European Journal of Human Genetics (2004) 12, 478–482

www.nature.com/ejhg

ARTICLE

Candidate gene approach in association studies: would the factor V Leiden mutation have been found by this approach?

Astrid van Hylckama Vlieg[1,2], Lodewijk A Sandkuijl*[,3], Frits R Rosendaal*[1,2], Rogier M Bertina[2], Hans L Vos[2]

[1]*Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands,* [2]*Hemostasis and Thrombosis Research Center, Leiden University Medical Center, Leiden, the Netherlands,* [3]*Department of Medical Statistics, Leiden University Medical Center, Leiden, the Netherlands*

A re-emerging strategy in the search for disease susceptibility genes is the evaluation of candidate genes, which are thought to play a role in disease pathogenesis. Candidate genes are screened for single nucleotide polymorphisms (SNPs) in a case–control study. The factor V Leiden (FVL) mutation (1691G→A in the *F5* gene) is an important risk factor for venous thrombosis. We asked ourselves whether the FVL mutation would have been found using the candidate gene approach in the absence of prior knowledge of the haplotype structure of the *F5* gene. We typed four SNPs in the *F5* gene in the Leiden Thrombophilia study, that is, promoter (99930G→A), exon 13 (55907A→G), exon 16 (42855A→G), and intron 19 (37833T→G). These SNPs were known to have different population frequencies, making their presence in distinct haplotypes likely. None of these SNPs has previously been associated with venous thrombotic risk. Subsequently we derived haplotypes. One haplotype was clearly more frequent in patients than controls (GAAT; 20 *versus* 9%), suggesting that a polymorphism in or near the *F5* gene in this haplotype is associated with an increased thrombotic risk. If we had sequenced the *F5* gene in patients homozygous for this haplotype, in order to locate the possible causal polymorphism, we would have found that 16 (76%) patients were homozygous or heterozygous for a missense mutation in exon 10 (1691G→A), which predicts the replacement of Arg506 by Gln in one of the cleavage sites for activated protein C, a mutation that we now know as the FVL mutation.
European Journal of Human Genetics (2004) **12**, 478–482. doi:10.1038/sj.ejhg.5201183
Published online 31 March 2004

Keywords: association study; factor V Leiden; haplotypes; candidate gene

Introduction

The detection of genes underlying complex diseases has proven to be far more complicated than the detection of genes associated with diseases with a Mendelian or near-Mendelian inheritance. Many studies that reported an association between a genotype and disease have not been replicated.[1,2] Reasons for this irreproducibility may be a reduced power of some of the studies because of a small sample size or the use of subgroup analysis, or multiple comparisons. Other problems are the small magnitude of the effect of these genes on disease or the absence of a hypothesis on a biologic mechanism, which could have resulted in a measurable intermediate phenotype.[3,4]

Nevertheless, the attributable risk associated with genes that have a moderate effect on disease, that is, the fraction

*Correspondence: Dr FR Rosendaal, Department of Clinical Epidemiology, Leiden University Medical Center, C9-P, PO Box 9600, 2300 RC Leiden, The Netherlands. Tel: +31 71 5264037; Fax: +31 71 5266994; E-mail: f.r.rosendaal@lumc.nl
*Deceased.
Received 16 October 2003; revised 7 January 2004; accepted 6 February 2004

of the disease risk caused by the allelic variation in this gene, may be relatively large when the disease-causing genotype is common. Risch and Merikangas[5] showed that population-based association studies had more power to detect such loci than linkage analysis.

Therefore, a re-emerging strategy in the search for disease susceptibility genes is the evaluation of candidate genes in population-based association studies. In this approach, the genotype of several markers in or around the candidate gene, most often single nucleotide polymorphisms (SNPs), are studied in unrelated patients and healthy control subjects. Identification of a candidate gene can be based on the results from positional cloning, but also on a hypothesis on the biochemical properties of the encoded protein and its role in the etiology of the disease. By typing several SNPs, one can estimate haplotypes of the candidate gene and establish all the common variants of a gene. The main hypothesis in the candidate gene approach is that a relatively common functional variant exists that is still mainly present in its founder haplotype and that increases the risk of disease. As a result, the frequency of the founder haplotype will be increased in the patient population. The magnitude of this increase is dependent on the risk associated with the causal polymorphism and on its frequency in the founder haplotype. Ideally, when information on haplotypes is already available for the gene under study, one can limit the exercise to those SNPs that are specific to the existing haplotypes (haplotype-tagging SNPs). However, when the information on the haplotype-structure of a gene is nonexistent or only very limited, as was the case for the *F5* gene, the decision of which SNPs to determine is dependent on their frequency in the general population, their degree of linkage disequilibrium and their position in the gene or, if all the rest is unknown, only on the latter. By choosing SNPs spread throughout the gene, of which the rare alleles have nonidentical frequencies between 15 and 50%, the chance of distinguishing most of the major haplotypes is high. This will increase the likelihood of finding the haplotype(s) that contain(s) the causal polymorphism. After identifying a disease-associated haplotype, one should subsequently sequence (preferably homozygous) carriers of this haplotype to identify additional SNPs that are specific to this haplotype or that contribute to a subhaplotype. By selecting homozygous patients, the chance of finding such SNPs is maximized.

Deep venous thrombosis is a multicausal disease. Several genetic as well as acquired risk factors have been described that contribute to the risk of venous thrombosis, which is thought to result from specific interaction between genes and environment.[6] Until 1993, only a minority of thrombosis patients was found to carry one of the genetic risk factors known at that time. In 1993, Dahlbäck *et al*[7] described a new mechanism for thrombosis that was characterized by the laboratory phenotype of a poor anticoagulant response to activated protein C (APC resistance). In 1994, we described a mutation in the factor V gene (G→A substitution at nucleotide position 1691 in the cDNA), which predicts the synthesis of a factor V molecule that is less efficiently downregulated by APC, and thus explains the laboratory phenotype of APC resistance.[8] The factor V Leiden (FVL) mutation has been associated with a three- to eight-fold increased risk of venous thrombosis. The prevalence of this mutation ranges between 3 and 7% in Caucasian populations with large regional differences. Several studies reported on a haplotype in the factor V gene that was associated with the FVL mutation, thus providing evidence for a single origin of this mutation.[9–13]

We now asked ourselves whether the FVL mutation would have been found by evaluating *F5* as a candidate gene in a population-based association study using a case–control design. This study was therefore performed without knowing beforehand who carried the FVL mutation, and who did not, and without taking the limited haplotype information on this gene into account.

Patients and methods

Selection of a candidate gene can be based on the biochemical properties of the encoded protein. Coagulation factor V, which plays an essential role in blood coagulation, is as such an obvious candidate in an investigation of the etiology of venous thrombosis. It is together with factor VIII, the only physiological substrate of APC. The gene for human factor V (*F5*) is localized on chromosome 1q23–24, and consists of 25 exons and 24 introns.[14] The size of the factor V gene is 75 kb (Figure 1).

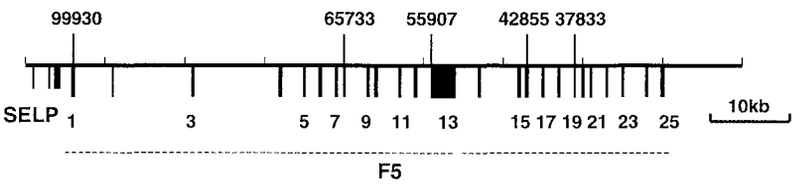

Figure 1 Schematic representation of the genomic region of the factor V gene. The gene is shown approximately to scale; the relative sizes of the exons, shown below the line, are indicated by the thickness of the lines. The approximate position of the SNPs and their exact coordinates in GenBank: Z99572 are shown above the line. F5 = factor V gene; SELP = P-selectin gene. Numbers below the line refer to exon numbers.

Since, during the performance of this study, very little information was available with respect to the general haplotype structure of the factor V gene, SNPs were chosen based on their high frequency in the Dutch population (15–50% for the rare allele) and their position in the gene. By choosing SNPs spread throughout the gene, of which the rare alleles have nonidentical frequencies between 15 and 50%, the chance of distinguishing most of the major haplotypes is high. We determined a SNP in the promoter region (99930G→A), in exon 13 (55907A→G), in exon 16 (42855A→G), and in intron 19 (37833T→G) (numbering according to GenBank: Z99572).

This study was performed in 474 unselected patients with a first deep venous thrombosis and 474 unrelated controls from the Leiden Thrombophilia Study (LETS).[15] The control subjects came from the same geographical area as the patients.

The risk of venous thrombosis was calculated for each SNP individually by calculating odds ratios (OR) and their 95% confidence intervals (95% CI). Furthermore, using Arlequin, a software program for population genetics,[16] we estimated the haplotype frequencies in patients and control subjects. With this program, maximum-likelihood haplotype frequencies are computed using an expectation-maximization (EM) algorithm.[17]

Results and discussion

The genotypes for all four SNPs could be determined by polymerase chain reaction (PCR) in 471 patients and 472 control subjects. Of these four polymorphisms, the SNPs in exons 13, 16 and intron 19 were in linkage disequilibrium (exons 13 and 16: $\Delta = 0.90$; exon 13 and intron 19: $\Delta = 0.67$; exon 16 and intron 19: $\Delta = 0.74$).

In Table 1 the venous thrombotic risk associated with the less frequent allele of each SNP is shown. The SNPs in the promoter region and exons 13 and 16 were itself not associated with an increased risk of venous thrombosis, that is, carriership of the rare allele of these SNPs is at most associated with a weakly protective effect. Individuals homozygous for the less frequent allele of the SNP in intron 19 (T allele) had a weakly increased risk of venous thrombosis compared with individuals homozygous for the G allele (OR = 1.4; 95% CI: 1.0–2.1).

Using Arlequin, a software program for population genetics,[16] we subsequently estimated the haplotype frequencies in patients and control subjects. In Table 2, the frequencies in both patients and control subjects are shown for all haplotypes with a frequency of more than 1% in the total population of both patients and control subjects (LETS). In total, 11 of the 16 possible haplotypes were found in this study population, seven of which had a frequency of more than 1%. The haplotype distribution in patients and control subjects was significantly different (χ^2: 26.0, df: 10, *P*-value: 0.004).

As shown in Table 2, there was one haplotype that was clearly more frequent in patients compared with control subjects (GAAT; 20% in patients *versus* 9% in control subjects). This finding suggests that a polymorphism in or near the factor V gene in this haplotype is associated with

Table 1 The risk of deep venous thrombosis associated with several SNPs in the factor V gene

	N patients (%)	*N controls (%)*	*OR (95%CI)*
FV Promoter G99930A			
GG	314 (66.7)	306 (64.8)	1[a]
AG	144 (30.6)	147 (31.1)	1.0 (0.7–1.3)
AA	13 (2.8)	19 (4.0)	0.7 (0.3–1.4)
Frequency A allele	18.0	19.6	
FV exon 13 A55907G			
AA	311 (66.0)	262 (55.5)	1[a]
AG	135 (28.7)	178 (37.7)	0.6 (0.5–0.8)
GG	25 (5.3)	32 (6.8)	0.7 (0.4–1.1)
Frequency G allele	19.6	25.6	
FV exon 16 A42855G			
AA	273 (58.0)	233 (49.4)	1[a]
AG	162 (34.4)	198 (41.9)	0.7 (0.5–0.9)
GG	36 (7.6)	41 (8.7)	0.7 (0.5–1.2)
Frequency G allele	24.8	29.7	
FV intron 19 T37833G			
GG	133 (28.2)	154 (32.6)	1[a]
GT	226 (48.0)	228 (48.3)	1.1 (0.9–1.5)
TT	112 (23.8)	90 (19.1)	1.4 (1.0–2.1)
Frequency T allele	47.8	43.2	

[a]Reference category.

Table 2 Haplotypes in the factor V gene with a frequency >0.01 in the total population (standard deviation)

Haplotype[a]	*All*	*Patients*	*Controls*
GAAG	0.44 (0.01)	0.41 (0.02)	0.47 (0.02)
GGGT	0.19 (0.01)	0.17 (0.01)	0.20 (0.01)
GAAT	**0.15** (0.01)	**0.20** (0.01)	**0.09** (0.01)
AAAG	0.10 (0.01)	0.11 (0.01)	0.10 (0.01)
AGGT	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)
GAGT	0.04 (0.005)	0.04 (0.01)	0.04 (0.01)
AAAT	0.03 (0.01)	0.03 (0.01)	0.04 (0.01)

[a]Order of the SNPs in the table is: promoter G99930A, exon 13 A55907G, exon 16 A42855G, intron 19 T37833G.

an increased risk of venous thrombosis, that is, could be the causal variant, and thus causes a higher frequency of this haplotype in patients. Consequently, the frequency of some of the other haplotypes is slightly lower in patients compared with the control subjects. This explains the weakly protective effect of some of the polymorphisms described in Table 1.

In all, 21 patients and seven control subjects were homozygous for the GAAT-haplotype. If we had sequenced the factor V gene in all these patients in order to locate the possible causal polymorphism, we would have found that 16 (76%) patients were homozygous ($n=7$) or heterozygous ($n=9$) for a so far unknown missense mutation in exon 10 (1691G→A), which predicts the replacement of Arg 506 by Gln in one of the cleavage sites for activated protein C, a mutation, that we now know as the FVL mutation. Probably, also other sequence variations would have been found during the sequencing of the factor V genes of these subjects. However, the fact that this mutation affects one of the cleavage sites for activated protein C would have made this missense mutation a likely candidate for being the functional mutation, and thus one of the first mutations that would be investigated in more detail.

The SNPs in exons 13 and 16 were in strong linkage disequilibrium. The construction of haplotypes using only the SNPs in the promoter, intron 19 and either exon 13 or 16 would also have led to the discovery of a haplotype that was clearly more frequent in patients than in control subjects (eg, Prom–exon 13–intron 19: GAT patients: 23.9%, controls subjects: 12.8%).

As mentioned above, in a multicausal disease such as deep venous thrombosis, in which more than one risk factor needs to be present to cause the disease, highly prevalent, weak to moderate risk factors may occur, as well as numerous rare, very strong risk factors. The latter are more likely to be found using linkage analysis.

Clearly, more genetic risk factors will be discovered in the future, especially since in about 30% of patients with a family history of the disease as yet no (genetic) risk factor has been found.

We now demonstrate that an earlier discovered common risk factor for venous thrombosis, that is, the FVL mutation, would indeed have been found using population-based association studies with a case–control design. Recently, in the study of risk factors for venous thrombosis, the pros and cons of association studies *versus* linkage analysis have been debated.[18,19] Following the candidate gene approach in an association study, prevalent risk factors associated with a moderate increase in the risk of thrombosis can be found that are unlikely to be found using linkage analysis due to a lack of power of the latter studies.

The chance of finding a causal mutation with the candidate gene approach depends on the number of patients (N_{pat}) and control subjects (N_{con}), the risk associated with the mutation (OR_{mut}), and the frequency of the mutation in the general population (f_{mut}). Under the assumption that a causal mutation exists only in one (founder) haplotype, one can calculate the maximum frequency (f_{max}) of this (founder) haplotype in the general population, with which it is still possible to detect an increased risk associated with this haplotype. In general, since it is unknown beforehand which haplotype contains the causal mutation, the frequency of the most common haplotype should not exceed f_{max}.

In this study (i.e., $N_{pat}=N_{con}=474$, $OR_{mut}=6$, $f_{mut}=4\%$), we can calculate that $f_{max}=64\%$. As shown in Table 2, the frequency of the most common haplotype equals 47%. Thus, we would have been able to find the FVL mutation in this study. Note that in a smaller study, more SNPs would be needed to split the most frequent haplotype into smaller groups (eg, when $N_{pat}=N_{con}=100$, $f_{max}=23\%$).

Several known genetic variations, for example, the FVL mutation and the prothrombin G20210A mutation, each appear to be the product of a unique mutational event in a founder haplotype,[9–13,20] validating the above assumption. Dominant gain-of-function mutations such as these are likely to be found using the candidate gene approach, and we consider it likely that many more such disease-causing mutations exist in the human gene pool.

Acknowledgements
The Leiden Thrombophilia study was supported by the Netherlands Heart Foundation (Grant No 89.063).

References

1 Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG Replication validity of genetic association studies *Nat Genet* 2001, **29** 306–309
2 Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K A comprehensive review of genetic association studies *Genet Med* 2002, **4** 45–61
3 Lane DA, Grant PJ Role of hemostatic gene polymorphisms in venous and arterial thrombotic disease *Blood* 2000, **95** 1517–1532
4 Cardon LR, Bell JI Association study designs for complex diseases *Nat Rev Genet* 2001, **2** 91–99
5 Risch N, Merikangas K The future of genetic studies of complex human diseases *Science* 1996, **273** 1516–1517
6 Rosendaal FR Venous thrombosis a multicausal disease *Lancet* 1999, **353** 1167–1173
7 Dahlback B, Carlsson M, Svensson PJ Familial thrombophilia due to a previously unrecognized mechanism characterized by poor anticoagulant response to activated protein C prediction of a cofactor to activated protein C *Proc Natl Acad Sci USA* 1993, **90** 1004–1008
8 Bertina RM, Koeleman BP, Koster T *et al* Mutation in blood coagulation factor V associated with resistance to activated protein C *Nature* 1994, **369** 64–67
9 Lunghi B, Iacoviello L, Gemmati D *et al* Detection of new polymorphic markers in the factor V gene association with factor V levels in plasma *Thromb Haemost* 1996, **75** 45–48
10 Cox MJ, Rees DC, Martinson JJ, Clegg JB Evidence for a single origin of factor V Leiden *Br J Haematol* 1996, **92** 1022–1025
11 Zivelin A, Griffin JH, Xu X *et al* A single genetic origin for a common Caucasian risk factor for venous thrombosis *Blood* 1997, **15** 397–402
12 Zoller B, Hillarp A, Dahlback B Activated protein C resistance caused by a common factor V mutation has a single origin *Thromb Res* 1997, **85** 237–243
13 Castoldi E, Lunghi B, Mingozzi F, Iannou P, Marchetti G, Bernardi F New coagulation factor V gene polymorphisms define a single and infrequent haplotype underlying the factor V Leiden mutation in Mediterranean populations and Indians *Thromb Haemost* 1997, **78** 1037–1041
14 Cripe DC, Moore KD, Kane WH Structure of the gene for human coagulation factor V *Biochemistry* 1992, **31** 3777–3785
15 Koster T, Rosendaal FR, de Ronde H, Briet E, Vandenbroucke JP, Bertina RM Venous thrombosis due to poor anticoagulant response to activated protein C Leiden Thrombophilia Study *Lancet* 1993, **342** 1503–1506
16 Schneider S, Roessli D, Excoffier L *A software for population genetics data analysis Version 2.000* Switzerland Genetics and Biometry Laboratory, University of Geneva, 2000
17 Excoffier L, Slatkin M Maximum likelihood estimation of molecular haplotype frequencies in a diploid population *Mol Biol Evol* 1995, **12** 921–927
18 Souto JC Genetic studies in complex disease the case prolinkage studies *J Thromb Haemost* 2003, **1** 1676–1678
19 Rosendaal FR Genetic studies in complex disease the case proassociation studies *J Thromb Haemost* 2003, **1** 1679–1680
20 Rosendaal FR, Doggen CJ, Zivelin A *et al* Geographic distribution of the 20210 G to A prothrombin variant *Thromb Haemost* 1998, **79** 706–708