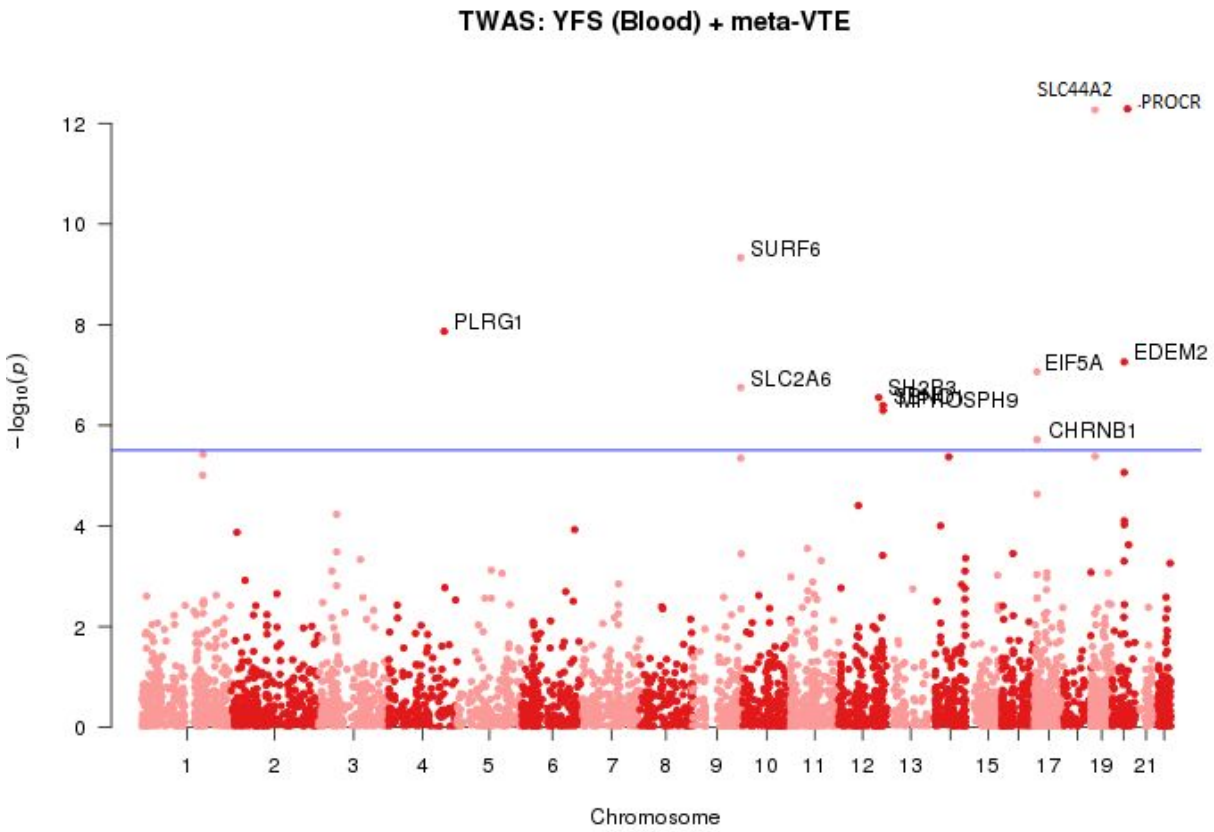

Figure 5. Manhattan plot of VTE TWAS generated from YFS (blood) expression prediction dataset.

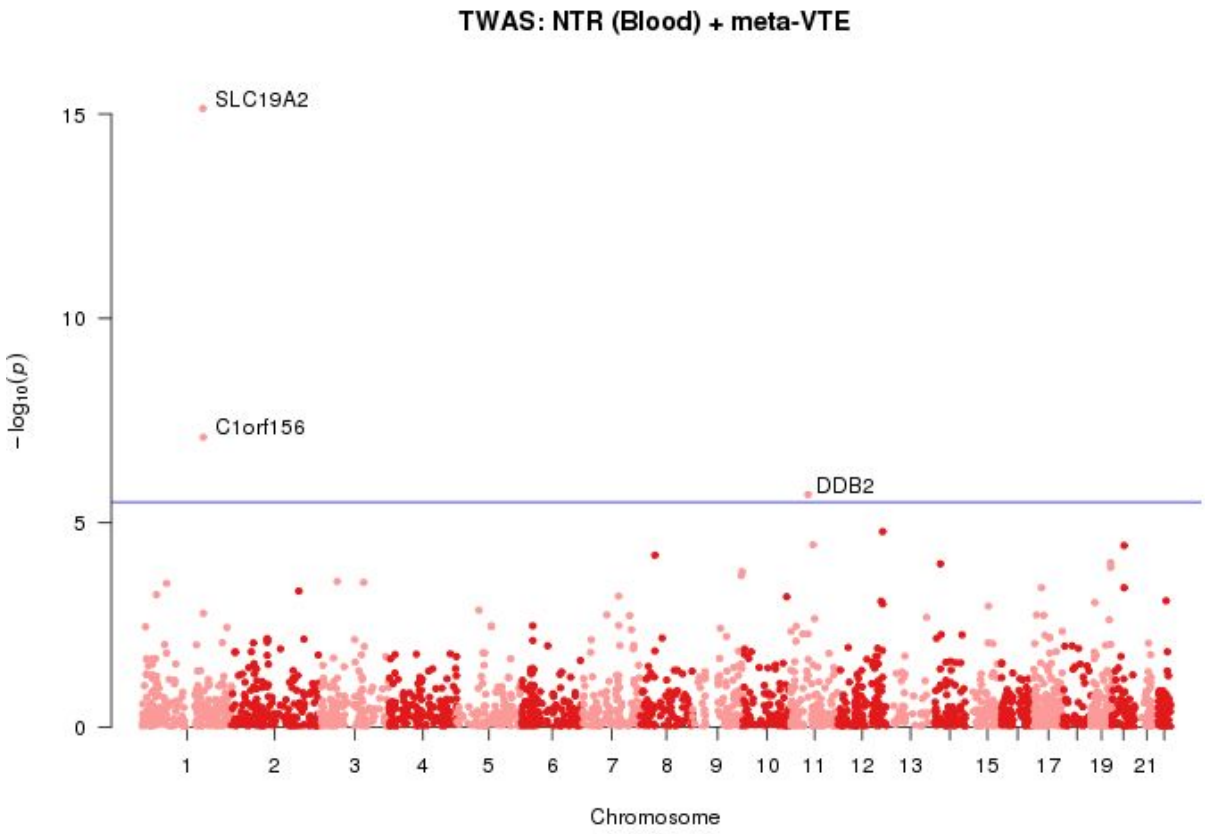

Figure 6. Manhattan plot of VTE TWAS generated from NTR (blood) expression prediction dataset.

Dataset	Gene	GWAS SNP	Chr.	Best GWAS p	TWAS Z	TWAS p	Pr(Colocalized)
GTEx7 Blood	SPSB1	rs11590326	1	1.99e-07	5.20	2.03e-07	0.997
GTEx7 Blood	RP11-747H7_3	rs10498632	14	5.80e-08	4.87	1.09e-06	0.945
GTEx7 Blood	RP4-737E23_2	rs6137866	20	2.13e-07	-5.27	1.34e-07	0.995
YFS Blood	ERAP1	rs27039	5	7.71e-06	-4.90	9.49e-07	0.349
YFS Blood	SBNO1	rs641760	12	3.04e-07	5.07	4.03e-07	0.940
GTEx7 Liver	SH2B3	rs3184504	12	2.96e-06	5.14	2.80e-07	0.156

Table 5. Significant TWAS signals, after permutation filtering, that do not overlap with GWAS signals. GWAS SNP: identity of most significant GWAS SNP within TWAS prediction model;

Chr.: chromosome; Best GWAS p: GWAS p value of GWAS SNP; TWAS Z: TWAS Z score for gene; TWAS P: p value for corresponding TWAS Z score; Pr(Colocalized): probability of colocalization between eQTL and GWAS signal

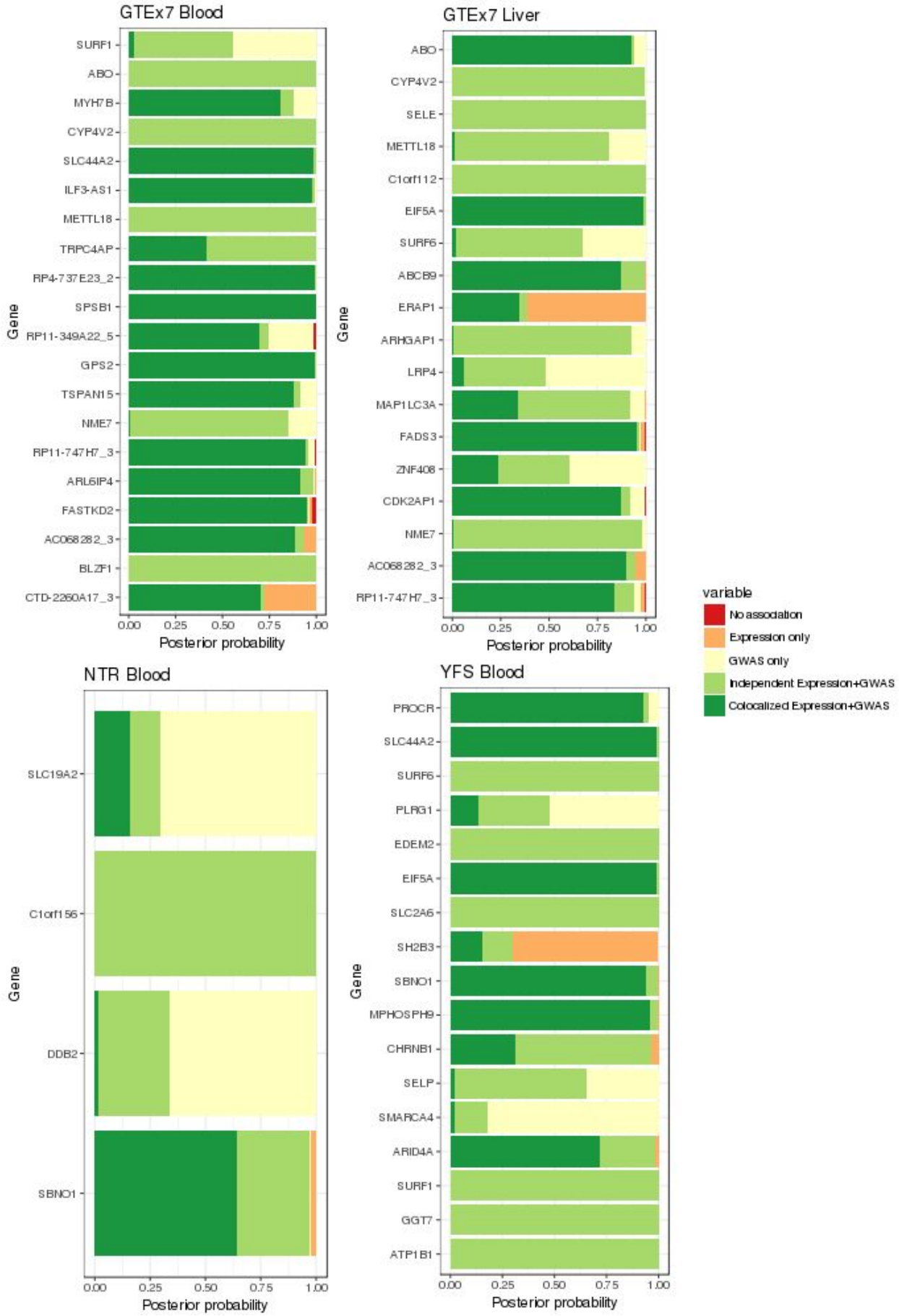
GTEx7 Blood
SURF1
ABO
MYH7B
CYP4V2
SLC44A2
ILF3-AS1
METTL18
TRPC4AP
RP4-737E23_2
SPSB1
RP11-349A22_5
GPS2
TSPAN15
NME7
RP11-747H7_3
ARL6IP4
FASTKD2
AC068282_3
BLZF1
CTD-2260A17_3
Gene
0.00
0.25
0.50
0.75
1.00
Posterior probability
GTEx7 Liver
ABO
CYP4V2
SELE
METTL18
C1orf112
EIF5A
SURF6
ABCB9
ERAP1
ARHGAP1
LRP4
MAP1LC3A
FADS3
ZNF408
CDK2AP1
NME7
AC068282_3
RP11-747H7_3
Gene
0.00
0.25
0.50
0.75
1.00
Posterior probability
variable
No association
Expression only
GWAS only
Independent Expression+GWAS
Colocalized Expression+GWAS
NTR Blood
SLC19A2
C1orf156
DDB2
SBNO1
Gene
0.00
0.25
0.50
0.75
1.00
Posterior probability
YFS Blood
PROCR
SLC44A2
SURF6
PLRG1
EDEM2
EIF5A
SLC2A6
SH2B3
SBNO1
MPHOSPH9
CHRNB1
SELP
SMARCA4
ARID4A
SURF1
GGT7
ATP1B1
Gene
0.00
0.25
0.50
0.75
1.00
Posterior probability

Figure 8. Colocalization probabilities for significant TWAS signals

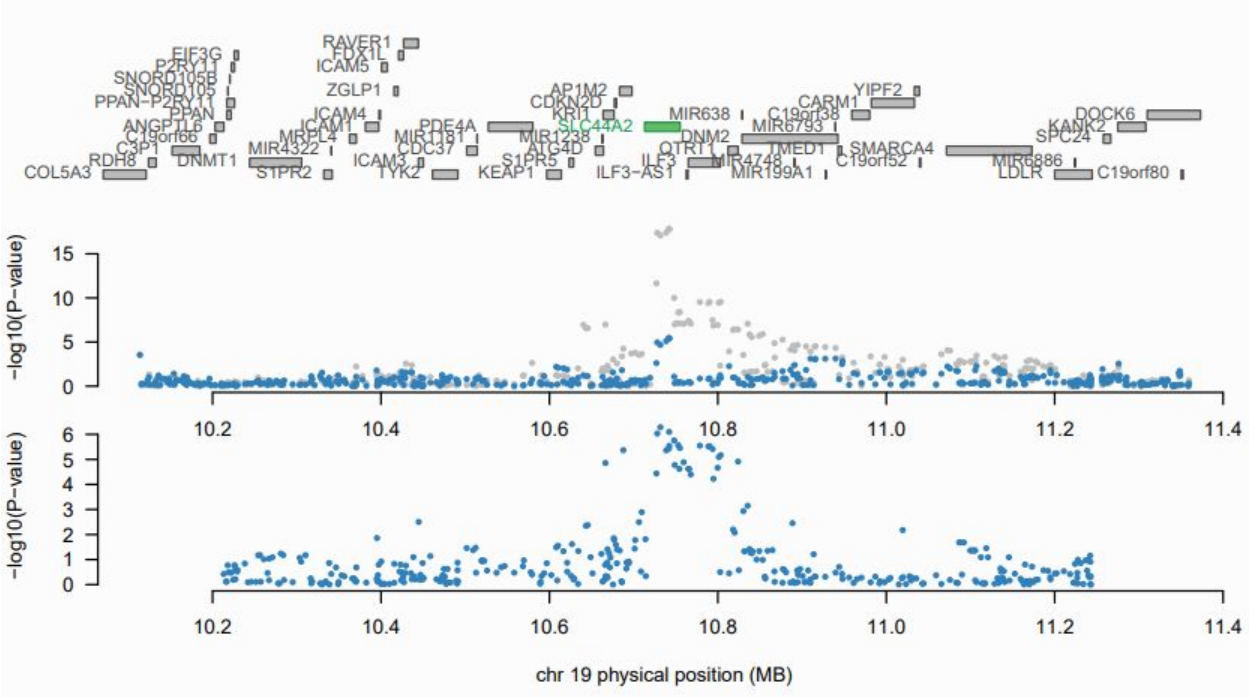

Figure 9. Upper panel: GWAS associations for the *SLC44A2* neighborhood before (grey points) and after (blue points) conditioning on predicted expression of (GTEx v.7 Whole Blood). Lower panel: eQTL results for gene expression of *SLC44A2*.

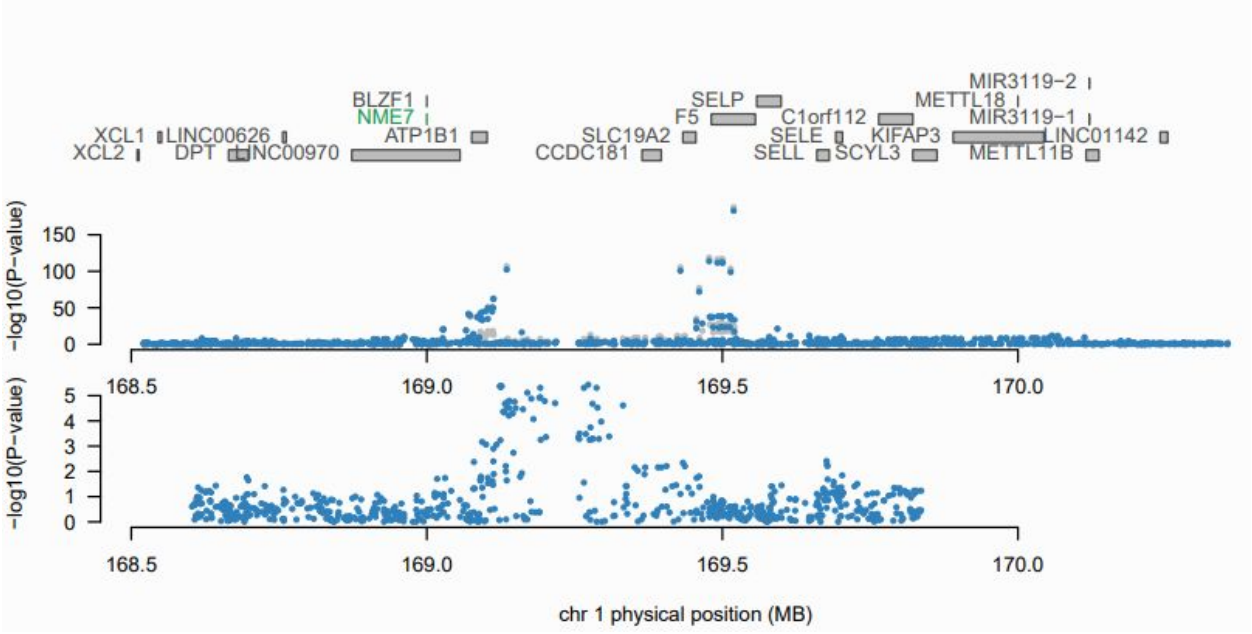

Figure 10. Upper panel: GWAS associations for the *NME7* neighborhood before (grey points) and after (blue points) conditioning on predicted expression of (GTEx v.7 Whole Blood). Lower panel: eQTL results for gene expression of *NME7*.

ID	Chr.	GWAS ID	TWAS p	Pr(Coloc.)	Perm. p	Dataset	GWAS Overlap?
RP11-747H7_3	14	rs10498632	1.09e-06	0.945	0.00040	GTEx7 Blood	No
SPSB1	1	rs11590326	2.03e-07	0.997	0.00271	GTEx7 Blood	No
RP4-737E23_2	20	rs6137866	1.34e-07	0.995	0.00263	GTEx7 Blood	No
SLC44A2	19	rs12972963	3.51e-15	0.987	0.00010	GTEx7 Blood	Yes
ILF3-AS1	19	rs12972963	1.15e-14	0.981	0.00350	GTEx7 Blood	Yes
SLC44A2	19	rs12972963	5.37e-13	0.991	0.00060	YFS Blood	Yes
EIF5A	17	rs4796399	1.73e-08	0.990	0.00216	GTEx7 Liver	Yes
ABO	9	rs612169	3.83e-148	0.926	0.00148	GTEx7 Liver	Yes

Table 6. Best Candidate TWAS Genes, result of filtering significant TWAS genes by Pr(Coloc.) > 0.9 and Perm. p < 0.01. ID: gene Identity; GWAS ID: SNP Identity of most significant GWAS

SNP within TWAS prediction model; TWAS p: significance of gene's TWAS score; Pr(Coloc.): probability of colocalization between eQTL and GWAS signal; Perm. p: p value generated from permutation-based empirical null distribution of p values; Dataset: expression data used for TWAS prediction modeling; GWAS Overlap?: does this TWAS signal contain a known VTE GWAS variant?

References

1000 Genomes Project Consortium. "A global reference for human genetic variation." Nature 526.7571 (2015): 68.

Barbeira, Alvaro, et al. "MetaXcan: summary statistics based gene-level association method infers accurate PrediXcan results." bioRxiv (2016): 045260.

Bertina, R. M., A. W. Broekmans, and C. Es Krommenhoek-van. "The use of a functional and immunologic assay for plasma protein C in the study of the heterogeneity of congenital protein C deficiency." Thrombosis and haemostasis 51.1 (1984): 1-5.

Bonferroni, C. E., Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936

Brionez, Tamar F., and John D. Reveille. "The contribution of genes outside the major

histocompatibility complex to susceptibility to ankylosing spondylitis." Current opinion in rheumatology 20.4 (2008): 384-391.

Broekmans, André W., Jan J. Veltkamp, and Rogier M. Bertina. "Congenital protein C deficiency and venous thromboembolism: a study of three Dutch families." New England Journal of Medicine 309.6 (1983): 340-344.

CIDR Pricing. (n.d.). Retrieved May 31, 2018, from http://www.cidr.jhmi.edu/services/pricing.html

Comp, Philip C., and Charles T. Esmon. "Recurrent venous thromboembolism in patients with a partial deficiency of protein S." New England Journal of Medicine 311.24 (1984): 1525-1528.

Desch, Karl C. "Dissecting the genetic determinants of hemostasis and thrombosis." Current opinion in hematology22.5 (2015): 428.

De Stefano, Valerio, and Giuseppe Leone. "Resistance to activated protein C due to mutated factor V as a novel cause of inherited thrombophilia." Haematologica 80.4 (1995): 344-356.

Devallière, Julie, and Béatrice Charreau. "The adaptor Lnk (SH2B3): an emerging regulator in vascular cells and a link between immune and inflammatory signaling." Biochemical pharmacology 82.10 (2011): 1391-1402.

Doray, D., D. Patton, and C. T. Esmon. "An abnormal plasma distribution of protein S occurs in functional protein S deficiency." Blood 67.2 (1986): 504-508.

Egeberg, O. "Inherited antithrombin deficiency causing thrombophilia." Thrombosis et diathesis haemorrhagica 13 (1965): 516-30.

Fasel, Jean HD, et al. "Textbook of hepatology: from basic science to clinical practice." (2007).

Germain, Marine, et al. "Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism." The American Journal of Human Genetics 96.4 (2015): 532-542.

Giambartolomei, Claudia, et al. "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics." PLoS genetics 10.5 (2014): e1004383.

Grosse, Scott D., et al. "The economic burden of incident venous thromboembolism in the United States: a review of estimated attributable healthcare costs." Thrombosis research 137 (2016): 3-10.

GTEx Consortium. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans." Science 348.6235 (2015): 648-660.

Gusev, Alexander, et al. "Integrative approaches for large-scale transcriptome-wide association studies." Nature genetics 48.3 (2016): 245.

Gusev, Alexander, et al. "Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights." Nature genetics 50.4 (2018): 538.

He, Xin, et al. "Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS." The American Journal of Human Genetics 92.5 (2013): 667-680.

Heestermans, Marco, et al. "Circulating nucleosomes and elastase α1-antitrypsin complexes and the novel thrombosis susceptibility locus SLC44A2." Thrombosis research 142 (2016): 8-10.

Heit, John A. "Epidemiology of venous thromboembolism." Nature Reviews Cardiology 12.8 (2015): 464.

Hormozdiari, Farhad, et al. "Colocalization of GWAS and eQTL signals detects target genes." The American Journal of Human Genetics 99.6 (2016): 1245-1260.

INVENT, INVENT-VTE, 2016, www.invent-vte.com/.

Kaiser, Annette. "Translational control of eIF5A in various diseases." Amino acids 42.2-3

(2012): 679-684.

Kampf, Caroline, et al. "The human liver-specific proteome defined by transcriptomics and antibody-based profiling." The FASEB Journal 28.7 (2014): 2901-2914.

Kong, Minyoung, Younyoung Kim, and Chaeyoung Lee. "Functional investigation of a venous thromboembolism GWAS signal in a promoter region of coagulation factor XI gene." Molecular biology reports 41.4 (2014): 2015-2019.

Levy, Daniel, et al. "Genome-wide association study of blood pressure and hypertension." Nature genetics 41.6 (2009): 677.

Mailman, Matthew D., et al. "The NCBI dbGaP database of genotypes and phenotypes." Nature genetics 39.10 (2007): 1181.

Martinelli, Ida, et al. "The risk of venous thromboembolism in family members with mutations in the genes of factor V or prothrombin or both." British journal of haematology 111.4 (2000): 1223-1229.

Martinelli, Ida. "Risk factors in venous thromboembolism." Thrombosis and haemostasis 86.01 (2001): 395-403.

McCarthy, Shane, et al. "A reference panel of 64,976 haplotypes for genotype imputation." Nature genetics 48.10 (2016): 1279.

Morange, Pierre-Emmanuel, Pierre Suchon, and David-Alexandre Trégouët. "Genetics of venous thrombosis: update in 2015." Thrombosis and haemostasis 114.05 (2015): 910-919.

Ochoa, Eguzkine, et al. "Thrombotic antiphospholipid syndrome shows strong haplotypic association with SH2B3-ATXN2 locus." PLoS One 8.7 (2013): e67897.

Perez-Garcia, Arianne, et al. "Genetic loss of SH2B3 in acute lymphoblastic leukemia." Blood 122.14 (2013): 2425-2432.

Raitakari, Olli T., et al. "Cardiovascular risk factors in childhood and carotid artery intima-media thickness in adulthood: the Cardiovascular Risk in Young Finns Study." Jama 290.17 (2003): 2277-2283.

Rosendaal, Frits R. "Venous thrombosis: the role of genes, environment, and behavior." ASH Education Program Book 2005.1 (2005): 1-12.

Sano, Rie, et al. "Epithelial expression of human ABO blood group genes is dependent upon a downstream regulatory element functioning through an epithelial cell-specific transcription factor, Elf5." Journal of Biological Chemistry 291.43 (2016): 22594-22606.

Simioni, Paolo, et al. "Incidence of venous thromboembolism in families with inherited thrombophilia." Thrombosis and haemostasis 82.02 (1999): 198-202.

Stegle, Oliver, et al. "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses." Nature protocols 7.3 (2012): 500.

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.

Tragante, Vinicius, et al. "Gene-centric meta-analysis in 87,736 individuals of European ancestry identifies multiple blood-pressure-related loci." The American Journal of Human Genetics 94.3 (2014): 349-360.

Trégouët, David-Alexandre, 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 113.21 (2009): 5298-5303.

Wainberg, Michael, et al. "Vulnerabilities of transcriptome-wide association studies." bioRxiv (2017): 206961.

Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews genetics 10.1 (2009): 57.

Wen, Xiaoquan, Roger Pique-Regi, and Francesca Luca. "Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization." PLoS genetics 13.3 (2017): e1006646.

White, Richard H. "The epidemiology of venous thromboembolism." Circulation 107.23 suppl 1 (2003): I-4.

Willemsen, Gonneke, et al. "The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection." Twin Research and Human Genetics 16.1 (2013): 271-281.

Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, Lloyd-Jones LR, Marioni RE, Martin NG, Montgomery GW, Deary IJ, Wray NR, Visscher PM, McRae AF & Yang J (2018) Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nature Communications, 9: 918.

Zou, Fanggeng, et al. "Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants." PLoS genetics 8.6 (2012): e1002707.

Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal

of the Royal Statistical Society: Series B (Statistical Methodology) 67.2 (2005): 301-320.