IJEEC

International Journal of Electrical Engineering and Computing
Vol. 9, No. 1 (2025)

Original research paper
UDC 007.52:[616-005.1:617-089-055.26
DOI 10.7251/IJEEC2501010Z

Population Genetic Analysis of Allele Frequencies for Hereditary Thrombophilia in Pregnant Women Using Artificial Intelligence

Katarina Živojinović[1],Marko M. Živanović[2], Stefan Erčić[3] and Vanja Luković[5]

[1]Family Medica, polyclinics & private practices, The Republic of Serbia
[2] Faculty of Information Technology, University Metropolitan, Belgrade, The Republic of Serbia
[3]Grammar school (Savremena gimnazija, Zemun, Belgrade), The Republic of Serbia
[5]University of Kragujevac, Faculty of Technical Sciences Čačak, The Republic of Serbia

E-mail address: marko.zivanovic@metropolitan.ac.rs, stefercic@gmail.com, kzivojinovic@familymedica.rs, nbogdanovic@fsu.edu, vanja.lukovic@ftn.kg.ac.rs

Abstract— **Thrombophilia is an inherited condition characterized by an increased tendency for blood clot formation, affecting approximately 8-11% of the European population. It is typically inherited in an autosomal dominant manner and includes approximately 10 different subtypes, classified based on genetic factors. This condition often leads to complications in pregnant women, including spontaneous miscarriage, fetal deformities, and an increased risk of heart attack, stroke, pulmonary embolism, and deep vein thrombosis. The goal of this study was to conduct a detailed genetic analysis of thrombophilia using population genetics methods (Hardy–Weinberg equilibrium, Shannon index, genotype frequencies) and to develop an AI-based predictive model for high-risk genotypes, evaluated via confusion matrices. The research focused on assessing allele frequencies, genetic diversity, and deviations from Hardy-Weinberg equilibrium in a cohort of 2,760 pregnant women to enhance the understanding of the genetic basis of this condition. Our analysis revealed significant correlations between coagulation factor genes and identified distinctive patterns of genetic diversity across 12 thrombophilia-associated markers. The results provide key insights into genetic variations and their potential implications for pregnancy complications. This study opens new perspectives for improving early detection and more effective risk management of thrombophilia during pregnancy.**

Keywords – Population Genetics; Allele Frequencies; Thrombophilia; Genotypic frequency; Pregnancy; Shannon index; Hardy-Weinberg Equilibrium (HWE); Artificial intelligence;

I. INTRODUCTION

Thrombophilia refers to a group of disorders that predispose individuals to an increased risk of blood clot formation. These conditions can be either inherited or acquired. Acquired forms of thrombophilia may arise from secondary factors such as obesity, smoking, oral contraceptive use, surgery, neoplasia, antiphospholipid antibody syndrome, or heparin-induced thrombocytopenia. On the other hand, genetic or primary thrombophilia is caused by mutations or deficiencies in specific proteins, including Factor V Leiden mutation, prothrombin gene mutation (FII), as well as deficiencies in proteins such as antithrombin III, protein C, protein S, and histidine-rich glycoprotein.

The increased clotting tendency associated with thrombophilia significantly raises the risk of developing deep venous thrombosis (DVT) and venous thromboembolism (VTE). Thrombosis in individuals with this condition may also occur in atypical areas, such as the splanchnic, cerebral, and retinal veins. However, the clinical manifestation of hereditary thrombophilia can vary significantly among individuals. Some may never develop thrombosis, others could remain asymptomatic until adulthood, and there are cases where recurrent thromboembolic events occur in individuals before reaching 30 years of age. For individuals who carry heterozygous mutations in Factor V Leiden or FII (Prothrombin G20210A), the risk of thrombosis is relatively mild, with these individuals being 3.8 and 4.9 times more likely, respectively, to experience a first clot. When both heterozygous mutations are present in a patient, the risk of developing thrombosis is significantly higher, increasing up to 20 times. Homozygous mutations in these genes are extremely rare in the general population, highlighting the unique nature of these genetic predispositions[1].

Factor V Leiden thrombophilia is the most widespread genetic form of thrombophilia. In regions such as the United States and Europe, about 3-8% of people carry a single copy of the Factor V Leiden mutation. However, individuals who inherit two copies of this mutation, making them homozygous, are exceptionally rare, with an estimated prevalence of just 1 in every 5,000 individuals [2]. A moderate deficiency in protein S

This paper is a revised and expanded version of the paper presented at the XX International Symposium INFOTEH-JAHORINA 2021

is thought to occur in about 1 in 500 individuals, while severe deficiency is uncommon, and its exact prevalence remains unidentified [3]. Moderate protein C deficiency is found in approximately 1 in 500 individuals, while severe deficiency is rare, affecting about 1 in 4,000,000 newborns [4]. Due to its relatively high prevalence (1.7–3% in the European and U.S. populations), prothrombin-related thrombophilia is an important risk factor considered in the diagnosis of coagulation disorders [5]. Antithrombin III deficiency, an inherited condition, affects approximately 1 in 500 to 5,000 individuals in the general population, making it a relatively rare but clinically significant thrombophilia [6]. Depending on the specific mutation, thrombophilia may be passed down through autosomal dominant, recessive, or X-linked inheritance mechanisms [7].

These findings also have significant implications in the context of population policy. A high proportion of mutated alleles—particularly in the homozygous mutant (mut/mut) form—among key markers indicates a widespread genetic predisposition to thrombophilia within the analyzed population [8]. Such a genetic profile can greatly influence public health strategies, especially in the areas of prenatal and postnatal care. Identifying at-risk groups through genetic screening can enable targeted preventive measures and better individualization of therapeutic approaches. In the context of demographic dynamics and modern public health challenges, monitoring genetic markers may become a vital tool for improving population health—particularly in terms of family planning, reducing pregnancy-related complications, and supporting sustainable natural population growth.

This study aimed to: 1) characterize the distribution and frequency of thrombophilia-associated genetic markers in pregnant women; 2) assess genetic diversity and potential linkage disequilibrium among these markers; 3) identify correlation patterns between different thrombophilia-associated genes; and 4) develop an AI-based predictive model for identifying high-risk genotype combinations.

II. MATERIALS AND METHODS

This study is based on an dataset comprising 2,760 individual samples, primarily collected from pregnant women aged between 18 and 40 years, who represent approximately 90% of the total sample size. The data spans a six-year period, specifically from 2018 to 2023. For statistical analysis and data visualization, a combination of the Python programming language and Microsoft Power BI tool was used, enabling detailed data processing as well as clear graphical representation. Key findings, along with analytical methods related to genotype distribution, genetic diversity, and allele frequency, are thoroughly discussed in a dedicated section titled *Population Genetics.*

Of the total number of samples for all allele variants of all thrombophilia types, 26.2% are homozygotes, while 73.3% are heterozygotes.

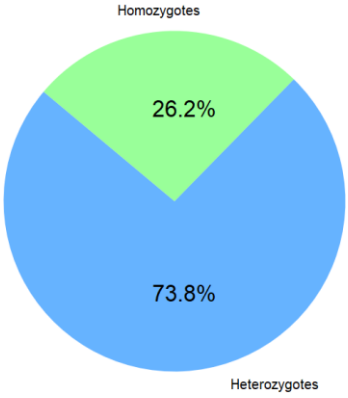

Figure 1. **Overall distribution of homozygotes and heterozygotes across all markers:** Pie chart showing the aggregate proportion of individuals carrying two mutant alleles (homozygotes, 26.2 %) versus one mutant allele (heterozygotes, 73.8 %) in a cohort of 2,760 pregnant women genotyped for 12 thrombophilia-associated markers. Homozygotes (light green slice) represent those with mut/mut genotypes across any marker; heterozygotes (sky-blue slice) represent wt/mut genotypes. Percentages are displayed to one decimal place. This distribution highlights that three-quarters of the study population carry exactly one risk allele, whereas just over one-quarter carry two risk alleles, underscoring the predominance of heterozygous variants in this cohort.

A pie chart illustrates the aggregate proportion of individuals carrying two mutant alleles (homozygotes, 26.2 %) versus one mutant allele (heterozygotes, 73.8 %) in a cohort of 2,760 pregnant women genotyped for 12 thrombophilia-associated markers. Homozygotes, represented by the light green slice, correspond to individuals with mut/mut genotypes across any marker, while heterozygotes, shown in sky-blue, represent those with wt/mut genotypes. Percentages are displayed to one decimal place. This distribution reveals that three-quarters of the study population carry exactly one risk allele, whereas just over one-quarter carry two, emphasizing the predominance of heterozygous variants in this cohort. The high heterozygosity rate suggests that a single risk allele is significantly more common in pregnant women than carriage of two mutant alleles, which may contribute to a moderated overall population risk. Data were pooled across all 12 genetic markers (N = 2,760), and for visual clarity, colors were chosen for high contrast with slices labeled by both percentage values and a legend.

III. POPULATION GENETICS

This paper focuses on the processing, analysis, and visualization of genotype data for markers associated with thrombophilia, using various statistical and bioinformatics methods. The central part of the analysis involves calculating genotype frequencies for each marker. These values serve as the foundation for many other analyses in population genetics, as they indicate the structure of genetic variability within the sample.

In addition, a Hardy-Weinberg equilibrium (HWE) analysis was conducted. In addition, a Hardy-Weinberg equilibrium (HWE) analysis was conducted independently for each of the 12 genetic markers. No correction for multiple testing was applied. This equilibrium represents the theoretical expectation of genotype distribution in an ideal population, under the conditions of random mating and the absence of selection, mutations, migration, and genetic drift [9]. Using the function

the expected genotype frequencies were calculated based on the allele frequencies p and q = 1 - p. The result is the values p^2, 2pq, and q^2, which represent the theoretical distribution of the 'wt/wt', 'wt/mut', and 'mut/mut' genotypes. An important indicator in the analysis of these genetic data is the Shannon diversity index, which measures genetic diversity. A high value of this index indicates the presence of diverse genotypes in the sample, while a lower value may suggest homogeneity or the dominance of a single genotype. The shannon diversity function uses logarithmic entropy to quantify this diversity, which is particularly useful in population studies where diversity is associated with health, adaptability, and the evolutionary potential of a population.

A significant aspect of genetic analysis is the examination of correlations between different markers, which can indicate the presence of linkage disequilibrium or functional connectivity between genes. A correlation matrix was calculated using the data.corr() method and visualized using the seaborn.heatmap function. Darker shades indicate a stronger correlation (positive or negative), which may suggest a shared genetic basis or co-regulation among the markers.

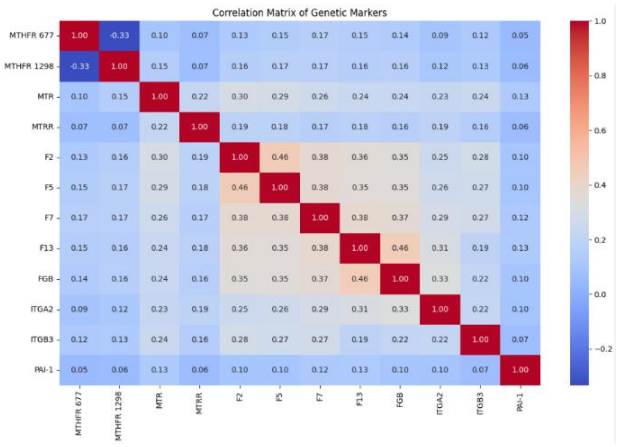

Figure 2. **Corelation Matrix of Genetic Markers**: Correlation Matrix of Thrombophilia Genetic Markers: Heatmap displaying Pearson correlation coefficients between genotype frequencies of 12 thrombophilia-associated markers (MTHFR 677, MTHFR 1298, MTR, MTRR, F2, F5, F7, F13, FGB, ITGA2, ITGB3, PAI-1) in 2,760 pregnant women. Cells are colored from blue (negative correlation) through white (no correlation) to red (positive correlation). Strong positive correlations (r ≥ 0.45) appear between F2–F5, F5–F13, and F13–FGB, indicating linked coagulation factors, while MTHFR 677 and MTHFR 1298 show a modest negative correlation (r = –0.33). Numeric values are annotated in each cell.

A. Hardy-Weinberg Equilibrium (HWE)

The Hardy-Weinberg principle (also known as the Hardy-Weinberg theorem) is a fundamental concept in population genetics that describes how allele and genotype frequencies behave across generations in the absence of evolutionary forces. This theorem provides the expected proportions of genotypes in the next generation based on current allele frequencies, assuming that certain conditions are met [10].

For a population to be in Hardy-Weinberg equilibrium, certain conditions must be met: mating must be random, meaning that each individual has an equal chance of mating with any other in the population; there must be no mutations that alter alleles; migration must be absent, i.e., there should be no individuals entering or leaving the population; there must be no selection, as all genotypes must have the same probability of survival and reproduction; and finally, the population must be sufficiently large—ideally infinite—to eliminate the effects of genetic drift. When all these conditions are satisfied, allele and genotype frequencies in the population remain constant across generations, keeping the population in Hardy-Weinberg equilibrium [11].

The mathematical formulation of Hardy-Weinberg equilibrium begins with the consideration of a single locus with two alleles: allele A, often referred to as the reference or "wild type" (wt), and allele a, which represents the mutated form. The frequency of allele A in the population is denoted as p, while the frequency of allele a is denoted as q. Since these are the only two alleles at the same locus, the sum of their frequencies must equal 1, meaning that the basic relation

p + q = 1 holds true [12]. This formulation serves as the foundation for further analysis of genotype frequencies in an equilibrium population.

TABLE I. GENOTYPE FREQUENCIES AND THEIR MATHEMATICAL REPRESENTATION

Genotype	Mathematical notation	Probability / Frequency
AA (wt/wt)	p^2	p^2
Aa (wt/mut)	2pq	2pq
aa (mut/mut)	q^2	q^2

Thus, the genotype frequencies are:

$$ft(wt/wt) = p^2 \quad (1)$$
$$ft(wt/mut) = 2pq \quad (2)$$
$$ft(mut/mut) = q^2 \quad (3)$$
$$p = 0.7q = 0.3 \quad (4)$$

For illustration purposes, example values of p=0.7 and q=0.3 are used to demonstrate the expected genotype frequencies under Hardy-Weinberg equilibrium. In our actual data analysis, the observed allele frequencies varied for each marker as detailed in Section IV.

$$wt/wt = 0.7^2 = 0.49 \quad (5)$$
$$wt/mut = 2 \cdot 0.7 \cdot 0.3 = 0.49 \quad (6)$$
$$mut/mut = 0.3^2 = 0.09 \quad (7)$$

The application of Hardy-Weinberg equilibrium (HWE) in the analysis of genetic data, as presented in this scientific work, represents a crucial step in population genetics. This equilibrium serves not only as a theoretical framework for understanding the relationship between alleles and genotypes, but also as a practical tool for assessing the validity of genetic data [13]. By using HWE, it is possible to identify potential technical errors in genotyping, detect selective pressures that may influence allele frequency, and gain insight into the structure and dynamics of a population. Moreover, deviations from HWE may indicate the presence of migration, mutations, or improper sampling, making HWE a key reference point in the interpretation and quality control of genetic analyses [14].

B. Genotypic frequency

Genotypic frequency refers to the occurrence of a specific genotype within a population and represents a fundamental concept in population genetics, statistical analysis of genetic data, and the study of trait inheritance [15]. A genotype is the combination of alleles an individual possesses for a particular gene (or marker) – for example, an individual may have two copies of the normal allele (wt/wt), one normal and one mutated copy (wt/mut), or two mutated copies (mut/mut). Genotypic frequency measures the percentage of individuals that carry a specific genotype within a population or sample.

Let **N** be the total number of valid (non-missing) observations for a given genetic marker (after removing NaN values), i.e.:

N = the number of observed individuals without missing values.

Let n1, n2, ..., nk be the frequencies of each unique genotype category in the marker (e.g., "wt/wt", "wt/mut", "mut/mut") [16]. Then, the relative frequency (proportion) of each genotype is calculated as:

$$f_i = \frac{n_i}{N} for i = 1,2,3, \dots, k \tag{8}$$

Since the frequency function further returns results in percentages, the following is obtained:

$$f_i^{(\%)} = \left(\frac{n_i}{N}\right) \cdot 100 \tag{9}$$

$f_i^{(\%)}$ – Genotypic frequency expressed as a percentage for genotype i.

n_i – Number of individuals with genotype i

N – Total number of valid individuals for the analyzed marker

Finally, all values are rounded to two decimal places:

$$f_i^{(\%)} = round\left(\left(\frac{n_i}{N}\right) \cdot 100,2\right) \tag{10}$$

C. Shannon index

The Shannon diversity index, often denoted as H′, is one of the most well-known and widely used measures for assessing biological diversity within a population [17]. This index is used in various disciplines, including ecology, economics, and population genetics, to quantify diversity in a set of objects (e.g., genotypes, species, resources, etc.). In the context of population genetics, the Shannon index measures genetic diversity within a population based on the frequency of genotypes or alleles [18].

The mathematical formulation of the Shannon diversity index is based on the probability of occurrence of a certain type in the set. For genotypic frequencies, the formula is as follows:

$$H' = -\sum_{i=1}^{k} p_i log_2(p_i) \tag{11}$$

Where:

- H′ is the Shannon diversity index.
- p_i is the proportion (or frequency) of the i-th genotype or allele in the population, i.e., $p_i = \frac{n_i}{N}$, where is the number of individuals with genotype i and **N** is the total number of individuals.
- **k** is the number of different genotypes or alleles in the population.
- log2 represents the logarithm with base 2, as the Shannon index uses the binary logarithm in its formulation, which allows the amount of information to be expressed in bits.

The Shannon diversity index is based on the probability of occurrence of each genotype or allele in the population. Each value **pi log2 pi** represents the probability of selecting an individual that possesses genotype **i** [19]. This index combines two key aspects: first, the proportion of individual genotypes/alleles (frequency), where a higher frequency of a specific genotype or allele in the population means that this type will have a smaller contribution to the Shannon index, as this frequency increases, reducing overall diversity; second, the balance among genotypes/alleles, as the index favors a more even distribution of genotypes in the population [20]. Therefore, if there are very few genotypes with high frequencies, the Shannon index will be lower, indicating lower diversity.

For a simple example, imagine we have a population with three genotypes:

30 individuals with genotype A,
50 individuals with genotype B,
20 individuals with genotype C,

The total number of individuals N is 30 + 50 + 20 = 100. The genotype frequencies are:

$$p_A = \frac{30}{100} = 0.30 \tag{12}$$

$$p_B = \frac{50}{100} = 0.50 \tag{13}$$

$$p_C = \frac{20}{100} = 0.20 \tag{14}$$

Now we can calculate the Shannon diversity index:

$$H' = -((0.30)log_2(0.30) + (0.50)log_2(0.50) + (0.20)log_2(0.20))$$

(15)

$$0.30log_2(0.30) = 0.30 \cdot (-1.737) = -0.5211 \quad (16)$$

$$0.50log_2(0.50) = 0.50 \cdot (-1) = -0.5000 \quad (17)$$

$$0.20log_2(0.20) = 0.20 \cdot (-2.322) = -0.4644 \quad (18)$$

So:

$$H' = -(-0.5211 - 0.5000 - 0.4644) = 1.4855 \quad (19)$$

D. Correlation matrix

A correlation matrix is a key tool in statistics and data analysis used to display the relationships between multiple variables [21]. It is used to determine whether there is a linear connection between different variables in a dataset. Mathematically, the correlation between two variables X and Y measures the strength and direction of their linearity [22].

The correlation between two variables X and Y, denoted as ρ(X,Y) or r(X,Y), is defined as:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \quad (20)$$

where Xi and Yi are the individual values of variables X and Y, and $\bar{X}$ and $\bar{Y}$ are the mean values of those variables.

σX and σY are the standard deviations of the variables X and Y, which are calculated as:

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad (21)$$

$$\sigma_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \quad (22)$$

$$\rho(X,Y) = Cov(X,Y)/(\sigma X \cdot \sigma Y) \quad (23)$$

A correlation matrix is a square matrix that shows the correlation between all variables in a dataset [23]. Each element of the matrix **rij** represents the correlation between variables **Xi** and **Xj**, where:

$$r_{ij} = \rho(X_i, X_j) \quad (24)$$

The diagonal elements of the matrix are always 1, because the correlation of any variable with itself is always perfect. The correlation between variables Xi and Xj can range between -1 and 1. A value of rij = 1 indicates a perfect positive linear correlation, rij = -1 denotes a perfect negative linear correlation, while rij = 0 indicates that there is no linear correlation between the variables Xi and Xj.

If the correlation rij is close to 1, it means that the variables Xi and Xj are positively related, i.e., as one variable increases, the other also increases.

If the correlation rij is close to -1, the variables are negatively related, i.e., as one variable increases, the other decreases.

A correlation close to 0 means that the variables are not linearly related, although other forms of relationships may exist.

IV. Genetic Markers and Genotype Proportion

In this chapter, we analyzed the distribution of genotypes for several genetic markers in the population, focusing on the frequency of different genotypes for each marker, as well as the proportional distribution between heterozygous and homozygous variants [24]. The presented results relate to 12 different genetic markers that play a significant role in various biological processes.

A. MTHFR 677

For the MTHFR 677 marker, the genotype distribution showed that the largest proportion was in the homozygous mutant variant mut/mut (46.94%). Approximately 39% of the samples were heterozygous (wt/mut), while only 0.94% of the samples had the wt/wt genotype.

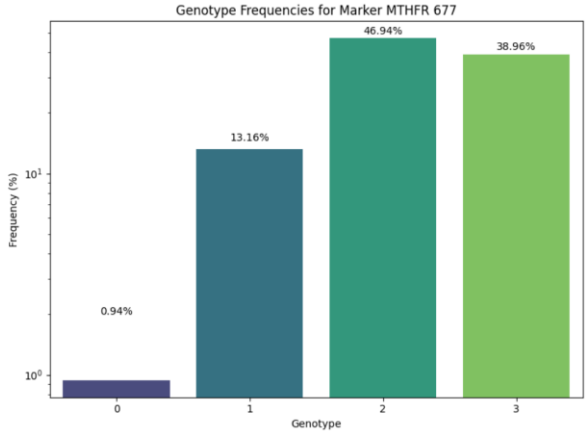

Figure 3. **Genotype Frequencies for Marker MTHFR 677** Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

B. MTHFR 1298

Similar to MTHFR 677, the MTHFR 1298 marker has a dominant distribution with the highest percentage of samples in the mut/mut variant (50.24%). The heterozygous wt/mut variant accounts for 39.69%, while the wt/wt genotype is present in only 0.94% of cases.

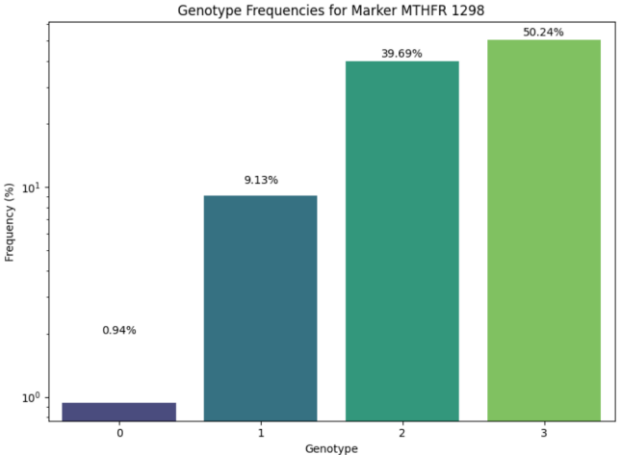

Figure 4. **Genotype Frequencies for Marker MTHFR 1298** Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

C. MTR

For the MTR marker, the majority of the samples fall into the mut/mut variant with 66.29%, while smaller percentages are present in wt/mut (28.49%) and wt/wt (1.27%).

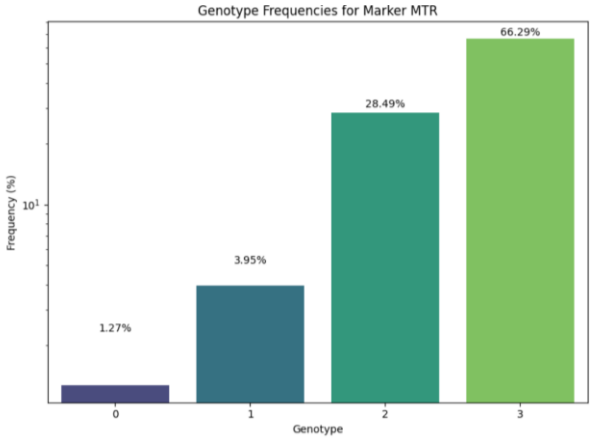

Figure 5. **Genotype Frequencies for Marker MTR:** Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the MTR marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

D. MTRR

For the MTRR marker, the mut/mut genotype dominates (43.39%), while the heterozygous variant makes up 30.45%. The proportion of wt/wt is 1.27%, indicating a very low frequency of this variant in the analyzed population.

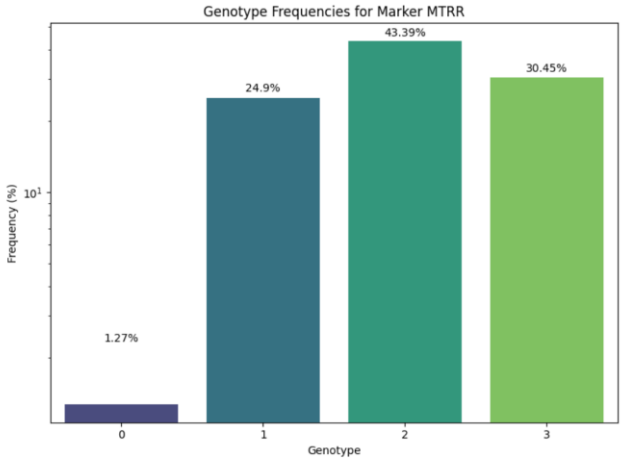

Figure 6. Genotype Frequencies for Marker MTRR: Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the MTRR marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

E. F2, F5, F7, F13, FGB, ITGA2, ITGB3, PAI-1

For the other markers (F2, F5, F7, F13, FGB, ITGA2, ITGB3, PAI-1), the majority of the samples also belong to the mut/mut variant, with percentages ranging from 62.52% to 95.98%. In all of these markers, the wt/wt variant is present in extremely small percentages (less than 1%), while the heterozygous variants are present in lower percentages than the mut/mut variant.

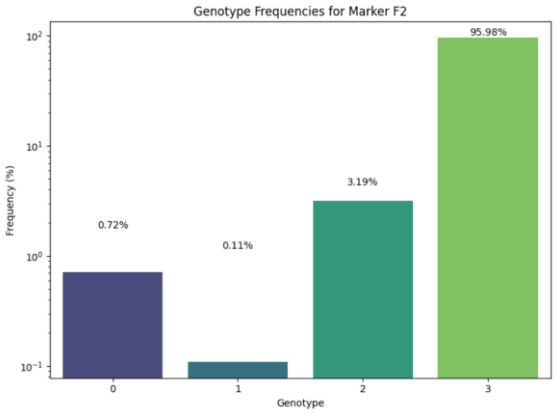

Figure 7. **Genotype Frequencies for Marker F2:**Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the F2 (prothrombin) marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

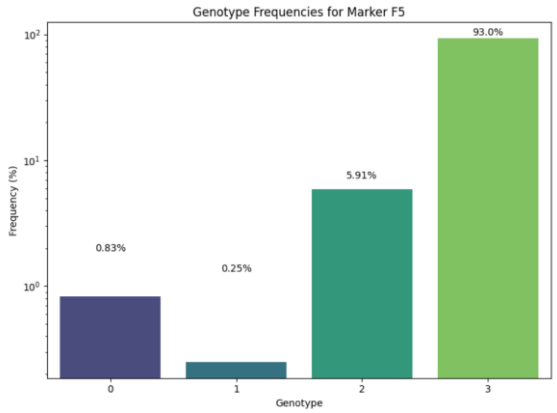

Figure 8. **Genotype Frequencies for Marker F5:** Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the F5 (Factor V Leiden) marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

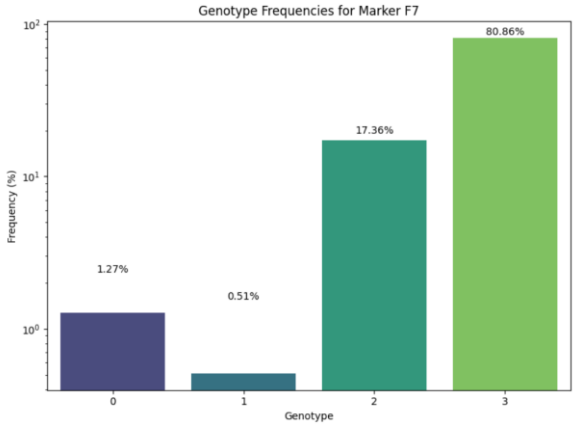

Figure 9. **Genotype Frequencies for Marker F7**: Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the F7 (Factor VII) marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

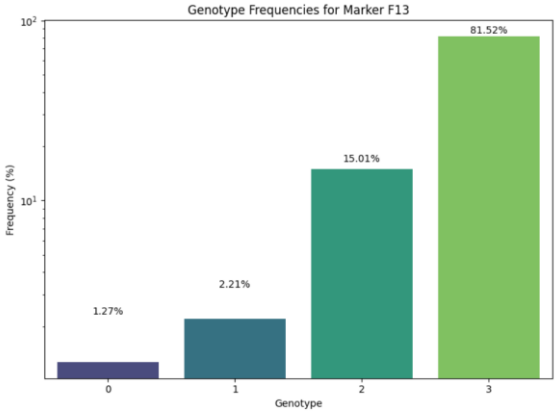

Figure 10. **Genotype Frequencies for Marker F13 -** Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the F13 (Factor XIII) marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

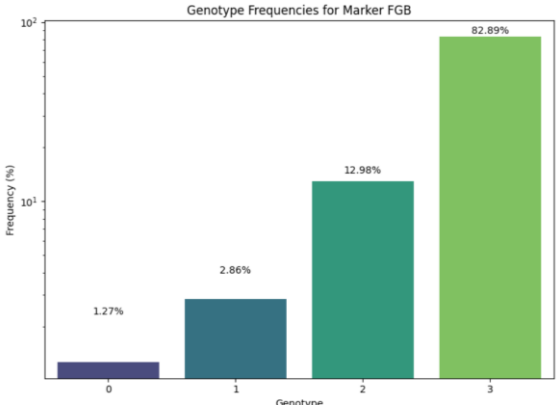

Figure 11. **Genotype Frequencies for Marker FGB** Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the FGB (fibrinogen β-chain) marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

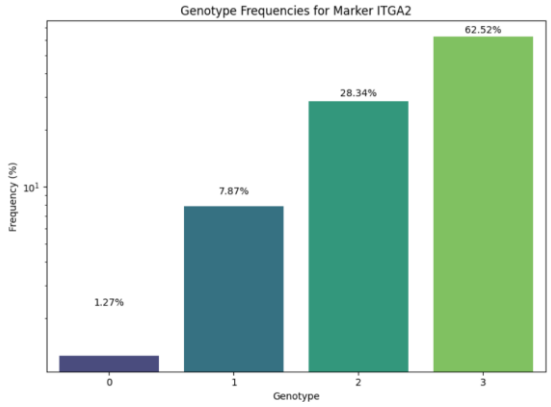

Figure 12. **Genotype Frequencies for Marker ITGA2**: Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the ITGA2 marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

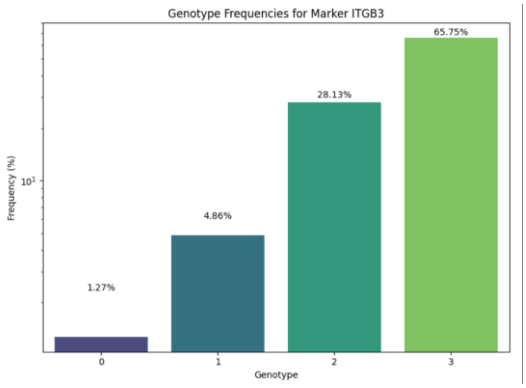

Figure 13. **Genotype Frequencies for Marker ITGB3:**Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the ITGB3 marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

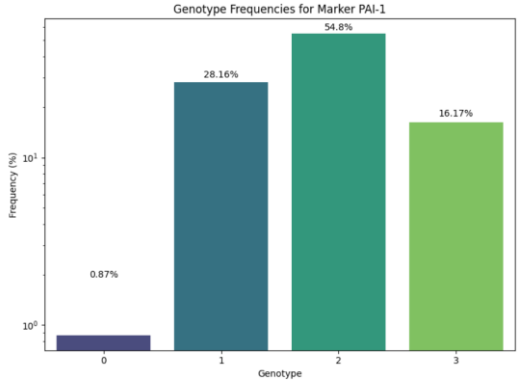

Figure 14. **Genotype Frequencies for Marker PAI-1:**Bar plot showing the percentage frequency of genotypes 0 (wt/wt), 1 (wt/mut), 2 (mut/mut), and 3 (other variants) for the PAI-1 marker in a cohort of N = 2,760 pregnant women. Each bar is annotated with its exact percentage value.

V. SHANNON INDEX FOR MARKERS

The Shannon Index values are presented as absolute values, as by definition, this index ranges from 0 (no diversity) to some positive maximum value. The original calculation yielded negative values due to the base-2 logarithm, but these have been converted to their absolute values for correct interpretation.

On the X-axis of the chart, genetic markers are displayed, specifically genes such as MTHFR 677, MTHFR 1298, MTR, F2, F5, PAI-1, and others. On the Y-axis, the values of the Shannon index are shown, all of which are negative—an unusual result, as the Shannon index is by definition a positive number or zero, which will be further discussed in the following analysis.

Interpretation:

- A higher Shannon index indicates greater genetic diversity.
- A lower index suggests lower diversity, which may indicate the dominance of a single variant.

The values of the Shannon index on the chart are all negative, which is not consistent with the standard interpretation of this index, as it is by definition always a positive number or zero—zero in cases where only one allelic variant exists. This negativity likely results from the use of a base-10 or base-2 logarithm without applying the negative sign (i.e., omitting the minus in front of the log), or it may simply be a visual convention intended to improve the orientation of the graph, where higher diversity is displayed lower on the Y-axis and lower diversity higher.

Looking at the specific values, the F5 and F2 markers show the lowest diversity (with the most negative index values), which may indicate the dominance of a single allelic variant in the population. On the other hand, the MTHFR 677 marker has an index closest to zero, suggesting the highest genetic diversity among the markers presented, indicating a more balanced distribution of allelic variants.

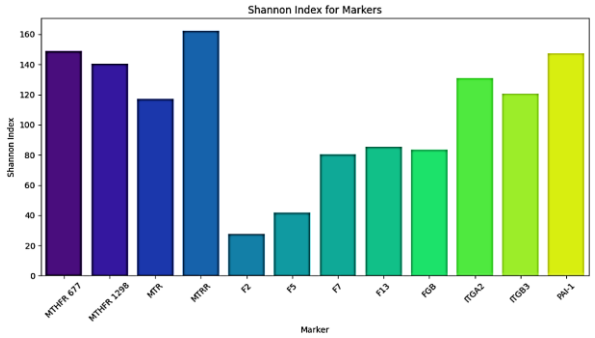

Figure 15. **Genotype Frequencies for Marker PAI-1:**Bar chart displaying the Shannon diversity index (H') for each of the 12 thrombophilia-associated markers—MTHFR 677, MTHFR 1298, MTR, MTRR, F2, F5, F7, F13, FGB, ITGA2, ITGB3, and PAI-1—calculated from genotype frequencies in N = 2,760 pregnant women. Higher H′ values indicate greater genetic diversity (more balanced genotype distribution), with MTR showing the lowest diversity and MTHFR 1298 the highest. All indices are non-negative, reflecting correct application of the formula $H' = -\sum_{i=1}^{k} p_i log_2(p_i)$

VI. DISCUSSION OF THE CORRELATION MATRIX

The general impression of the correlation matrix is that the coefficient values range from -1 to 1. A value of 1.00 indicates a perfect positive correlation, meaning that the two variables increase or decrease simultaneously. Conversely, a value of -1.00 represents a perfect negative correlation, where an increase in one variable is associated with a decrease in the other. A value of 0.00 suggests no correlation between the two variables. Visually, the use of color enhances interpretation: red indicates a stronger positive correlation, blue indicates a negative correlation, while white and light shades suggest a weak or nearly nonexistent relationship between the markers being observed. In the correlation matrix, several moderate to strong correlations were identified.Particularly notable were correlations between F2–F5 (0.46), F5–F13 (0.46), and F13–FGB (0.46). While these values fall short of what is traditionally considered strong correlation ($r > 0.7$), they nevertheless suggest important functional or genetic associations among these factors, which is expected considering their collective involvement in the blood coagulation process. Additionally, moderate correlations ranging from 0.25 to 0.38 were identified between pairs such as MTR – F2 (0.30), F7 – F13 (0.38) and F2 – F7 (0.38), which may indicate partial functional associations or shared contributions in the regulation of coagulation.

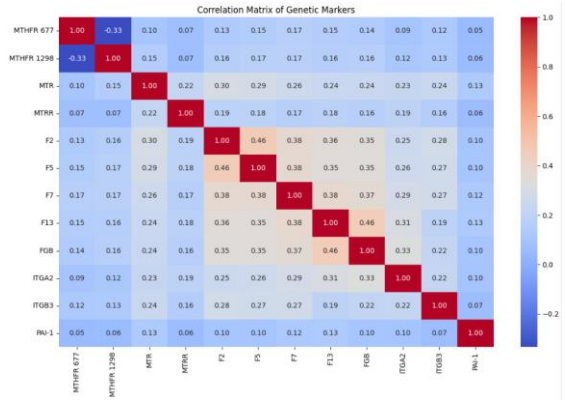

Figure 16. Correlation Matrix of Genetic Markers

VII. CONCLUSIONS

This scientific paper provides valuable insights into the genetic basis of thrombophilia, with a specific focus on genotype distributions, allele frequencies, genetic diversity, and the relationships between various genetic markers. The analysis of 12 genetic markers highlighted the prevalence of homozygous mutant (mut/mut) variants in the population, with varying levels of heterozygous (wt/mut) and wild-type (wt/wt) variants across different markers. The markers examined play significant roles in biological processes, particularly in blood coagulation. The Shannon index revealed varying levels of genetic diversity across the markers, with some markers indicating a higher dominance of specific allelic variants, while others suggested a more balanced distribution. The negative values observed in the Shannon index chart were an anomaly in the interpretation, possibly resulting from the use of logarithmic transformations or a visual representation convention. The correlation matrix further demonstrated strong positive correlations among certain markers, such as F2, F5, F13, and FGB, which are closely linked due to their shared involvement in coagulation pathways. Conversely, the negative correlation between MTHFR 677 and MTHFR 1298 indicated a genetic exclusivity between the two markers, which could reflect the exclusion of one variant by the other within the same gene. Overall, this scientific paper enhances our understanding of the genetic foundation of thrombophilia, providing new avenues for early detection and risk management, particularly in pregnant women. The findings also offer important implications for future genetic studies and could help refine predictive models to address thrombophilia-related complications during pregnancy.

ACKNOWLEDGEMENT

This study was supported by **Aqualab Plus d.o.o (Kneginje Zorke 30a, Belgrade)**. The authors would like to express their gratitude for the technical support, laboratory resources, and expert guidance provided throughout the research process. Special thanks also go to **Family Medica polyclinics and private practices** for their collaboration and contribution. The authors express their sincere gratitude to **Dr. Nebojša Bogdanović**, a structural biologist, for the constructive discussion and critical reading of the manuscript. We also extend our thanks to **Dr. Nemanja Erčić**, a specialist in gynecology and obstetrics, for his valuable advice and suggestions during the preparation of this paper.

REFERENCES

[1] A. Hatzaki et al., "The impact of heterozygosity for the factor V Leiden and factor II G20210A mutations on the risk of thrombosis in Greek patients," International Angiology: A Journal of the International Union of Angiology, vol. 22, no. 1, pp. 79-82, 2003.

[2] J. Kvasnicka et al., "Prevalence trombofilních mutací FV Leiden, protrombinu G20210A a PAI-1 4G/5G a jejich vzájemných kombinací v souboru 1450 zdravých osob stredního veku v regionu Praha a strední Cechy (výsledky real-time PCR analýzy FRET)," Casopis lekaru ceskych, vol. 151, no. 2, pp. 76–82, 2012.

[3] P. Hundsdoerfer et al., "Homozygous and double heterozygous Factor V Leiden and Factor II G20210A genotypes predispose infants to thromboembolism but are not associated with an increase of foetal loss," Thrombosis and Haemostasis, vol. 90, no. 4, pp. 628–635, 2003, doi: 10.1160/TH03-02-0096.

[4] J. L. Kujovich, "Prothrombin-related thrombophilia," GeneReviews, University of Washington, Seattle, WA, 2006.

[5] A. Dautaj et al., "Hereditary thrombophilia," Acta Biomedica, vol. 90, Suppl. 10, pp. 44–46, 2019.

[6] M. M. Patnaik and S. Moll, "Inherited antithrombin deficiency: A review," Haemophilia, vol. 14, no. 6, pp. 1229–1239, 2008.

[7] N. Aračić et al., "The impact of inherited thrombophilia types and low molecular weight heparin treatment on pregnancy complications in women with previous adverse outcomes," Yonsei Medical Journal, vol. 57, no. 5, pp. 1230–1236, 2016.

[8] SACHER, RONALD A. Thrombophilia: a genetic predisposition to thrombosis. Transactions of the American Clinical and Climatological Association, 1999, 110: 51.

[9] MOONESINGHE, Ramal, et al. A Hardy-Weinberg equilibrium test for analyzing population genetic surveys with complex sample designs. American journal of epidemiology, 2010, 171.8: 932-941.

[10] RYCKMAN, Kelli; WILLIAMS, Scott M. Calculation and use of the Hardy-Weinberg model in association studies. Current protocols in human genetics, 2008, 57.1: 1.18. 1-1.18. 11.

[11] HANSEN, Thomas F.; WAGNER, Günter P. Modeling genetic architecture: a multilinear theory of gene interaction. Theoretical population biology, 2001, 59.1: 61-86.

[12] ROHLFS, R. V.; WEIR, B. S. Distributions of Hardy–Weinberg equilibrium test statistics. Genetics, 2008, 180.3: 1609-1616.

[13] GRAFFELMAN, Jan; MORENO, Victor. The mid p-value in exact tests for Hardy-Weinberg equilibrium. Statistical applications in genetics and molecular biology, 2013, 12.4: 433-448.

[14] GRAFFELMAN, Jan; WEIR, B. S. Testing for Hardy–Weinberg equilibrium at biallelic genetic markers on the X chromosome. Heredity, 2016, 116.6: 558-568.

[15] HEDRICK, Philip W. Genetics of populations. Jones & Bartlett Publishers, 2009.

[16] CLANCY, Damian; PEARCE, Christopher J. The effect of population heterogeneities upon spread of infection. Journal of Mathematical Biology, 2013, 67.4: 963-987.

[17] KONOPIŃSKI, Maciej K. Shannon diversity index: a call to replace the original Shannon's formula with unbiased estimator in the population genetics studies. PeerJ, 2020, 8: e9391.

[18] Kosman, E. (2014). Measuring diversity: from individuals to populations: mini-review. European Journal of Plant Pathology, 138, 467-486.

[19] SHERWIN, William B. Entropy and information approaches to genetic diversity and its expression: genomic geography. entropy, 2010, 12.7: 1765-1798.

[20] Sherwin, W. B. (2010). Entropy and information approaches to genetic diversity and its expression: genomic geography. entropy, 12(7), 1765-1798.

[21] COHEN, Jacob, et al. Applied multiple regression/correlation analysis for the behavioral sciences. Routledge, 2013.

[22] ASUERO, Agustin Garcia; SAYAGO, Ana; GONZÁLEZ, A. G. The correlation coefficient: An overview. Critical reviews in analytical chemistry, 2006, 36.1: 41-59.

[23] KUMAR, Dheeraj, et al. clusiVAT: A mixed visual/numerical clustering algorithm for big data. In: 2013 IEEE International Conference on Big Data. IEEE, 2013. p. 112-117.

[24] UITDEWILLIGEN, Jan GAML, et al. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PloS one, 2013, 8.5: e62355.

Mcs Katarina Živojinović holds a degree in Economics and brings over seven years of professional experience across the banking and healthcare sectors. She currently serves as the General Director of a network of polyclinics and medical practices, in addition to her role as Deputy CEO for a prominent laboratory chain. Demonstrating a strong commitment to continuous professional growth, Ms. Živojinović is presently pursuing two master's programs: an international Executive MBA and a domestic degree specializing in Business Analytics. She possesses well-developed leadership and organizational capabilities, with a proven track record in strategic management and team coordination.

Mcs Marko Živanović is a doctoral student at the Faculty of Technical Sciences at the University of Čačak, specializing in Information Technology. He has completed both vocational and academic undergraduate and master's studies. His research interests include artificial intelligence, machine learning, and soft computing. Marko has written several scientific papers and works on industry projects. He is currently employed at the Faculty of Information Technology at Metropolitan University in Belgrade and also teaches part-time at a modern high school.

Msc Stefan Erčić graduated from BSc and MSc studies at the Faculty of Biology at the University of Belgrade. After graduation, he worked in various fields of fundamental and applied biology. He gained his laboratory experience in the field of physiology, microbiology and genetics both in the country, the Directorate for National Reference Laboratories (Ministry of Agriculture) in Serbia and abroad by working in Eurofins Scientific in Germany. He currently works in the education sector.

PhD Vanja Luković is an Associate Professor in the Department of Computer and Software Engineering at the Faculty of Technical Sciences in Čačak, University of Kragujevac, Serbia.

She earned her Master's degree in Electrical Engineering from the Faculty of Electrical Engineering in Belgrade in 2007 and completed her PhD in Technical Sciences at the Faculty of Technical Sciences in Čačak in 2015. Since 2001, she has been a dedicated member of the faculty. Her research interests include digital circuits, intelligent systems and ontologies, machine learning algorithms, neural networks, information systems, and object-oriented programming in C++, C#, and Java. She has authored or co-authored more than 80 scholarly papers.