TÉCNICO LISBOA

A new value proposition for a genetic test of hereditary thrombophilia

Beatriz Melo Silveira

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Prof. Ana Teresa Correia de Freitas
Dr. José Miguel Ranhada Vellez Caldas

Examination Committee

Chairperson: Prof. João Pedro Estrela Rodrigues Conde
Supervisor: Dr. José Miguel Ranhada Vellez Caldas
Member of the Committee: Susana de Almeida Mendes Vinga Martins

November 2018

Acknowledgments

First of all, I would like to thank my supervisors, Professor Ana Teresa Freitas, for giving me the opportunity to develop my thesis at Heartgenetics, Doctor José Caldas, for his constant availability and willingness to help throughout the last six months, and Doctor João Carriço, for accepting to be part of my supervising committee . I would also like to thank the rest of the team at Heartgenetics for being very welcoming, and particularly Hugo Loureiro, for being always available to clarify any doubts I had and for kindly replying to the ridiculous amount of emails sent by me.

Additionally, I would like to thank my family for the endless support and patience, throughout the last five years, and my friends for the laughs, dinners, and trips that kept us all sane.

Lastly, I would like to thank my boyfriend, João, for the endless support and for lovingly showing me that no matter what problems you may have, most of the times, it's really not that bad.

Resumo

A predisposição genética para a formação de coágulos sanguíneos – trombofilia hereditária – acarreta consequências potencialmente fatais, sendo um dos fatores de risco do tromboembolismo venoso (TEV). O TEV caracteriza-se pela formação patológica de coágulos nas veias que podem eventualmente entrar na circulação sanguínea e obstruir outro vaso do corpo - embolia. As hipóteses de sobrevivência após um evento de TEV diminuem com o tempo, sendo apenas 41.5% oito anos após a primeira embolia pulmonar [1]. A taxa de incidência anual de TEV em indivíduos com ascendência europeia varia entre 104 e 183 por cada 100,000 pessoas-ano [1].

Considerando o impacto negativo do TEV na saúde, uma ferramenta de diagnóstico que estime a propensão genética à coagulação anormal torna-se bastante útil. Esta dissertação teve como objetivo o desenvolvimento de uma nova proposta de valor para o TromboGene Kit (TRB kit), um teste genético de trombofilia hereditária desenvolvido pela Heartgenetics, Genetics Biotechnology, S.A.. Primeiramente efetuou-se uma revisão da literatura sobre a trombofilia hereditária. Seguiu-se uma análise estatística envolvendo dois coortes de pacientes com TEV e abortos recorrentes, também uma manifestação clínica de interesse, visando averiguar a associação entre o painel genético do teste e essas manifestações. Os passos anteriores conduziram à proposta de alterações ao painel atual, cujo impacto no modelo de risco do produto foi posteriormente analisado.

Os resultados deste trabalho representam importantes contribuições para o contínuo desenvolvimento de uma proposta de valor sólida para o TRB kit.

Palavras-chave: Trombofilia Hereditária, Tromboembolismo Venoso, Aborto Recorrente, Testes Genéticos, Modelo de Previsão de Risco

Abstract

Having a genetic predisposition to the formation of blood clots – hereditary thrombophilia – comes with potentially fatal consequences, being one of the risk factors for developing venous thromboembolism (VTE). VTE is characterized by the pathological formation of blood clots in veins, which can possibly enter blood circulation and obstruct another vessel in the body – embolism. The chances of survival after a VTE event tend to decrease over time, being 41.5% eight years after the first pulmonary embolism (PE) [1]. The annual incidence rate of VTE among people of European ancestry varies between 104 and 183 per 100,000 person-years [1].

Given the negative impact of VTE in health, a diagnostic tool which estimates the genetic predisposition to abnormal coagulation is of great value. The goal of this dissertation was to develop a new value proposition for the TromboGene Kit (TRB kit), a genetic test of hereditary thrombophilia developed by Heartgenetics, Genetics Biotechnology, S.A.. First, a literature review about hereditary thrombophilia was performed. A statistical analysis followed involving two cohorts of VTE patients and of recurrent miscarriage (RM) patients, a clinical manifestation also of interest, aiming to investigate the association between the test's genetic panel and those manifestations. The previous steps led to the proposal of alterations to the TRB kit's current panel, whose impact was then assessed in the product's future risk prediction model.

This work paved the way for the development of a new solid value proposition for the TRB kit.

Keywords: Hereditary Thrombofilia, Venous Thromboembolism, Recurrent Miscarriage, Genetic Tests, Risk Prediction Model

Contents

List of Tables

List of Figures

Acronyms

VTE Venous thromboembolism

VT Venous thrombosis

PE Pulmonary embolism

TRB TromboGene

RM Recurrent miscarriage

DVT Deep vein thrombosis

SNP Single nucleotide polymorphism

MAF Minor allele frequency

LD Linkage Disequilibrium

HWE Hardy-Weinberg equilibrium

GWAS Genome-wide association study

PCA Principal component analysis

OR Odds ratio

CI Confidence interval

RR Relative risk

AT Antithrombin

PC Protein C

PS Protein S

PAI Plasminogen activator inhibitor

TFPI Tissue factor pathway inhibitor

FVL Factor V Leiden

APC Activated protein C

CVD Cardiovascular disease

IS Ischemic stroke

MI Myocardial infarction

PPO Poor pregnancy outcome

ATE Arterial thromboembolism

CHD Coronary heart disease

CAD Coronary artery disease

ACS Acute coronary syndrome

MTHFR Methylenetetrahydrofolate reductase

CVT Cerebral venous thrombosis

BIC Bayesian information criterion

ROC Receiver operating characteristics

AUC Area under the curve

vWF von Willebrand factor

Chapter 1

Introduction

Hemostasis is a vital set of physiological mechanisms that, in normal conditions, stops bleeding at the site of an injury while simultaneously keeping normal blood flow elsewhere in the circulation [2].

In the absence of damage, the internal layer of blood vessels - endothelium - maintains an anticoagulant surface that allows blood to remain in a fluid state. However, upon damage to a vessel, subendothelial components are exposed to the blood, which then activate two main hemostatic processes: primary and secondary hemostasis. Both processes take place simultaneously and lead to the formation of a blood clot, ultimately stopping the bleeding [2, 3].

Primary hemostasis is characterized by the aggregation of a type of blood cells called platelets. Platelets adhere to the site of injury and to each other, thus creating a platelet plug. Secondary hemostasis, on the other hand, is characterized by the deposition of an insoluble protein, namely fibrin, a product of a set of chain reactions known as the coagulation cascade. The deposition of fibrin in and around the platelet plug creates a mesh that strengthens and stabilizes the clot [2, 3].

Besides the processes that lead to the formation of blood clots, also the mechanism by which they are dissolved - fibrinolysis - is very important to maintain hemostasis. When the delicate balance between these mechanisms is disrupted, abnormal bleeding can occur or, on the contrary, abnormal formation of blood clots, also known as thrombi [2, 3].

Thrombosis is characterized by pathological thrombus formation in blood vessels. When veins are affected, it is called venous thromboembolism (VTE). Thus, besides thrombus formation, an embolism can occur if the thrombus dislodges from its original site and enters the circulation, potentially blocking a vessel anywhere in the circulatory system [2, 3]. There are a number of risk factors for the develop-

ment of VTE. These can be acquired, including major surgery or trauma, malignancy, the use of oral contraceptives, pregnancy, or extended immobility, with varying degrees of severity, among many other factors. Furthermore, an inherent predisposition to forming blood clots - thrombophilia - is also a risk factor for VTE [4]. Thrombophilia can be inherited or acquired. Throughout this work, the focus will be on hereditary thrombophilia and, as explained later on, a variety of inherited genetic factors determines this propensity to thrombotic events, hence the importance of studying the human genome.

Genetics has allowed the investigation of the underpinnings of human diseases and paved the way for personalized medicine which, despite still not being generally part of current medical practice, holds great promise for the future. The ability to forecast the odds of a given disease, based on an individual's genetic profile, provides the chance of incorporating appropriate prophylactic measures that wouldn't be possible otherwise. Plus, the same genetic profile may also provide some valuable insights on how an individual would respond to a given drug, therefore allowing the design of a personalized course of treatment. Thus, one may say that genetics has brought a new light to how we view health care.

1.1 Motivation

VTE is a multifactorial disorder, resulting from the complex interaction between genetic and environmental factors. Moreover, family-based studies have suggested that over 60% of the variation in the susceptibility to thrombosis is explained by inherited factors [5].

The epidemiology of VTE has been extensively studied in European-origin populations. It is estimated that the average annual incidence rate of overall VTE among people of European ancestry varies between 104 and 183 per 100,000 person-years [1]. Despite the lack of epidemiological data from non-Western countries, there are different sources of evidence that indicate that VTE has its highest incidence in individuals of African ancestry, followed by Caucasians, Hispanics, and being the lowest in Asians [1, 5].

The risk of VTE increases with age and the condition is rare prior to adolescence. Incidence rates appear to be higher in women during childbearing years, though they are generally higher in men after the age of 45 years [1]. Overall, the age-adjusted incidence rate is slightly higher for men than for women, the male:female ratio being 1.2:1 [1].

Survival after a VTE event depends on the nature of the event itself. If a thrombus enters the circulation and ends up blocking a vessel in the lungs - pulmonary embolism (PE) - the survival rate is much worse compared with that of deep vein thrombosis (DVT) alone, meaning a thrombus forming in a deep vein of the leg, for instance. Immediately after a DVT event, the chance of survival is 97%. For PE, it is about 76.5% [1]. The chances of survival tend to decrease as time passes, being 65.2% and 41.5%, respectively, eight years after the first event [1]. The risk of early death in patients with PE is eighteen-fold

higher compared with patients who have just suffered a DVT event. Plus, for almost one-quarter of PE patients, the initial and only clinical presentation of the condition is sudden death [1]. On the other hand, a thrombus entering the circulation also includes, for instance, the possibility of eventually blocking a vessel irrigating the brain, consequently compromising blood circulation and oxygen delivery to this vital organ, an event known as ischemic stroke (IS), also with potentially fatal consequences.

Bearing in mind the negative impact of VTE in human health, having ways of knowing whether one is prone to developing the disease is certainly valuable. Since genetics plays a role in VTE's etiology, testing for the presence of specific genetic markers that are indicative of a genetic predisposition to forming blood clots - hereditary thrombophilia - can make a difference regarding prophylaxis. Heartgenetics, Genetics Biotechnology, S.A., is a company that has developed genetic testing kits in the cardiovascular, pharmacogenetics, and wellness fields. One of the company's products in the cardiovascular field is the TromboGene Kit (TRB kit), which performs a genetic study of hereditary thrombophilia. It is currently a qualitative *in vitro* diagnostic (IVD) device that, by genotyping a set of genetic variants, provides an estimate of the genetically determined propensity to thrombotic events. Based on the results of the test, the product provides a detailed description regarding the susceptibility to different thrombotic events, as well as the condition's impact on pregnancy, which will also be explored in this dissertation.

Diagnostic devices such as the TRB kit provide a valuable contribution to the management of conditions such as VTE, hence the importance of continuously improving this sort of technology, so that patients are informed much earlier about their disease susceptibility and can incorporate changes to their lifestyles in accordance.

1.2 Objectives and thesis outline

The goal of this dissertation was to develop a new value proposition for the TRB kit, which essentially translated into proposing a number of alterations to the current version of the test, aiming to improve its risk-predicting ability. This work was integrated in the company's initiative to add a quantitative risk measure - a risk score - to a currently qualitative diagnostic device.

This dissertation is divided into five chapters: chapter two provides a brief review about relevant concepts of genetics and bioinformatics, as well as a review about hereditary thrombophilia, including important topics such as its genetic basis and clinical manifestations; chapter three provides a detailed description of the methodology followed, whereas chapter four features all of the results obtained following that methodology; the discussion of the results obtained, resulting conclusions, and future work are the focus of chapter five.

Chapter 2

Background

2.1 Genetics and bioinformatics background

The focus of this section is to briefly review and introduce basic concepts of genetics and bioinformatics, some of which will be referred to later on in other chapters.

2.1.1 Genetics background

Deoxyribonucleic acid (DNA) is the organic material, capable of replication, that stores the genetic information in nearly every organism. This information is stored as a code consisting of four chemical bases, namely adenine (A), guanine (G), cytosine (C), and thymine (T). Each base is attached to a sugar molecule and a phosphate molecule, thus creating a nucleotide. These bases pair up with each other – A with T and C with G – to form base pairs. Moreover, the pairing of the nucleotides gives rise to the widely recognized ladder-resembling structure of the DNA molecule, in which nucleotides are arranged in two strands coiled clockwise around each other, hence forming a double helix [6, 7].

A sequence of DNA forms a gene, the basic unit of heredity that encodes for a functional molecule, a protein. It is estimated that humans have between 20,000 and 25,000 genes, with variable sizes that can go from just a few hundred bases to more than two million bases [6]. A gene generally comprises protein-encoding segments called exons alternating with noncoding segments called introns. A human gene features other additional sequences such as two flanking regions, namely the five prime (5') and three prime (3') flanking regions. DNA containing the sequences of different genes is organized in chromosomes, which are thread-like structures present in the nucleus of each cell. The DNA molecule in a chromosome is tightly coiled around supporting proteins, namely histones. Human cells normally

contain 23 pairs of chromosomes, yielding a total of 46 chromosomes. Twenty two of these pairs are autosomal chromosomes, given that they are the same regardless of gender. The remaining pair is composed by the sex chromosomes, therefore differing between males and females - while females have two copies of the X chromosome, males have one X and one Y chromosome [6, 7].

Every individual inherits two copies of each gene, one from each parent. Despite the majority of genes being the same in all people, each individual still displays a set of unique physical features. This is due to a small number of genes, representing less than 1% of the total number, that harbor small differences between individuals [6]. These small differences in the DNA sequence of a gene lead to different forms called alleles. Thus, for a given trait, the genotype refers to the pair of alleles inherited, whereas the phenotype refers to the trait itself, *i.e.*, the resulting observable or measurable characteristic. Further, an individual with two copies of the same allele is said to be homozygous, whereas an heterozygous individual has two different alleles. A phenotype can be dominant, when the trait is expressed in the heterozygous state, or recessive, when the trait is expressed only in the homozygous state. Plus, a co-dominant phenotype results from the expression of both alleles in the heterozygous state [6, 8].

As already mentioned, genetic differences between individuals partially explain the diversity of traits in humans, which is also true when considering disease susceptibility, for example. Genetic variants, such as mutations and Single Nucleotide Polymorphisms (SNPs) have contributed to intra- and interpopulation variation over time [7].

A genetic mutation is a permanent alteration in a DNA sequence that can range in size - it can affect just a single base pair or a large segment of a chromosome comprising multiple genes. Mutations can be inherited - germline mutations - therefore being present in every cell of the individual, or acquired during life - somatic mutations - which affects only certain cells. Somatic mutations usually arise from unrepaired DNA damage or replication errors, for example, and do not necessarily cause phenotypic changes. There are different types of DNA mutations. Point mutations, for instance, imply that only a single nucleotide is added, deleted, or substituted. In the case of a single substitution, a missense mutation occurs when this event causes the change of one amino acid, the building block of proteins, in the final protein synthesized. On the other hand, whole chromosomal regions can be flipped, deleted, duplicated, or rearranged between chromosomes belonging to different pairs (nonhomologous), a process known as translocation [7, 9].

SNPs are single base pair variations in a DNA sequence resulting, for instance, from unrepaired replication errors. Despite potentially happening in any location throughout the genome, SNPs are frequently found in flanking regions of protein-coding genes, these regions being recognized as critical for the regulation of gene/protein expression. The concepts of DNA mutation and SNP have become, to some extent, blurred over time, both terms being sometimes used interchangeably for the same event. Despite both constituting DNA variants, i.e., differences in comparison to a reference, mutations are detectable in less

than 1% of the population, while SNPs' prevalence is considered to be greater than 1% [7, 9]. In other words, one can use the minor allele frequency (MAF), i.e., the frequency of the less common allele at a variable site, as the distinguishing factor between rare genetic variants (or mutations) - MAF less than 1% - and common genetic variants (or SNPs) - MAF greater than 1%. Further, given that it is estimated that the human genome contains around eleven million SNPs, seven million of which with a MAF greater than 5%, this class of genetic variation is considered the most prevalent among individuals [8].

2.1.1.1 Population genetics

Population genetics aims to study the genetic variation at the level of a whole population. In this context, SNPs play an important role, as there are a number of reasons that explain why SNPs are preferable to other types of genetic variation when investigating a population's genetic predisposition to certain traits or diseases. The fact that SNPs are the most common and can be found throughout the genome is one of those reasons. Other reasons include the fact that differences in their allele frequencies between populations can be useful in population-based genetic studies, as well as their greater stability comparatively to other types of genetic variation, hence allowing for more consistent estimates of genotype-phenotype associations [8, 9].

Linkage disequilibrium The concept of linkage refers to the tendency of genes or DNA sequences at specific locations in a chromosome, also known as loci, to be inherited together due to their physical proximity in a given chromosome. This concept is tightly related to the process of genetic recombination, in which two homologous chromosomes randomly exchange parts of their DNA, producing new combinations of alleles to be passed on to the next generation. Thus, recombination is very likely to happen between two segments of DNA separated by a considerable distance, unlike two close sequences, which tend do stay together rather than to be split. In the latter case, the sequences are considered to be linked. Moreover, sequences that are far enough from each other and consequently have high recombination frequencies are said to be in linkage equilibrium. On the other hand, when there is a nonrandom association of sequences or alleles in adjacent loci these are considered to be in linkage disequilibrium (LD). In this case, these loci are found together more often than expected if they were segregating independently in a population [8, 10].

The two most commonly used statistics to measure the amount of LD are the D' and r^2 parameters, the r^2 parameter being more often used out of the two mentioned. While r^2 generally measures the correlation between two variables, D measures the deviation of haplotype frequencies from the equilibrium state, an haplotype being a group of linked alleles in a chromosome inherited together from a single parent. D' will, in turn, be computed as the absolute ratio of D compared with its maximum value, when $D \geq 0$, or compared with its minimum value, when $D < 0$. In both statistics, the higher the value the lower the possibility that recombination occurred between two loci. Moreover, having $D' = 1$ or $r^2 = 1$ means that the two loci have not been separated by a recombination event. In this case of perfect LD, in which both

sequences are completely linked, observations about one sequence provide accurate predictions about the other [8, 10].

Hardy-Weinberg equilibrium The Hardy-Weinberg principle is a concept in population genetics that relates genotype frequencies to allele frequencies in a population. When a population meets all the requirements to be in Hardy-Weinberg equilibrium (HWE), alleles segregate randomly in a population from generation to generation, therefore allowing expected genotype frequencies to be computed from allele frequencies [8, 11].

Assuming that a population is in HWE, considering a biallelic locus with two alleles, A and B, with known frequencies (allele A = p; allele B = q) adding up to one, the possible genotypes will be AA, AB, and BB. Thus, the probability of having an AA individual will be given by $p \times p = p^2$. Similarly, the probability of producing a BB individual will be given by $q \times q = q^2$. Moreover, the probability of having an AB individual will be given by $2 \times p \times q = 2pq$, as one must take into account that if sperm and eggs meet randomly, an AB genotype may arise from the combination of an egg containing the A allele and a sperm containing the B allele or vice versa [8, 10]. The general formula that describes the HWE in a population is given by the following expression [8]:

$$p^2 + 2pq + q^2 = 1 \tag{2.1}$$

There are five assumptions on which HWE is based. A population is no longer under HWE when one or more of these assumptions is violated. These are random selection, since allele frequencies may change from one generation to the next due to individuals with certain genotypes surviving better than others; no mutation, given that new alleles produced by mutations, for instance, may cause allele frequencies to change from one generation to the next; no migration, as the movement of individuals in or out of a population is expected to change allele frequencies; no chance events, as allele frequencies may change due to some individuals contributing, by chance, with more alleles than others to the next generation; random selection of mates, which if violated also changes allele frequencies [8].

As human populations do not meet all the criteria of HWE exactly, populations are able to evolve [8]. Further, it is possible to assess how far a population deviates from HWE by using the Pearson chi-square test (χ^2), or the ”goodness-of-fit” test - to be addressed in the next subsection - of the observed genotype counts to their expectation under HWE [8, 11].

2.1.2 Bioinformatics background

The continuous technological progress in genomics brought new DNA sequencing technologies, the major recent breakthrough being next generation sequencing (NGS). Unlike the traditional Sanger sequencing technology, NGS made possible to sequence an entire human genome within a single day and at a lower cost [12]. Advancements such as NGS created an exciting opportunity for investigators to further understand the genetic underpinnings of disease, for instance of complex diseases, in which both genetic and environmental factors contribute to the susceptibility risk [13].

In order to uncover the genetic contributors to disease, population genetic association studies, which involve unrelated individuals, play a key role. Genetic association studies aim to identify candidate genes or genomic regions that contribute to a given disease status by assessing the correlation between the disease status and genetic variation [10, 13]. As already mentioned, SNPs are the most widely used markers in genetic association studies. A higher frequency of a SNP allele or genotype in a set of individuals with a disease can potentially mean that the tested variant confers an increased risk to such disease. Plus, in the case of common complex diseases, such as diabetes or heart disease, among others, genetic studies over the years have confirmed that many distinct genetic variants influence disease risk. On the other hand, each variant involved seems to have only a subtle effect size, meaning the risk contribution of a genetic variant to a disease [13]. This is in line with the common disease/common variant hypothesis, which states that common genetic variation in the population is likely to underlie the genetic architecture of common disorders, though one should not assume that the entire genetic component of any common disorder is due to common variants only [14].

Genetic association studies can be direct or indirect. A direct association study assesses the association between a known functional variant and a disease. On the other hand, an indirect association study tests such association using a marker that is located closely to the disease locus and is in LD with it. The latter is the more common approach [8].

2.1.2.1 Genetic association study designs

Candidate gene studies A candidate gene study is an indirect association study that usually involves multiple SNPs within a single gene. The chosen SNPs, which may or may not be functional variants, are thought to capture information about the genetic variability of the gene being studied. In this context, measures of LD are a useful tool when selecting SNPs presumed to be physically close to a single true disease-causing locus [8, 10].

Candidate polymorphism studies Candidate polymorphism studies aim to assess if there is an association between a given SNP, or set of SNPs, and a disease, for example. These studies usually are

backed by prior scientific evidence supporting the relevance of the SNPs under investigation and the primary hypothesis is that the variants being tested are functional [8, 10].

Fine mapping studies Fine mapping studies aim to identify, with a greater level of precision, the location of a genetic variant, such as a disease-causing variant, in the genome. LD is also very useful in these studies as it aids the process of finding such locations [8, 10].

Genome-wide association studies (GWAS) These studies involve performing whole or partial genome-wide scans with the goal of identifying associations between SNPs and a given trait. Unlike candidate gene or candidate polymorphism approaches, GWAS are less hypothesis-driven and usually imply characterizing a large number of SNPs. GWAS also require more data preprocessing, as well as a greater computational burden compared to other study designs [8, 10].

Even though GWAS gained considerable popularity in recent years, candidate gene studies still play an important role when it comes to validating the findings of a GWAS. In the context of disease, candidate gene studies also allow to further investigate the biological and clinical interactions between genes and different traditional risk factors or patient-level characteristics [10].

2.1.2.2 Population-based studies: case-control study

A case-control study is a widely used population-based study design for binary traits - meaning that there are only two phenotypic values (affected or unaffected) - regardless of the focus being on candidate genes, regions, or the entire genome [8]. In a case-control study, genotyping information is collected from a set of individuals who have been diagnosed with the disease under investigation - cases - as well as from a set of individuals either known to be unaffected by the condition, or selected randomly from the population - controls [13]. One must ensure that there is a good match between the genetic backgrounds of cases and controls, so that any genetic difference identified is related to the disease under study and not because of biased sampling. Cases and controls are usually selected independently from the same ethnic population. To guard against subtle genetic differences, controls can be selected from the same geographical area as cases, for example. An alternative approach to selecting cases and controls can be to sample from a cohort of individuals, meaning a group of individuals that share a characteristic such as age, for instance, being followed up prospectively - prospective cohort study [8, 13].

Once the cases and controls have been selected and genotyping has been performed, the genotyping information is compared between both sets of individuals. Thus, an increased frequency of a SNP allele or genotype in cases comparatively to controls indicates that the presence of such allele potentially increases the risk of disease [8, 13].

Association tests In a study involving multiple SNPs, tests of association are usually performed separately for each SNP. Thus, for a SNP with alleles A and B, the six total counts of the number of possible genotypes (AA, AB, and BB) in cases and controls can be represented in a 2×3 table, also known as a contingency table [8, 13].

Table 2.1: Contingency table for a case control study. *a*,*b*, *c*, *d*, *e*, and *f* represent genotype counts in cases and controls. Adapted from Lewis, 2002 [13].

Genotypes	**AA**	**AB**	**BB**
Cases	*a*	*b*	*c*
Controls	*d*	*e*	*f*

To analyze the previous table, an observed-expected test statistic - a Pearson chi-square test - can be used to test for deviation from the expected values across genotype counts. In this case, the test statistic follows a chi-square distribution with 2 degrees of freedom, the degrees of freedom being given by $(r-1)(c-1)$, r and c referring to the number of rows and columns of the contingency table, respectively [8, 13, 15].

Considering, for instance, the observed counts for the AA genotype in cases, $O_1 = a$, its expected value will be given by the total number of AA genotypes, n_{AA}, multiplied by the proportion of cases, n_{cases}, out of the total sample, n [13]. This translates into the following expression:

$$E_1 = \frac{n_{AA} \times n_{cases}}{n} \tag{2.2}$$

Using expression (2.2), the full test statistic will be given by:

$$X = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(2) \tag{2.3}$$

where the summation includes the difference between the observed value, O_i, and the expected value, E_i, for every cell in the contingency table [13]. $\chi^2(2)$ indicates that the distribution has two degrees of freedom.

It is noteworthy that, in this case, no order or trend of the genotype-phenotype relationship is assumed, that is, each genotype is assumed to be independently associated with the trait (or disease). The conventional statistical significance level (α) when performing this test is 0.05, α being the probability of rejecting the null hypothesis when it is true. The null hypothesis is the premise being tested and

states that there is no difference between genotype frequencies of cases and controls, which can be interpreted as there being no association between genotypes and the disease. If the resulting test statistic is larger than the critical value from the chi-square distribution, for the corresponding number of degrees of freedom, it means that the null hypothesis of no association should be rejected, therefore indicating an association between the SNP being tested and the disease [8, 15]. In other words, if the resulting *p*-value of the test is below 0.05, it indicates that there is evidence that supports the rejection of the null hypothesis.

Besides the chi-square test, other statistical tests can be applied to a contingency table such as the Fisher exact test, for example. Despite both tests assessing the null hypothesis of independence between two variables, the chi-square test assumes that the sample is large and performs an approximation, whereas the Fisher exact test is more suitable for small-sized samples and performs an exact test. Another aspect that differs from the chi-square test is that the Fisher test applies a hypergeometric distribution [8, 15].

This approach enables the testing of alternative models of association, meaning that the data may be analyzed assuming a genetic model [8, 13]. For instance, supposing that carrying the A allele increases the risk of disease implies that the AA and AB genotypes must be grouped and their respective counts summed, giving rise to a 2×2 contingency table - dominant model. On the contrary, under a recessive model, only homozygosity for the A allele is associated with an increased risk of disease. This implies that the new 2×2 table will result from grouping the AB and BB genotypes [8, 13].

An alternative method to analyzing case control data is based on allele frequencies, in which the total number of A and B alleles is compared between cases and controls, regardless of genotypes. As a result, AA and BB genotypes will contribute with twice the counts to the new 2×2 table. This method implies assuming a multiplicative genetic model, where the risk of developing a disease increases by a factor r for each copy of the risk allele - considering the A allele, the risk will be r for AB genotype and r^2 for AA genotype. An allelic association test using a chi-square test with one degree of freedom is considered to be more reliable than a genotypic test with two degrees of freedom. However, this is only the case when the penetrance of the heterozygous genotype lies between the penetrance of the two homozygous genotypes [16]. Further, selecting this genetic model comes with the assumption that both case and control genotypes are in HWE, which can be tested for, as previously mentioned [13].

Table 2.2: Contingency table under a multiplicative model. Adapted from Lewis, 2002 [13].

Allele	**A**	**B**
Cases	*2a + b*	*b + 2c*
Controls	*2d + e*	*e + 2f*

A fourth possible model is the additive genetic model, where individuals carrying one copy of the A

allele will have an increased disease risk of r, whereas individuals carrying two copies of the A allele will have a $2r$ increased disease risk. In this case, genotype frequencies are analyzed instead of allele frequencies and the contingency table will be again table 2.1 [13]. Under an additive genetic model, the Cochran-Armitage test for trend is a preferable choice to analyze genotype frequencies comparatively to the chi-square test [16]. Unlike the chi-square test, the Cochran-Armitage trend test is more conservative and does not rely on an assumption of HWE [8, 16]. Instead, it tests the null hypothesis that a line that best fits the three genotypic risk estimates has zero slope [8].

Besides contingency table techniques, case-control traits can also be analyzed using logistic regression [16]. In a logistic regression model, the outcome (0/1) is a case or control and the possible genotypes (three, for instance) are the levels of the explanatory variable of genotype, meaning the variable that is being tested for association with the outcome [13].

Logistic regression is an adaptation of linear regression in which the use of the logit transformation makes possible the analysis of a binary outcome [11]. This method allows the inclusion of multiple variables, which can be other SNPs, epidemiological risk factors, or patient-level characteristics, such as age or gender [13]. For instance, considering the simultaneous analysis of the effect of two SNPs in a disease, the logistic regression model will be represented by the following expression:

$$logit(p) = ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{2.4}$$

where p is the probability of having the disease and x_1 and x_2 contain information concerning the genotypes of the two SNPs. β_0 is the intercept term, β_1 and β_2 represent the effects of each marker on the disease, and β_3 models the interaction between SNPs [11].

Population stratification One of the major aspects to take into account in the design of a case control association study is the influence of confounding factors such as population stratification. Population stratification results from the sampling, in different proportions, of cases and controls from genetically different populations, causing found associations to be due to sampling differences, rather than related to the disease under study. In other words, confounding occurs when there are allele frequency differences between cases and controls due to systematic ancestry differences [8, 13, 17].

There are several methods to prevent population stratification. During the study design stage, it can be minimized by ensuring that cases and controls are ethnically matched or by restricting sampling to a specific ethnic group. Plus, if gender has an effect in disease prevalence, stratification may also be reduced by matching cases and controls based on gender [8].

One of the common methods to test for population stratification in a genetic association study is to perform principal component analysis (PCA). PCA is a statistical method that, given a matrix composed of m observations (samples, for instance) and n variables or features (genetic markers, for instance) - $m \times n$ matrix - finds directions of maximum variance (of the data), which are mutually orthogonal. In other words, PCA finds another basis of n axes and enables the rotation of the data that maximizes the variance in those new axes. The set of n new variables or axes are called principal components, the first one accounting for as much of the variance in the data as possible, the second one for as much of the remaining variance and so on. This property makes PCA a commonly used dimensionality reduction or feature selection technique, as ranking the principal components, in terms of their explained variance, allows one to discard dimensions or features that don't account for a significant amount of variance, without losing too much information [17, 18].

Since a significant feature is one that exhibits differences between groups of observations, and PCA captures those differences, when plotting the rotated data using the first two principal components, for example, one may observe the formation of clusters, which may possibly be due to the existence of different subpopulations within the group being investigated. Furthermore, PCA may be used to adjust for population stratification in association testing by, for example, adding the principal components to a logistic regression model as covariates [17, 18].

Another approach to adjust for population stratification can be through genomic control. Genomic control assumes that an association test statistic is uniformly inflated by a constant factor (λ), whose magnitude is estimated by comparing the observed median of a set of markers with the expected median in the absence of stratification. Population stratification is said to exist when λ has a value greater than one, which can be corrected by dividing the association test statistic values by λ. This control must be restricted to SNPs or genomic markers that are unlinked to the SNP(s) being tested for an association with a given disease status [8, 13].

Risk estimation: odds ratio Besides assessing the departure from the null hypothesis of equal SNP allele frequencies in cases and controls, contingency table methods can also provide estimates of disease risk conferred by a given allele [13]. An odds ratio (OR) is obtained by computing the ratio of the odds of developing an outcome (e.g. a disease status) given an exposure, (e.g. carrying a risk allele, also known as the effect allele) and the odds of developing the same outcome in the absence of that exposure [13, 19]. Taking as an example table 2.1, and the A allele as a supposedly dominantly acting allele, the increased disease risk for A allele carriers compared to non-carriers will be given by:

$$OR = \frac{\frac{a+b}{d+e}}{\frac{c}{f}} = \frac{f}{c}\frac{a+b}{d+e} \tag{2.5}$$

Regarding expression (2.5), if the OR is equal to one, it means that the odds of disease are not affected by the presence of the allele. On the other hand, if the OR is greater than one, it indicates that carrying the allele is associated with higher odds of having the disease, whereas an OR smaller than one indicates that carrying the allele is associated with lower odds of having the disease [19].

An OR is usually presented with an associated confidence interval (CI), typically at the confidence level of 95%. A 95% CI reflects the precision of the OR. If the level of precision of the OR is low, the associated CI will be large. On the contrary, an OR with a higher precision level will have a small CI. Moreover, a 95% confidence level means that if a given experiment were to be repeated numerous times for a certain population, the resulting CI of every experiment would include the true parameter of the population 95% of the times [19].

Besides allelic ORs, genotypic ORs can also be computed by comparing the odds of disease when having a specific genotype to the odds of disease when having another genotype. [8] On the other hand, when adopting a univariate logistic regression model, ORs are simply obtained by applying the exponential function to the β parameter [10]. However, in multivariate logistic regression, it is noteworthy that the β parameters are often similar but not strictly equal to the logarithm of the corresponding ORs.

Another risk estimate is the relative risk (RR), which can be determined in prospective cohort studies [20]. Supposing that table 2.1 would be from a cohort study, the RR would be given by the ratio of the risks of developing the disease in the group carrying the A allele and in the group not carrying the A allele, as given by the following expression:

$$RR = \frac{\frac{a+b}{a+b+c}}{\frac{d+e}{d+e+f}} = \frac{a+b}{d+e}\frac{d+e+f}{a+b+c} \tag{2.6}$$

The fundamental difference between both estimates is precisely the fact that while the first one yields a ratio of odds, the second one yields a ratio of probabilities. The RR also usually comes with an associated CI and can be directly estimated from a prospective cohort study, since in this study design the overall prevalence of the outcome is known, contrarily to what happens in a case control study, where the OR is used instead [8, 20]. Further, if the RR equals one, then the risk is the same for both groups. When the RR is greater than one, it indicates that the risk of disease in greater for those who carry the allele, whereas the opposite is said when the RR is smaller than one [20].

Statistical power of association studies The statistical power of an association study is defined as the probability of the study rejecting the null hypothesis of no association between a genetic marker and a phenotype when there is indeed a true association - true-positive association. There are 2 types of error worth distinguishing. A type I error (α) is the probability of rejecting the null hypothesis when it is

true - false-positive association - being also known as the already mentioned statistical significance level. On the other hand, a type II error (β) is the probability of not rejecting the null hypothesis when is it false - false-negative association. Thus, the statistical power will be the complement of the probability of making a type II error, or $1 - \beta$ [8, 21]. A study with 80% power is considered to have adequate statistical power already. Moreover, having 80% power means that if a study were to be performed repeatedly over time, it would produce statistically significant results 8 times out of 10 [21].

There are several factors that can affect the power of a genetic association study. In the case of a GWAS, for example, power depends on sample size, allele frequency, effect size, and on the genotyping platform used [8]. When it comes to determining the effective sample size, meaning the minimum number of samples required to achieve adequate statistical power, measures of LD are a useful tool. The r^2 parameter is inversely proportional to the sample size required to detect an association between a variant and a disease. For instance, if a genotyped variant is in LD (r^2 = 0.5) with a non genotyped functional variant, the effective sample size would have to be doubled to detect an association [8] Another advantage of using LD as a tool is that if a genomic region has multiple SNPs that are in strong LD (conventionally for $r^2 > 0.8$), the number of included SNPs for analysis can be reduced with small losses in power [8].

A fundamental quality control step to ensure that the inferences of an association study are not biased is checking for deviations from HWE in the control group. This group is intended to represent the general population of the region where the cases are sampled from, and departure from HWE may reflect important problems that can potentially compromise study power. Controls not being in HWE can arise from genotyping problems or population stratification, the latter potentially leading to false-positive associations or to missing true associations, therefore being very important that the appropriate adjustments are made when identifying such shortcomings [13, 22].

2.2 Hereditary thrombophilia background

This section focuses on the characterization of hereditary thrombophilia, of its genetic basis, and clinical manifestations. It also addresses some of the controversies in the literature surrounding this condition.

2.2.1 Hereditary thrombophilia

Thrombophilia can be defined as an abnormality in blood coagulation, leading to a predisposition to form blood clots which, in turn, increases the risk of thrombotic events. Such predisposition can result from genetic factors – hereditary thrombophilia – acquired changes in the clotting mechanisms or, more commonly, from the interaction between both [23].

The genetic basis of this hypercoagulable state has been attributed to changes in the amount and/or

function of proteins involved in the coagulation process. One can say that there are different main inherited thrombophilias. Loss-of-function mutations of the genes encoding natural anticoagulant proteins, namely antithrombin (AT), protein C (PC), and its cofactor, protein S (PS), as well as gain-of-function mutations in blood coagulation factors II (FII) and V (FV) belong to this group [24].

Carrying a homozygous abnormality, the combination of two or more heterozygous abnormalities, or milder heterozygous traits result in different phenotypic consequences. Clinically evident thrombotic disorders can occur at an early age in the presence of homozygous abnormalities, or when carrying the combination of two or more heterozygous abnormalities. On the other hand, the detection of milder heterozygous traits, when isolated, tends to require laboratory investigation [23]. Thus, while homozygous genotypes are rare and are associated with spontaneous thrombotic events, heterozygous genotypes are more often observed and can increase the risk of thrombosis concomitantly with the presence of other genetic or environmental risk factors, for example [25].

2.2.2 The pathogenesis of thrombosis

Injury to the vessel wall, blood stasis, and hypercoagulability have been originally identified by Rudolf Virchow as the core of the etiology of thrombosis – Virchow's triad (1856). Regarding VTE, stasis and hypercoagulability have been pinpointed as the key contributors to this condition [23, 24].

The main components of venous thrombi are fibrin and red blood cells, contrarily to arterial thrombi, in which blood platelets assume a pivotal role. Bearing in mind the composition of venous thrombi, the concept of hemostatic balance between fibrin formation and dissolution has been implicated in the etiology of VTE. Thus, both the coagulation pathway, leading to fibrin formation, and the fibrinolysis pathway, leading to fibrin dissolution, are worth understanding in order to determine the factors underlying an imbalance in hemostasis [23, 24].

The coagulation and fibrinolytic systems are two separate, yet linked, enzyme cascades whose combined functions aim to regulate the formation and breakdown of fibrin [23].

The coagulation pathway is a proteolytic cascade where each enzyme involved is present in the plasma as a zymogen, an inactive form or precursor molecule. Each zymogen, upon activation, undergoes proteolytic cleavage and releases its active factor. This pathway comprises a series of positive and negative feedback loops that control the activation process, its ultimate goal being the production of thrombin. Thrombin can, in turn, convert fibrinogen into fibrin, the latter forming a clot. Besides the prothrombotic cleavage of soluble fibrinogen to insoluble fibrin, and the activation of coagulation factors V, VIII, XI and XIII, thrombin can also exert an anticoagulant effect by forming an enzyme complex with cofactor protein thrombomodulin to activate the natural anticoagulant PC [23].

Among the enzymes involved in fibrinolysis, plasmin is one of the most important. Plasmin is responsible for the degradation of fibrin and results from the activation of plasminogen, its zymogen. The activation of plasminogen is mediated by two serine enzymes, tissue-type plasminogen activator (t-PA) and urokinase plasminogen activator (u-PA). In turn, these serine proteases are regulated by protease inhibitors, namely plasminogen activator inhibitor (PAI) – 1 and PAI-2 [23, 26].

2.2.3 Classic inherited thrombophilias

The main genetic determinants of hereditary thrombophilia previously mentioned will be described in greater detail in the following sections.

2.2.3.1 Loss-of-function mechanisms

AT deficiency AT is a single chain glycoprotein synthetized by the liver and a member of the serpin superfamily, a family of serine protease inhibitors. Thus, it exerts a major inhibitory effect on serine proteases such as activated factor X (FXa) and thrombin, decreasing both the production and the half-life of the latter. This natural anticoagulant also has a heparin-binding site which, in turn, enhances the molecule's ability to inactivate activated coagulation factors [24].

Over 250 loss-of-function mutations have been identified in the AT gene, *SERPINC1* [24]. Most of these mutations cause AT plasma levels to decrease, or compromise the glycoprotein's ability to interact with either heparin or activated coagulation factors. AT deficiency is transmitted as an autosomal dominant trait whose estimated prevalence is extremely low in the Caucasian population, ranging between 0.02 and 0.2% [24]. On the other hand, its penetrance - the proportion of individuals carrying a particular variant that also expresses its associated phenotype - is very high, as the risk of developing VTE is increased more than 50-fold when compared with non-carriers, hence being considered the most severe inherited thrombophilia [24]. Depending on the mutations' nature, two types of AT deficiency can unfold. Type I, as a result of a wide variety of mutations, is characterized by reduced functional levels and is typically associated with premature recurrent VTE. On the other hand, type II deficiency is caused by missense mutations and is characterized by normal protein levels, however with impaired inhibitory activity, due to the production of a variant protein. In this case, the risk of thrombosis depends on the position of the amino acid substitution relatively to the protein's reactive site [24].

PC deficiency PC is a vitamin-K-dependent glycoprotein synthesized in the liver in an inactive form, which is further activated in the presence of thrombin. The activation process is accelerated by a complex formed by thrombin with the endothelial protein C receptor (EPCR) and thrombomodulin. Activated PC (APC), together with its cofactor, PS, reduces thrombin production through the inactivation of activated coagulation factors V (FVa) and VIII (FVIIIa) [24].

More than 200 loss-of-function mutations in the PC gene, *PROC*, have been identified as causing this protein's deficiency [24]. Hereditary PC deficiency is transmitted as a dominant autosomal trait, being also very rare in the Caucasian population, accounting for an estimated prevalence around 0.2% [24]. Despite the disease's penetrance being lower than that of AT deficiency, heterozygotes still have a 15-fold increased risk of developing premature VTE [24]. Similarly to AT deficiency, two types of PC deficiency can also be distinguished based on the nature of the mutations taking place. Type I is the most common and is typically caused by missense mutations, consequently leading to the reduced synthesis of the functional protein. In turn, the rarer type II deficiency also arises usually from missense mutations and is characterized by normal levels of a dysfunctional protein [24].

PS deficiency PS is a vitamin-K-dependent protein synthetized in the liver which circulates in the plasma both in an active (approximately 40%) and inactive (approximately 60%) form. Besides acting as a cofactor of APC, PS also functions as a cofactor of tissue factor pathway inhibitor (TFPI) protein in the inhibition of FXa [24].

Close to 200 loss-of-function mutations have been identified in the PS gene, *PROS1*, most of them being missense mutations, short deletions, or insertions [24]. Similarly to the previous natural anticoagulants, hereditary PS deficiency is transmitted as an autosomal dominant trait with a prevalence of 0.03-0.1% in the Caucasian population [24]. Carriers have a 10-fold increased risk of VTE compared to non-carriers [24]. Three types of PS deficiency can be distinguished: type I, characterized by decreased plasma levels of functional total and free protein; type II, characterized by impaired cofactor activity but normal total and free protein levels; type III, in which there is a reduced functional activity, as well as of free protein levels, while total protein levels remain normal [24].

2.2.3.2 Gain-of-function mechanisms

Factor V Leiden (FVL) A gain-of-function mutation in the *F5* gene was first identified in 1994 and described as the cause of the majority of cases of APC resistance. This mutation results from the substitution of guanine to adenine at position 1691 (G1691A) of the gene. As this event occurs in a region of the gene encoding one of the factor's cleavage sites, key to its inactivation, the mutant molecule – FVL – becomes resistant to inactivation by APC and continues to exert its full procoagulant activity. Other FV variants, namely FV Hong Kong and FV Cambridge, result from point mutations in a different APC cleavage site and are responsible for varying degrees of APC resistance, ranging between that of a wild-type factor and of a mutant one [24].

The FVL mutation is transmitted as an autosomal dominant trait which, in heterozygosity, is considered the most common prothrombotic mutation in the Caucasian population, with a prevalence of 5% [24]. Considering a positive gradient from Southern to Northern Europe, its prevalence increases from 2 to

10% [24]. While heterozygous carriers have a 7-fold increased VTE risk, compared to non-carriers, the same risk is increased by 80-fold in homozygous carriers [23, 24].

Prothrombin G20210A This mutation was identified soon after the FVL mutation, in 1996. It results from the substitution of guanine to adenine at nucleotide 20210 of the prothrombin gene (*F2*), consequently leading to an increase of around 30% in prothrombin plasma levels. It has an autosomal dominant transmission and constitutes the second most common prothrombotic mutation in the Caucasian population. In heterozygosity, its prevalence is about 2-3%, and unlike FVL, it is more common in Southern than in Northern Europe [23, 24].

Heterozygosity for this mutation is associated with a relatively low risk of VTE – around 3- to 4-fold increased risk – and most carriers do not develop a premature thrombotic event. However, homozygosity, which is rarer, confers a 30-fold increased risk [24].

The following table summarizes the prevalence, in the Caucasian population, of the thrombophilia markers addressed, as well as the relative risk for the development of a thrombotic event, namely VTE, followed by the risk of recurrence.

Table 2.3: Prevalence and VTE relative risk (RR) of the main inherited thrombophilias described. Adapted from Mannucci *et al.*, 2015 [24].

Marker	Prevalence (%)	First VTE (RR%)	Recurrent VTE (RR%)
AT deficiency	0.02-0.2	50	2.5
PC deficiency	0.2-0.4	15	2.5
PS deficiency	0.03-0.1	10	2.5
FVL (heterozygous)	5	7	1.5
FVL (homozygous)	0.02	80	-
Prothrombin G20210A (heterozygous)	2	3-4	1.5
Prothrombin G20210A (homozygous)	0.02	30	-

2.2.4 Clinical manifestations

The previous markers – AT, PC, PS deficiency, and APC resistance – can also be referred to as defects in the naturally occurring anticoagulant systems. So far, the main clinical manifestation mentioned, as a result of carrying these defects, was VTE. De Stefano *et al.* were able to describe other clinical manifestations and even distinguish between the clinical picture of heterozygotes and homozygotes [23, 27].

2.2.4.1 Clinical manifestations in heterozygotes

Individuals with heterozygosity for AT, PC or PS deficiency, as well as APC resistance, display similar clinical manifestations. In these individuals, VTE is typical. In fact, this condition is estimated to develop in 60% to 80% of carriers, usually occurring before the age of 40 to 45 years old [27]. Moreover, recurrence tends to happen in approximately half of the patients [27]. VTE often leads to DVT of the lower limbs, with or without PE. VTE can also manifest itself as an inflammatory-thrombotic disorder, namely superficial thrombophlebitis, whose effects become visible in the skin. The latter tends to be more frequent in patients with PC deficiency, PS deficiency, and APC resistance, comparatively to AT-deficient patients [27].

2.2.4.2 Clinical manifestations in homozygotes

Homozygous AT deficiency is extremely rare and is associated with severe thrombotic events occurring at a young age, often affecting arteries – arterial thrombosis [27].

Homozygous PC deficiency is also rare and is associated with peculiar phenotypes. While in patients with very low but measurable PC (5% to 20%) clinical manifestations are similar to those with heterozygous deficiency, in patients with unmeasurable PC, neonatal purpura fulminans may occur. This condition, a severe thrombotic disorder characterized by blood spots, bruising, and discoloration of the skin, may also develop during the first year of life. The very rare PS-deficient homozygotes also present a more severe clinical picture with neonatal purpura fulminans [24, 27].

2.2.4.3 Other clinical manifestations

Throughout the years, hereditary thrombophilia has been systematically implicated in other conditions in the literature. Among the most common are events or complications related with cardiovascular disease (CVD), such as IS or myocardial infarction (MI), for example. Thus, as thrombosis is involved in the etiology of such events, the search for a link between these and a hypercoagulable state, meaning thrombophilia, seemed only logical. Furthermore, also the link between poor pregnancy outcome (PPO) and thrombophilia has been a topic of extensive research. However, since both CVD and PPO are multifactorial conditions, proving an association, let alone causality, between the former and thrombophilia is not straightforward, therefore giving rise to contradictory evidence in some cases.

The following sections report the findings of recent evidence assessing the role of hereditary thrombophilia in different conditions or complications, which have repeatedly been associated with this inherited hypercoagulable state.

Hereditary thrombophilia and cardiovascular disease (CVD) Thromboembolic events, whether occurring in veins (VTE) or arteries – arterial thromboembolism (ATE) – are often addressed in the context of CVD. CVD encompasses a group of diseases affecting both the heart and blood vessels. It includes coronary heart disease (CHD), coronary artery disease (CAD) and acute coronary syndrome (ACS), a subcategory of CAD. CAD usually refers to disease affecting coronary arteries, frequently due to atherosclerosis. In turn, CHD is said to result from CAD and includes the diagnoses of MI and silent myocardial ischemia, for example. Despite not being the same in their core, health professionals tend to use the terms CAD, CHD and ACS interchangeably, CAD being often referred to as CHD [28].

Mahmoodi *et al.* investigated the influence of hereditary thrombophilia and traditional cardiovascular risk factors on the risk of developing ATE [29]. Previous studies have suggested that the role of thrombophilic defects in ATE is weak, which is supported by the much higher relative risk of individuals, with such inherited defects, developing VTE, comparatively to the risk of developing ATE [30]. In fact, ATE is usually attributed to endothelial damage followed by atherosclerosis, the latter associated with traditional cardiovascular risk factors, such as hypertension, diabetes mellitus, hyperlipidemia, hypercholesterolemia, smoking, and obesity [29]. Given that MI or IS, for example, is often caused by the rupture of an atherosclerotic plaque, with subsequent thrombus formation, the authors hypothesized a synergistic interaction between hereditary thrombophilic defects and traditional cardiovascular risk factors. Five defects were considered for analysis, namely FLV, prothrombin G20210A, PC, PS, and AT deficiency. From the results obtained, it was concluded that hereditary thrombophilia was indeed associated with a higher risk of ATE. Such association was overall stronger in the presence of traditional cardiovascular risk factors, particularly in the presence of diabetes mellitus. Furthermore, it also tended to be stronger in females when compared to males, as well as in individuals before the age of 55 years old comparatively to those over such age [29].

The link between younger age and the influence of hereditary thrombophilic defects in CVD is also supported by Dragoni, who studied the effects of inherited and acquired factors in patients under the age of 55 with ACS or IS. The author reinforces the former idea by stating that the combination of thrombophilic risk factors, inherited or acquired, with common cardiovascular risk factors appears to significantly increase the risk of younger patients developing ACS or IS. Dragoni further explains that the reason why hereditary thrombophilic defects, with an emphasis on FVL and prothrombin G20210A, could have a greater impact in younger individuals is because atherosclerosis has had less time to progress, at that point in life, thus not being always clinically evident. Also, traditional cardiovascular risk factors are less frequent when compared to older patients [31].

Regarding IS, Pahus *et al.* investigated the implication of inherited and acquired thrombophilia as risk factors for IS, as well as for transient ischemic attack (TIA), in patients under the age of 50. The authors found no strong association between thrombophilia and IS in these patients, and only FVL appeared to increase the risk of TIA [32]. On the other hand, Sharma *et al.* did recognize the role of thrombophilia in IS,

stating, however, that atherosclerosis and cardioembolism still constitute more likely causes. The authors further added that thrombophilic disorders account for 10% to 15% of the cases of IS in younger patients [33]. Considering cardioembolism, cerebral thromboembolic events in patients with atrial fibrillation may be potentiated by inherited thrombophilic defects, given the stasis of blood in the left atrium during this event [29].

In light of the above, one can argue that having a predisposition to forming blood clots, determined by the presence of well-established thrombophilic defects, such as FVL, prothrombin G20210A, and the deficiency of natural anticoagulants, may have some impact in the risk of IS, especially in individuals under the age of 55. Nevertheless, it is not considered the primary cause when investigating the etiology of IS [34].

Hereditary thrombophilia and poor pregnancy outcome (PPO) Pregnancy is a prothrombotic state by itself in which all three factors of the Virchow's triad are present: the hormonal changes of pregnancy cause the pooling of venous return in the lower extremities, leading to vasodilation and subsequent venous stasis; endothelial damage is present in both the implantation process and later profoundly during the delivery; there is a natural increase in the levels of prothrombotic factors and decrease in the levels of antithrombotic factors. This state of hypercoagulability is thought to be a protection mechanism developed during this period, given the bleeding risks associated with childbirth or miscarriage [35, 36]. According to Croles *et al.*, the inherent hypercoagulability of pregnancy means that the risk of VTE in women increases 5- to 6-fold compared with age matched controls, the risk being higher in the six weeks following childbirth [37]. The same authors conducted a recent systematic review and meta-analysis, aiming to provide scientific evidence supporting prophylaxis guidelines for the management of pregnant women with hereditary thrombophilia. It was concluded that AT deficiency, PC deficiency, PS deficiency, and homozygous FVL were considered high risk thrombophilias for a first VTE event. As a result, the authors recommended that women with such defects should be considered for antepartum or postpartum thrombosis prophylaxis, or even both. These results were consistent with the ones of other sources stating that hereditary thrombophilia further increases the risk of the mother developing pregnancy associated VTE [37, 38].

Even though there appears to be consensus about the fact that hereditary thrombophilia has a compounding effect on the inherent hypercoagulability of pregnancy, hence putting pregnant women at a higher risk of developing VTE, some controversy remains regarding its connection with PPO [35, 39–41].

PPO can result from placenta-mediated pregnancy complications such as recurrent miscarriage (RM) [42]. Despite the definition of RM often changing, it can commonly be defined as the occurrence of two or more consecutive fetal losses before twenty weeks of gestation [36]. RM affects 1 to 5% of couples and has a multifactorial nature [40]. It can be due to chromosomal abnormalities incompatible with life, a common cause of PPO in the first trimester, as well as due to anatomical, hormonal, immune, genetic,

and unexplained causes. In fact, the cause of RM cannot be unraveled in 30 to 40% of cases [39, 43]. Growing evidence throughout the years has suggested that inherited thrombophilias, besides augmenting the inherent prothrombotic state of pregnancy, compromise adequate fetomaternal circulation, through the formation of microthrombi in the placental vessels, for instance, and even the process of placentation in the developing embryo, ultimately leading to a failed pregnancy [36, 42, 43]. Genetic defects linked to RM include classic thrombophilic gene variants, namely FVL and prothrombin G20210A. Moreover, also variants of the methylenetetrahydrofolate reductase (*MTHFR*) gene have been linked to RM [36, 42].

Chatzidimitriou *et al.* conducted a study in order to identify thrombophilic genetic polymorphisms and their correlation to RM in a small sample of Greek women. From the results obtained, FVL, MTHFR C677T, and MTHFR A1298C genetic variants were identified as risks factors for RM. Also, polymorphisms of the gene encoding PAI-1, in particular SERPINE1 rs1799889, showed a correlation with RM [36]. Moreover, Gao *et al.* published a systematic review and meta-analysis of 37 case-control studies assessing the role of prothrombin G20210A in RM. The main conclusion was that this mutation increased the risk of RM, particularly in European women over the age of 29 years old [44].

Though many studies, with different sample sizes and methodologies, can be found in the literature featuring results that support a connection between hereditary thrombophilia and PPO, in particular RM similarly to the ones above, many others can also be found reaching contrary conclusions. For instance, Unterscheider *et al.* published a review featuring recent evidence regarding the role of hereditary thrombophilia in PPO resulting from placenta-mediated complications such as RM. This review also took into account the benefits of thromboprophylaxis for women suffering from such complications. The authors stated that more recent studies have failed to confirm the otherwise extensively described association between hereditary thrombophilia and PPO. The authors also added that, considering current knowledge and evidence, inherited thrombophilias appear to be a weak contributor to PPO in a diverse set of underlying causes. Furthermore, it is argued that even if causation is established, effective intervention following a positive test result needs to be provided. Current thromboprophylaxy approaches, involving the administration of low-molecular-weight heparin (LMWH) or low-dose aspirin, for example, have not shown to be successful in improving pregnancy outcomes. Thus, the authors question the clinical relevance of hereditary thrombophilia testing, since this approach has, so far, not yielded a positive net effect [42].

The contradictory results of the previous sources somewhat reflect what is seen in the literature regarding this topic - some authors support a connection while others do not, and therefore do not recommend screening or thromboprophylaxis aiming to improve pregnancy outcome. Another example of the latter is shown by Pritchard *et al.* in an article reviewing the impact of classic inherited thrombophilias in RM. The authors concluded that the inability to generalize results, leading to inconsistent evidence on the matter, currently discourages anticoagulation among women with hereditary thrombophilia, hoping to improve pregnancy outcome, adding, however, that there are sufficient evidence supporting VTE prevention treatment in the same group [35]. Moreover, the lack of consensus in the definition of RM itself may

contribute to this inability to generalize results.

Hereditary thrombophilia and combined oral contraceptives (OCs) Combined OCs contain both ethinylestradiol and progestogens, and since their introduction in the market, in 1960, have been associated with an increased risk of VTE. This association is explained by the changes such OCs induce in coagulation, anticoagulation, and fibrinolysis mechanisms towards a prothrombotic state [45–47]. Moreover, the risk of suffering a thrombotic event is aggravated by the presence of inherited thrombophilic defects, namely the deficiency of AT, PC, and PS, as well as FVL and prothrombin G20210A. In fact, the World Health Organization (WHO) currently acknowledges that the intake of these contraceptives by women with the former defects implies an unacceptable health risk [45].

Vlijmen *et al.* published a recent systematic review and meta-analysis assessing the additional VTE risk conferred by combined OCs use in women with hereditary thrombophilia. Their results indicated that in the presence of severe thrombophilia, meaning carriers of natural anticoagulant deficiencies, double heterozygosity or homozygosity of FVL and prothrombin G20210A, the risk of VTE increased 7-fold. On the other hand, a nearly 6-fold increased VTE risk was estimated in the presence of mild thrombophilic defects, meaning carriers of heterozygous FVL or prothrombin G20210A. Absolute VTE risk estimates further clarified that the presence of severe thrombophilia added a considerably higher risk (4.3% – 4.6% per 100 pill-years), comparatively to the additional risk conferred by mild thrombophilia (0.49% – 2.0% per 100 pill-years). Thus, the authors concluded that the use of combined OCs is discouraged in the presence of severe thrombophilia, whereas in the presence of mild thrombophilia it can be offered to women when no other risk factors are present (family history, for instance) and alternative contraceptives are not an option [45].

Chapter 3

Methodology

3.1 Genetic variants of hereditary thrombophilia - database search

Besides the classic inherited thrombophilias previously described, several other genetic variants have been linked to hereditary thrombophilia in the literature. Thus, the first stage was to perform an extensive search for studies reporting other variants, in an attempt to list all the genetic variants associated with this condition. To do so, PubMed and Web of Science databases were the main search engines used, both platforms being widely used in research.

As it was challenging to find studies focusing explicitly on hereditary thrombophilia, the clinical manifestations linked to this condition had to become the focus instead when searching. Thus, the majority of the studies selected investigated the genetic basis of VTE (including DVT and PE) and RM. A few studies concerning IS and cerebral venous thrombosis (CVT), a rare manifestation of VTE leading to stroke, were also included. All of these studies were published from 2010 onward and were predominantly genetic association studies (case-control study design) and meta-analyses conducted in different populations.

As VTE, RM, IS, and CVT, are all conditions or events of multifactorial nature, thus not being necessarily caused by abnormalities in blood coagulation, the genetic panels featured in these studies often included genetic variants that had no known association with the coagulation process. Moreover, when reviewing each study, only genetic variants known to be related to blood coagulation were selected, hence in accordance with the definition of thrombophilia *per se*. However, some genetic variants without a role in coagulation, namely *MTHFR* genetic variants, were also included, since these have been repeatedly linked to hereditary thrombophilia in the literature, thus being noteworthy.

3.2 Statistical analysis

After listing the genetic variants associated with hereditary thrombophilia in the literature, a statistical analysis (or, in other words, a genetic association analysis) was performed using a set of internal samples made available by the company. All analyses were performed using R software (www.r-project.org). The different steps of the statistical analysis will be described in the following subsections.

3.2.1 Cohorts description

The samples used for this analysis were of patients who took Heartgenetics' TRB test and gave a research consent. Plus, only patients whose medical reasons behind taking the test were VTE (including DVT, PE, and thrombophlebitis) or RM were selected, thus creating two separate analysis cohorts - a cohort of VTE patients and a cohort of RM patients. These two traits were the focus of the analysis for being the most relevant considering the company's interest.

Both analyses focused on the Caucasian ethnicity. To do so, only samples processed in European labs were first selected, and from that set, samples with specified Caucasian ethnicity were included in the cohorts, as well as samples with unknown ethnicity - given the scarcity of samples specifying a Caucasian ethnicity, samples with unknown ethnicity were also considered as Caucasian as an approximation. Regarding gender, the VTE cohort was composed of 57 female cases and 21 male cases, whereas the RM cohort was composed of 128 female cases. In both analysis, the control samples were retrieved from the 1000 Genomes Project (www.1000genomes.org), which was the first initiative to sequence the genomes of a large group of individuals (about 2504 individuals) from different populations - African, European, East Asian, South Asian, and American. The resulting data on human variation was then made freely accessible to the worldwide scientific community. In the VTE analysis, the control group was composed of 503 Caucasian samples, 263 of which were female and 240 were male. In the RM analysis, only the 263 Caucasian female samples were used. Overall, the VTE cohort was comprised of a total of 581 samples and the RM cohort of a total of 391 samples.

3.2.2 Genetic panel

The genetic variants up for analysis included the fifteen variants featured in the TRB kit, which are listed in table 3.1. The variants presented in the table below were selected based on scientific evidence curated by the company's scientific team, as well considering experts opinions, when developing the test. These variants are mostly of genes affecting blood coagulation - variants affecting coagulation factors, fibrinogen, platelets, and natural anticoagulant proteins, for example. From this genetic panel, four variants were excluded from the analysis because there was no genotypic data on them in the 1000 Genomes Project,

resulting in null genotype counts, for the control samples, for the following variants: GP1BA CM061054, SERPINE1 rs1799889, PROS1 CD066393, and SERPINC1 rs121909564. Plus, two other variants, namely F2 rs202003146 and SERPINC1 rs121909548 were also excluded as all cases and controls were homozygous for the common or 'normal' allele - wild type - for both variants. This was also the case for GP1BA CM061054, PROS1 CD066393, and SERPINC1 rs121909564, since besides not having genotypic data for the control samples, also all case samples had the wild type genotype. Finally, the F12 rs1801020 was not considered for analysis either, since it was recently included in the genetic panel of the test and, as a result, there were very few samples for which the variant had been genotyped.

Table 3.1: Genetic panel of the TRB kit. Most variants are identified by an rsID, according to the Ensembl annotation system (www.ensembl.org).

Gene	Chromosome	SNP (rsID)	Risk Allele/Wild Type Allele
F13A1	6	rs5985	T/G
F2	11	rs202003146	A/G
F2	11	rs1799963	A/G
F5	1	rs6025	A/G
FGB	4	rs1800790	A/G
GP1BA	17	CM061054*	G/T
MTHFR	1	rs1801131	C/A
MTHFR	1	rs1801133	T/C
SERPINE1	7	rs2227631	A/G
SERPINE1	7	rs1799889	Deletion/G
PROCR	20	rs867186	G/A
PROS1	3	CD066393*	Deletion/T
SERPINC1	1	rs121909548	T/G
SERPINC1	1	rs121909564	T/C
F12	5	rs1801020	T/C

*As these variants do not have an rsID, they are identified according to The Human Gene Mutation Database (HGMD) annotation system (www.hgmd.cf.ac.uk).

3.2.3 HWE of controls

As it is an important quality control step, deviations from HWE were assessed in the control population. To do so, the R function "HWExact" was used. Thus, instead of performing a traditional chi-square test assessing the goodness-of-fit of observed genotype counts to their expected values under HWE, the "HWExact" function performs an exact test analogous to the Fisher exact test on a contingency table. Moreover, an exact HWE test is considered to have more power than a test statistic with an asymptotic chi-square distribution, since this asymptotic test can give rise to unreliable results when having small sample sizes or rare alleles [48]. Bearing in mind the relatively small number of control samples, an exact test appeared to be the better option.

A two-sided test, meaning that both the excess and scarcity of heterozygotes count as evidence against HWE, was performed for each variant. In each test, the R function "HWExact" function received as argument a vector containing the genotype counts (AA, AB, and BB) of the variant being tested in the control population. The function then returned a p-value that, if smaller than 0.05, indicated a departure from HWE.

3.2.4 Population stratification

To ensure that there was no population stratification, a clustering-based method was employed. Thus, the K-means clustering algorithm was used, which aims to find the cluster centroids that minimize the distance between data points and the nearest centroid. In the standard algorithm, having defined a number of k clusters *a priori*, as well as chosen k centroids (which may be k data points chosen randomly, for instance), each data point is assigned to the cluster whose centroid has the least squared Euclidean distance. Once all points have been assigned to a cluster, the cluster centroids are updated by computing the means of their respective elements. This step is followed by the new assignment of each data point considering the updated centroids. The cluster assignment and centroid updating steps are repeated until the assignments no longer change, meaning that the algorithm has converged [18, 49].

The goal of the K-means algorithm is to minimize the within-cluster sum of squares, given by the sum of the (squared) distances of each point to its centroid. The within-cluster sum of squares, corresponding to the algorithm's objective function, J, translates into the following expression:

$$J = \sum_{i=1}^{k} \sum_{j \in C_i} ||x_j - c_i||_2^2 \tag{3.1}$$

where x is a data point belonging to cluster C_j with centroid c_j.

The Bayesian Information Criterion (BIC) is a criterion for model selection. Given a set of models, the model with the lowest BIC value should be selected as the best model. This criterion is given by the following expression [18]:

$$BIC = ln(n)k - 2ln(\hat{L}) \tag{3.2}$$

where n is the number of observations, or the sample size, k is the number of parameters estimated by the model, and $\hat{L}$ is the maximized value of the likelihood function of a model M, $\hat{L} = p(x|\hat{\theta}, M)$. Moreover, x is the set of observations and $\hat{\theta}$ represents the parameters that maximize the likelihood

function.

Thus, this criterion allows selecting the set of parameters of the model that best fits x, hence maximizing $\hat{L}$. On the other hand, it prevents overfitting by adding a penalty term for the number of parameters in the model [18].

In this case, this criterion was used to infer the number of K clusters of the K-means algorithm, therefore indicating how many subpopulations existed in both cohorts. To do so, the algorithm was run for different K values (using the R function ”kmeans”), ranging from one to twenty, and the model whose K value had the lowest associated BIC value was selected, thus being the number of clusters that best fits the data, and consequently indicating whether or not there is population stratification.

As a pre-processing step, the genetic data was subjected to PCA using the R function ”prcomp”. This is an important step as the centroids learned by the K-means algorithm may be biased by the existence of correlations in the data [49]. To solve this problem, the data must go through a whitening transformation - a process by which a set of variables, with a known covariance matrix, is transformed into a new set of uncorrelated variables, now having the identity matrix as the covariance matrix. This transformation is achieved by performing PCA on the data, therefore being good practice. Furthermore, PCA was only used as a pre-processing step and not with the goal of selecting components [18, 49].

3.2.5 Association tests

To assess the association between the genetic panel and the two main traits of interest - VTE and RM - a Fisher exact test was performed, being again the better option comparatively to the chi-square test, given the sample size. To do so, the R function ”fisher.test” was used, which tests the null hypothesis of independence between the rows and columns of a contingency table. Thus, for each variant being tested, the function received as argument a 2$\times$3 matrix containing the genotype counts for cases and controls. Plus, a two-sided test was performed, meaning that the possibility of a relationship in both directions is tested - the ORs of disease may be smaller or greater than one.

After these first association tests, in order to determine which genetic model suited best each variant, five genetic models were tested: full (corresponding to the 2$\times$3 matrix first tested), dominant, recessive, allelic, and additive. The model’s respective contingency tables were built taking into account the information, provided by the company, regarding the risk alleles of each variant. Moreover, the R function ”fisher.test” was again applied to the new tables, except for the additive model, in which case the R function ”CochranArmitageTest” was used, which performs a Cochran-Armitage test for trend, considered a more suitable test for the additive model. A two-sided test was performed when using both functions. After testing the five genetics models for each variant, the one with the smallest associated p-value was

selected as the appropriate genetic model.

3.2.6 Model building and performance evaluation

3.2.6.1 Logistic regression model

After the association tests, a logistic regression model was built having as explanatory variables the genotypes of the final set of genetic variants:

$$logit(p) = ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_8 x_8 \tag{3.3}$$

where p is the probability of having the disease (or trait) and $x_1...x_8$ contain information concerning the genotypes of the eight SNPs included in the model - the genotypes were converted into the number of risk alleles, hence in accordance with the additive model. β_0 is the intercept term and $\beta_1...\beta_8$ represent the effects of each marker on the trait.

To train the model and estimate parameters $\beta_0...\beta_8$ for each cohort, a leave-one-out cross-validation method was employed. Cross-validation is a frequently used resampling method when building and evaluating the generalization ability of a model, that is, the model's ability to predict an outcome in new cases. Moreover, though it may become time consuming and computationally intensive when having large amounts of data, it is a useful technique when having small data sets, as it is the case for both cohorts. k-fold cross-validation implies partitioning the data set into k folds containing approximately the same number of samples and then fitting k models, each on a different training set. To do so, one fold is used for testing while the rest is used for training and the test fold rotates k times. Furthermore, the model's overall performance will be the average of the performances of the k models fitted. When k equals the total number of samples, n_s, the term used is leave-one-out cross-validation. In this case, each model is fitted using $n_s - 1$ samples as the training set, and then tested using the remaining sample [50].

For both cohorts, in each step of their respective leave-one-out cross-validation loops, the R function "glm" was used to fit a model. This function is used to fit generalized linear models, which include logistic regression models. The function receives as argument a description of the model to be fitted - a symbolic description of the model (a formula), the type of model to be fitted (in this case, a logistic regression model), and the training set, among other optional arguments. Once the model was created, the R function "predict" was used to predict the outcomes of the test set, composed by a single sample. In the end of both loops, the predictions for all the samples were stored in two separate data frames, together

with their respective true classifications (1 for case and 0 for control).

3.2.6.2 Scores model

A second model was also taken into consideration in this analysis. This model, developed internally by the company to be included in a future version of the TRB kit, includes all fifteen genetic variants of table 3.1. Based on a subject's genotypes for the fifteen variants included, the model returns a risk score which represents the overall, genetically determined, susceptibility to thrombotic events. This risk is graphically represented through a risk bar featured in the test's report. Since the model described returns a risk score, it will, from this point forward, be referred to as the scores model.

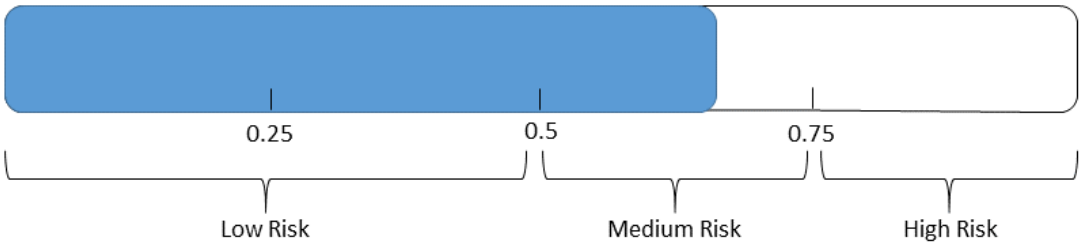

Figure 3.1: Graphical representation resembling the risk bar to be featured in the future version of the TRB kit's report.

Given that the scores model is completely specified, thus not requiring any training to fit model parameters, it was readily used to obtain the risk scores of all the samples of both cohorts. These risk scores were then stored in two separate data frames together with their respective true classifications. It is noteworthy that, although the model originally includes fifteen genetic variants, two of them - SERPINE1 rs1799889 and F12 rs1801020 - were excluded from the following steps for the reasons already explained in subsection 3.2.2.

3.2.6.3 Performance comparison

To compare the performance of both models in the VTE and RM cohorts, a receiver operating characteristics (ROC) curve was used. ROC curves are a useful way of visualizing a model's performance - the x axis of a ROC curve represents the false positive rate (FP_{rate}, given by the number of negatives incorrectly classified divided by the total number of negatives, being in turn the sum of the true negatives and the false positives). On the other hand, the y axis depicts the true positive rate (TP_{rate}, given by the number of positives correctly classified divided by the number of total positives, being in turn the sum of the true positives and false negatives). TP_{rate} is also known as sensitivity, while specificity is given by the number of true negatives divided by the total number of negatives, therefore being $1 - FP_{rate}$ [51].

ROC curves usually feature a diagonal line, $y = x$, representing the random guessing of a class - random performance. For instance, if a model or classifier randomly guesses the positive class half the time, it is

expected to guess correctly half the positives (TP_{rate}) and incorrectly half the negatives (FP_{rate}), thus corresponding to the point (0.5, 0.5) in the ROC curve. Also, if it randomly guesses the positive class 70% of the time, it is expected to correctly classify 70% of the positives (TP_{rate}) and incorrectly 70% of the negatives (FP_{rate}), therefore yielding the point (0.7, 0.7). Thus, a model with a random performance will be represented by a ROC curve whose points slide back and forth in the diagonal line. On the other hand, a better model will have its ROC curve in the upper triangular region - lower FP_{rate} and higher TP_{rate}. Another important parameter of ROC curves is the area under the curve (AUC), which provides a quantitative measure of a model's performance - the closer to one the better the performance [51].

Since both models return a score (or a probability, in the case of the logistic regression model) and not a discrete binary response, a ROC curve is obtained by first sorting the scores in a decreasing manner. Then each score is compared to a different threshold that also decreases as each instance is processed. If the score is greater than the current threshold, it is classified as a positive and otherwise as a negative. By comparing the predicted classifications with their respective true classifications, it is possible to update the TP_{rate} and FP_{rate}, and therefore obtain the points of the ROC curve [51]. To build the ROC curves, the R function "roc" was used, which essentially receives the model predictions (scores, later sorted) and corresponding true classifications to build a ROC curve from them.

Once the ROC curves were built, the R function "roc.test" was used in order to compare both curves in terms of their AUCs. This function performs a statistical test to assess if the AUCs of the curves being tested are significantly different or not, which is indicated by the p-value returned. A two-sided test was performed, meaning that the alternative hypothesis included the AUC of the first ROC curve being significantly smaller or greater than the AUC of the second ROC curve. Furthermore, the function being used employs a bootstrap method, in which n replicates or permutations are drawn from the original data and, for each replicate, the AUC of both curves is computed and the difference between them is stored. Then, the difference between the original AUCs of the two ROC curves is divided by the standard deviation of the bootstrap differences and the resulting statistic is compared to the normal distribution.

3.3 Development of a new value proposition

Based on the results of the database search and of the statistical analysis, a number of alterations to the current genetic panel of the TRB kit was suggested. A resulting new set of genetic variants was then proposed as the genetic panel of the future version of the test.

3.3.1 Scores model impact assessment

The previous step was followed by the assessment of the impact of the new genetic panel on the scores model.

Since the new genetic panel included the addition of a new variant that had not been genotyped in the case samples of both cohorts, it had to be simulated. Thus, as there were genotypic data on the variant available in the 1000 Genomes Project, the genotypes of the control samples were retrieved from its database. Given that the genetic model chosen for the variant added was the additive genetic model, these genotypes were then converted to the number of risk alleles - 0, 1, or 2. Moreover, only the genotypes of the case samples were simulated. To do so, a number of wildtype (0), heterozygous (1), and homozygous (2) genotypes were generated and randomly attributed to the case samples so that the total number of cases was in accordance with the variant's genotype frequencies indicated in the 1000 Genomes Project - 60.8% wild type, 34.2% heterozygous, and 5% homozygous - similarly to the controls. Considering, for example, the 78 cases of the VTE cohort, $0.608 \times 78 \approx 47$ (0) samples, $0.342 \times 78 \approx 27$ (1) samples, and $0.05 \times 78 \approx 4$ (2) samples were created. The same procedure was followed in the RM cohort.

The assessment of the impact of the new panel on the final risk scores produced was performed by computing the relative difference (%), for each sample, between the risk score produced by the model featuring the new panel and the risk score resulting from the original panel - $RelativeDifference(\%) = \frac{NewRiskScore - OldRiskScore}{OldRiskScore} \times 100$. As the cases for the new variant were simulated, the procedure involving randomly attributing (0), (1), and (2) to the samples and computing the relative differences (%) was repeated 100 times. Then, for each sample, the mean of the resulting 100 relative differences (%) was computed. The means of the relative differences (%) for the low risk samples - score equal or smaller than 0.45 - medium risk samples - score greater than 0.5 and equal or smaller than 0.7 - and high risk samples - score greater than 0.75 - were represented in separate histograms.

Chapter 4

Results

4.1 Genetic variants of hereditary thrombophilia - database search

From the set of 27 studies selected, a total of 55 genetic variants were reported. Detailed information about the studies reviewed, the genetic variants, and their respective genes are available in Appendix A.

The variants reported are of genes predominantly encoding proteins related to blood coagulation, such as anticoagulant proteins, coagulation factors, subunits of fibrinogen, among others. Variants of other potentially interesting genes with a role in coagulation - for instance, *TFPI* or the von Willebrand factor (*vWF*) gene, which encodes the vWF protein that binds to factor VIII (FVIII) while inactive and also plays an important role in platelet adhesion - were not reported due to the lack of relevant sources found.

The list of variants reported includes, as expected, well-established genetic defects. Finding sources reporting genetic variants of genes encoding natural anticoagulants, or leading to their deficiency, posed some challenges, which is actually addressed in some of the few studies found. According to Gindele *et al.*, AT deficiency belongs to the group of rare diseases, therefore making hardly possible to conduct large clinical studies concerning it. [52] Also Di Minno *et al.* stated that despite the association between these inherited deficiencies and VTE being widely recognized, meta-analytical data providing overall info about the risk of VTE is lacking. [53] Thus, when searching for studies, most of those addressing natural anticoagulant deficiencies assessed such conditions through the measurement of the serum levels and activities of these proteins, instead of performing genetic analyses. On the other hand, most studies performing genetic analyses presented results highlighting the identification of a new mutation or SNP in one family, or in a small group of families, rather than involving a large group of individuals. On the contrary, as expected, several sources reported the well-known FVL (F5 rs6025) and prothrombin

G20210A (F2 rs1799963) mutations, given their higher prevalence, compared to natural anticoagulant deficiencies, and their status as two of the most frequently assessed thrombophilic variants.

4.1.1 *ABO* genetic variants

Even though the *ABO* gene does not exert a direct effect in blood coagulation, thirteen *ABO* genetic variants were still included in the list of thrombophilic variants. The *ABO* gene encodes proteins related to the ABO blood group system. The ABO blood group system classifies human blood depending on the presence or absence of A and B antigens on the surface of red blood cell membranes. Thus, an individual may have type A, type B, type O, or type AB blood. Besides antigens A and B, antigen H also belongs to the group of ABO antigens. These molecules are complex carbohydrates that are widely expressed in human tissues and cells besides red blood cells. While the codominant A and B alleles of the ABO locus encode glycosyltransferases that convert the common precursor, the H determinant, into A or B antigens, the recessive O allele does not encode a functional enzyme, causing OO carriers to keep the unaltered H structure [24].

The presence of ABH antigenic structures on circulating vWF, which influences the molecule's half-life in plasma, explains the significant role of the ABO blood group system in hemostasis. This differentiated expression causes the vWF plasma half-life to be 10 hours for group O, whereas for non-O groups it rises to 25 hours. As a result, individuals of non-O blood groups have vWF levels, and consequently FVIII levels, approximately 25% higher when compared to those of O blood group individuals [24]. Given that increased vWF and FVIII levels are considered important risk factors for the development of thrombotic events, in particular VTE, non-O blood type has been linked to the development of this condition. In fact, non-O blood type has been shown to increase the risk of VTE by approximately 2-fold and has also been associated with recurrence. Comparatively to the main inherited thrombophilia markers, this marker carries a lower associated risk and is much more frequent in the Caucasian population (55-57%) [24, 54].

The list of variants of the *ABO* gene includes variants related to blood type, such as the ABO rs8176719 variant, whose C allele is linked to the non-O blood group [55], as well as other variants whose association with VTE is independent of blood type, as it is the case of the ABO rs2519093 variant, therefore suggesting that there may be other mechanisms underlying the connection between *ABO* genetic variants and VTE besides blood type.

4.1.2 *MTHFR* genetic variants

Folate, the salt of folic acid, is an essential vitamin for the body's daily functions. Folate deficiency can lead to gastrointestinal lesions, anaemia, and poor growth, among other conditions, and is required in higher doses during pregnancy to prevent the formation of neural tube defects [56].

The primary form of circulating folate in the blood, 5-methyltetrahydrofolate, is converted from its precursor by the MTHFR enzyme. 5-methyltetrahydrofolate is then involved in the remethylation of homocysteine to methionine, which, in turn, is converted to S-adenosylmethionine. S-adenosylmethionine acts as a methyl donor in several cellular reactions in the body [56, 57].

The MTHFR enzyme is encoded by the *MTHFR* gene, which has two main polymorphisms, namely MTHFR C677T (rs1801133) and MTHFR A1298C (rs1801131), both reported following the database search. These variants have repeatedly been associated in the literature with hyperhomocysteinemia, a condition characterized by elevated plasma levels of homocysteine. MTHFR C677T (rs1801133) in homozygosity is said to lead to a decrease in MTHFR enzyme activity. On the other hand, MTHFR A1298C (rs1801131) alone does not lead to hyperhomocysteinemia, regardless of being in heterozyosity or homozygosity, though it may affect enzyme activity when inherited with MTHFR C677T (rs1801133). Thus, hyperhomocysteinemia occurs because a reduction in enzyme activity disrupts the balance of the conversion of homocysteine to methionine [57].

Hyperhomocysteinemia may have different consequences, such as abnormalities in cellular proliferation or apoptosis, an increase in oxidative stress, and an impact in thrombus formation. The latter occurs primarily because elevated levels of homocysteine in the blood may cause vascular endothelial injury, one of the components of Virchow's triad that explain the etiology of thrombosis. In other words, the tendency of homocysteine to promote cellular toxicity through oxidative stress contributes to the onset of an endothelial lesion [58]. This explains the relevance of *MTHFR* genetics variants in hereditary thrombophilia, despite the *MTFHR* gene not having a specific role in coagulation *per se*.

4.2 Statistical analysis

The results of the statistical analysis, featuring the eight remaining genetic variants of the TRB kit, will be presented in the following subsections.

4.2.1 HWE of controls

The p-values resulting from the HWE tests performed for each variant are presented in the following table:

Table 4.1: Results of the 'HWExact' test for testing for HWE deviations in the control population.

Gene	SNP	'HWExact' p-value
F13A1	rs5985	0.39
F2	rs1799963	1
F5	rs6025	0.064
FGB	rs1800790	0.0035
MTHFR	rs1801131	0.76
MTHFR	rs1801133	0.85
SERPINE1	rs2227631	0.78
PROCR	rs867186	0.088

The results of the table above indicate that all variants, except for FGB rs1800790, are in HWE, as the p-value of the test is greater than 0.05.

4.2.2 Population stratification

Figures 4.1 and 4.2 represent the genetic data of both cohorts projected on their respective first two principal components resulting from the PCA:

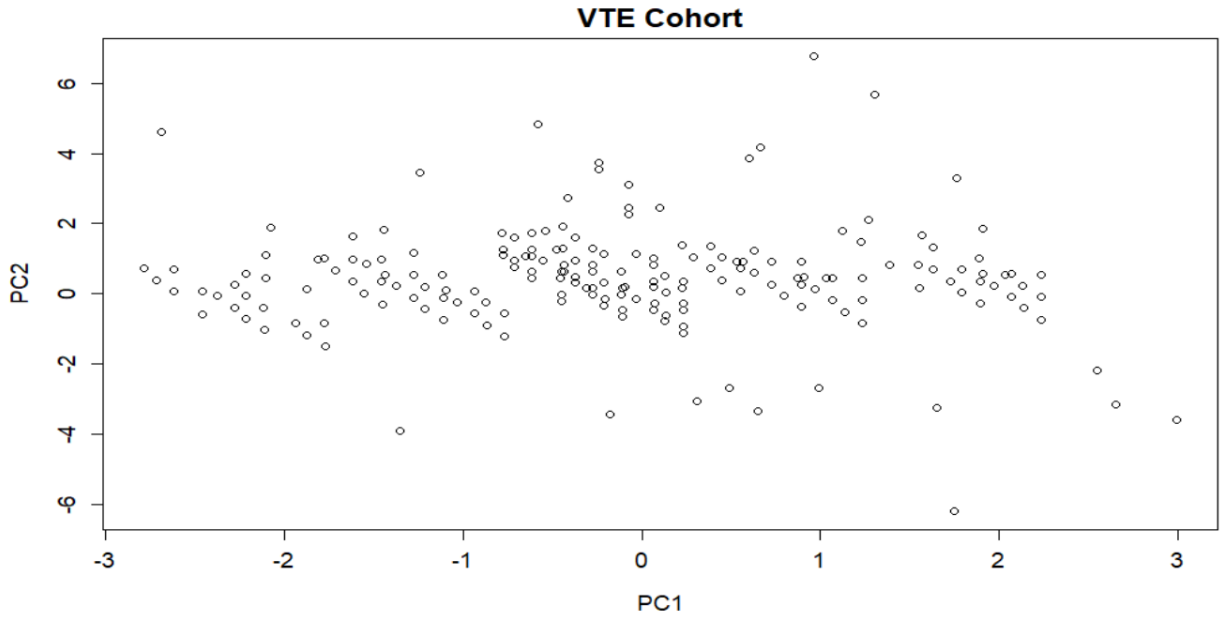

Figure 4.1: Graphical representation of the projection of the genetic data, from the VTE cohort, in the first two principal components.

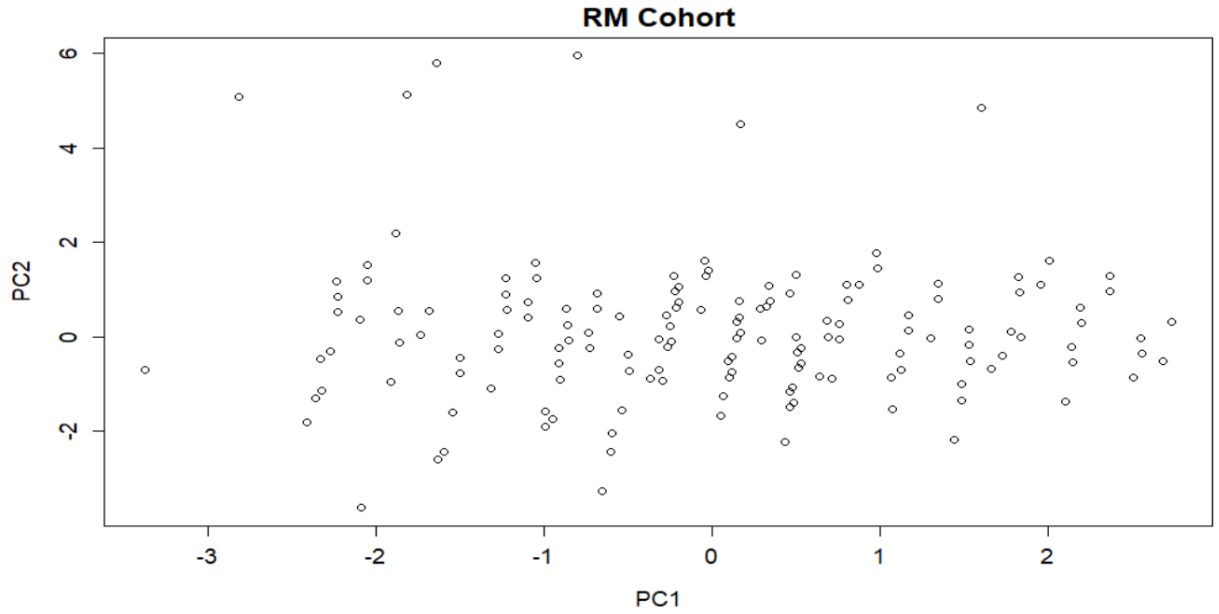

Figure 4.2: Graphical representation of the projection of the genetic data, from the RM cohort, in the first two principal components.

From the figures above, one can observe that that the data projected doesn't look completely homogeneous and that is even possible to identify some clusters, particularly in 4.1. This could indicate the existence of population stratification. However, the lowest BIC value obtained for both cohorts correspond to $K = 1$, therefore confirming that there is only one population in both cohorts.

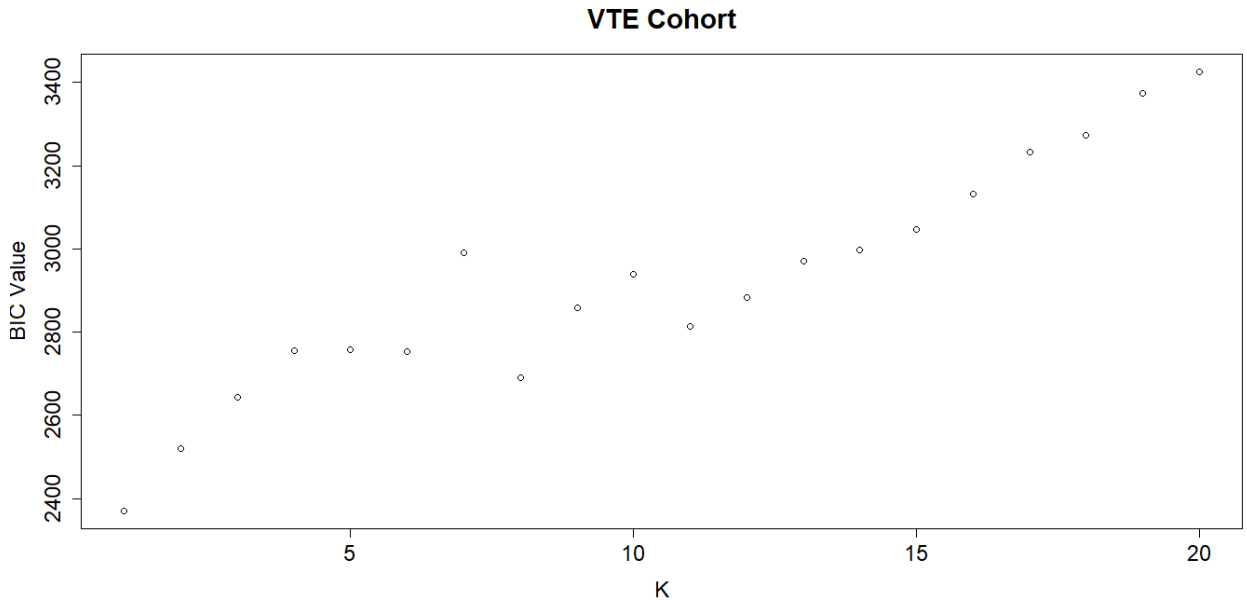

Figure 4.3: BIC values for each K tested, ranging from one to twenty, for the K-means algorithm in the VTE cohort.

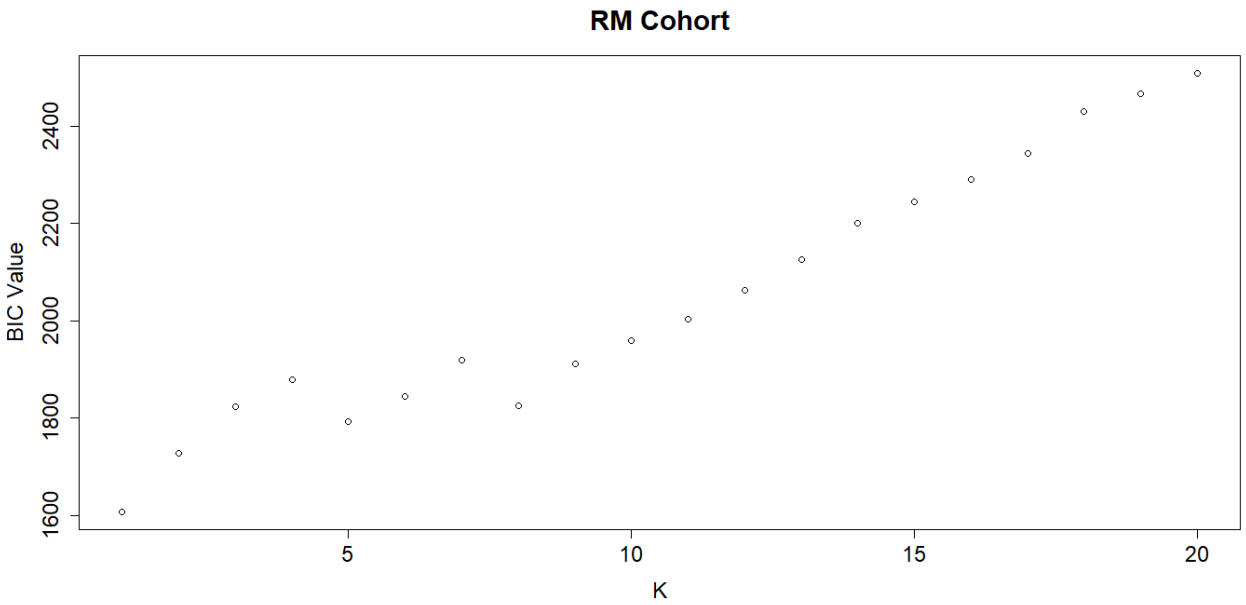

Figure 4.4: BIC values for each K tested, ranging from one to twenty, for the K-means algorithm in the RM cohort.

The figures above indicate that the model that best fits the data, in both cohorts, is the model whose K equals 1.

4.2.3 Association tests

Confidential content not presented in this version.

4.2.4 Models' performance evaluation

Confidential content not presented in this version.

4.3 Development of a new value proposition

Confidential content not presented in this version.

4.4 Impact of the new genetic panel proposed in the scores model

Confidential content not presented in this version.

Chapter 5

Discussion and conclusions

Hereditary thrombophilia is a condition with potentially serious health consequences, therefore being of relevance to study its genetic basis in order to come up with tools, such as Heartgenetics' TRB kit, that help estimate an individual's genetic predisposition to conditions such as VTE.

The aim of this dissertation was essentially to develop a new value proposition for the TRB kit, a genetic test that provides an estimate of the genetic predisposition to thrombotic events, based on genotypic data of set of genetic markers.

The first step was to attempt to list all the genetic variants involved in this condition. As a result, 55 genetic variants were reported, the majority of them associated with the two main traits of interest mentioned throughout this work, namely VTE and RM. Despite the large number of variants reported, it is worth noting that variants of other potentially interesting genes with a role in coagulation, such as *TFPI* or the (*vWF*) gene, for example, are not part of the list due to the lack of relevant sources found.

The second step was to perform a statistical analysis, or a genetic association analysis, involving eight genetic variants of the test's current panel, in order to investigate the association between that set of variants and the two traits of interest, VTE and RM.

When testing for deviations from HWE in the control group, it was observed that only FGB rs1800790 appeared not to be in HWE. The HWE test tends to be sensitive to the existence of homozygous individuals for rare alleles, which are unlikely to exist in a single, randomly mating population [59]. This may explain the low p-value obtained for F5 rs6025, though still under HWE. However, this is not the case for FGB rs1800790, since the variant has a MAF of 22% in the European population, according to the information available in the 1000 Genomes Project . Since it was demonstrated that there was no

population stratification in both cohorts, the low p-value obtained for FGB rs1800790 may possibly be due to low levels of genotyping errors, since no genetic data set is completely free from errors [59]. Thus, the significant deviation from HWE observed can be considered as an artifact with no consequences in the interpretation of the results of the association tests involving this variant.

The results obtained in the population stratification analysis appeared to be contradictory at first, given the lack of homogeneity in the graphical representation of the genetic data projected in the first two principal components, and the optimal number of clusters being one for both cohorts, according to BIC. Thus, the clustering shown in figures 4.1 and 4.2 may be due to the presence of two variants slightly correlated, namely MTHFR 1801131 and MTHFR 1801133 - $R^2 \approx 0.26$, according to LDMatrix, a tool belonging to LDlink, a set of online tools freely accessible to assess LD between genetic variants (ldlink.nci.nih.gov). Even though these two variants are in low LD, it could still be enough to cause PCA to overemphasize their contribution, hence explaining the lack of homogeneity in figures 4.1 and 4.2. Plus, MTHFR rs1801133 has the highest value in the first principal component, in both cohorts. Hence, the clustering in figures 4.1 and 4.2 can also be interpreted as an artifact, rather than a reflection of population structure.

Based on the results of the database search and of the statistical analysis, as well as taking into account the input from the company's scientific team, the third step was to suggest alterations to the TRB kit's current genetic panel, which included the addition of a new variant.

The fourth and final step was to assess the impact of the panel being proposed on the TRB kit's risk prediction model. When analyzing the mean relative differences between the final risk scores produced by both panels, it was observed that the alterations proposed did not impact the risk scores produced by the model severely, as expected, given the small magnitude of the alterations proposed.

This work has some limitations that are worth addressing. The small sample size of both cohorts is one of them and certainly had an impact on the overall results of the statistical analysis. Plus, two relevant variants - SERPINE1 rs1799889 and F12 rs1801020 - had to be excluded from the statistical analysis. It would have been interesting to see their respective results in the association tests, or how the affected the logistic regression model and the scores model.

To conclude, the results of this work represent interesting contributions to the continuous development of a new solid value proposition for the TRB kit.

5.1 Future work

Confidential content not presented in this version.

Bibliography

[1] John A. Heit *et al.* The epidemiology of venous thromboembolism. *J Thromb Thrombolysis*, 41, 2016.

[2] Andrew J. Gale. Current understanding of hemostasis. *Toxicol Pathol.*, 39(1), 2011.

[3] Ewelina M. Golebiewska *et al.* Platelet secretion: From haemostasis to wound healing and beyond. *Blood Reviews*, 29, 2015.

[4] Fatemeh Moheimani *et al.* Venous thromboembolism: Classification, risk factors, diagnosis, and management. *ISRN Hematology*, 2011.

[5] Liang Tang *et al.* Ethnic diversity in the genetics of venous thromboembolism. *Thrombosis and Haemostasis*, 2015.

[6] Help me understand genetics. cells and dna. Genetics Home Reference, May 2018.

[7] Kathleen M. Murphy *et al. Handbook of Pharmacogenomics and Stratified Medicine*, chapter The Human Genome, Gene Regulation, and Genomic Variation, pages 41–56. Elsevier, 2014.

[8] Jahad Alghamdi*et al. Handbook of Pharmacogenomics and Stratified Medicine*, chapter Fundamentals of Complex Trait Genetics and Association Studies, pages 235–256. Elsevier, 2014.

[9] R. Karki *et al.* Defining 'mutation' and 'polymorphism' in the era of personal genomics. *BMC Med. Genomics*, 8(1):1–7, 2015.

[10] Andrea S. Foulkes. *Applied Statistical Genetics with R For Population-based Association Studies*. Springer, 2009.

[11] Cathryn M. Lewis *et al.* Introduction to genetic association studies. *Cold Spring Harbor Laboratory Press*, 2012.

[12] Sam Behjati *et al.* What is next generation sequencing? *Arch Dis Child Educ Pract Ed*, 98:236–238, 2013.

[13] Cathryn M. Lewis. Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, 3(2):146–153, 2002.

[14] William S. Bush *et al.* Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, 8, 2012.

[15] Kim Hae-Young. Statistical notes for clinical researchers: Chi-squared test and fisher's exact test. *Restorative Dentistry and Endodontics*, 2017.

[16] Mansi Ghodsi *et al.* An enhanced version of cochran-armitage trend test for genome-wide association studies. *Meta Gene*, (9):225–229, 2016.

[17] Alkes L Price *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 2006.

[18] Chih Lee *et al.* Pca-based population structure inference with generic clustering algorithms. *BMC Bioinformatics*, 10, 2009.

[19] Magdalena Szumilas. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*, (19), 2010.

[20] Anthony J. Viera. Odds ratios and risk ratios: What's the difference and why does it matter? *Southern Medical Journal*, 101(7), 2008.

[21] Eun Pyo Hong *et al.* Sample size and statistical power calculation in genetic association studies. *Genomics Informatics*, 10(2):117–122, 2012.

[22] Atefeh Namipashaki *et al.* The essentiality of reporting hardy-weinberg equilibrium calculations in population-based genetic association studies. *Cell J*, 17(2), 2015.

[23] S. Khan *et al.* Hereditary thrombophilia. *Thromb. J.*, 4:1–17, 2006.

[24] P. M. Mannucci *et al.* Classic thrombophilic gene variants. *Thromb. Haemost.*, 2015.

[25] M. B. Fallon *et al.* Inherited thrombophilia and the risk of portal vein thrombosis: Progress towards individualized anticoagulation in cirrhosis? *Clinical Gastroenterology and Hepatology*, 12:1813–1814, 2014.

[26] J. Wang *et al.* Association between the plasminogen activator inhibitor-1 4g/5g polymorphism and risk of venous thromboembolism: A meta-analysis. *Thromb. Res.*, 134(6):1241–1248, 2014.

[27] V. De Stefano *et al.* Inherited thrombophilia: Pathogenesis, clinical syndromes, and management. *J Am Soc Hematol.*, 87(9), 1996.

[28] F. Sanchis-Gomar *et al.* Epidemiology of coronary heart disease and acute coronary syndrome. *Ann. Transl. Med.*, 4(13), 2016.

[29] B. K. Mahmoodi *et al.* Interaction of hereditary thrombophilia and traditional cardiovascular risk factors on the risk of arterial thromboembolism. *Circ. Cardiovasc. Genet.*, 9(1):79–85, 2016.

[30] D. J. Ye Z *et al.* Seven haemostatic gene polymorphisms in coronary disease: meta-analysis of 66 155 cases and 91 307 controls. *Elsevier*, 367, 2006.

[31] Dragoni F. Inherited and acquired thrombophilic factors in young patients with acute coronary syndrome or ischemic stroke. *Austin J Clin Cardiol.*, 2(2), 2015.

[32] S. H. Pahus *et al.* Thrombophilia testing in young patients with ischemic stroke. *Thromb. Res.*, 2016.

[33] K. W. P. Ng *et al.* Role of investigating thrombophilic disorders in young stroke. *Stroke Res. Treat.*, 2011.

[34] D. Green. Thrombophilia and stroke. *Top. Stroke Rehabil.*, 2003.

[35] Ashley M. Pritchard *et al.* Hereditary thrombophilia and recurrent pregnancy loss. *Clinical Obstetrics and Gynecology*, 59(3), 2016.

[36] M. Chatzidimitriou *et al.* Thrombophilic gene polymorphisms and recurrent pregnancy loss in greek women. *Int. J. Lab. Hematol.*, 2017.

[37] F. N. Croles *et al.* Pregnancy, thrombophilia, and the risk of a first venous thrombosis: systematic review and bayesian meta-analysis. *BMJ*, 2017.

[38] S. M. Bates *et al.* Guidance for the treatment and prevention of obstetric-associated venous thromboembolism. *J. Thromb. Thrombolysis*, 41(1):92–128, 2016.

[39] C. Karadag *et al.* Obstetric outcomes of recurrent pregnancy loss patients diagnosed wıth inherited thrombophilia. *Ir J Med Sci*, 186, 2017.

[40] Ana-Luisa Stefanski *et al.* Maternal thrombophilia and recurrent miscarriage – is there evidence that heparin is indicated as prophylaxis against recurrence? *GebFra Science*, 78, 2018.

[41] N. Roozbeh *et al.* Potential role of factor v leiden mutation in adverse pregnancy outcomes: An updated systematic review. *Biomed Res Ther*, 4(12):1832–1846, 2017.

[42] J. Unterscheider *et al.* The role of thrombophilia testing in women with adverse pregnancy outcomes. *Obstet. Gynaecol.*, 2017.

[43] Nadja Bogdanova *et al.* Hereditary thrombophilic risk factors for recurrent pregnancy loss. *J Community Genet*, 1, 2010.

[44] Hui Gao *et al.* Prothrombin g20210a mutation is associated with recurrent pregnancy loss: A systematic review and meta-analysis update. *Thrombosis Research*, 135, 2015.

[45] E. F. W. van Vlijmen *et al.* Combined oral contraceptives, thrombophilia and the risk of venous thromboembolism: a systematic review and meta-analysis. *J. Thromb. Haemost.*, 14(7):1393–1403, 2016.

[46] O. Lidegaard *et al.* Hormonal contraception and risk of venous thromboembolism: national follow-up study. *BMJ*, 339, 2009.

[47] A. Van Hylckama Vlieg *et al.* The venous thrombotic risk of oral contraceptives, effects of oestrogen dose and progestogen type: Results of the mega case-control study. *BMJ*, 339(7720):561, 2009.

[48] William R. Engels. Exact tests for hardy–weinberg proportions. *Genetics*, 183, 2009.

[49] Adam Coates *et al. Neural Networks: Tricks of the Trade*, chapter Learning Feature Representations with K-means. Springer, 2012.

[50] Rosa J. Meijer *et al.* Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155, 2013.

[51] Tom Fawcett. An introduction to roc analysis. *Elsevier*, 2005.

[52] R. Gindele *et al.* Clinical and laboratory characteristics of antithrombin deficiencies: A large cohort study from a single diagnostic center. *Thromb. Res.*, 160:119–128, 2017.

[53] M. N. D. Di Minno *et al.* Natural anticoagulants deficiency and the risk of venous thromboembolism: A meta-analysis of observational studies. *Thromb. Res.*, 135(5):923–932, 2015.

[54] M. F. Francesco Dentali *et al.* Non-o blood type is the commonest genetic risk factor for vte: Results from a meta-analysis of the literature. *Thromb. Haemost.*, 2012.

[55] Licínio Manco *et al.* Venous thromboembolism risk associated with abo, f11 and fgg loci. *Blood Coagulation and Fibrinolysis*, 0(0), 2018.

[56] S. Long *et al.* Mthfr genetic testing: Controversy and clinical implications. *AFP*, 45(4), 2016.

[57] R. Stefanov *et al.* Application of single-nucleotide polymorphism-related risk estimates in identification of increased genetic susceptibility to cardiovascular diseases: A literature review. *Frontiers in Public Health*, 5, 2018.

[58] Gilberto Vizcaíno *et al.* Homocisteína: bases genéticas y sus implicaciones cardiovasculares y cognitivas como factor de riesgo. *Invest Clin*, 58(4), 2017.

[59] Phillip A. Morin *et al.* Significant deviations from hardy–weinberg equilibrium caused by low levels of microsatellite genotyping errors. *Molecular Ecology Resources*, 9, 2009.

[60] W. Zeng *et al.* Recurrent mutations in a serpinc1 hotspot associate with venous thrombosis without apparent antithrombin deficiency. *Oncotarget*, 8(48):84417–84425, 2017.

[61] J. Navarro-Fernández *et al.* Antithrombin dublin (p.val30glu): A relatively common variant with moderate thrombosis risk of causing transient antithrombin deficiency. *Thromb. Haemost.*, 116(1): 146–154, 2016.

[62] R. Daneshjou *et al.* Population-specific single-nucleotide polymorphism confers increased risk of venous thromboembolism in african americans. *Mol. Genet. Genomic Med.*, 4(5), 2016.

[63] M. Bruzelius *et al.* Predicting venous thrombosis in women using a combination of genetic markers and clinical risk factors. *J. Thromb. Haemost.*, 13(2), 2015.

[64] Z. Xu *et al.* Polymorphisms of f2, proc, proz, and f13a1 genes are associated with recurrent spontaneous abortion in chinese han women. *Clinical and Applied Thrombosis/Hemostasis*, 2018.

[65] Zengliang Wang *et al.* Genetic association of proc variants with pulmonary embolism in northern chinese han population. *Springer Plus*, 5(147), 2016.

[66] David A. Hinds *et al.* Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. *Hum. Mol.Genet.*, 25(9):1867–1874, 2016.

[67] M. Germain *et al.* Meta-analysis of 65,734 individuals identifies tspan15 and slc44a2 as two susceptibility loci for venous thromboembolism. *Am. J. Hum. Genet.*, 96(4):532–542, 2015.

[68] J. A. Heit *et al.* Genetic variation within the anticoagulant, procoagulant, fibrinolytic and innate immunity pathways as risk factors for venous thromboembolism. *J. Thromb. Haemost.*, 9(6):1133–1142, 2011.

[69] D. Klarin *et al.* Genetic analysis of venous thromboembolism in uk biobank identifies the zfpm2 locus and implicates obesity as a causal risk factor. *Circ. Cardiovasc. Genet.*, 10(2), 2017.

[70] H. de Haan *et al.* Genome-wide association study identifies a novel genetic risk factor for recurrent venous thrombosis. *Circ Genom Precis Med.*, 11, 2018.

[71] X. Shi *et al.* Maternal genetic polymorphisms and unexplained recurrent miscarriage: a systematic review and meta-analysis. *Clin. Genet.*, 91(2):265–284, 2017.

[72] Kamelia Farahmand *et al.* Thrombophilic genes alterations as risk factor for recurrent pregnancy loss. *J Matern Fetal Neonatal Med*, 2015.

[73] Thomas Marjot *et al.* Genes associated with adult cerebral venous thrombosis. *Stroke*, 42, 2011.

[74] Mandy N. Lauw *et al.* Cerebral venous thrombosis and thrombophilia: A systematic review and meta-analysis. *Semin Thromb Hemost*, 39, 2013.

[75] Marine Germain *et al.* Genetics of venous thrombosis: Insights from a new genome wide association study. *PLoS One*, 6(9), 2011.

[76] Baijia Jiang *et al.* Prothrombin g20210a mutation is associated with young-onset stroke. the genetics of early-onset stroke study and meta-analysis. *Stroke*, 45, 2014.

[77] R. Malik *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.*, 14:1–14, 2018.

[78] Frances M. K. Williams *et al.* Ischemic stroke is associated with the abo locus: The euroclot study. *ANN NEUROL*, 73, 2012.

[79] Regina Komsa-Penkova *et al.* Rs5918 itgb3 polymorphism, smoking, and bmi as risk factors for early onset and recurrence of dvt in young women. *Clinical and Applied Thrombosis/Hemostasis*, 23 (6), 2017.

[80] Mahmood Jeddi-Tehrani *et al.* Analysis of plasminogen activator inhibitor-1, integrin beta3, beta fibrinogen, and methylenetetrahydrofolate reductase polymorphisms in iranian women with recurrent pregnancy loss. *American Journal of Reproductive Immunology*, 66, 2011.

[81] M. D. Salazar Garcia *et al.* Plasminogen activator inhibitor-1 4g/5g polymorphism is associated with reproductive failure: Metabolic, hormonal, and immune profiles. *Am. J. Reprod. Immunol.*, 76(1): 70–81, 2016.

[82] V. Rai. Methylenetetrahydrofolate reductase c677t polymorphism and recurrent pregnancy loss risk in asian population: A meta-analysis. *Indian J. Clin. Biochem.*, 2016.

[83] Yunlei Cao *et al.* Association study between methylenetetrahydrofolate reductase polymorphisms and unexplained recurrent pregnancy loss: A meta-analysis. *Gene*, 514, 2012.

[84] Y. Cao *et al.* The association of idiopathic recurrent pregnancy loss with polymorphisms in hemostasis-related genes. *Gene*, 530(2):248–252, 2013.

[85] W. Cohen *et al.* Risk assessment of venous thrombosis in families with known hereditary thrombophilia: The marseilles-nimes prediction model. *J. Thromb. Haemost.*, 12(2):138–146, 2014.

Appendix A

Supplementary material

The tables presented in the following sections, namely sections A.1, A.2, and A.3, contain the data collected about the 55 thrombophilic variants selected after the database search. Plus, section A.4 contains more detailed information about the variants' respective genes, and the studies from which the data of the previous tables was collected.

A.1 Natural anticoagulants

Table A.1: Genetic variants found for the *SERPINC1* gene (chr 1).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs201381904	Missense	VTE	Chinese	nr	nr	13.6(1.7-107.1)	0.013	Logistic regression; genetic model nr	[60]
rs2227624	Missense	VT	European	T(0.001)	-	2.94(1.07–8.09)	0.037	Logistic regression; genetic model nr	[61]

Table A.2: Genetic variants found for the *PROS1* gene (chr 3).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs138925964	Missense	VTE	African Americans	T(nr)	-	4.62(1.51-15.20)	0.0041	Logistic regression; additive genetic model	[62]

Table A.3: Genetic variants found for the *PROC* gene (chr 2).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs1799810	5'-UTR	VTE	European	nr	nr	1.15 (1.03–1.28)	0.01	Logistic regression; additive genetic model	[63]
rs2069906	Intronic	RM	Chinese Han	nr	nr	0.114 (0.014-0.902)	0.021	Chi-square test; recessive genetic model	[64]
rs199469469	Inframe deletion	PE	Northern Chinese Han	-	del(nr)	5.34(1.47–19.39)	0.011	Logistic regression; dominant genetic model	[65]

Table A.4: Genetic variants found for the *PROCR* gene (chr 20).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs34234989	Intronic	VTE	European	-	I(0.7148)	0.885 (0.849, 0.922)	$6.7x10^{-9}$	Logistic regression; additive genetic model	[66]*
rs6087685	Intronic	VTE	European	-	C(0.302)	1.15 (1.10–1.21)	$1.65x10^{-8}$	Logistic regression; additive genetic model	[67]

A.2 Coagulation factors

Table A.5: Genetic variants found for the *F5* gene (chr 1).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs6025	Missense	VTE	European	A (0.066)	-	3.40 (2.65–4.35)	$3.07x10^{-22}$	Logistic regression; additive genetic model	[68]
rs6025	Missense	VTE	European	nr	nr	3.51 (2.77–4.45)	$4.4x10^{-25}$	Logistic regression; additive genetic model	[63]
rs6025	Missense	VTE	European	-	T(0.0253)	2.927 (2.715, 3.154)	$3.6x10^{-137}$	Logistic regression; additive genetic model	[66]*
rs6025	Missense	VTE	European	-	T (0.02)	3.5 (2.96–4.11)	$7.10x10^{-50}$	Logistic regression; additive genetic model	[69]
rs6025	Missense	VTE	European	-	T(0.033)	3.25 (2.91–3.64)	$1.10x10^{-96}$	Logistic regression; additive genetic model	[67]
rs6025	Missense	Recurrent VT	European	T(0.096)	-	2.4 (1.75–3.15)	$1.29x10^{-8}$	Logistic regression; additive genetic model	[70]
rs6025	Missense	RM	Overall**/ Caucasian	nr	nr	5.20 (2.81–9.64)	<0.00001	Random-effects meta-analysis; recessive genetic model	[71]
rs6025	Missense	RM	Iranian	-	A(nr)	3.15 (1.5-6.59)	$2.00x10^{-3}$	Chi-square test; genetic model nr	[72]
rs6025	Missense	CVT	European	-	A(nr)	2.4 (1.75-3.30)	<0.00001	Random-effects meta-analysis; dominant genetic model	[73]
rs6025	Missense	CVT	Mostly mixed	nr	nr	2.89 (2.10-3.97)	<0.001	Random-effects meta-analysis	[74]
rs4525	Missense	VTE	European	G (0.245)	-	0.77 (0.68–0.87)	$2.34x10^{-5}$	Logistic regression; additive genetic model	[68]
rs4524	Missense	VTE	European	G (0.245)	-	0.77 (0.68–0.87)	$2.51x10^{-5}$	Logistic regression; additive genetic model	[68]
rs4524	Missense	VTE	European	-	T(0.736)	1.20 (1.14–1.26)	$2.65x10^{-11}$	Logistic regression; additive genetic model	[67]
rs10158595	Intronic	VTE	European	A (0.218)	-	0.76 (0.67–0.87)	$3.03x10^{-5}$	Logistic regression; additive genetic model	[68]
rs6032	Missense	VTE	European	G (0.245)	-	0.77 (0.68–0.87)	$3.35x10^{-5}$	Logistic regression; additive genetic model	[68]
rs2213867	Intronic	VTE	European	G (0.245)	-	0.78 (0.69–0.88)	$4.37x10^{-5}$	Logistic regression; additive genetic model	[68]
rs2420371	Intronic	VT	European	G(0.115)	-	2.49(nr)	$4.02x10^{-26}$	Logistic regression; additive genetic model	[75]

Table A.6: Genetic variants found for the *F2* gene (chr 11).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs1799963	3'-UTR	VTE	European	A (0.025)	-	2.46 (1.70–3.54)	$1.69x10^{-6}$	Logistic regression; additive genetic model	[68]
rs1799963	3'-UTR	VTE	European	nr	nr	1.86 (1.27–2.73)	0.0015	Logistic regression; additive genetic model	[63]
rs1799963	3'-UTR	VTE	European	-	G(0.9855)	0.512 (0.456, 0.576)	$1.3x10^{-24}$	Logistic regression; additive genetic model	[66]*
rs1799963	3'-UTR	VTE	European	-	A(0.01)	2.63 (2.03–3.40)	$4.90x10^{-13}$	Logistic regression; additive genetic model	[69]
rs1799963	3'-UTR	VTE	European	-	A(0.010)	2.29 (1.75–2.99)	$1.73x10^{-9}$	Logistic regression; additive genetic model	[67]
rs1799963	3'-UTR	RM	Overall**	nr	nr	1.97 (1.59–2.45)	$<$0.00001	Random-effects meta-analysis; overdominant genetic model	[71]
rs1799963	3'-UTR	RM	Caucasian	nr	nr	1.83 (1.35–2.47)	$<$0.0001	Random-effects meta-analysis; overdominant genetic model	[71]
rs1799963	3'-UTR	IS	Mostly European	-	A(nr)	1.5 (1.1-2.0)	0.005	Fixed-effect meta-analysis; genetic model nr	[76]
rs1799963	3'-UTR	CVT	European	-	A(nr)	5.37 (3.78-7.63)	$<$0.00001	Random-effects meta-analysis; dominant genetic model	[73]
rs1799963	3'-UTR	CVT	Mostly mixed	nr	nr	6.05 (4.12-8.90)	$<$0.001	Random-effects meta-analysis; genetic model nr	[74]
rs3136516	Intronic	VTE	European	-	G(0.48)	1.10 (1.06–1.13)	$7.60x10^{-9}$	Logistic regression; additive genetic model	[69]
rs3136520	Intronic	RM	Chinese Han	nr	nr	0.986 (0.976-0.997)	0.031	Chi-square test; recessive genetic model	[64]

Table A.7: Genetic variants found for the *F11* gene (chr 4).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs2036914	Intronic	VTE	European	nr	nr	1.15 (1.04–1.28)	0.009	Logistic regression; additive genetic model	[63]
rs2036914	Intronic	VTE	European	nr	nr	1.16(nr)	$1.20x10^{-8}$	Logistic regression; additive genetic model	[69]
rs2289252	Non coding transcript exon	VTE	European	nr	nr	1.19 (1.07–1.32)	0.002	Logistic regression; additive genetic model	[63]
rs2289252	Non coding transcript exon	VTE	European	nr	nr	1.17(nr)	$1.90x10^{-9}$	Logistic regression; additive genetic model	[69]
rs4253399	Intronic	VTE	European	C (0.409)	-	1.28 (1.15–1.43)	$6.33x10^{-6}$	Logistic regression; additive genetic model	[68]
rs4444878	Intronic	VTE	European	-	C(0.5988)	0.81(0.780, 0.841)	$7.0x10^{-28}$	Logistic regression; additive genetic model	[66]*
rs4253416	Intronic	VTE	European	-	T(0.45)	1.18 (1.12–1.24)	$2.0x10^{-10}$	Logistic regression; additive genetic model	[69]
rs4253417	Intronic	VTE	European	-	C(0.405)	1.27 (1.22–1.34)	$1.21x10^{-23}$	Logistic regression; additive genetic model	[67]

Table A.8: Genetic variants found for the *F8* gene (chr X).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs114209171	Non coding transcript exon	VTE	European	-	T(0.7756)	1.153 (1.108 - 1.200)	$7.0x10^{-13}$	Logistic regression; additive genetic model	[66]*

Table A.9: Genetic variants found for the *F13A1* gene (chr 6).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs5985	Missense	RM	Overall**	nr	nr	1.42 (1.11–1.83)	0.005	Random-effects meta-analysis; overdominant genetic model	[71]

A.3 Other genes

Table A.10: Genetic variants found for the *FGG* gene (chr 4).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs2066865	3-'UTR	VTE	European	nr	nr	1.38 (1.22–1.55)	$8.6x10^{-8}$	Logistic regression; additive genetic model	[63]
rs2066865	3-'UTR	VTE	European	-	A(0.24)	1.21 (1.15–1.29)	$3.10x10^{-11}$	Logistic regression; additive genetic model	[69]
rs2066865	3-'UTR	VTE	European	-	A(0.244)	1.24 (1.18–1.31)	$1.03x10^{-16}$	Logistic regression; additive genetic model	[67]
rs2066865	3-'UTR	VT	European	A(0.250)	-	1.53(nr)	$2.286x10^{-13}$	Logistic regression; additive genetic model	[75]
rs7654093	5'	VTE	European	-	T (0.2322)	1.216 (1.166 - 1.267	$2.0x10^{-19}$	Logistic regression; additive genetic model	[66]*

Table A.11: Genetic variants found for the *FGA* gene (chr 4).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs6825454	3'	VT	European	C(0.268)	-	1.47(nr)	$4.32x10^{-12}$	Logistic regression; additive genetic model	[75]
rs6825454	3'	IS	Mixed	C(0.31)	-	1.06(1.04–1.08)	$7.43x10^{-10}$	Logistic regression; additive genetic model	[77]

Table A.12: Genetic variants found for the *KLKB1* gene (chr 4).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs3087505	3'-UTR	VTE	European	A (0.099)	-	0.63 (0.52–0.75)	$4.34x10^{-7}$	Logistic regression; additive genetic model	[68]

Table A.13: Genetic variants found for the *ABO* gene (chr 9).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs2519093	Intronic	VTE	European	A (0.243)	-	1.68 (1.48–1.91)	$8.08x10^{-16}$	Logistic regression; additive genetic model	[68]
rs2519093	Intronic	VTE	European		T(0.19)	1.41 (1.32–1.50)	$6.00x10^{-26}$	Logistic regression; additive genetic model	[69]
rs505922	Intronic	VTE	European	G (0.402)	-	1.49 (1.33–1.66)	$1.52x10^{-12}$	Logistic regression; additive genetic model	[68]
rs505922	Intronic	VT	European	C(0.430)	-	1.92(nr)	$1.386x10^{-34}$	Logistic regression; additive genetic model	[75]
rs505922	Intronic	IS	Mostly European	-	C(nr)	1.07 (1.03-1.11)	0.0006	Fixed-effect meta-analysis; genetic model nr	[78]
rs687289	Intronic	VTE	European	A (0.403)	-	1.48 (1.33–1.65)	$3.03x10^{-12}$	Logistic regression; additive genetic model	[68]
rs8176719	Frameshift	VTE	European	G (0.419)	-	1.47 (1.32–1.64)	$5.68x10^{-12}$	Logistic regression; additive genetic model	[68]
rs8176719	Frameshift	VTE	European	nr	nr	1.51 (1.35–1.69)	$4.1x10^{-13}$	Logistic regression; additive genetic model	[63]
rs643434	Intronic	VTE	European	A (0.421)	-	1.44 (1.30–1.61)	$3.39x10^{-11}$	Logistic regression; additive genetic model	[68]
rs630014	Intronic	VTE	European	A (0.421)	-	0.75 (0.67–0.84)	$2.67x10^{-7}$	Logistic regression; additive genetic model	[68]
rs630014	Intronic	VT	European	A(0.424)	-	0.62(nr)	$2.537x10^{-19}$	Logistic regression; additive genetic model	[75]
rs660340	nr	VTE	European	A (0.408)	-	0.77 (0.69–0.85)	$1.13x10^{-6}$	Logistic regression; additive genetic model	[68]
rs659104	nr	VTE	European	A (0.408)	-	0.77 (0.69–0.85)	$1.28x10^{-6}$	Logistic regression; additive genetic model	[68]
rs529565	Intronic	VTE	European	-	T(0.6588)	0.723 (0.697 - 0.751)	$7.1x10^{-63}$	Logistic regression; additive genetic model	[66]*
rs529565	Intronic	VTE	European	-	C(0.354)	1.55 (1.48–1.63)	$4.23x10^{-75}$	Logistic regression; additive genetic model	[67]
rs8176645	Intronic	VTE	European	-	A(0.33)	1.28 (1.22–1.35)	$4.40x10^{-21}$	Logistic regression; additive genetic model	[69]
rs657152	Intronic	VT	European	A(0.454)	-	1.78(nr)	5.462×10^{-28}	Logistic regression; additive genetic model	[75]
rs495828	5’	VT	European	T(0.316)		1.68(nr)	4.002×10^{-21}	Logistic regression; additive genetic model	[75]
rs635634	5’	IS	European	T(0.19)	-	1.08 (1.05–1.11)	$9.18x10^{-9}$	Logistic regression; additive genetic model	[77]

Table A.14: Genetic variants found for the *ANXA5* gene (chr 4).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs11575945	5'-UTR	RM	Overall**	nr	nr	1.90 (1.27–2.85)	0.002	Random-effects meta-analysis; overdominant genetic model	[71]
rs28651243	5'-UTR	RM	Overall**	nr	nr	1.77 (1.20–2.62)	0.004	Random-effects meta-analysis; overdominant genetic model	[71]
rs28717001	5'-UTR	RM	Overall**	nr	nr	1.82 (1.22–2.71)	0.003	Random-effects meta-analysis; overdominant genetic model	[71]
rs112782763	5'-UTR	RM	Overall**	nr	nr	1.72 (1.16–2.55)	0.007	Random-effects meta-analysis; overdominant genetic model	[71]
rs1050606	5'-UTR	RM	Overall**	nr	nr	0.52 (0.42–0.66)	<0.00001	Random-effects meta-analysis; dominant genetic model	[71]

Table A.15: Genetic variants found for the *ITGB3* gene (chr 17).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs5918	Missense	RM	Overall**	nr	nr	0.12 (0.04–0.43)	0.001	Random-effects meta-analysis; recessive genetic model	[71]
rs5918	Missense	DVT	Caucasian	-	C(nr)	2.289 (1.260 - 4.160)	0.008	Fisher exact test; dominant genetic model	[79]
rs5918	Missense	RM	Iranian	-	C(nr)	0.303 (0.159–0.579)	<0.001	Logistic regression; dominant genetic model	[80]

Table A.16: Genetic variants found for the *SERPINE1* gene (chr 7).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs1799889	5 (I/D)	RM	Overall**	nr	nr	0.71 (0.62–0.82)	<0.00001	Random-effects meta-analysis; allelic genetic model	[71]
rs1799889	5 (I/D)	RM	Caucasian	nr	nr	2.10 (1.46–3.01)	<0.0001	Random-effects meta-analysis; recessive genetic model	[71]
rs1799889	5 (I/D)	RM	Caucasian	-	4G (0.5)	2.7(1.21-6.03)	0.01	Logistic regression; recessive genetic model	[81]
rs1799889	5 (I/D)	VTE	Caucasian	-	4G (nr)	1.31 (1.10-1.56)	0.003	Random-effects meta-analysis; dominant genetic model	[26]
rs1799889	5 (I/D)	VTE	Asian	-	4G (nr)	2.08 (1.29-3.35)	0.003	Random-effects meta-analysis; dominant genetic model	[26]

Table A.17: Genetic variants found for the *MTHFR* gene (chr 1).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs1801131	Missense	RM	Overall**	nr	nr	0.59 (0.46–0.76)	<0.0001	Random-effects meta-analysis; allelic genetic model	[71]
rs1801131	Missense	RM	Caucasian	nr	nr	0.43 (0.29–0.65)	<0.0001	Random-effects meta-analysis; allelic genetic model	[71]
rs1801131	Missense	RM	Asian	nr	nr	1.83 (1.36–2.46)	<0.0001	Random-effects meta-analysis; recessive genetic model	[71]
rs1801131	Missense	RM	Iranian	nr	C(nr)	22.9 (13.99-37.51)	<0.0001	Chi-square test; genetic model nr	[72]
rs1801133	Missense	RM	Overall**	nr	nr	1.60 (1.38–1.86)	<0.00001	Random-effects meta-analysis; recessive genetic model	[71]
rs1801133	Missense	RM	Caucasian	nr	nr	0.77 (0.67–0.87)	0.0001	Random-effects meta-analysis; allelic genetic model	[71]
rs1801133	Missense	RM	Asian	nr	nr	1.93 (1.51–2.46)	<0.00001	Random-effects meta-analysis; recessive genetic model	[71]
rs1801133	Missense	RM	Asian	nr	nr	1.44 (1.14–1.82)	0.006	Random-effects meta-analysis; dominant genetic model	[82]
rs1801133	Missense	RM	Asian	-	T(nr)	2.11 (1.40-3.19)	0.0004	Random-effects meta-analysis; recessive genetic model	[83]
rs1801133	Missense	RM	Mixed	-	T(nr)	3.47 (2.31-5.20)	<0.0001	Random-effects meta-analysis; recessive genetic model	[83]
rs1801133	Missense	RM	Iranian	nr	T(nr)	1.59 (1.17-2.17)	0.003	Chi-square test; genetic model nr	[72]
rs1801133	Missense	CVT	European	-	T(nr)	2.30(1.20-4.42)	0.02	Random-effects meta-analysis; recessive genetic model	[73]

Table A.18: Genetic variants found for the *PROZ* gene (chr 13).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs3024731	Intronic	RM	Chinese Han	nr	nr	1.479 (1.098-1.994)	0.01	Chi-square test; dominant genetic model	[64]

Table A.19: Genetic variants found for the *THBD* gene (chr 20).

SNP	Description	Trait	Population	MA(MAF)	EA(EAF)	OR (95% CI)	p-value	Statistical Analysis	Source
rs1042579	Missense	RM	Chinese	-	T(nr)	1.83(1.10-3.06)	0.02	Logistic regression; dominant genetic model	[84]

nr: not reported

MA(MAF): Minor Allele (Minor Allele Frequency)

EA(EAF): Effect Allele (Effect Allele Frequency)

* The genes reported from this source are the nearest genes related to coagulation in the region.

** Refers to the total sample of the study (Caucasians, Asians, and a minority of mixed ethnic populations).

A.4 More information about the genes and studies included

Table A.20: Role in blood coagulation of the genes featured in tables A.1 to A.19.*

Gene	Chromossome	Role In Blood Coagulation
SERPINC1	1	Encodes the anticoagulant AT.
PROS1	3	Encodes the anticoagulant PS.
PROC	2	Encodes the anticoagulant PC.
PROCR	20	Encodes a receptor for APC.
F5	1	Encodes the coagulation factor FV.
F2	11	Encodes the coagulation factor FII.
F11	4	Encodes the coagulation factor XI, FXI.
F8	X	Encodes the coagulation factor VIII, FVIII.
F13A1	6	Encodes the A subunit of the coagulation factor FXIII.
FGG	4	Encodes the gamma subunit of fibrinogen.
FGA	4	Encodes the alpha subunit of fibrinogen.
KLKB1	4	Encodes prekallikrein, a protein that participates in the surface-dependent activation of blood coagulation and fibrinolysis.
ABO	9	Encodes glycosyltransferases that determine blood group, which in turn can influence the circulating plasma levels of vWF.
ANXA5	4	Encodes the annexin 5 protein. One of its functions is to act as an anticoagulant (vascular and placental).
ITGB3	17	Encodes a subunit of a platelet receptor protein, integrin alphaIIb/beta3 (IIb3).
MTHFR	1	Encodes the MTHFR enzyme, which has no known direct effect on blood coagulation.
SERPINE1	7	Encodes the plasminogen activator inhibitor 1 (PAI-1) protein, which is involved in blood coagulation by affecting the process of fibrinolysis.
PROZ	13	Encodes the anticoagulant protein Z (PZ).
THBD	20	Encodes the thrombomodulin protein. This protein binds with thrombin to activate PC.

*The description of the genes presented were based on the information available in U.S National Library of Medicine (NLM).*Genetics Home Reference* [database]. Retrieved from https://ghr.nlm.nih.gov/.

Table A.21: Supplementary information about the studies selected.

Source	Cases	Controls	Population (Countries)	Type of study	Age of cases (SD)	Age of controls (SD)	Sex
[68]	1488	1439	European ancestry (midwestern USA)	Candidate polymorphism (case-control)	54.7 (16.3)	55.5 (15.7)	50.5% of cases female; 52.4% of controls female
[63]	1433	1402	European (Sweden)	Development and assessment of a risk prediction model (case-control)	46 (12.9)	46.9 (12.6)	Female
[85]	1201 subjects from 430 families with inherited thrombophilia		European (France)	Family study; development and assessment of a risk prediction model	34.26 (17.9) [total sample]		60.8% Female
[66]	6135	252827	European ancestry (countries nr)	GWAS (case-control)	58.2% >=60 (nr)	57.5% >=45(nr)	56.9% of cases female; 46.6% of controls female
[69]	3290	116868	European (UK)	GWAS (case-control)	59.5 (7.2)	56.8 (7.9)	43.3% of cases male; 47.4% of controls male
[75]	1542	1110	European (France)	GWAS (case-control)	nr	>65 (nr)	nr
[67]	7507	52632	European (French, Dutch, and USA subjects with European ancestry)	Meta-analysis of GWAS	nr	nr	nr
[71]	369 studies; sample sizes nr		Mainly Caucasian, followed by Asian, and a minority of mixed ethnic populations (countries nr)	Systematic review and meta-analysis	nr	nr	Female
[77]	67,162	454,450	European, East Asian, South Asian, African, mixed Asian, and Latin American (countries nr)	Meta-analysis of GWAS	nr	nr	nr
[62]	306	370	African Americans	Candidate polymorphism (case-control)	<50 (nr)	>70 (nr)	nr
[70]	447	832	Northwest-Europe (countries nr)	GWAS (case-control)	50.2 (12.7)	47.0 (12.8)	64.2 % of cases male; 40.7 % of controls male

Table A21: Continued.

Source	Cases	Controls	Population (Countries)	Type of study	Age of cases (SD)	Age of controls (SD)	Sex
[82]	3573	4257	Asian (Behrain, China, Egypt, India, Iran, Israel, Japan, Palestine, SriLanka, Turkey)	Meta-analysis	nr	nr	Female
[64]	426	444	Chinese	Candidate polymorphism (case-control)	29.26 (4.294)	34.50 (4.895)	Female
[60]	1304	1334	Chinese	Candidate polymorphism (case-control)	52 (nr)	52 (nr)	52% females in the total sample
[61]	1520	2594	European (Spain, Denmark)	Candidate polymorphism (case-control)	nr	nr	nr
[81]	208	92	North American	Candidate polymorphism (retrospective cohort study)	35.8 (5) [total sample]		Female
[84]	94	169	Chinese	Candidate polymorphism (case-control)	28.370 (3.742)	28.07 (3.611)	Female
[65]	101	279	Chinese	Candidate gene (case-control)	63 (nr)	65 (nr)	60.4% of cases male; 67% of controls male
[26]	3561	5693	Caucasian (Turkey, Croatia, Slovenia, France, USA, Spain, Italy, Egypt, Germany, Portugal, Sweden, UK)	Meta-analysis	nr	nr	nr
[78]	8,884	55,254	Mostly European (UK, Iceland, Italy, Netherlands) but also included individuals from the USA and Australia	Candidate polymorphism (case-control)	67.1 (10.5)	59.4 (9.9)	48.2% females in the total sample
[76]	2305	5977	Mostly European (Spain, France, Italy, Netherlands, Poland, Germany, UK) and also subjects from Brazil and the USA	Meta-analysis	$<$50 (nr)	$<$50 (nr)	$\sim$30.7% males and $\sim$69.3% females in the total sample
[73]	1183	5189	European ancestry (countries nr)	Meta-analysis	nr	nr	nr
[74]	1695	5804	Mixed populations (countries nr)	Meta-analysis	The only information about ages is that there is a wide range		nr

Table A.21: Continued.

Source	Cases	Controls	Population (Countries)	Type of study	Age of cases (SD)	Age of controls (SD)	Sex
[83]	1729	2060	Caucasians of European origin, East Asian (Chinese, Japanese and Korean) and a mixed subgroup (Indian, Brazilian, Bahrainese, Mexican and Egyptian)	Meta-analysis	The only information about age is that there is a wide range		Female
[72]	330	350	Iranian	Candidate polymorphism (case-control)	30.37 (5.05)	29.88 (4.09)	Female
[79]	224	216	Caucasian (Bulgaria)	Candidate polymorphism (prospective cohort study)	48.8 (15.05)	45.92 (13.47)	52.72% male cases; 50% male controls
[80]	100	100	Iranian	Candidate polymorphism (case-control)	$<$35 (nr)	nr	Female

A.5 Definitions of medical terms used

Anaemia: a condition in which either the level of red blood cells or the level of haemoglobin is lower than normal, commonly caused by iron deficiency.

Antigen: any substance capable of triggering an immune response.

Aspirin: a commonly used drug to treat mild to moderate pain, reduce fever or inflammation, acting also as an anticoagulant.

Atherosclerosis: a condition in which plaque made up of fat, cholesterol, among other substances, builds up inside the arteries, hardening over time and narrowing the vessels.

Atrial fibrillation: a condition in which the upper chambers of the heart - the atria - beat irregularly.

Cardioembolism: an embolism of cardiac origin.

Diabetes mellitus: a disorder characterized by elevated levels of sugar in the blood due to the inadequate production or action of insulin.

Ethinylestradiol: estrogen medication widely used in birth control pills.

Heparin: medication used as an anticoagulant.

Hypercholesterolemia: a condition characterized by elevated levels of cholesterol in the blood.

Hyperlipidemia: a condition characterized by increased levels of lipids or fat proteins in the blood.

Hypertension: a condition characterized by high blood pressure.

Myocardial infarction (MI): medical term for heart attack. Characterized by the decrease, or even complete stop, of blood flow to a part of the heart, causing damage to the heart muscle.

Neural tube defects: birth defects of the brain, spine, or spinal cord, usually happening in the first month of pregnancy.

Progestogens: synthetic forms of progesterone - a naturally occurring sex hormone - widely used in birth control pills.

Silent myocardial ischemia: myocardial ischemia in the absence of symptoms such as chest pain, for instance.

Transient ischemic attack (TIA): an acute event similar to a stroke, but usually lasting only for a few minutes and causing no permanent damage.