Published in final edited form as:
Thromb Res. 2009 ; 123(Suppl 4): S25–S29. doi:10.1016/S0049-3848(09)70139-2.

Discovering Novel Risk Factors for Venous Thrombosis: a Candidate-Gene Approach

Nicholas L. Smith, PhD[1,5], **Kenneth M. Rice, PhD**[2], **Thomas Lumley, PhD**[2], **Susan R. Heckbert, MD, PhD**[1], and **Bruce M. Psaty, MD, PhD**[1,3,4]

[1]Department of Epidemiology, University of Washington, Seattle Washington

[2]Department of Biostatistics, University of Washington, Seattle Washington

[3]Department of Medicine, University of Washington, Seattle Washington

[4]Department of Health Services, University of Washington, Seattle Washington

[5]Seattle Epidemiologic Research and Information Center of the Department of Veterans Affairs Office of Research and Development, Seattle, Washington

Abstract

The candidate-gene approach can be used to locate and identify genetic variations that are associated with a particular phenotype. This gene-centric approach assumes that there exists important genetic variation within genes that can influence health. Identifying known genes which are candidates for the phenotype of interest can be accomplished using existing knowledge about biology or using findings from genome-wide association studies. Genetic variation can be characterized locally by single nucleotide polymorphisms (SNPs) or insertiondeletions, or it can be characterized more broadly in terms of haplotypes and diplotypes, which usually need to be inferred statistically. As an example, we present a candidate-gene approach to identify novel associations between variation in 24 clotting genes and the risk of incident venous thrombosis.

Candidate-Gene Approach

Several approaches can be used to locate and identify genetic variations that are associated with a particular phenotype. In this paper, we describe the candidate-gene approach, in which genetic variations at the level of the gene are hypothesized to be associated with a phenotype. We focus on the phenotype of venous thrombosis (VT), a clinical event, but the methods described can be applied to other clinical phenotypes or intermediate phenotypes, such as plasma hemostasis factor levels. Our candidate-gene approach assumes that there exists important genetic variation within genes that can influence health. This is a reasonable assumption as gene-level variation within several genes, such as factor V, prothrombin, and protein C, has already been associated with VT risk.[1] The approach, however, basically ignores genomic variation that exists between genes throughout the chromosome and also disregards macro-level genomic variation across genes, such as copy-number variants.2

Identifying known genes which are candidates for the phenotype of interest can be accomplished in several ways. The biologic approach uses existing knowledge about biology to identifying genes whose variation may be important in the etiology of disease. This approach, however, is limited by our current understanding of the disease and is not likely to lead to new insights into pathophysiology. In contrast, the more recent genome-wide approach, which ignores biology, uses markers throughout the genome to identify regions where genetic variation is observed to be associated with disease risk.[3] Regions can include 1 or more novel genes that may be culprits in disease causation. To date, this approach has not been applied to the VT phenotype but for other complex phenotypes, there has been modest success identifying new associations with a genome-wide approach.[4]

Genetic Variation

Two types of genetic variation are of primary interest in a candidate-gene approach. The most common is the single-nucleotide polymorphism (SNP). A SNPs is a single base-pair substitution of a nucleotide in the DNA strand of the gene.[5] Most SNPs are bi-allelic meaning that variation is limited to a single substitution at a locus. For example a guanine (G) nucleotide may be substituted with a thymine (T) nucleotide. These substitutions can occur anywhere in the genome. When they occur in exon regions of a gene, regions which code for the transcription of proteins, the nucleotide substitution may lead to a change in the amino acid for which the segment codes. The other type of genetic variation is the insertion-deletion, which is a small-scale deletion—or insertion—of 1 or more nucleotides. These also occur throughout the genome but are less common than SNPs.5 When occurring in a gene, insertion-deletions can shift the reading frame for the gene. Current high-throughput, genotyping technology efficiently captures individual's variants at SNPs but is not able to capture insertion-deletions.

Both SNPs and insertion-deletions can affect the biologic function of the gene. Variation in the promoter region, which initiates transcription, can influence gene synthesis. Variation within exons and splicing regions can lead to changes in amino acids and to alterations in splicing, often resulting in a gene-product with less function than its non-variant counterpart. Variation that increases the function of the gene-product is less common but include the important factor V Leiden mutation.[6] Either scenario can result in down-stream biologic changes depending on the physiologic role of the gene-product. Most SNP variants in genes, however, do not appear to have consequences for the structure and function of the gene.

Haplotypes and Diplotypes

Genetic variation can be characterized locally by SNPs or insertion-deletions, or it can be characterized more broadly in terms of haplotypes and diplotypes. For all chromosomes, except the X and Y sex chromosomes in males, human beings have 2 strands of DNA. One strand, or haplotype, is inherited from the mother and the other from the father. For any particular region in a chromosome these 2 DNA strands may carry 1 or more SNPs or insertion-deletions. For a candidate-gene investigation, we are interested in gene-specific haplotypes. Figure 1 presents 7 segments of DNA that include 4 bi-allelic SNPs. The polymorphic sites are represented in bold font and the variant nucleotides are in gray font. The configuration of SNPs is such that 6 unique haplotypes are characterized by the 4 SNPS. For example, haplotype I does not carry any variant at the 4 polymorphic sites whereas haplotype II carries 1 variant at the 2nd polymorphic site. Note that the 3rd and 4th SNPs carry redundant information for the haplotype for this segment of DNA. The correlation between genotypes at these 2 SNPs would be very high and would be described as having strong linkage disequilibrium (LD). On a larger scale, we can characterize the complete haplotype for a gene by considering all variants along the strand that encompasses the gene. In most situations, SNP-level data are not strand-specific so haplotype structure is inferred statistically.[7] Since each person has 2 haplotypes for each gene, the gene-level genetic variation on both haplotypes can be considered simultantaneously when characterizing the gene. This is the diplotype approach. For example, a person can be a carrier of haplotypes I and IV or a carrier of 2 copies of haplotype IV. As the number of SNPs increases in a gene, the number of haplotypes and diplotypes characterizing the gene increases. For example, 1 SNP produces 2 haplotypes (1 haplotype defined by presence of variant, 1 haplotype defined by absence of variant), which produced 3 diplotypes (carrier of 2 haplotypes with the variant, carrier of 2 haplotypes without the variant, and carrier of 1 haplotype with and 1 haplotype without the variant). A gene with 3 SNPs can theoretically produce 8 haplotypes and 36 diplotypes. However, fewer haplotypes and diplotypes are often observed since there is a large amount of LD (correlation) between SNPs and not all possible combinations of SNPs on a haplotype and haplotype-pairs exist.

Example using 24 Clotting Candidate Genes

We used a candidate-gene approach to identify novel associations between variation in 24 clotting genes and the risk of incident VT. Based on the biology of clotting, we selected 24 genes that code for proteins affecting coagulation (factors II, V, VII, VIII, IX, X, XI, XII, XIIIa1, and XIIIb; α-, β-, and γ-fibrinogen; and tissue factor), anti-coagulation (antithrombin, proteins C and S, endothelial protein C receptor, thrombomodulin, and tissue factor pathway inhibitor [TFPI]), fibrinolysis (plasminogen and tissue-type plasminogen activator [TPA]), and antifibrinolysis (type 1 plasminogen activator inhibitor [PAI-1] and thrombin activatable fibrinolysis inhibitor [TAFI]).

Source of Genetic Variation

We used publicly available data on gene-wide variation derived from sequencing efforts. These data came from the Seattle SNPs Program for Genomic Applications (PGA) that conducted complete sequencing of all 24 genes in 23 individuals (46 alleles) of European ancestry and 24 individuals (48 alleles) of African ancestry (http://pga.gs.washington.edu/). These data include all SNPs and insertion deletions identified in the 94 alleles (haplotypes). From these data we chose SNPs that had a minor allele frequency of at least 5% in either population and used the computer program ldSelect (http://droog.gs.washington.edu/ldSelect.html), which considers correlation information between all SNPs in a region, to select a set of SNPs, known as tag SNPs, which efficiently characterize the common haplotypes.[8]

As an example, Figure 2 provides a diagram of the genetic structure and variation in the factor IX gene (F9). The Seattle SNPs PGA sequenced 100% of the 33,000 base pairs that encompass the F9 gene and identified 77 variants in the 2 populations. Using ldSelect, we chose 10 SNPs to tag haplotypes for F9 in subjects of European ancestry and 20 SNPs to tag haplotypes for F9 in subjects of African ancestry. For some genes, such as factor VIII (F8), Seattle SNPs was not able to sequence the full gene and certain gene segments, such as exons and introns, were omitted and did not contribute to SNP selection or haplotype estimation.

Study Population

The study population came from a population-based, case-control study of 349 women with incident events of deep vein thrombosis or pulmonary embolism and 1,680 control subjects. All women were 30–89 years of age, post-menopausal, and not related to one another.[9] All events were verified by medical record review. Genomic DNA was collected from whole blood and genotyping was performed using a custom Illumina GoldenGate panel (Illumina Inc., San Diego, California).

We used the PHASE computer program to infer full-gene haplotypes for each subject (http://www.stat.washington.edu/stephens/software.html) based on our tag-SNP data. Haplotypes with a minor allele frequency (MAF) of less than 2% of the study population were combined into 1 ‘rare haplotype’ category. The factor V (F5) gene was large and there were few common haplotypes with a MAF of 2% or greater. For this reason, we split the F5 gene in 2 at a locus of historical recombination when creating haplotypes.

Our approach was to test global variation in a gene using haplotype information and to test for SNP and haplotype associations independently. We conducted tests on 25 genes (F5 was split), 170 SNPs, and 173 haplotypes. Additive genetic models were used to test the association between additional SNP and haplotype copies and VT risk. Global testing compared the haplotype model that characterizes all common haplotypes within the gene with a model that included no haplotype information. We excluded from hypothesis testing the 2 well-characterized SNPs, FV Leiden and prothrombin (F2) 20210A, and their tagged haplotypes.[10, 11] To deal with the issue of multiple testing, we used the false discovery rate (FDR) q-statistic to guide statistical significance.[12] Briefly, FDR methods output a set of hypothesis-testing decisions, among which the rate of false-positives is controlled below a specified value, in expectation. Use of the FDR is appropriate when the number of expected true positives is not expected to be small. Our threshold of significance was set at 0.2, meaning that no more than 20% of reported results are expected to be false positives.[13]

The results from our candidate-gene approach for incident VT have been published elsewhere.[9] Briefly, of the 25 gene-wide variation tests performed, 1 was associated with q-value <0.2 (TFPI) and 2 other genes were associated with a p-value <0.05 (F5 upper half and protein C). Of the 170 SNPs tested, 5 SNPs had a q-value <0.2 among the 21 (12%) SNPs that had a p-value <0.05. Of the 173 haplotypes tested, 1 haplotype was associated with a q-value <0.2 among the 20 (12%) haplotypes that had a p-value <0.05. Among the 5 SNPs with an associated q-value <0.2, FV K858R (rs4524, MAF=27%, OR=0.7, p-value = 0.003) and FXI 22771 (rs2289252, MAF=40%, OR=1.3, p-value = 0.002) were novel findings. The 3 remaining SNPs were in the protein C gene (PROC 2583: rs1799810, MAF=42%, OR=1.3, p-value = 0.001; PROC 4919: rs2069915, MAF=41%, OR=0.8, p = 0.005; and PROC 11310: rs5937, MAF=32%, OR=1.4, p < 0.001), the latter being a novel discovery.[14, 15] The haplotype associated with a q-value <0.2 was in the PROC gene and was uniquely marked by the 11310 variant.

In summary, 1 gene (TFPI) was globally associated with risk although neither its haplotypes or SNPs reached our threshold for significance. Five SNPs in 3 genes (F5, F11, and PROC) were significantly associated with risk. Overall, few of the associations suggested more than doubling or halving of risk and most variants were commonly occurring. Of note, few haplotype effects were detected. Diplotype analyses were not conducted.

Limitations

For candidate-gene approaches to discovering novel risk factors for VT, gene coverage is only as good as the source of the variation data. Genetic regions or variants that are rare, that are not genotyped, or are not in sufficiently high LD with SNPs that are genotyped would be overlooked. Publicly available sequence data, like those available on the Seattle PGA, are unlikely to contain most rare SNPs, such as those with a MAF less than 2%. For example the F2 20210A variant (MAF 0.02) was not detected in the 2 panels of subjects sequenced by the Seattle PGA. When investigating genes with a large amount of genetic variation and multiple genes, issues of multiple testing need to be considered and accounted for when interpreting p-values. Like all association studies, the candidate-gene approach does not provide functional information or definitive conclusions on causality.

Conclusions

Candidate-gene studies use knowledge from biology to identify and investigate gene-centric variation associated with novel genetic risk factors. This approach can be used to take advantage of findings from genome-wide association studies of VT and provide further insight into new candidate genes that influence the risk of thrombosis.

Acknowledgments

This paper was supported by the symposium entitled, “Venous Thromboembolism: From Bench to Bedside” sponsored by the University of North Carolina at Chapel Hill; by the National Health Lung and Blood Institute grants HL73410, HL60739, HL68639, HL43201, HL74745, HL68986; by a grant from the Leducq Foundation, Paris, France for the development of Transatlantic Networks of Excellence in Cardiovascular Research; and by the Seattle Epidemiologic Research and Information Center of the Department of Veterans Affairs Office of Research and Development.

References

1. Franco RF, Reitsma PH. Genetic risk factors of venous thrombosis. Hum Genet 2001;109:369–384. [PubMed: 11702218]
2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet 2006;7:85–97. [PubMed: 16418744]
3. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008;118:1590–1605. [PubMed: 18451988]
4. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–678. [PubMed: 17554300]
5. Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet 2001;27:234–236. [PubMed: 11242096]
6. Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, van der Velden PA, Reitsma PH. Mutation in blood coagulation factor V associated with resistance to activated protein C. Nature 1994;369:64–67. [PubMed: 8164741]
7. French B, Lumley T, Monks SA, Rice KM, Hindorff LA, Reiner AP, Psaty BM. Simple estimates of haplotype relative risks in case-control data. Genet Epidemiol 2006;30:485–494. [PubMed: 16755519]
8. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004;74:106–120. [PubMed: 14681826]
9. Smith NL, Hindorff LA, Heckbert SR, Lemaitre RN, Marciante KD, Rice K, Lumley T, Bis JC, Wiggins KL, Rosendaal FR, Psaty BM. Association of genetic variations with nonfatal venous thrombosis in postmenopausal women. JAMA 2007;297:489–498. [PubMed: 17284699]
10. Rosendaal FR, Koster T, Vandenbroucke JP, Reitsma PH. High risk of thrombosis in patients homozygous for factor V Leiden (activated protein C resistance). Blood 1995;85:1504–1508. [PubMed: 7888671]
11. Poort SR, Rosendaal FR, Reitsma PH, Bertina RM. A common genetic variation in the 3'-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis. Blood 1996;88:3698–3703. [PubMed: 8916933]
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 1995;57:289–300.
13. Benjamini Y, Yekutieli D. Quantitative trait Loci analysis using the false discovery rate. Genetics 2005;171:783–790. [PubMed: 15956674]
14. Aiach M, Nicaud V, Alhenc-Gelas M, Gandrille S, Arnaud E, Amiral J, Guize L, Fiessinger JN, Emmerich J. Complex association of protein C gene promoter polymorphism with circulating protein C levels and thrombotic risk. Arterioscler Thromb Vasc Biol 1999;19:1573–1576. [PubMed: 10364092]
15. Spek CA, Koster T, Rosendaal FR, Bertina RM, Reitsma PH. Genotypic variation in the promoter region of the protein C gene is associated with plasma protein C levels and thrombotic risk. Arterioscler Thromb Vasc Biol 1995;15:214–218. [PubMed: 7749828]

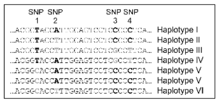
SNP 1
SNP 2
SNP 3
SNP 4
...ACCCTACCATTGGAGTCCTCCCCCTCA... Haplotype I
...ACCCTACCTTTGGAGTCCTCCCCCTCA... Haplotype II
...ACCCTACCTTTGGAGTCCTCGCCTTCA... Haplotype III
...ACGGTACCATTGGAGTCCTCGCCTTCA... Haplotype IV
...ACGGCACCATTGGAGTCCTCCCCCTCA... Haplotype V
...ACGGCACCATTGGAGTCCTCCCCCTCA... Haplotype V
...ACGGCACCTTTGGAGTCCTCCCCCTCA... Haplotype VI

Figure 1.

Genetic variation across 7 strands of DNA. Variant loci are depicted in bold font and variant nucleotide in gray font.

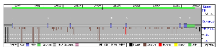

Figure 2.

Genetic structure and variation across the factor IX gene (F9), Seattle Program for Genomic Applications (http://pga.gs.washington.edu/data/f9/).