HHS Public Access

Author manuscript

Thromb Res. Author manuscript; available in PMC 2016 April 01.

Published in final edited form as:

Thromb Res. 2015 April ; 135(4): 659–665. doi:10.1016/j.thromres.2015.02.003.

Whole blood gene expression profiles distinguish clinical phenotypes of venous thromboembolism☆

Deborah A. Lewis[a], **Sunil Suchindran**[b], **Michele G. Beckman**[c], **W. Craig Hooper**[c], **Althea M. Grant**[c], **John A. Heit**[d], **Marilyn Manco-Johnson**[e], **Stephan Moll**[f], **Claire S. Philipp**[g], **Kristy Kenney**[c], **Christine De Staercke**[c], **Meredith E. Pyle**[c], **Jen-Tsan Chi**[h], and **Thomas L. Ortel**[a,*]

[a]Thrombosis and Hemostasis Center, Division of Hematology, Duke University Medical Center, Durham, NC

[b]Center for Applied Genomics, Duke University School of Medicine, Durham NC

[c]Centers for Disease Control and Prevention, Atlanta, GA; Division of Cardiovascular Diseases

[d]Mayo Clinic, Rochester, MN

[e]University of Colorado and Children's Hospital, Aurora, CO

[f]University of North Carolina, Chapel Hill, NC

[g]Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ

[h]Department of Molecular Genetics and Microbiology and Center for Genomic and Computation Biology, Duke University Medical Center, Durham NC

Abstract

Introduction—Recurrent venous thromboembolism (VTE) occurs infrequently following a provoked event but occurs in up to 30% of individuals following an initial unprovoked event. There is limited understanding of the biological mechanisms that predispose patients to recurrent VTE.

Objectives—To identify whole blood gene expression profiles that distinguished patients with clinically distinct patterns of VTE.

Patients/Methods—We studied 107 patients with VTE separated into 3 groups: (1) 'low-risk' patients had one or more provoked VTE; (2) 'moderate-risk' patients had a single unprovoked VTE; (3) 'high-risk' patients had ≥2 unprovoked VTE. Each patient group was also compared to twenty-five individuals with no personal history of VTE. Total RNA from whole blood was isolated and hybridized to Illumina HT-12 V4 Beadchips to assay whole genome expression.

☆Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

*Corresponding author at: Duke Thrombosis and Hemostasis Center, Box 3422 DUMC, Durham, NC 27710 USA. Tel.: +1 919 660 4363; fax: +1 919 681 1177. thomas.ortel@duke.edu (T.L. Ortel).

Conflict of Interest Disclosures

The authors declare no competing financial interests for this study.

Results—Using class prediction analysis, we distinguished high-risk patients from low-risk patients and healthy controls with good receiver operating curve characteristics (AUC = 0.81 and 0.84, respectively). We also distinguished moderate-risk individuals and low-risk individuals from healthy controls with AUC's of 0.69 and 0.80, respectively. Using differential expression analysis, we identified several genes previously implicated in thrombotic disorders by genetic analyses, including *SELP*, *KLKB1*, *ANXA5*, and *CD46*. Protein levels for several of the identified genes were not significantly different between the different groups.

Conclusion—Gene expression profiles are capable of distinguishing patients with different clinical presentations of VTE, and genes relevant to VTE risk are frequently differentially expressed in these comparisons.

Keywords

venous thrombosis; deep-vein thrombosis; gene expression profiling

Introduction

Deep vein thrombosis (DVT) or pulmonary embolism (PE), referred to collectively as venous thromboembolism (VTE), affects approximately 350,000 to 600,000 individuals in the United States each year, and up to 100,000 will die from the thromboembolic event [1]. VTE may occur after transient exposures such as a surgical procedure, prolonged immobilization, or with the use of certain therapies, such as oral contraceptives and hormone replacement therapy, which is referred to as a provoked event [2]. VTE can also occur in the absence of any acquired risk factors, which is referred to as an unprovoked, or idiopathic, event [1,3,4]. Other factors that may increase an individual patient's risk for VTE include increased age, the presence of a thrombophilia [5], race/ethnicity, and a variety of medical conditions [6].

The current standard of care for patients with provoked VTE consists of therapeutic anticoagulation for three months [7]. In contrast, for patients with an unprovoked VTE, up to 30% will sustain a recurrent event within ten years of completing a standard course of therapy [3,8]. Consequently, it is recommended to consider an extended course of therapy for patients with an initial unprovoked event [7]. Continued anticoagulant therapy has been shown in several studies to significantly decrease the risk for recurrent VTE [6–8] but the risk of major bleeding in individuals after the first three months of therapy ranges from a baseline of 0.3% to ≥2.5% per year [7].

Determining which patients with VTE have a high risk for recurrent events, and balancing this risk with the potential for bleeding if anticoagulation is continued, is an important health concern. Multiple studies have investigated biomarkers to help predict which patients are at a higher risk for recurrent VTE [9]. Current evidence suggests that inherited thrombophilic disorders are not helpful to predict which patients with a first unprovoked VTE are at an increased risk for recurrent events [10]. In contrast, elevated D-dimer levels obtained after completing a standard course of anticoagulant therapy are associated with an increased risk for recurrent VTE [11]. Other biomarkers that have been associated with recurrent VTE include elevated levels of soluble p-selectin [12] and elevated thrombin generation [13].

Whole blood gene expression studies have been used in a variety of disorders including myocardial infarction and systemic lupus erythematosus [14,15]. We previously used whole blood gene expression profiles to distinguish patients with a single VTE from patients with recurrent VTE [16], but this study combined patients with provoked and unprovoked events. Here we extend this initial study by using clinically well-defined patient groups with the objectives of comparing individuals based on the type of VTE (provoked versus unprovoked) as well as by the number of events (single versus multiple). We used two distinct analytical approaches, class prediction analysis [17–20] and differential expression analysis [21] to identify means to distinguish among these patient groups. A group of healthy individuals was included to look for genes and pathways that are differentially expressed in healthy individuals compared to individuals with different types of VTE.

Material and Methods

Patient Population

Participants were enrolled in 2009 and 2010 at 4 sites participating in the Thrombosis and Hemostasis Centers Research and Prevention Network supported by the Centers for Disease Control and Prevention (CDC): Duke University Medical Center, Durham NC; Mayo Clinic, Rochester MN; University of North Carolina, Chapel Hill NC; and Rutgers Robert Wood Johnson Medical School, New Brunswick NJ. This Network consisted of Thrombosis and Hemostasis Centers that provided comprehensive specialty care to patients with thrombophilia and thrombotic disorders [22]. Study protocol and consent forms were approved by Institutional Review Boards at each site and at the CDC.

Patients with at least one VTE, defined as either PE or DVT of the leg or arm, with the first event occurring at age 18 years or older, and who were, at the time of enrollment, greater than 10 weeks from their most recent VTE, were approached for participation. The diagnosis of VTE was reviewed and objectively confirmed by the site investigator, based on clinical history and imaging data. Individuals with no prior history of VTE or known inherited clotting disorder and similar in age, gender, and race to the VTE case were identified at each site and approached to participate as controls. Patients with known antiphospholipid syndrome, active or prior malignancy (excluding skin cancer) at the time of VTE diagnosis, infection within the past two weeks of enrollment or currently pregnant were not included in this study.

Consenting VTE patients were allocated to 3 groups: (1) low-risk, defined as patients who had sustained 1 or more provoked VTE with no history of an unprovoked VTE; (2) moderate-risk, defined as patients who had sustained a single unprovoked VTE (with or without additional provoked VTE); and (3) high-risk, defined as patients who had sustained 2 or more unprovoked VTE (with or without additional provoked VTE). A provoked event was defined as a VTE occurring in a patient with a clear transient acquired risk factor for VTE, i.e. VTE occurring within 3 months after trauma, hospitalization, prolonged immobilization, or surgery and the post-operative setting; or in patients taking oral contraceptives or hormone replacement therapy; or during pregnancy or the post-partum period. Unprovoked events were defined as VTE occurring in the absence of any of these transient risk factors.

Patients with more than one VTE (provoked or unprovoked) had distinct clinical events that occurred at different points in time. Thromboembolic events affecting more than one vascular bed but occurring at the same time were considered to be a single event (*e.g.*, a patient presenting with PE and DVT).

Data and Sample Collection

Demographic and clinical information was collected from each participant through chart abstraction or in-person interview. Citrated plasma and serum samples were collected for each participant, processed, and stored at −80 °C at each site. Blood was simultaneously collected in PAXgene RNA tubes and stored according to the manufacturer's instructions. De-identified samples were shipped to the CDC Division of Blood Disorders' Molecular and Hemostasis Laboratories for analysis.

RNA Isolation and Microarray Hybridization

Total RNA was isolated from whole blood drawn into PAXgene tubes using the PAXgene Blood RNA kit (PreAnalytiX; Qiagen GmbH-USA). The quality and quantity of the RNA was confirmed using the Nanodrop 1000 Spectrophotometer (Thermo Fisher Scientific, Wilmington, DE). Samples with an A260/A280 ratio >2.19, or ≤70 ng of RNA, were excluded from the final analyses. RT PCR using probes for IL-1beta and CD141 was used to check RNA expression levels for several of the initial samples from each of the sites, to confirm comparable yields. RNA was amplified and the cRNA was biotinylated using the Ambion Illumina TotalPrep RNA Amplification Kit (Life Technologies, Carlsbad, CA). Following labeling, cRNA samples were hybridized to Illumina HT-12 V4 Beadchips to assay whole genome gene expression with over 47,000 probes against human transcripts.

Microarray Data Processing

A comprehensive quality control process was performed on all arrays using the lumi package in Bioconductor in the R environment for statistical computing [23,24]. Quality of the raw data was assessed using the percent of probes present, MA plots, boxplots of the expression distribution, and heatmaps to visualize the correlation between samples. Samples in which the percent of probes present was 15% or less were excluded, and all probes that were not detected in greater than 95% of the remaining samples were removed (21,174 out of 47,304). The lumi package was also used to perform background corrections, expression value log-transformation, and quantile normalization. The data was then filtered without regard to phenotype to include only the top 10,000 probes that varied the most among all of the samples [25,26] The microarray data files were submitted to NCBI's Gene Expression Omnibus and are accessible through GEO series accession number GSE48000 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48000) [27].

Statistical Analyses

Descriptive statistics of the demographics of the study participants, including means and frequencies, were performed using Excel (Microsoft, Redmond, WA). The Mann–Whitney nonparametric test in Prism 6 (Graphpad Software Inc., La Jolla, CA) was used to compare BMI, the ages of the participants at the time of enrollment, at the time of first VTE, and time

since last VTE. Fisher's exact test (Graphpad Prism 6) was used to compare the gender, race and proportion of PE and anticoagulant therapy among the groups. The mean concentrations and 95% confidence intervals for the biomarker levels were determined using Graphpad Prism 6.

Class prediction was done on the six possible pair-wise comparisons of the four study groups using penalized binary regression. Bayesian Factor Regression was used to model correlation structure in the expression data [17]. This served as a dimension reduction step in which a large number of expression vectors were expressed as a smaller number of factors. In addition to reducing the dimension of the data, the factors were closer to being independent when compared to the original data, which facilitates model building. Factor analysis is unsupervised and, therefore, does not use phenotype labels. The resulting factor scores were then used as predictors to build classification models. Penalized regression implemented in the glmnet R package was used to build the classification models [18,19]. Estimates of prediction accuracy were obtained through leave-one-out cross-validation, which provides unbiased estimates of accuracy, and is appropriate for samples sizes in the range of the current study [28–30]. Receiver Operating Characteristic curves and the corresponding area under the curve (AUC) were generated using Graphpad Prism 6.

Differential expression was also performed on the six possible pair-wise comparisons of the four study groups using the limma package in Bioconductor [21]. A false discovery rate of 0.005 was used to determine which genes were significantly differentially expressed in each comparison.

Functional Analysis of Gene Lists Obtained from Differential Expression

DAVID, a program which aides in the functional interpretation of large lists of genes using information from a variety of public bioinformatics databases, was used to understand the gene ontologies, biological function and pathways associated with the genes identified in the differential expression analysis [31,32]. The functional annotation chart tool in DAVID was used to determine the top enriched ontologies in the gene lists from the differential expression analysis and functional annotation clustering tool in DAVID was used to look for ontologies specific to VTE.

Biomarker Testing

Factor XI, annexin A5, sP-selectin, endothelin, and CD46 levels in plasma or serum samples collected at the same time as the whole blood RNA sample were determined by ELISA. Serum was used for annexin A5, and citrated plasma was used for the other four biomarkers. Optical density was measured with a Vmax Kinetic Microplate Reader (Molecular Devices, Sunnyvale, CA). The following kits were used: Total Human Coagulation Factor XI Antigen Assay (Molecular Innovations MI, USA); Human Annexin V Platinum ELISA kit (eBioscience, San Diego CA); Human sP-selectin/CD62P ELISA kit (R&D Systems, Minneapolis, MN); Quantitative ELISA Endothelin-1 (ET-1) Immunoassay kit (R&D Systems); and the MCP (CD46) ELISA Kit from UscnLife-Science Inc. (Wuhan EIAab Science Co, China).

Results

Study Participants

One hundred and seventy-eight participants were enrolled, but 3 were subsequently excluded because they did not meet enrollment criteria (1 in the low-risk group and 2 in the moderate-risk group). Of the 175 participants meeting enrollment criteria, 43 participants were excluded because ≤15% of the probes were detected in samples from these individuals (9 in the low-risk group, 10 in the moderate-risk group, 5 in the high-risk group, and 19 in the healthy control group).

Characteristics of the 132 remaining participants with genomic expression data that passed quality assessment are shown in Table 1. Individuals in the high-risk and moderate-risk groups were older at the time of enrollment, but the age of the individuals at the time of their first VTE event did not differ by risk group (Table 1). Five of the low-risk patients had sustained more than one provoked VTE, and twelve of the moderate-risk patients had sustained one or more provoked events in addition to a single, unprovoked VTE (Table 1). A significantly higher proportion of individuals were on anticoagulant therapy at the time of enrollment in the high-risk and moderate-risk groups compared to the low-risk group. Fifteen individuals in the low-risk group were on warfarin therapy for more than six months, for various reasons, including recurrent provoked thrombotic events and ongoing exposure to identified risk factors.

Class Prediction Analysis

Class prediction analysis was done on each of the six possible pairwise comparisons (Table 2). Bayesian Factor Regression Modeling was used to estimate factors based on a given signature. The factor scores were used to predict the phenotype, and leave-one-out cross-validation was used to assess the success of the predictive model (Table 2). The best results were obtained for the comparisons between the high-risk and low-risk groups, the high-risk group and the healthy controls, and the low-risk group and healthy controls, where the AUCs were 0.81, 0.84 and 0.80 respectively. The comparison between the moderate-risk group and the healthy controls had an AUC of 0.69, but the comparisons between the moderate-risk group and the high-risk and low-risk groups had AUC values of 0.50 and 0.58, respectively.

Since the individuals in the high-risk and low-risk groups, and the healthy controls, differed in age and gender (Table 1), we performed the class prediction analysis using only age and gender. These two parameters were not good predictors of the phenotypes, resulting in AUC's of 0.60 or lower suggesting that the age and gender differences were not contributing to the class prediction analysis (data not shown). To exclude the possibility that the use of an anticoagulant might influence the expression profiles, we also performed these analyses excluding those participants in the high-, moderate-, and low-risk groups who were not taking warfarin. This did not significantly alter the results (data not shown) suggesting that warfarin use was not significantly contributing to our class prediction analysis.

Differential Expression Analysis

To explore the biologic differences between the groups, we used differential expression analysis to determine which genes were differentially expressed in each of the 6 comparisons. Since there was a significant difference in age, gender, and BMI among several of the groups (Table 1), these parameters were factored into this analysis. Using a false discovery rate of 0.005, we found that 3111 gene probes were differentially expressed when the high-risk group and healthy controls were compared, and 446 gene probes were differentially expressed when the high-risk and low-risk groups were compared. These two comparisons had 177 gene probes in common (Table S1).

Only 1 gene (*MGC4677*, long intergenic non-protein coding RNA 15) was differentially expressed in the comparisons between the low-risk group and the healthy controls, and the moderate-risk group and the healthy controls. No genes were differentially expressed when comparing the moderate-risk group to either the high-risk or the low-risk groups, even when false discovery rates of up to 0.25 were used.

Top Significantly Enriched Gene Ontologies

We used the functional annotation chart tool in DAVID to determine the top differentially expressed gene ontologies in the comparisons between the high-risk and low-risk groups, and the high-risk group and healthy controls. The top ontologies for the high-risk compared to the low-risk groups included extrinsic to membrane (GO:0019898) as well as other membrane-related and transport-related ontologies (Table S2). The top ontologies for the high-risk group compared to healthy controls included intracellular organelle lumen (GO: 0070013) as well as several mitochondrial-related ontologies (Table S2).

Gene Ontologies Relevant to VTE

We next used the functional annotation clustering tool in DAVID to look at all of the ontologies of the differentially expressed genes in these two comparisons. Functional annotation clustering groups genes with similar annotation terms including ontologies and pathways, providing a way to look at biological mechanisms. Clustering revealed several categories of gene ontologies with potential relevance to VTE, including blood coagulation (Table 3), immune response (Table S3), and vascular biology (Table S4).

In the coagulation-related category, three genes, *CD46* (complement regulatory protein), *F2RL1* (coagulation factor II receptor-like 1 (PAR2)), and *RAB27A* (Rab27A, member RAS oncogene family) were differentially expressed in the comparisons between the high-risk group and the low-risk group as well as the high risk group and healthy controls (Table 3). For each of these genes, expression is lower in the high-risk group, compared to either the low-risk group or the healthy controls. One additional gene is differentially expressed in the high risk vs. low risk comparison, and 21 genes are differentially expressed in the high risk vs. healthy controls comparison (Table 3). Several of the genes differentially expressed in these two comparisons have been previously identified as being of potential clinical relevance in patients with VTE, including, *SELP*, *ANXA5*, *KLKB1*, and *F11* [12,33,34] (Table 3).

In the immune-response related category, 14 genes were differentially expressed in both comparisons (Table S3). An additional 6 genes were differentially expressed in the comparison between the high-risk and low-risk groups, and 122 genes were differentially expressed in the comparison between the high-risk group and healthy controls (Table S3). Multiple genes differentially expressed in these two comparisons have been previously identified as being of potential clinical relevance in patients with VTE, including *SELP*, *IL4*, and *TF*.

In the vascular biology-related category, 6 genes were differentially expressed in both comparisons (Table S4). An additional 2 genes were expressed in the comparison between high-risk and low-risk groups, and 44 genes were differentially expressed in the comparison between the high-risk group and healthy controls (Table S4) Several genes unique to this category have been associated with VTE, including *ANXA2*, in patients with antiphospholipid syndrome, and *HIF1A*.

Differentially Expressed Genes in Pathways Relevant to VTE

Twelve genes in the KEGG complement and coagulation cascades pathway (hsa04610) were differentially expressed in the high-risk group versus healthy controls comparison (Table 3). One of these genes, *CD46*, is also differentially expressed in the high-risk versus low-risk comparison.

Correlation Between Gene Expression and Protein Expression

To investigate whether there was any relationship between gene expression and protein levels, we selected five genes relevant to VTE that are differentially expressed in at least one of the comparisons (*F11, ANXA5, EDN1, SELP, and CD46*) and measured the corresponding protein levels in plasma or serum (Fig. 1). Factor XI levels were significantly higher in the healthy controls compared to the high-risk group (mean plasma protein level 3761 versus 2707 ng/ml, $p = 0.003$), and CD46 levels were significantly higher in the high-risk group compared to the moderate-risk group (mean plasma protein level 1467 versus 1183 pg/ml, $p = 0.042$). All other pairwise comparisons were not significantly different (Fig. 1). The mean concentration and 95% confidence interval of each protein in the 4 study groups is shown in Table S5.

Discussion

We used gene expression profiling as an unbiased approach to explore the relationship between RNA expression levels and the different clinical phenotypes of VTE. Applying class prediction analysis to the gene expression profiles, we obtained the best discrimination between patients with recurrent unprovoked VTE (high-risk group) and healthy controls as well as individuals with provoked VTE only (low-risk group). We obtained reasonable levels of discrimination between patients with a single unprovoked VTE (moderate-risk group) and those with provoked VTE only compared to the healthy controls, but discrimination was poor between individuals with a single unprovoked VTE and the other two VTE groups (Table 2). The moderate-risk group would be expected to be the most heterogeneous of the three patient groups in this study. More than 90% of the patients in the

moderate-risk group were on anticoagulant therapy at the time of enrollment, and the average time from the most recent VTE for this group was 2.17 years (range 0.23 to 7.3). If these patients had discontinued anticoagulant therapy, it would be expected that up to a third of them would sustain a recurrent, unprovoked VTE within ten years, which would then place them in the high-risk group [3,8]. Prospective studies will be necessary to determine whether gene expression profiles can identify which patients with a single unprovoked VTE are at highest risk for developing a recurrent event after a standard course of anticoagulant therapy.

Using differential expression analysis, we found several genes previously identified by alternative strategies as potentially having a role in thrombotic disorders (Tables 3, S3, and S4). Single nucleotide polymorphisms within *F11*, *SELP* and *KLKB1* have been found to be associated with VTE [34]. The *ANXA5* M2 haplotype has been found to be significantly and independently associated with the occurrence of DVT [33]. Upregulation of HIF-1a has been reported to stimulate recanalization of venous thrombus [35]. Our results confirm that these genes are contributing to VTE risk, and that this contribution can be detected at the level of RNA expression in whole blood. Correlations between genotype and RNA expression, and between RNA and protein expression, will be important to understand the relationship between these findings and the risk of recurrent VTE.

In addition to coagulation-related genes, we also found that immune-response genes were frequently differentially expressed in our analyses (Table S3). Crosstalk between the complement and coagulation cascades has been well established. Proteins in the complement cascade can increase the thrombogenicity of blood and coagulation proteins can activate components of the complement cascade [36,37]. We identified 12 genes in the coagulation and complement cascades that are differentially expressed in the high risk group compared to the healthy controls, and one of these genes, CD46, is also differentially expressed in the high risk group compared to the low risk group. Seven genes (CD46, CR1, CR2, C5, CFH, C1QB and SERPING1) are involved in complement activation. Three genes (CR1, C5, and C1QB) were also found to be differentially expressed in peripheral blood mononuclear cells in an independent study comparing patients with pulmonary embolism to patients with ischemic heart disease [38]. Two of the differentially expressed genes we identified in the complement cascade have been linked to thrombotic disorders. Gene mutations in CD46 and CFH have been identified in atypical hemolytic uremic syndrome (aHUS), a complement-mediated form of renal thrombotic microangiopathy [39]. In addition, eculizumab, a humanized monoclonal antibody to C5, reduces the rate of thrombotic events in patients with paroxysmal nocturnal hemoglobinuria [40].

To assess whether plasma levels of several of the proteins expressed by the genes identified by differential expression might be informative, we measured the corresponding levels of five proteins that have been previously associated with risk for VTE. Factor XI levels were significantly, albeit slightly, higher in the healthy control group compared to the high-risk group (Fig. 1), which mirrored the relationship observed in the comparison of RNA expression (Table 3). Prior reports have observed a higher risk of VTE in patients with elevated levels of Factor XI [41], however, and this observation needs to be replicated in a larger patient population with concomitant determination of genetic variants, gene

expression and protein levels. Recent reports in the literature have demonstrated that mRNA levels cannot be relied on to predict protein abundance [42].

We previously used gene expression profiles to evaluate individuals with single versus recurrent VTE [16]. The two patient groups in that study included individuals with provoked as well as unprovoked events, resulting in a more heterogeneous mix of phenotypes. Nevertheless, a 50 gene probe model could distinguish individuals in the two groups with an AUC of 0.75 (95% confidence interval, 0.60 to 0.90). Two genes involved in platelet aggregation (IGF1R and PPARD) were included in that model, as well as ten genes involved in immune and inflammatory responses. There is one gene in common with our current study, *SNRK* an immune response related gene (Table S3). That study used a different platform (Affymetrix), however, which limits the ability to compare results from the two analyses. More recently, Wang, et al. [43] used gene expression profiling of peripheral blood mononuclear cells on an Agilent microarray to compare 20 patients with PE with 20 age and gender matched individuals with ischemic heart disease but without pulmonary embolism. They observed increased mRNA expression of L-selectin, ITGAL, and ICAM-1 in participants with pulmonary embolism [43].

There are several limitations to this study that merit consideration. First, the individual cohorts were clinically heterogeneous, differing by the proportion of patients with PE, the time since their most recent event, and the proportion on anticoagulant therapy at the time of enrollment (Table 1). Patients were identified as belonging in the individual cohorts by the site investigators using pre-defined criteria, however, and were representative of patients encountered in clinical practice at the study sites. A second limitation is that we did not enroll sufficient patients for an independent validation set. For studies with moderate sample sizes, it has been shown that resampling the data provides a more accurate estimate of prediction error than splitting the samples into training and validation sets [28–30]. We chose this approach, estimating the prediction error using leave-one-out cross-validation, an approach that iteratively evaluates each sample and its contribution to the overall model. Our final sample size was smaller than our target, primarily due to the fact that almost a quarter of the participants (43 of 175) were excluded from the final analysis on the basis of sample quality. Sample collection and processing at multiple sites most likely contributed to this outcome.

In summary, we have used gene expression profiling to characterize patients with different clinical phenotypes of VTE. The profiles obtained distinguish patients with recurrent, unprovoked VTE from healthy controls and patients with provoked VTE only, and provide insights into approaches that might be useful in the identification of individuals with a single thrombotic event who are at highest risk for a recurrent VTE after completing a standard course of therapy. Prospective studies are needed to determine the prognostic value of gene expression analyses in identifying these high-risk patients and guiding duration of anticoagulant therapy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by grants from the Centers for Disease Control and Prevention (DD000014 to T. L. Ortel; DD000016 to M. Manco-Johnson; DD000017 to C. S. Philipp; DD000235 to J. A. Heit; and DD000292 to S. Moll). The Centers for Disease Control and Prevention had a collaborative role in the entire study.

We acknowledge the contributions of our clinical research coordinators for screening and enrolling participants into this study, our laboratory staff for initial processing and shipping of blood samples, and the participants who agreed to be included in the study. We thank J. E. Lucas for helpful discussion, advice on analyzing and interpreting the data and critically reading the manuscript.

Abbreviations

ANXA2	annexin A2
ANXA5	annexin A5
AUC	area under the curve
C1QB	complement component 1, q subcomponent, B chain
C5	complement component 5
CD46	complement regulatory protein
CDC	Centers for Disease Control and Prevention
CFH	complement factor H
CR1	complement component receptor 1
CR2	complement component receptor 2
DVT	Deep vein thrombosis
EDN1	endothelin 1
F2RL1	coagulation factor II receptor-like 1
F11	coagulation factor XI
HIF1A	hypoxia inducible factor 1, alpha subunit
ICAM-1	intercellular adhesion molecule 1
IGF1R	insulin-like growth factor 1 receptor
IL4	interleukin 4
ITGAL	integrin alpha L chain
KEGG	Kyoto Encyclopedia of Genes and Genomes
KLKB1	kallikrein B
PE	pulmonary embolism
PPARD	peroxisome proliferator-activated receptor delta
SELP	selectin P
SERPING1	Serpin peptidase inhibitor Glade G (C1 inhibitor)

SNPs	single nucleotide polymorphisms
TF	transferrin
VTE	venous thromboembolism

References

1. Beckman MG, Hooper WC, Critchley SE, Ortel TL. Venous thromboembolism: a public health concern. Am J Prev Med. 2010; 38:S495–501. [PubMed: 20331949]
2. Heit JA, Silverstein MD, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ III. Risk factors for deep vein thrombosis and pulmonary embolism: a population-based case–control study. Arch Intern Med. 2000; 160:809–15. [PubMed: 10737280]
3. Heit JA, Mohr DN, Silverstein MD, Petterson TM, O'Fallon WM, Melton LJ III. Predictors of Recurrence After Deep Vein Thrombosis and Pulmonary Embolism: A Population-Based Cohort Study. Arch Intern Med. 2000; 160:761–8. [PubMed: 10737275]
4. Baglin T, Luddington R, Brown K, Baglin C. Incidence of recurrent venous thromboembolism in relation to clinical and thrombophilic risk factors: prospective cohort study. Lancet. 2003; 362:523–6. [PubMed: 12932383]
5. Crowther MA, Kelton JG. Congenital thrombophilic states associated with venous thrombosis: a qualitative overview and proposed classification system. Ann Intern Med. 2003; 138:128–34. [PubMed: 12529095]
6. Keenan CR, White RH. The effects of race/ethnicity and sex on the risk of venous thromboembolism. Curr Opin Pulm Med. 2007; 13:377–83. [PubMed: 17940480]
7. Kearon C, Akl EA, Comerota AJ, Prandoni P, Bounameaux H, Goldhaber SZ, et al. Antithrombotic therapy for VTE disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012; 141:e419S–94S. [PubMed: 22315268]
8. Schulman S, Lindmarker P, Holmstrom M, Larfars G, Carlsson A, Nicol P, et al. Post-thrombotic syndrome, recurrence, and death 10 years after the first episode of venous thromboembolism treated with warfarin for 6 weeks or 6 months. J Thromb Haemost. 2006; 4:734–42. [PubMed: 16634738]
9. Pabinger I, Ay C. Biomarkers and venous thromboembolism. Arterioscler Thromb Vasc Biol. 2009; 29:332–6. [PubMed: 19228607]
10. Kyrle PA, Rosendaal FR, Eichinger S. Risk assessment for recurrent venous thrombosis. Lancet. 2010; 376:2032–9. [PubMed: 21131039]
11. Verhovsek M, Douketis JD, Yi Q, Shrivastava S, Tait RC, Baglin T, et al. Systematic review: D-dimer to predict recurrent disease after stopping anticoagulant therapy for unprovoked venous thromboembolism. Ann Intern Med. 2008; 149:481–90. w94. [PubMed: 18838728]
12. Kyrle PA, Hron G, Eichinger S, Wagner O. Circulating P-selectin and the risk of recurrent venous thromboembolism. Thromb Haemost. 2007; 97:880–3. [PubMed: 17549288]
13. Hron G, Kollars M, Binder BR, Eichinger S, Kyrle PA. Identification of patients at low risk for recurrent venous thromboembolism by measuring thrombin generation. JAMA. 2006; 296:397–402. [PubMed: 16868297]
14. Kim J, Ghasemzadeh N, Eapen DJ, Chung NC, Storey JD, Quyyumi AA, et al. Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. Genome Med. 2014; 6:40. [PubMed: 24971157]
15. Higgs BW, Liu Z, White B, Zhu W, White WI, Morehouse C, et al. Patients with systemic lupus erythematosus, myositis, rheumatoid arthritis and scleroderma share activation of a common type I interferon pathway. Ann Rheum Dis. 2011; 70:2029–36. [PubMed: 21803750]
16. Lewis DA, Stashenko GJ, Akay OM, Price LI, Owzar K, Ginsburg GS, et al. Whole blood gene expression analyses in patients with single versus recurrent venous thromboembolism. Thromb Res. 2011; 128:536–40. [PubMed: 21737128]

17. Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. J Am Stat Assoc. 2008; 103:1438–56. [PubMed: 21218139]
18. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005; 67:301–20.
19. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33:1–22. [PubMed: 20808728]
20. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A. 2001; 98:11462–7. [PubMed: 11562467]
21. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004; 3:Article 3.
22. Dowling NF, Beckman MG, Manco-Johnson M, Hassell K, Philipp CS, Michaels LA, et al. The U.S. Thrombosis and Hemostasis Centers pilot sites program. J Thromb Thrombolysis. 2007; 23:1–7. [PubMed: 17111206]
23. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008; 24:1547–8. [PubMed: 18467348]
24. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statisical Computing; 2012. [http:www.R-project.org]
25. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci. 2010; 107:9546–51. [PubMed: 20460310]
26. Hackstadt AJ, Hess AM. Filtering for increased power for microarray data analysis. BMC Bioinf. 2009; 10:11.
27. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013; 41:D991–5. [PubMed: 23193258]
28. Simon, R. Guidelines for the Design of Clinical Studies for the Development and Validation of Therapeutically Relevant Biomarkers and Biomarker-Based Classification Systems. In: Gasparini, G.; Hayes, D., editors. Biomarkers in Breast Cancer. Humana Press; 2006. p. 3-15.
29. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005; 21:3301–7. [PubMed: 15905277]
30. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54:774–81. [PubMed: 11470385]
31. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57. [PubMed: 19131956]
32. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1–13. [PubMed: 19033363]
33. Grandone E, Tiscia G, Colaizzo D, Vergura P, Pisanelli D, Margaglione M. The haplotype M2 within the ANXA5 gene is independently associated with the occurrence of deep venous thrombosis. Thromb Haemost. 2010; 103:1102–3. [PubMed: 20174767]
34. Heit JA, Cunningham JM, Petterson TM, Armasu SM, Rider DN, De Andrade M. Genetic variation within the anticoagulant, procoagulant, fibrinolytic and innate immunity pathways as risk factors for venous thromboembolism. J Thromb Haemost. 2011; 9:1133–42. [PubMed: 21463476]
35. Evans CE, Humphries J, Mattock K, Waltham M, Wadoodi A, Saha P, et al. Hypoxia and upregulation of hypoxia-inducible factor 1{alpha} stimulate venous thrombus recanalization. Arterioscler Thromb Vasc Biol. 2010; 30:2443–51. [PubMed: 20930171]
36. Markiewski MM, Nilsson B, Ekdahl KN, Mollnes TE, Lambris JD. Complement and coagulation: strangers or partners in crime? Trends Immunol. 2007; 28:184–92. [PubMed: 17336159]
37. Lupu F, Keshari RS, Lambris JD, Mark Coggeshall K. Crosstalk between the coagulation and complement systems in sepsis. Thromb Res. 2014; 133(Suppl 1):S28–31. [PubMed: 24759136]

38. Lv W, Wang L, Duan Q, Gong Z, Yang F, Song H, et al. Characteristics of the complement system gene expression deficiency in patients with symptomatic pulmonary embolism. Thromb Res. 2013; 132:e54–7. [PubMed: 23726092]
39. Zipfel PF, Heinen S, Jozsi M, Skerka C. Complement and diseases: defective alternative pathway control results in kidney and eye diseases. Mol Immunol. 2006; 43:97–106. [PubMed: 16026839]
40. Van Bijnen ST, Van Heerde WL, Muus P. Mechanisms and clinical implications of thrombosis in paroxysmal nocturnal hemoglobinuria. J Thromb Haemost. 2012; 10:1–10. [PubMed: 22077430]
41. Cushman M, O'Meara ES, Folsom AR, Heckbert SR. Coagulation factors IX through XIII and the risk of future venous thrombosis: the Longitudinal Investigation of Thromboembolism Etiology. Blood. 2009; 114:2878–83. [PubMed: 19617576]
42. Nie L, Wu G, Culley DE, Scholten JC, Zhang W. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. Crit Rev Biotechnol. 2007; 27:63–75. [PubMed: 17578703]
43. Wang H, Duan Q, Wang L, Gong Z, Liang A, Wang Q, et al. Analysis on the pathogenesis of symptomatic pulmonary embolism with human genomics. Int J Med Sci. 2012; 9:380–6. [PubMed: 22811612]

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.thromres.2015.02.003.

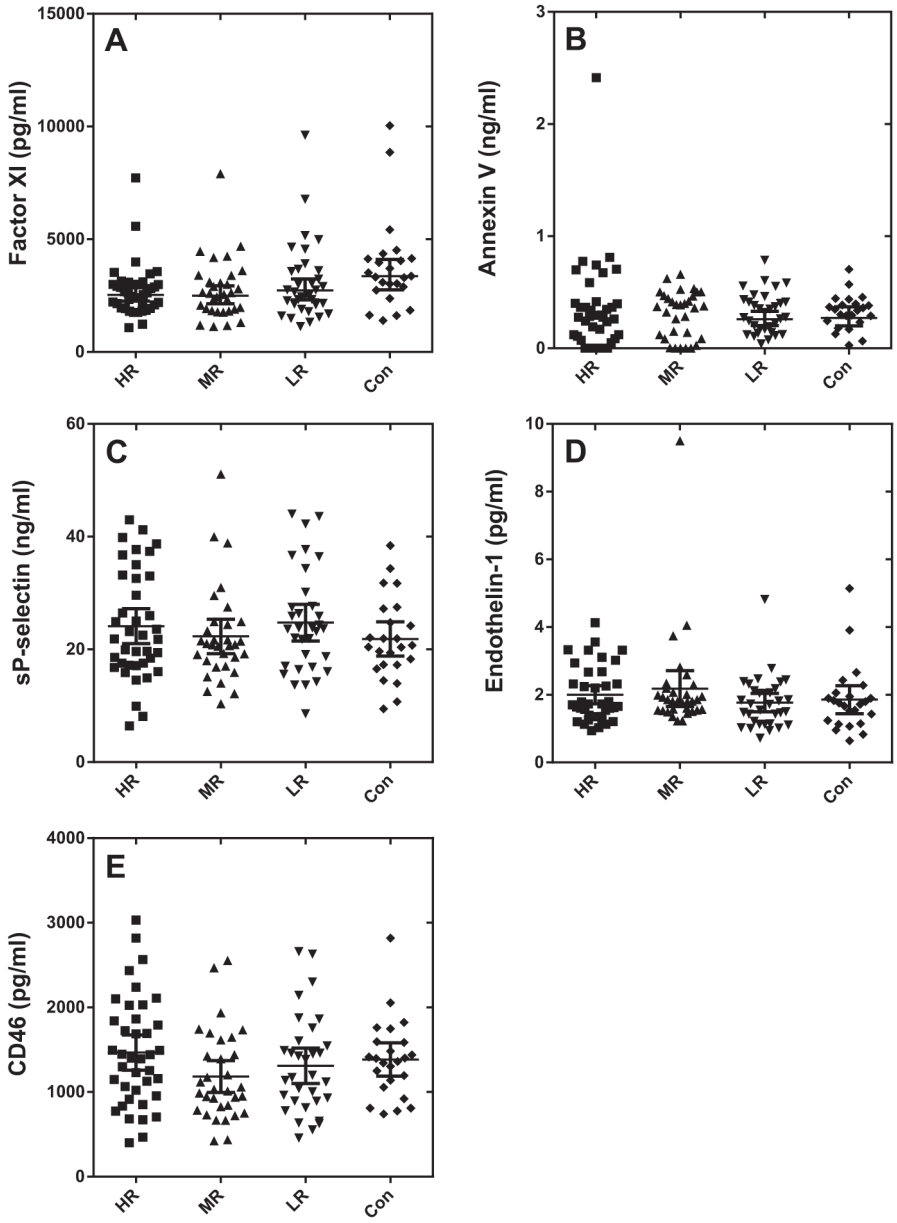

Fig. 1.

Biomarker Levels in the Plasma or Serum of Study Participants. (A) factor XI, (B) annexin A5, (C) sP-selectin, (D) endothelin-1 and (E) CD46 levels were measured in plasma or serum samples collected at the same time as the PAXgene tubes as described in the Material and Methods section. The abbreviations used include: HR, high-risk group; MR, moderate-risk group; LR, low-risk group; and Con, Healthy controls. The mean and 95% confidence intervals are indicated on each graph.

Table 1

Demographics of the Study Participants.[a]

Variable	High-risk (n = 40)	Moderate-risk (n = 33)	Low-risk (n = 34)	Healthy Controls (n = 25)
Age (years), *mean (range)*	56 (27–81)	54(21–84)	48 (24–89)	46 (29–70)
Female, *n (%)*	13 (32)	18 (54)	20 (59)	16 (64)
BMI, mean (range)	33.0 (19.1–47.0)	31.3 (17.1–48.6)	31.9 (19.2–54.0)	28.8 (19.9–43.8)
Race, *n (%)*				
White	33 (82)	27 (82)	30 (88)	21 (84)
Black	7 (17)	6 (18)	3 (9)	3 (12)
Other	-	-	1 (3)	1 (4)
VTE events per subject, *n (%)*				
One	0	21 (64)	29 (85)	
Two	25 (63)	10 (30)	5 (15)	N/A
Three	9 (22)	2 (6)	0	
≥Four	6 (15)	0	0	
Pulmonary embolism, *n (%)*	20 (50)	27 (82)	11 (32)	N/A
Age at first event (years), *mean (range)*	44 (9–74)	50 (17–84)	44 (19–88)	N/A
Time since last VTE (years), *mean (range)*	4.76 (0.24–20.98)	2.17 (0.23–7.30)	2.61 (0.22–13.62)	N/A
Anticoagulant therapy, *n (%)*				
Warfarin	35 (88)	29 (88)	21 (62)	-
Other	5 (12)	2 (6)	1 (3)	-
None	-	2 (6)	12 (35)	25 (100)

[a] Pair-wise comparisons between the study groups were significantly different for the following comparisons. Age: Individuals in the high-risk group were significantly older than those in the healthy control ($p = 0.01$) and the low-risk group ($p = 0.03$), and individuals in the moderate-risk group were significantly older than those in the healthy control group ($p = 0.04$). BMI: Individuals in the high-risk group were significantly larger than those in the healthy control group ($p = 0.017$). Sex: There were fewer females in the high-risk group than the low-risk group ($p = 0.03$) or the healthy controls ($p = 0.02$). Type of VTE: More individuals in the moderate-risk group had PE compared to the high-risk ($p = 0.006$) and low-risk groups ($p < 0.0001$). Time since the last VTE was significantly longer for individuals in the high-risk group compared to the moderate-risk ($p = 0.05$) and low-risk groups ($p = 0.03$). Anticoagulant therapy: More individuals were on anticoagulant therapy in the high-risk and moderate-risk groups compared to the low-risk group ($p = <0.0001$ and $p = 0.005$, respectively).

Table 2

Leave-one-out Cross Validation of Class Prediction using Factors with Penalized Regression.

Comparison	Total Participants in Each Set	Number Classified Correctly	Class error rate	AUC of ROC curve
High-risk vs.	40	37	0.07	0.50
Moderate-risk	33	1	0.96	
High-risk vs.	40	31	0.22	0.81
Low-risk	34	24	0.29	
High-risk vs.	40	36	0.10	0.84
Healthy Controls	25	19	0.24	
Moderate-risk vs.	33	18	0.45	0.58
Low-risk	34	18	0.47	
Moderate-risk vs.	33	29	0.11	0.69
Healthy controls	25	16	0.36	
Low-risk vs.	34	33	0.03	0.80
Healthy Controls	25	12	0.25	

Table 3

Differentially Expressed Coagulation Related Genes.[a]

Gene Symbol	HR vs. LR[b,c]	Expressed higher in	HR vs. Con[b,c]	Expressed higher in
Differentially expressed in both comparisons				
F2RL1	Yes	Low-risk	Yes	Controls
RAB27A	Yes	Low-risk	Yes	Controls
CD46[d]	Yes	Low-risk	Yes	Controls
Differentially expressed in HR vs. LR only				
STXBP3	Yes	Low-risk	No	----------
Differentially expressed in HR vs. Con only				
KLKB1[d]	No	----------	Yes	High-risk
GP1BA	No	----------	Yes	High-risk
SERPINA1[d]	No	----------	Yes	High-risk
ANXA5	No	----------	Yes	Controls
ANXA2	No	----------	Yes	Controls
EDN1	No	----------	Yes	High-risk
F11[d]	No	----------	Yes	Controls
SELP	No	----------	Yes	High-risk
SERPING1[d]	No	----------	Yes	High-risk
EFEMP2	No	----------	Yes	High-risk
DTNBP1	No	----------	Yes	High-risk
PLAUR[d]	No	----------	Yes	High-risk
KIAA1715	No	----------	Yes	High-risk
HPS5	No	----------	Yes	High-risk
C1QB[d]	No	----------	Yes	High-risk
CR1[d]	No	----------	Yes	Controls
CR2[d]	No	----------	Yes	Controls
C5[d]	No	----------	Yes	High-risk
CFH[d]	No	----------	Yes	High-risk
BDKRB1[d]	No	----------	Yes	High-risk
VWA3B	No	----------	Yes	High-risk

[a] The following gene ontology terms were used to identify genes related to coagulation in the two comparisons: GO:0050817 ~ coagulation, GO: 0007596 ~ blood coagulation, and GO:0030193 ~ regulation of blood coagulation and the interpro term, IPR002035:von Willebrand factor, type A.

[b] "Yes" indicates that the gene is differentially expressed; "No" indicates that it is not differentially expressed.

[c] Abbreviations used in this table: HR, high-risk and LR, low-risk; and Con; healthy controls.

[d] Indicates that the gene is also in the KEGG pathway hsa04610: coagulation and complement cascades.