Genetics of Venous Thrombosis: update in 2015

Pierre-Emmanuel Morange1,2,3, Pierre Suchon1,2,3, David-Alexandre Trégouët4,5,6
1 Laboratory of Haematology, La Timone Hospital, Marseille, France; 2 Institute National pour la Santé et la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1062, Nutrition Obesity and Risk of Thrombosis, Marseille, France; 3 Aix-Marseille University, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, Marseille, France; 4 INSERM UMR_S 1166, Paris, France; 5 Sorbonne Universités, Université Pierre et Marie Curie (UPMC Univ Paris 06), UMR_S 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, Paris, France; 6 Institute for Cardiometabolism and Nutrition (ICAN), Paris, France

Summary
Venous thrombosis (VT) is a common multifactorial disease with a genetic component that was first suspected nearly 60 years ago. In this review, we document the genetic determinants of the disease, and update recent findings delivered by the application of high-throughput genotyping and sequencing technologies. To date, 17 genes have been robustly demonstrated to harbour genetic variations associated with VT risk: ABO, F2, F5, F9, F11, FGG, GP6, KNG1, PROC, PROCR, PROS1, SERPINC1, SLC44A2, STXBPS, THBD, TSPAN15 and VWF. The common polymorphisms are estimated to account only for a modest part (~5%) of the VT heritability. Much remains to be done to fully disentangle the exact genetic (and epigenetic) architecture of the disease. A large suite of powerful tools and research strategies can be deployed on the large collections of patients that have already been assembled (and additional are ongoing).

Keywords
Venous thrombosis, single nucleotide polymorphisms, gene mutations

Introduction
Venous thrombosis (VT) is the pathologic expression of the development of a blood clot (i.e. thrombus), usually in the deep veins of the leg which can further obstruct venous circulation (deep-vein thrombosis, DVT) or subsequently embolise, travel to the lung and lead to pulmonary embolism (PE). VT is a multifactorial disease with a major contribution of genetic factors. The recognition that VT has a strong genetic risk component dates back to the 1960s (1, 2). Modern methodologies in genetic epidemiology estimate that the heritability of the disease is around 50% (3–5) and characterised by a sibling relative risk of ~2.50 (6).

Jordan and Nandorff (1) were the first, in 1956, to describe a familial clustering for VT. But this was only a decade later, in 1965, that the first genetic cause of VT was discovered: a partial antithrombin (AT) deficiency in a Norwegian family presenting familial VT aggregation (2). A few years later, in 1969, the ABO locus coding for the ABO blood group was mentioned for the time as a susceptibility gene for VT (7). But the impact of ABO on disease risk will be really acknowledged 40 years later (see thereafter).

In the early 1980s, deficiencies in the other two natural coagulation inhibitors, protein C (PC) (8) and protein S (PS) (9), were also identified as additional culprit origins of VT. These deficiencies are the consequence of private mutations in their structural genes (SERPINC1, PROC and PROS1 for AT, PC and PS deficiencies, respectively) that increase the risk of VT 10-fold in heterozygous carriers. These plasmatic defects hamper the two main regulatory pathways of coagulation: the inhibition of serine proteases by antithrombin and that of the non-enzymatic cofactors VIIIa and Va by activated protein C (APC) and its cofactor protein S.

Candidate gene approach
The 1990s were the advent of the candidate gene approach (10). Of course, genes coding molecules of the yin/yang coagulation/fibrinolysis pathways were natural candidates and subject to intensive research for identifying genetic polymorphisms associated with VT. Compared to the number of candidate genes assessed, unquestionable successes were relatively few. In 1994, the Q506 allele of the F5 R506Q mutation (rs6025), also known as the Factor V Leiden (FVL) mutation, was discovered as a rather common (~5% allele frequency in the general population of European descent origin) susceptibility allele (11, 12). This allele is associated with a ~3-fold increased risk and influences VT risk through resistance to APC. More recently, another non-synonymous F5 variant, K858R (rs4524), was also found to influence VT risk independently of the rs6025 variant (13, 14). Its risk allele has frequency ~0.75 and associates with an increase of the odds ratio (OR) for VT of ~1.20 (13–15). The FVL mutation’s discovery was quickly followed in 1996 by the identification of the F2 G20210A polymorphism (rs1799963) as another genetic risk factor for VT (16). This
polymorphism is located in the 3’UTR region of the prothrombin gene and the risk allele, rs1799963-A, increases plasma prothrombin concentrations (16). The risk allele is rare, ~2% of the general population, and is associated with a ~2.5 fold increased risk of developing VT. About 10 years then passed before the discovery of a novel genetic risk factor for VT. The PROCR gene coding for the endothelial protein C receptor (EPCR), a major regulator of PC activation (17), was an obvious candidate to investigate through the genetic prism. In 2004, Saposnik et al. (18) were the first to report the association of PROCR Ser219Gly (rs867186) with VT risk and the variability of plasma soluble EPCR. In a recent systematic review, the rs867186-G allele whose frequency is ~8% in the European ancestry is associated with an OR for VT of 1.22 (19). The last candidate gene that provided robust positive findings was the fibrinogen gamma gene (FGG) belonging to the fibrinolysis cascade. The T allele of the rs2066865 polymorphism, with frequency ~0.25, was found in 2005 to reduce gamma’ fibrinogen plasma levels and to increase VT risk by a factor of ~1.50 (20).

### Common and low-frequency susceptibility alleles for VT – contribution of high-throughput genotyping

The early 2000s witnessed the development of micro-array based high throughput genotyping technologies, allied to a deeper understanding of the pattern of human sequence variations. This offered the opportunity to perform association studies with hundreds of thousand of single nucleotide polymorphisms (SNPs) expected to cover the whole genome without any biological and functional a priori on the tested SNPs. Genome-wide association studies (GWAS) represented an important advance compared to ‘candidate gene’ studies in which sample size were generally smaller and the variants assayed were limited to a selected few, often on the basis of imperfect understanding of biological pathways and often yielding associations that are difficult to replicate (21). Genome-wide association studies (GWAS) and meta-analysis of GWAS have identified lots of loci that contribute to complex diseases in humans. The National Human Research Institute Catalog of published GWAS indicates that by 2013, 1751 curated publications reported 11, 912 SNPs associations with $P<1 \times 10^{-5}$ (22).

The first attempt to perform agnostic association scan on large number of SNPs in the field of VT genetics was the work of Bezemer et al. in 2008 (23). It concentrated on approximatively 20,000 SNPs, mainly non-synonymous of known genes with minor allele frequency (MAF) >5%, that were assessed for their association with VT through a multi-stage study including up to ~3,000 cases and 5,000 controls. Two novel susceptibility loci were identified using this approach, GP6 and CYP4V2. GP6 encodes glycoprotein V1 that has a major role in collagen signalling and the identified VT-associated SNP was the Pro219Ser (rs1613662) variant. The common form, Pro219 (rs1613662-A), with frequency ~0.80, was associated with a 1.15 fold increased risk. This association was further confirmed in four independent populations of French and American origins (24, 25). The association observed at the CYP4V2 locus was later refined and explained by the presence of linkage disequilibrium (LD) between the CYP4V2 gene and the nearby F11 gene coding for Factor XI (26). Even though the true functional variants at the F11 locus remain to be identified, strong evidence suggest that they influence VT risk through a modulation of FXI levels (26).

### Insights from HapMap2 based imputation

Following this first proof-of-concept study, several GWAS projects were launched to genotype between 300K and 600K SNPs and test their association with VT risk (24, 27–29). The number of SNPs was bolstered through statistical inference of variants not molecularly typed on the genotyping array but inferred through phasing and imputation guided by publicly available reference panels (30). These first GWAS imputed data were derived from the HapMap2 dataset containing 2,557,252 autosomal SNPs. This genetic resource was particularly efficient for inferring and assessing the impact on a phenotype of all untyped SNPs of the genome with MAF greater than 5% (31). No new robust VT-associated locus emerged from these studies that mainly confirmed the effects of the common alleles at F5, F11, FGG, GP6 and PROCR genes discussed above. However, these studies permitted to highlight and to revisit the major role of the ABO locus in VT susceptibility that was let aside since the pioneer work of Jick (7). These GWAS studies not only confirmed that the A1 and B ABO blood groups were at a 1.5-fold increased risk of VT, but also suggested that a novel common SNP located in intron 1, rs2519093 could associate with VT (OR ~1.60) independently of the major ABO blood groups (28, 29). However, others (32) have suggested that ABO rs2519093, in strong LD (r2 = 0.96) with ABO rs579459 , is tagging for ABO A1 blood group. Further studies are still mandatory to dissect whether the effect on VT of ABO rs2519093 is really independent of the ABO blood group effect.

The application of GWAS strategy to quantitative intermediate traits for the disease has also brought several new findings to the field. The general idea of such strategy is to apply GWAS principles to identify loci that associate with plasma levels of procoagulant proteins and then to look at the influence of the identified SNPs on the risk of VT. Genome wide association scans for SNPs associated with quantitative risk factors are usually performed in large samples of unrelated healthy individuals. But patients and family members can also be handled. Quantitative traits-associated SNPs that reached genome-wide significance (i.e. association $p$-value < 5 × $10^{-8}$) can then be considered as good candidates for association with the disease. Since few SNPs are generally tested at this step, there is no need to apply stringent genome-wide significance level but just correct for the number of tested SNPs, enhancing the power to detect disease-SNP associations compared to a complete scan over the genome. The first major successes were brought by the application of HapMap2 imputation based GWAS on von Willebrand Factor (VWF) and Factor VIII (FVIII) plasma levels (33). From the eight loci that were found influencing VWF and/or
FVIII plasma levels, two apart from ABO were further observed to robustly associate with the risk of VT, VWF and STXBP5. VWF is the structural gene for VWF and the Ala789 allele of its coding Thr789Ala variant (rs1063856-G) associates with both increased VWF levels and VT risk (OR ~ 1.15) (34). While the association of VWF with VT-susceptibility was not surprising due to the key role of VWF in the coagulation cascade, that observed with STXBP5 was rather unexpected. The STXBP5 encodes syntaxin binding protein 5 (STXBP5; also known as Tomosyn-1). The less common STXBP5 rs1039084-G allele of the coding Asn436Ser variant that was found to decrease VWF plasma levels was observed to be associated with an OR for VT of 0.90 (34). Of note, the same allele that is protective for VT was observed associated with an increased bleeding phenotype in women (35). Another interesting result is the recent observation that the STXBP5 rs1039084 also associates with t-PA levels (36) suggesting a role of STXBP5 in the regulation of the secretory status of endothelium. Two studies have recently addressed the role of STXBP5 on the secretion of granular contents from platelets and endothelial cells (37, 38). These studies demonstrate different functions for STXBP5 in these two cell types. While STXBP5 facilitates granule release from platelets, it inhibits secretion from the Weibel-Palade bodies of endothelial cells. Two other genes identified in the GWAS analysis for VWF, TC2N and STAB2, also exhibited suggestive evidence for association with VT risk (27, 39) but additional support is needed to definitively declare them as susceptibility genes for the disease.

Another successful example of this GWAS approach on quantitative traits that further translates to the disease is the identification of the KNG1 gene as a new VT disease locus. KNG1 gene encodes high-molecular-weight kininogen (HK), an obvious candidate for VT physiopathology. Among three SNPs that were reported to influence activated partial thromboplastin time (aPTT) in a GWAS conducted by Houlihan et al. (40), only one, the KNG1 Ile581Thr (rs710446), was further confirmed to associate with VT (41). The less frequent rs710446-C allele, with frequency ~0.45, is associated with decreased aPTT levels and increased risk (OR ~1.15) of the disease, as expected from the known traditional biology relating aPTT to coagulation and thrombosis.

In addition to aPTT, FVIII, VWF phenotypes aforementioned, other quantitative intermediate phenotypes to VT have been investigated through the GWAS strategy relying on HapMap2 imputed data. These include fibrinogen (42), D-dimers (43), PC (44), PS (45), AT (45) and PAI-1 (46) levels. None of the novel phenotype-SNP associations discovered in these investigations led so far to novel genetic associations with respect to VT risk.

**Insights from 1000 Genomes-based imputation**

The HapMap2 reference database has been a valuable resource for assessing the impact of SNPs with MAF > 5% in numerous GWAS projects. However, the tagging for common SNPs was far from being complete even in known functional regions (47–50). The resequencing of 500 genes had revealed that a relatively high proportion of common SNPs, ranging from 50% to 20% according to populations, could not be tagged by SNPs from HapMap (47). The same study estimated that only 30% of non-synonymous SNPs were in high LD with any HapMap SNPs. This, with the increasing recognition of the importance of SNPs with lower MAF but larger genetic effects (50) allied to the development of new high-throughput sequencing technologies led to the constitution of international projects, such as the 1000 genomes projects (http://www.1000genomes.org) aiming at characterising all human polymorphisms with very low MAF. To overcome the HapMap2 limitation, the 1000G database (51) has become a standard reference for imputation based association analyses and make possible the inference of low-frequency SNPs with MAF between 1 and 5%. VT was one of the first traits that benefited from the application of the 1000G based imputation strategy (52). The analysis of 1,961 VT cases and 2,338 controls well imputed for ~6.7 millions with MAF higher than 1% enabled to re-discover the strong association at the F2 rs1799963. This variant, with MAF ~2% in the control population, was missed by the previous GWAS because it could not be tagged by more frequent SNPs. This work nicely illustrated the relevance and the accuracy of the method to extend the spectrum of genetic variants that can be assessed in a GWAS context.

Following this work, all VT GWAS datasets that were previously imputed for HapMap2 (28, 29) were re-imputed, together with some additional newly set up GWAS studies, using the 1000G reference as part of the International Network on VENous Thrombosis (INVENT) consortium. Altogether, 12 GWAS studies for VT totalling 7,507 cases and 52,632 controls with imputed genotyped data at ~7 millions of SNPs were assembled and entered a meta-analysis of study-specific GWAS results with the aim to discover novel variants in novel or known genes that associate with VT risk (15). The INVENT analysis has produced a wealth of results whose in-depth investigation is still at its beginning. While no new low-frequency disease-associated allele was discovered, this meta-GWAS yielded new findings at known and unknown genes with respect to common alleles. First, it suggested that the association of PROCOr and F11 loci with VT is more complex than initially thought. A second PROCOr SNP, rs6088735, independently added to the rs867186 discussed above to influence VT risk. The rs867186 has been previously reported to influence PC levels (44). The association at the F11 locus was even more complex with three SNPs, rs1218625, rs2036914 and rs2289252, observed at this locus to additively influence VT risk. A preliminary haplotype analysis revealed that a unique at-risk haplotype can be derived from these three SNPs. Its frequency was ~0.40 and it associated with an increased OR of 1.40 (15, 27). As no low-frequency variant demonstrated suggestive statistical evidence at the F11 locus, it is unlikely that the observed common haplotype tags for an uncommon variant and suggests that the functionality of this haplotype should be envisaged. More novel, and then opening novel perspectives with respect to VT pathophysiology, are the robust associations detected at the TSPAN15 and SLC44A2 loci for which no link with VT has ever been reported. Two common polymorphisms, TSPAN15 rs78707713 with MAF ~0.10 and SLC44A2...
rs2288904 with MAF ~0.20, were found to associate, and replicate in independent samples of ~3,000 cases and 2,600 controls, with VT. In both cases, the common alleles were those that associated with an increase of VT risk, OR ~ 1.30 for TSPAN15 and OR ~1.20 for SLC44A2, respectively. No association with known haemostatic plasma markers was observed for these variants. The challenge is now to understand the functional consequences of these SNPs and to accurately elucidate the biological mechanism by which these genes and SNPs act on the risk of VT. The lead rs78707713 SNP at the TSPAN15 gene, that codes for tetraxspanin 15, a member of the tetraxspanin superfamily acting scaffolding proteins, anchoring multiple proteins to the cell membrane (53), is intronic. No evidence has been collected so far to suggest that it could functionally influence pathophysiological process of VT. Conversely, the SLC44A2 rs2288904 SNP is a non-synonymous (Arg154Gln) variant previously shown to associate with transfusion related acute lung injury (TRALI), a life-threatening complication of blood transfusion. This study identified the SLC44A2 could represent a new binding partner of VWF (55).

Similar to what has been done with the HapMap2 imputation strategy, one could expect that the application of the 1000G based imputation to quantitative risk factors for VT would lead to novel discoveries. So far, only one study reporting the results of a 1000G based GWAS on such a trait has been published (56) and it related to thrombin generation. This study identified the ORM1 gene which encodes orosomucoid as a new gene contributing to thrombin generation. However the lead ORM1 rs150611042 SNP did not demonstrate any evidence for association with the disease per se. Of note, in this study, the two main genetic factors influencing thrombin generation were the F2 rs1799963 and rs3136516 SNPs. While the association of the rs1799963 variant with VT is well established and was discussed above, it is worthy of note that rs3136516 demonstrates strong, but not genome-wide significance, statistical association (p =5.65 10^-6) with VT in the INVENT meta-analysis. Larger sample size would be needed to confirm the association observed with the second SNP, that is relatively common and associated with a modest increase VT risk (OR ~1.10). This association would be independent of that observed at rs1799963 and would suggest that two SNPs could exerce additive effects at the F2 locus to impact on VT risk. Several projects are ongoing for revisiting previous HapMap2 GWAS on quantitative traits using 1000G reference. The impact on VT risk of the SNPs that will emerge from these analyses would deserve further attention. For this purpose, the data and results assembled by the INVENT consortium could be of high value.

To sum up, 12 genes have been robustly demonstrated to harbour common and low-frequency susceptibility alleles for VT: ABO, F11, F2, F5, FGG, GP6, KNG1, PROCR, SLC44A2, STXB5, TSPAN15 and VWF (Table 1). Even though the following estimates need to be revised in view of the INVENT results, common polymorphisms are expected to account only for a modest part (~5 %) of the VT heritability (27) and of the sibling relative risk (~1.11) (29) mentioned in the introductory paragraph. They cannot therefore explain the whole genetic susceptibility to VT nor its familial aggregation. Family history still remains a strong risk factor for VT after adjusting for all common VT SNPs and coagulation defects (57, unpublished data). The increased risk for VT in relatives is incompletely explained by the presence of known VT defects, as the risk for VT in first degree relatives is increased even if patients do not have a detectable defect (58, 59). The question of

<table>
<thead>
<tr>
<th>Locus</th>
<th>SNP</th>
<th>Alleles</th>
<th>Frequency</th>
<th>OR</th>
<th>Associated phenotype</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABO</td>
<td>[0, A2] vs [A1, B]</td>
<td>G/A</td>
<td>0.30</td>
<td>1.50</td>
<td>↑ WF, ↓ VIII</td>
<td>28, 29</td>
</tr>
<tr>
<td>F2</td>
<td>rs1799963</td>
<td>G/A</td>
<td>0.24</td>
<td>1.68</td>
<td>Unknown</td>
<td>28</td>
</tr>
<tr>
<td>F5</td>
<td>rs6025</td>
<td>G/A</td>
<td>0.02</td>
<td>2.50</td>
<td>↑ FII</td>
<td>16</td>
</tr>
<tr>
<td>F11</td>
<td>rs179963</td>
<td>G/A</td>
<td>0.05</td>
<td>3.00</td>
<td>Resistance to activated protein C</td>
<td>12</td>
</tr>
<tr>
<td>FGG</td>
<td>rs2036914</td>
<td>C/T</td>
<td>0.52</td>
<td>1.35</td>
<td>↑ FXI</td>
<td>26</td>
</tr>
<tr>
<td>GP6</td>
<td>rs1613662</td>
<td>A/G</td>
<td>0.82</td>
<td>1.15</td>
<td>↑ platelet activation and aggregation</td>
<td>23</td>
</tr>
<tr>
<td>KNG1</td>
<td>rs710446</td>
<td>A/G</td>
<td>0.45</td>
<td>1.20</td>
<td>↑ aPTT</td>
<td>41</td>
</tr>
<tr>
<td>PROCR</td>
<td>rs867186</td>
<td>A/G</td>
<td>0.07</td>
<td>1.22</td>
<td>↑ sEPCR, ↑ PC</td>
<td>18</td>
</tr>
<tr>
<td>SLC44A2</td>
<td>rs2288904</td>
<td>T/C</td>
<td>0.07</td>
<td>1.19</td>
<td>Still unknown</td>
<td>15</td>
</tr>
<tr>
<td>STXB5</td>
<td>rs1039084</td>
<td>A/G</td>
<td>0.46</td>
<td>1.11</td>
<td>↑ WF</td>
<td>34</td>
</tr>
<tr>
<td>TSPAN15</td>
<td>rs78707713</td>
<td>G/A</td>
<td>0.88</td>
<td>1.28</td>
<td>Still unknown</td>
<td>15</td>
</tr>
<tr>
<td>VWF</td>
<td>rs1063856</td>
<td>A/G</td>
<td>0.37</td>
<td>1.15</td>
<td>↑ WF</td>
<td>34</td>
</tr>
</tbody>
</table>

*Underlined are the at risk alleles. * Estimates of the risk allele frequency observed in References population. * Estimates of the odds ratio (OR) associated with the risk allele. * Phenotype associated with the at risk allele. * According to References numbering.

Table 1: VT-associated variants identified by GWAS conducted in individuals of Caucasian origin.
missing or hidden heritability (31, 60) that underlies complex diseases also holds for VT. An army of sources could contribute to this missing heritability including extremely rare variants (61).

**Rare variants for VT**

The 1000G based imputation strategy is often proposed as an efficient tool for discovering novel disease-associated variants with MAF as low as 1%. However, despite being based on the largest GWAS resources on VT, the INVENT meta-analysis failed to detect such associations. The 1000G INVENT analysis did discover two newly VT-associated SNPs, but the discovery of these two SNPs that were common was likely due to the increase in sample size compared to previous GWAS on VT rather than to the use of a more comprehensive panel of genetic variations used to impute non-typed variants. The INVENT meta-analysis had a 80% power to detect, at genome-wide significance level, the effect of a variant with MAF 0.01 as soon as it was associated with an increased risk of ~1.80. The power fell down to 25% for an OR of ~1.65. Without a dramatic increase in the number of VT patients included in GWAS, it is unlikely that low-frequency variants associated with VT risk could be discovered using a standard GWAS framework. Other strategies should then be envisaged.

**Genotyping arrays for low-frequency coding variants**

Novel original discoveries in the genetics of human complex traits are expected from the application to large epidemiological studies of DNA genotyping arrays dedicated to known coding variants including low-frequency ones with 0.5% < MAF< 5%, that may not be well imputed by statistical inference. The Illumina Exome chip is a popular array that is currently used by many projects (62). It has been developed to include on a dedicated array all rare variations identified through the sequencing of the coding part of the genome (i.e. exome) of ~12,000 individuals representing diverse populations and a range of common conditions such as asthma, stroke, myocardial infarction, extreme LDL, extreme blood pressure. The general idea under the use of such array is that common variations might be less important than rare variations because selective pressure has had more time to eliminate deleterious alleles and that coding variants can more easily pinpoint the causal gene in regions enriched for several loci. The analysis of such genotype data requires dedicated statistical tools that jointly test all variants of the same gene for association with the trait of interest rather than testing them individually as it is done in standard GWAS. This grouping is made necessary by the rarity of individuals carrying the low-frequency alleles. Because variants of the same gene are tested together, these methodologies are generally referred to as gene-based test approaches (63). The deployment of the exome array has already produced a few successes with the identification of novel variants at known and unknown loci associated with complex traits, such as psoriasis (64), insulin-related traits (65), glucose (66), haematological traits (67), lipid levels (68, 69), myocardial infarction (MI) (68, 69) and non-alcohol fatty liver disease (NAFLD) (70). However, not all discovered variants were with low frequency. For example, the same coding TM6FS2 rs58542926 (Gly167Lys) that associated with lipids, MI and NAFLD (68–70) was in fact a rather frequent SNP (MAF ~9%) whose associations with studied phenotypes were missed by previous GWAS analyses. As part of the French GENMED project (http://www.genmed.fr/), 2,700 VT patients and 3,000 healthy individuals have recently been typed with this array and similar international VT projects are ongoing. One can reasonably anticipate that the application of this array to VT will lead in the next coming years to new discoveries in the field.

**Next-generation sequencing for extremely rare variants**

By design, genotyping arrays are not able to assess unknown genetic variations. Sequencing is then the method of choice for identifying unknown, or extremely rare (<5%), mutations causing diseases. The next-generation sequencing (NGS) technologies have radically revolutionised the research landscape in the field of human disease genetics and offer a powerful tool to characterise the spectrum of very rare disease-causing variants. NGS of DNA samples can be applied to a set of candidate genes (target gene resequencing), to the whole coding part of the genome (the so-called exome) or in a complete agnostic manner by covering the whole genome. Only the latter two strategies afford the discovery of unsuspected genes and novel biological pathways related to the disease under study (71, 72).

**Targeted gene sequencing**

Four candidate genes for VT have been investigated with success through targeted gene re-sequencing. In a thrombophilic patient with factor IX (FIX) activity eight times higher than the normal physiological levels, a gain-of-function mutation (R338L) was discovered in the F9 gene (73). This mutation, named as the FIX Padua mutation in reference to the Italian city where the patient was hospitalised, is anticipated to be a private mutation as it has not been reported in the Exome Aggregation Consortium (ExAC) database (http://exac.broadinstitute.org) cataloguing all variants detected in the coding sequences of more than 60,000 individuals. Worthy of note, Finn et al used the sequence of the abnormally potent FIX for construction of improved gene therapy vectors to treat haemophilia B (74).

Two different extremely rare mutations at the same position of the F2 gene, R596L and R596Q, have recently been discovered causing inherited thrombophilia due to resistance to AT. The former, named the prothrombin Yukuashi mutation, was detected in a Japanese family (75) and the latter, known as the prothrombin Belgrade mutation, in two Serbian families (76).

By re-sequencing the entire genomic sequence of the THBD gene encoding thrombomodulin in 60 VT patients and 60 controls
from China, the c-151G>T variant in the 5'UTR region of the gene was discovered as a low-frequency SNP affecting VT risk (77). Its risk-allele frequency was estimated to be ~1% in the Chinese population and associated with a 2.8-fold increased risk of VT. It remains to be elucidated whether this genetic risk factor is specific to the Chinese population and/or what could be its prevalence in other populations.

The last example of the efficiency of the candidate gene resequencing strategy is the recent discovery of the FV Nara mutation (F5 W1920R) in a 13-year-old Japanese boy who developed recurrent VT during oral anticoagulant treatment (78). The patient has reduced FV levels and pronounced APC resistance due to homozygosity at the detected missense mutation. In line with the observations made in the patient, recombinant FV Nara showed reduced expression in conditioned media (~50% of wild-type FV) and conferred APC resistance to reconstituted FV-deficient plasma. Because the FV Nara mutation is located in the light chain of FV(a), far from the known APC-cleavage sites, the molecular mechanism by which it causes APC resistance remains unclear (79).

Altogether, these examples strongly advocate for extensive resequencing of the validated VT genes aforementioned and clearly demonstrate that both rare and common variants at a given gene can act to influence VT risk.

**Whole exome sequencing (WES)**

In presence of familial aggregation of VT that cannot be explained by a culprit mutation in known VT associated genes, the WES approach seems rather intuitive given the assumption that the causal mutation(s) lie(s) in (a) coding region(s). Numerous WES projects have demonstrated the power of the approach to identify very rare variants responsible for familial forms of human diseases. For instance, a major step forward was made during the last four years using this approach that resulted in the discovery of novel genes responsible for inherited platelet bleeding disorders (80–84). So far, its successful application to VT has never been reported. As part of the French INNOVTE network (http://www.fcrin.org/en/support-tools/innovte-thrombosis), a national initiative has recently been launched to make an inventory of all families with at least three family members who had unprovoked VT across three generations with the aim of performing WES in families excluded for mutations in known VT genes. The exome sequencing of several families has already started.

The WES strategy can of course be applied to unrelated idiopathic patients without any familial information. However, the outstanding wealth of rare single-nucleotide variants in individual human genomes calls for stringent criteria and large sample size to establish association, and hence a causal relationship, between rare single-nucleotide variants and a disease. So far, the WES strategy applied to large sample of independent disease patients have produced little success in identifying novel disease genes. It has rather led to the discovery of new rare mutations in genes already known to be involved in the pathophysiology of the disease under study (e.g. 85, 86).

**Whole genome sequencing (WGS)**

A large number of instances have been accumulated to highlight the role of non-coding variations in the molecular aetiology of complex diseases. WGS appears to be the method of choice for identifying extremely rare non-coding variants that cannot be properly addressed by WES. The WGS strategy has also the advantages over WES to be more sensitive to detect variations in coding regions where target capture is less efficient (87) and to detect a larger spectrum of genetic variants including structural variants (88). Both familial cases and unrelated disease patients can be assessed by WGS. But again, due to the complexity induced by the high number of detected variations, a family-based strategy relying on the information about the segregation of variants within affected and unaffected relatives substantially enhanced the power to detect the culprit variants (89). So far, WGS has been particularly efficient to detect causal variations in Mendelian disorders (e.g. 90) and few complex diseases such autism (91, 92) and some cancers (93, 94). At its early stage, WGS was preferentially advocated for families or patients for which WES had remained unfruitful. However, as the cost of the WGS dramatically decreased up to 1,300$ (for a mean coverage of 30X), and as WGS presents several merits over WES, the utility of WES (whose cost is ~450$ for the same coverage) starts to be questioned.

**Perspectives**

If we add to the list of genes with common and low-frequency susceptibility alleles for VT (see above) those known to carry extremely rare mutations responsible for VT (F9, PROC, PROS1, SERPINC1, and THBD), the total number of established VT associated genes raises to 17. Figure 1 displays a schematic network of the physical interactions between these 17 genes as produced by the GeneMANia (95) tool. GeneMANIA interrogates several protein-protein interactions databases to identify whether input proteins have been reported to be functionally related through direct physical interaction. Similar networks can also be derived using other functional information such as co-expression of genes/proteins. Figure 1 well illustrates how most of these genes are strongly interconnected. These connections can be direct or involve others partners among which one can easily notice some genes that have previously been suspected to associate with VT but without achieving sufficient level of evidence. Examples include ADAMTS13 (96), CBP2 (97) and C4BPA (98). It has been proposed that SNP x SNP interactions could contribute to the missing heritability of complex diseases (99). Rather than performing a SNP x SNP interaction scan over the whole genome which would require tremendous sample size to achieve statistical power while controlling for the multiplicity of the tests (100), the information of known physical interactions between genes as in Figure 1 could serve as an a priori to test for the presence of interaction only between predefined subsets of polymorphisms mapping to genes known to physically interact with each other.
Not all the missing heritability is hiding in DNA genetic variations and epigenetic mechanisms are also proposed as a possible source that could contribute to it (101). Although the role of epigenetic mechanisms in some rare developmental syndromes and in cancer is well established, systematic examination of their contribution to common non malignant disease is only beginning. For this purpose, specific high throughput technologies are available to quantify non coding RNAs expression and methylation profiles from cells, tissues and blood. A primary concern in this new era of epigenetic epidemiology is the tissue-specific nature of the epigenome as in large well-phenotype sample cohorts, DNA from whole blood and plasma or citrate are often the only source of biological material available. While assessment of DNA methylation in whole blood has been shown able to identify robust and biologically relevant epigenetic variation related to cardiometabolic quantitative risk factors (102, 103), it remains to be determined whether it can be relevant for specifically VT-related traits. A first attempt was undertaken in this spirit by searching for blood DNA methylation marks associated with thrombin generation markers using the dedicated Illumina HumanMethylation450 array in a sample of 350 individuals (104). No conclusive results were obtained from this pioneer work. This would suggest that DNA methylation measured in whole blood DNA may not be a good mirror of what happens in specific cell type more relevant to VT hemostasis regulation and/or that methylation marks in peripheral blood cells have tiny effects on plasma levels of thrombin generation that would require much larger sample size in order to be statistically detected.

High-throughput technologies also offered the opportunity to simultaneously quantify the expression levels of several thousand genes using microarrays. Comparing expression profiles between samples of disease patients and controls enables the identification of genes and/or set of genes simultaneously up or downregulated in patients, then pointing out functional modules that could be

Figure 1: Physical interaction map derived from established VT associated genes. Map produced by the GeneMANIA online tool (95) with the 17 established VT-associated genes (in black circles) as input. Linked circles represent pairs of genes that have been reported to physically interact with each other in the various protein-protein interaction sources implemented in GeneMANIA. The very recently suggested SLC44A2 – VWF physical interaction (discussed in the text) is not yet indicated in this map. Grey genes are additional genes identified by GeneMANIA to physically interact with VT-associated genes. The stronger evidence for interaction, the thicker the link. This map is not exhaustive and must be viewed as a simplified illustration of the main physical connections linking together, as much as possible, the 17 VT genes. The lack of connections at few genes does not preclude their co-regulation with other shown genes. For examples, ABO and STXBPS have been shown to influence vWF plasma levels (see text).
implicated in the disease aetiology. A recent application to VT has been proposed to identify in whole blood expression signatures differentiating different clinical phenotypes of VT (105). The profiles obtained distinguish patients with recurrent, unprovoked VT from healthy controls and patients with provoked VT only.

The annual incidence of VT is ~1 % (106) but it substantially varies across populations (107), and the prevalence of known genetic risk factors of VT can differ according to ethnicity such as FVL and F2 G20210A mutation that are very rare or absent among nonwhites. Until now all GWAS have been performed in individuals of European ancestry (whites) where it is known that LD can extend over large genomic regions compared to other populations, including Africans. Thoroughly designed epidemiological research (including GWAS) should be expanded to non-white populations with higher genetic diversity to facilitate the identification of additional genetic risk factors for VT.

Fifty percent of VT events occur without any precipitating conditions such as cancer, surgery, trauma, immobilisation, pregnancy, long haul flights, and are designed as unprovoked VT. The annual rate of recurrence is 10% in unprovoked patients stopping the anti coagulant treatment and ~10% of VT patients will die from fatal PE. Of survivors, 25–50 % will have lasting debilitating health problems such as post-thrombotic syndrome, severely hampering mobility and quality of life. Up to now, much efforts have been made to characterise the genetics of prevalent VT without focused attention on the distinction between DVT and PE. There are unmet needs for identifying (epi)-genetic factors predisposing more to PE than to DVT and those predicting the recurrence of the event (108). A few initiatives have recently been conducted to address the later issue. van Hylckama Vlieg et al. assessed the predictive value of multiple SNPs testing for the risk of recurrence using VT-associated polymorphisms at the ABO, F11, F2, F5 and FGG loci (109). Using a 5–SNPs genetic risk score (GRS), they were able to stratify patients on their recurrence risk: the six-year cumulative incidence of recurrence was high for individuals with ≥5 (20.3 %) and low for individuals with ≤1 risk allele (9.4 %) (110). The predictive ability was similar in individuals with first provoked or unprovoked event. The clinical utility of this GRS must be now assessed in a prospective randomised study where the duration of anticoagulant treatment would be individually adapted according to each patient’s score.

To conclude, VT is the third leading cause of cardiovascular mortality after coronary artery disease and stroke and is responsible for a large public health burden (105). Novel major genetic discoveries have recently been obtained using high-throughput genotyping and sequencing technologies but, as discussed in this review, much remains to be done to fully disentangle the exact genetic (and epigenetic) architecture of the disease. A large suite of powerful tools and research strategies can be deployed on the large collections of patients that have already been assembled (and additional are ongoing).

Acknowledgements

Pierre-Emmanuel Morange and David-Alexandre Trégouët’s work forms part of the research programs on venous thromboembolism genetics which are supported and funded by the the GenMed LABEX (ANR-10-LABX-0013) and the French INvestigation Network On Venous Thrombo-Embolism (INNOVTE).

Conflicts of interest
None declared.

References*


* Some earlier investigations on genetic variants of candidate genes are not cited due to space contraints.

© Schattauer 2015

Thrombosis and Haemostasis 114.5/2015

Genetic aspects of thrombotic disease


