

## RESEARCH ARTICLE

# A phenotype driven integrative framework uncovers molecular mechanisms of a rare hereditary thrombophilia

Noël Malod-Dognin<sup>1,2</sup>\*, Gaia Ceddia<sup>1</sup>, Maja Gvozdenov<sup>3</sup>, Branko Tomić<sup>3</sup>, Sofija Dunjić Manevski<sup>3</sup>, Valentina Djordjević<sup>3</sup>, Nataša Pržulj<sup>1,2,4</sup>

**1** Barcelona Supercomputing Center (BSC), Barcelona, Spain, **2** Department of Computer Science, University College London, London, United Kingdom, **3** Institute of Molecular Genetics and Genetic Engineering (IMGGE), University of Belgrade, Belgrade, Serbia, **4** ICREA, Barcelona, Spain

✉ These authors contributed equally to this work.

\* [noel.malod@bsc.es](mailto:noel.malod@bsc.es)



## OPEN ACCESS

**Citation:** Malod-Dognin N, Ceddia G, Gvozdenov M, Tomić B, Dunjić Manevski S, Djordjević V, et al. (2023) A phenotype driven integrative framework uncovers molecular mechanisms of a rare hereditary thrombophilia. PLoS ONE 18(4): e0284084. <https://doi.org/10.1371/journal.pone.0284084>

**Editor:** Achraf El Allali, Mohammed VI Polytechnic University, MOROCCO

**Received:** July 21, 2022

**Accepted:** March 23, 2023

**Published:** April 25, 2023

**Copyright:** © 2023 Malod-Dognin et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** We provide the germline datasets and scripts used for this analysis at [https://gitlab.bsc.es/gaiac/thrombophilia\\_datasets.git](https://gitlab.bsc.es/gaiac/thrombophilia_datasets.git).

**Funding:** N.P. was awarded by the European Research Council (ERC) Consolidator Grant 364770827 and by the Spanish State Research Agency AEI 10.13039/501100011033 grant number 365PID2019-105500GB-I00. B.T. was awarded by the Ministry of Education, Science and

## Abstract

Antithrombin resistance is a rare subtype of hereditary thrombophilia caused by prothrombin gene variants, leading to thrombotic disorders. Recently, the Prothrombin Belgrade variant has been reported as a specific variant that leads to antithrombin resistance in two Serbian families with thrombosis. However, due to clinical data scarcity and the inapplicability of traditional genome-wide association studies (GWAS), a broader perspective on molecular and phenotypic mechanisms associated with the Prothrombin Belgrade variant is yet to be uncovered. Here, we propose an integrative framework to address the lack of genomic samples and support the genomic signal from the full genome sequences of five heterozygous subjects by integrating it with subjects' phenotypes and the genes' molecular interactions. Our goal is to identify candidate thrombophilia-related genes for which our subjects possess germline variants by focusing on the resulting gene clusters of our integrative framework. We applied a Non-negative Matrix Tri-Factorization-based method to simultaneously integrate different data sources, taking into account the observed phenotypes. In other words, our data-integration framework reveals gene clusters involved with this rare disease by fusing different datasets. Our results are in concordance with the current literature about antithrombin resistance. We also found candidate disease-related genes that need to be further investigated. CD320, RTEL1, UCP2, APOA5 and PROZ participate in healthy-specific or disease-specific subnetworks involving thrombophilia-annotated genes and are related to general thrombophilia mechanisms according to the literature. Moreover, the ADRA2A and TBXA2R subnetworks analysis suggested that their variants may have a protective effect due to their connection with decreased platelet activation. The results show that our method can give insights into antithrombin resistance even if a small amount of genetic data is available. Our framework is also customizable, meaning that it applies to any other rare disease.

Technological 366 Development of the Republic of Serbia project 451-03-68/2022-14/200042. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

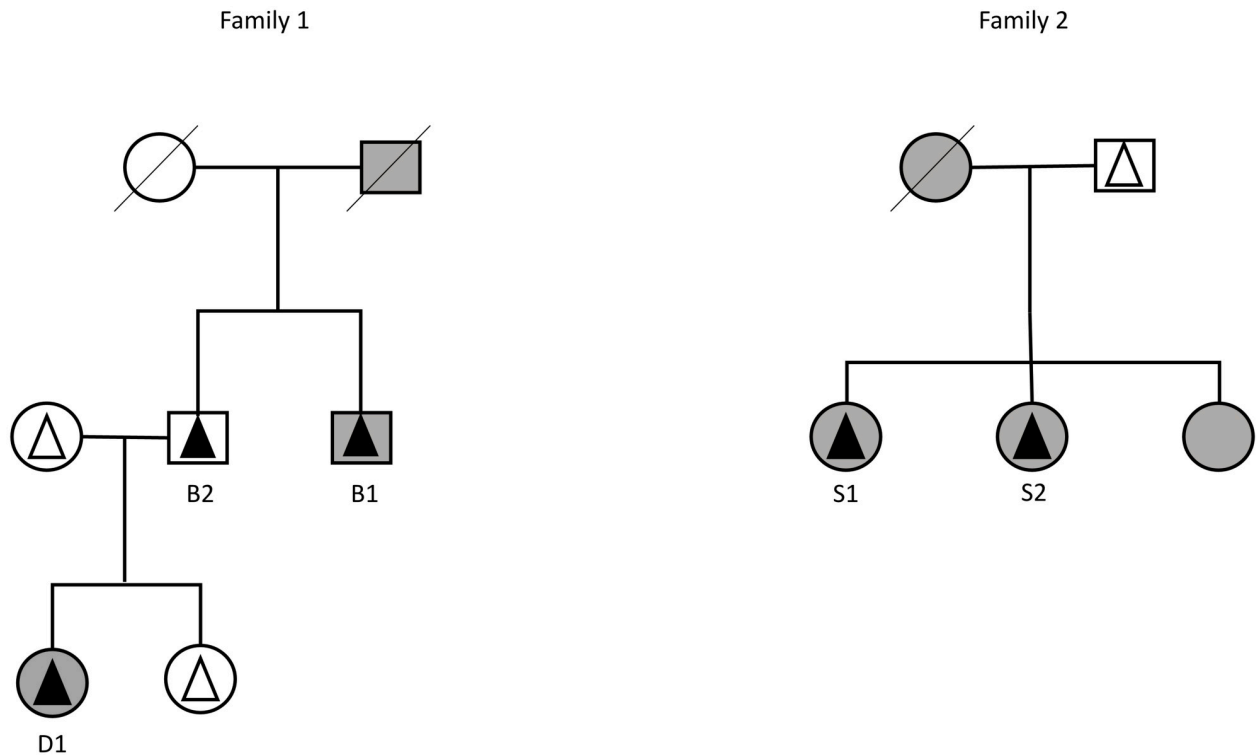
Familial thrombophilia (increased risk of developing thrombosis) is a multifactorial and polygenic disorder in which gene-gene and gene-environmental interactions contribute to complex phenotype manifestations. It is a common pathology underlying ischemic stroke, venous thromboembolism, ischemic heart disease and pregnancy loss [1–3]. Out of all deaths per day worldwide, one in four is caused by thrombosis, making it a leading global cause of death and disability [4]. Despite many clinical studies, the complex mechanisms of thrombophilia have not yet been elucidated [1, 4]. Considering the impact of thrombosis as a global cause of death, it is crucial to understand the mechanisms related to thrombophilia to improve the diagnosis, prevention and treatment further.

In this study, we focus on a rare subtype of hereditary thrombophilia, antithrombin resistance. Antithrombin resistance is caused by variants in the prothrombin gene (F2) located in the antithrombin binding region of the protein. The impaired thrombin inhibition by antithrombin leads to thrombotic disorders [5]. The Prothrombin Belgrade variant (c.1787G>A, p. Arg596Gln) is one of the antithrombin resistance-causing variants and it was detected for the first time in subjects from the Serbian population [6].

While genome-wide association studies (GWAS) have already demonstrated their effectiveness in finding disease-associated variants for common diseases, these approaches are not suited to rare diseases due to the heterogeneity and rarity of genetic data [7, 8]. On the other hand, it has been proven that Non-negative Matrix Tri-Factorization (NMTF) can deal with a low sample size by integrating prior knowledge into the small datasets [9]. NMTF-based data integration methodologies allow for combining different types of biological and medical data to provide a comprehensive view of biological systems [10]. Here, we propose an NMTF-based data integration (fusion) method that leverages different data sources to get molecular insights into antithrombin resistance. In other words, by applying our integrative framework to identify novel disease-related genes for antithrombin resistance, we can overcome the lack of genetic samples and make concrete progress toward possible treatments.

This study focuses on the full genome sequences of five heterozygous carriers (subjects) of the Prothrombin Belgrade variant from two Serbian families. These subjects have the following phenotypes. In family 1, the first brother,  $B_1$ , developed recurrent venous thromboembolism, while the second brother,  $B_2$ , is a healthy subject (asymptomatic carrier). Furthermore, the healthy brother has a daughter,  $D_1$ , who also had recurrent thrombotic events. In family 2, two sisters,  $S_1$  and  $S_2$ , experienced venous thromboembolism (illustrated in Fig 1).

To uncover genes for which our subjects possess germline variants (that we term “genes with variants”) associated with this rare subtype of hereditary thrombophilia, as well as the ones responsible for the healthy phenotype of brother  $B_2$ , we integrate the germline variant profiles of the subjects with three omics molecular datasets (protein-protein interactions, gene co-expressions and genetic interactions) using Simultaneous Orthogonal Non-negative Matrix Tri-Factorization (SONMTF), detailed in Section [Materials and methods](#). Also, we use the observed phenotypes as prior knowledge to enforce the subject stratification by separating healthy from diseased subjects in SONMTF. Then, we analyze the gene clusters resulting from SONMTF to find disease-related genes of antithrombin resistance. Our results demonstrate that integration and computational analysis of antithrombin resistance phenotypes and omics data using SONMTF has the potential to reveal novel insights into the biology of this rare genetic disease. Moreover, our framework’s straightforward applicability and effectiveness to a small number of genetic samples opens the door to its usage on other rare diseases.



**Fig 1. Observed phenotypes.** Squares stand for men and circles stand for women. The triangles mark the tested family members: black triangles—heterozygous carriers of prothrombin Belgrade mutations, white triangles—non-carriers of the prothrombin Belgrade mutation. Thrombophilia: grey—subjects suffering from thrombosis, white—healthy subjects (possibly asymptomatic carriers). B1: brother 1, B2: brother 2, D1: daughter 1, S1: sister 1, S2: sister 2.

<https://doi.org/10.1371/journal.pone.0284084.g001>

## Materials and methods

First, we describe the omics data used in this study, i.e., the molecular networks, the germline variants and the biological annotation datasets. Next, we explain our new data fusion framework based on Non-negative Matrix Tri-Factorization (NMTF) to integrate all the considered heterogeneous omics data. Finally, we extract clusters of subjects and clusters of genes and do the enrichment analysis in biological annotations on the resulting clusters of genes (detailed below) from which we find gene clusters specific to the healthy subject (henceforth termed *healthy specific*) and gene clusters specific to the diseased subjects (henceforth termed *disease specific*) (i.e., those gene subnetworks containing genes that are mutated only in the healthy subject and genes that are mutated in all diseased subjects, respectively; see below on the definitions of molecular networks).

## Datasets

**Molecular networks.** We collect three large-scale molecular interaction datasets for humans. From BioGRID 3.5.176 [11], we collect the experimentally validated protein-protein interactions, PPIs (270,554 PPIs between 17,104 proteins). From COXPRESdb v7.3 [12], we collect the co-expressions between genes (thresholded to preserve only the top 1% strongest co-expressions, following the methodology from Malod-Dognin *et al.* [13], leading to 4,172,940 interactions between 25,063 genes). Finally, we collect the genetic interaction data from Rausher *et al.* [14] (391,527 genetic interactions between 16,249 genes). Each molecular

**Table 1. Germline variant dataset.** For each subject (row), the table indicates the number of coding variants found in the subject's genome (column "#Genes with variants") and the number of these genes that are present in any of our molecular interaction datasets (column "#In the networks").

Subject	#Genes with variants	#In the networks
Brother 1	5,807	4,574
Brother 2	5,913	4,649
Daughter 1	6,218	4,924
Sister 1	6,327	4,996
Sister 2	6,446	5,089

<https://doi.org/10.1371/journal.pone.0284084.t001>

dataset is modeled as a network in which nodes represent genes and in which two nodes are connected by an edge if the corresponding genes interact in the dataset. As needed by our NMTF-based data-integration model, each network is represented by its adjacency matrix, in which a value in a given row and column is equal to one if the corresponding genes interact and zero otherwise. These matrices are all reordered so that a given row and column,  $i$ , represents the same gene in all matrices. Following the approach of Malod-Dognin *et al.* [13], and because PPI is the most direct evidence that two genes can interact, we exclude from our study (by removing the corresponding rows and columns from all matrices) the genes that have no PPI information at all (i.e., whose corresponding rows and columns in the PPI matrix are all zeros). However, we did not further filtered-out genes based on their variant status. Thus, while our molecular network matrices all have the same set of 17,104 genes (with variants or not), their edge sets (the ones in the matrices) differ.

**Germline variants.** The whole-genome sequences of five heterozygous carriers (subjects) of the Prothrombin Belgrade variant from two Serbian families are captured by using BGI/MGI DNBSEQ sequencing technology, 40x read coverage (Complete Genomics) and assembled using the human reference genome build 37 (GRCh37). From the whole genome sequences, we extract the protein-coding germline variants of each subject using the NCBI Homo Sapiens Annotation Release 104 [15] and Complete Genomics Analysis Tools, CGA Tools (with default settings) to filter out variants based on confidence scores and read coverages. We consider all the non-synonymous variants that map to coding regions, namely: missense, nonsense, nonstop, misstart, frameshift, insert, insert+, delete, delete+ and disrupt. We did not apply filters to remove common variants since the ones that protect the asymptomatic brother may be common. We term "genes with variants" the genes for which our subjects possess such non-synonymous variants.

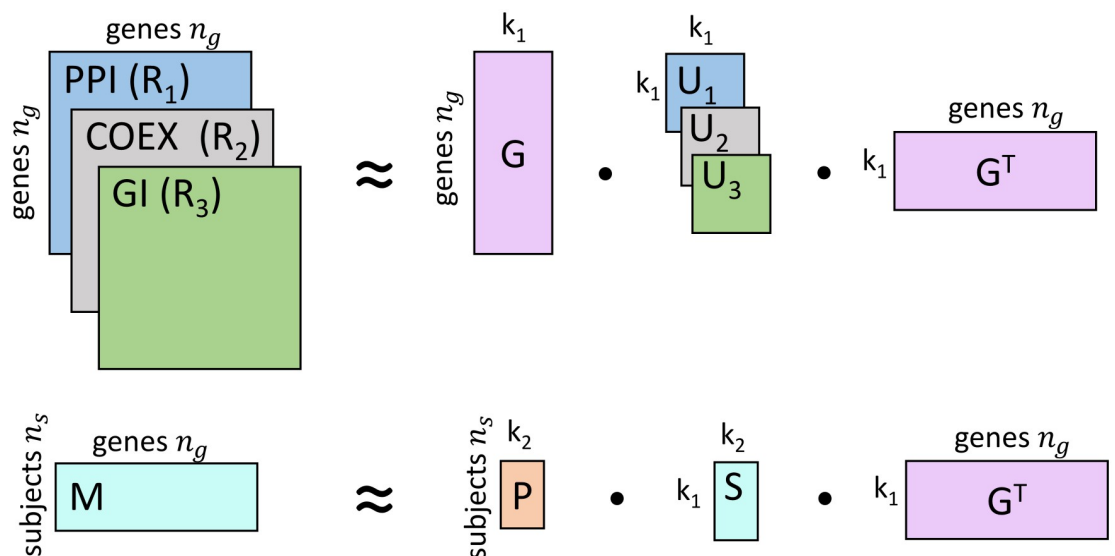
We verify that most of the genes with variants of the subjects are in the biological networks (about 79%, see Table 1). Also, to assess the relevance of these genes with variants to thrombophilia, we collect the variant-phenotype annotation dataset from DisGeNet v6.0 [16]. As detailed in S1 File, Section 3, we find that the genes with variants are highly related to thrombophilia and blood tests because 50.4% of their 492 associated phenotype annotations are either "Disease or Syndrome" annotations related to thrombophilia in PubMed publications, or "Laboratory Procedure" annotations related to blood tests, which is in accordance with thrombophilia-related tests and a risk of thrombosis (detailed in S1 File, Section 3).

**Biological annotations.** From Gene Ontology [17], we collect the experimentally validated biological process (BP), molecular function (MF) and cellular component (CC) annotations of the genes. Furthermore, from Reactome [18], we collect the pathways (RP) and reactions (RR) that the genes participate in. Finally, from DisGeNet v6.0 [16], we collect a list of 83 thrombophilia-related genes, out of which 76 are present in the networks.

### Data integration framework

To uncover genes with variants that are relevant to our cases of thrombophilia, we propose a data integration framework based on Non-negative Matrix Tri-Factorization (NMTF) [19], a machine learning technique originally proposed for co-clustering and dimensionality reduction that was recently used for data fusion [13, 20–24].

In our framework, illustrated in Fig 2, we consider the germline variant profiles of five subjects and also three molecular networks. Subjects and genes are related to each other by germline variant profiles, constructed for  $n_s$  subjects over  $n_g$  genes and captured in the relation matrix,  $M^{n_s \times n_g}$ . Its entries are binary values, with  $M[s][g] = 1$  if gene  $g$  is found to be mutated in subject  $s$ , and zero otherwise. Molecular networks, PPI, COEX and GI, are represented by their adjacency matrices,  $R_1^{n_g \times n_g}$ ,  $R_2^{n_g \times n_g}$  and  $R_3^{n_g \times n_g}$ , respectively. Their entries are binary values, with  $R_1[u][v] = 1$ ,  $R_2[u][v] = 1$ ,  $R_3[u][v] = 1$  if genes  $u$  and  $v$  interact in network PPI, COEX and GI, respectively, and zero otherwise. All these matrices are simultaneously decomposed as products of three matrix factors as  $M \approx P \cdot S \cdot G^T$  and  $R_i \approx G \cdot U_i \cdot G^T$  for  $1 \leq i \leq 3$ , where  $P^{n_s \times k_2}$  is interpreted as the *cluster membership indicator* matrix of subjects (grouping  $n_s$  subjects into  $k_2$  clusters),  $G^{n_g \times k_1}$  is interpreted as the *cluster membership indicator* matrix of genes (grouping  $n_g$  genes into  $k_1$  clusters) that is shared across all decompositions and hence allows learning from all data,  $S^{k_1 \times k_2}$  is interpreted as the compressed representation of the molecular profiles (that indicates how the  $k_2$  clusters of subjects relates to the  $k_1$  clusters of genes in  $M$ ), and  $U_i^{k_1 \times k_1}$  is interpreted as the compressed representation of each molecular network (that indicates how the  $k_1$  clusters of genes relate to each other in each molecular network). We also enforce the subject stratification (i.e., force the diseased subjects and the healthy subject to be in different clusters) by fixing the matrix factor,  $P$  (detailed in S1 File, Section 2). In other words, we take into account the observed phenotypes as prior knowledge to enforce the subject stratification with SONMTF. This leads to clusters of genes that better separate healthy-specific



**Fig 2. Our data-integration framework.** We used the NMTF method to integrate four different data types: germline mutations of the five subjects presented in  $M^{n_s \times n_g}$ , protein-protein interactions (PPIs) from BioGRID presented in  $R_1^{n_g \times n_g}$ , coexpressions (COEXs) from COXPRESdb presented in  $R_2^{n_g \times n_g}$  and genetic interactions (GIs) from [14] presented in  $R_3^{n_g \times n_g}$ .  $M$  is simultaneously decomposed into the product of three lower dimensional matrix factors,  $P$ ,  $S$  and  $G^T$ ; molecular networks,  $R_i$ , are simultaneously decomposed into the product of three factors,  $G$ ,  $U_i$  and  $G^T$ , as detailed in Materials and methods.  $k_1$  and  $k_2$  are the gene and the patient clusters, respectively.

<https://doi.org/10.1371/journal.pone.0284084.g002>

and disease-specific variants. We also demonstrate that our method is robust to data imbalance by comparing cluster stability of SONMTF runs, considering pairs of subjects as input (see [S1 File](#), Section 2).

This decomposition is done by solving the following minimization function, approximately solving the NMTF problem for simultaneous factorization of matrices  $M$  and  $R_i$  into three matrix factors each, in which we have made matrix factors  $P$  and  $G$  orthogonal to minimize dependencies and provide a *hard* clustering interpretation of  $P$  and  $G$  [25] as mentioned above (hence the name Simultaneous Orthogonal Non-negative Matrix Tri-Factorization, SONMTF):

$$\min_{P,S,G,U_i \geq 0} \|M - PSG^T\|_F^2 + \sum_{i=1}^3 \|R_i - GU_iG^T\|_F^2, \quad (1)$$

$$\text{s.t. } P^T P = I \text{ and } G^T G = I,$$

where  $F$  denotes the Frobenius norm. We heuristically minimize the above function using dedicated multiplicative update rules (see [S1 File](#), Section 1).

### Extracting clusters and subnetworks

The resulting matrices,  $P^{n_s \times k_2}$  and  $G^{n_g \times k_1}$ , are *cluster membership indicator* matrices for subjects and genes, respectively; based on their entries,  $n_s$  subjects are assigned to  $k_2$  subject clusters,  $n_g$  genes are assigned to  $k_1$  gene clusters, respectively. We extract subject and gene clusters from  $P$  and  $G$ , respectively, by using the hard clustering procedure as done by Brunet *et al.* [26]. In particular, for each row in  $P$ , we place subject  $s$  into cluster  $k$  if  $P[s][k]$  is the largest entry in row  $s$ . We apply the same procedure to extract gene clusters from  $G$ . Moreover, for each identified gene cluster, we create the corresponding *subnetwork* in which nodes represent genes in that specific cluster and in which two nodes are connected by an edge if the corresponding genes interact in any of the molecular interaction networks (since the edge set is different for each molecular interaction network).

### Enrichment in biological annotations

We identify the annotations that are significantly enriched in the clusters of genes as follows. The probability that an annotation is enriched in a cluster is computed using sampling without replacement strategy (also called the hypergeometric test) [27]:

$$p = 1 - \sum_{i=0}^{X-1} \binom{K}{i} \binom{M-K}{N-i} / \binom{M}{N}.$$

where  $N$  is the size of the cluster (only annotated genes from the cluster are taken into account),  $X$  is the number of genes in the cluster that are annotated with the annotation in question,  $M$  is the number of annotated genes in the network and  $K$  is the number of genes in the network that are annotated with the annotation in question. An annotation is considered to be significantly enriched if its enrichment  $p$ -value, after correction for multiple hypothesis testing [28], is lower than or equal to 5%. Then, we measure the quality of the clustering by the percentage of clusters having at least one enriched annotation.

### Ethics statement

This study was approved by the Ethic committee of the Institute of Molecular Genetics and Genetic Engineering, University of Belgrade (O-EO-004/2015/2). Written informed consent was obtained for all participants.



## Results

We report computational and biological results based on our data-integration framework. Computationally, we adapted our approach to consider subjects' phenotypes as detailed in [S1 File](#), Section 2. We used four different data types: germline variants, protein-protein interactions, co-expressions and genetic interactions. Each dataset is modeled as a network and integrated as detailed in Section [Materials and methods](#). By leveraging the co-clustering interpretation of the SONMTF, which simultaneously decomposes molecular interactions (protein-protein interactions, co-expressions and genetic interactions) and germline variants datasets, we retrieved interesting gene clusters and the corresponding subnetworks as detailed in Section [Extracting clusters and subnetworks](#) (see Section [Healthy-specific and disease-specific gene cluster analysis](#)). We also use the PhD-SNPg method [29] to compare our results with a baseline approach that predicts the impact of germline variants. The results presented in [S1 File](#), Section 3, show that the PhD-SNPg method identifies 1.01% of pathogenic variants in our datasets. In particular, two of the 44 reported gene clusters are significantly enriched in candidate pathogenic variants identified by PhD-SNPg ( $p$ -values  $<0.05$ ).

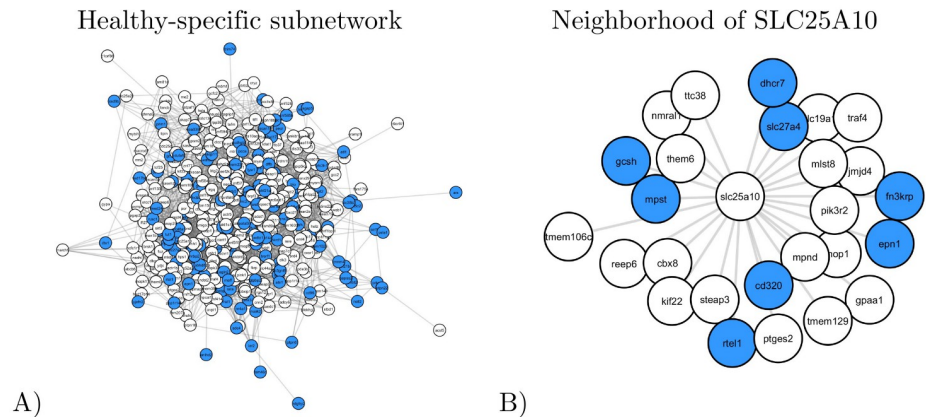
### Healthy-specific and disease-specific gene cluster analysis

The outputs of the SONMTF method are clusters and subnetworks of genes (see Section [Extracting clusters and subnetworks](#)), that best separate diseased subjects from the healthy carrier (detailed in [S1 File](#), Section 2). In particular, the resulting gene clusters are significantly enriched in GO biological processes (adjusted  $p$ -values  $<0.05$ ), i.e., they are biologically coherent compared to random clusters (see [S2 Fig](#) in [S1 File](#)). In the following Sections, we focus on two clusters having the largest percentages of *healthy-specific variants* (genes that are mutated only in the asymptomatic carrier) and *disease-specific variants* (genes mutated in all diseased subjects, but not mutated in the healthy one), respectively. Moreover, we analyze two other clusters: one containing ADRA2A and TBXA2R, the only two thrombophilia-related genes mutated in the healthy subject and the other containing F2, the main antithrombin resistance-causing gene, which is mutated in all subjects. For each subnetwork, we highlight well-known thrombophilia genes present in the DisGeNet database.

**Healthy-specific subnetwork.** We focus on the cluster containing the highest percentage of healthy-specific variants, which corresponds to the *healthy-specific subnetwork* of 461 nodes (genes) and 3,115 edges. Illustrated in [Fig 3](#), panel A, this subnetwork contains 29.1% of genes with healthy-specific variants (colored in blue) and does not contain any disease-specific variant genes (i.e., specific to diseased carriers). We hypothesize that this cluster contains specific variants that protect the healthy carrier from the disease.

Functionally, the healthy-specific subnetwork contains five protein-tyrosine phosphatases (PTPs) out of the 46 present in the entire dataset, and it is significantly enriched in "protein dephosphorylation" (GO:0006470). PTPs are known to be involved in the regulation of platelet activity, and thus, in thrombosis, validating our approach [30].

Within this subnetwork, SLC25A10 is the only gene known to be involved in thrombophilia, according to DisGeNet. SLC25A10 has 26 neighbors in the healthy-specific subnetwork ([Fig 3](#), panel B), out of which eight are mutated only in the healthy subject: CD320, DHCR7, EPN1, FN3KRP, GCSH, MPST, RTEL1 and SLC27A4. While these genes are not associated with thrombophilia, we find literature evidence that suggests that some of these genes may be relevant. For instance, there is a possible association between variants in the RTEL1 gene and stroke risk in the Chinese population [31] while CD320 (transcobalamin 2 receptor) is known to be associated with hyperhomocysteinemia, which is a risk factor for cardiovascular disease [32].



**Fig 3. An illustration of the healthy-specific subnetwork.** Nodes corresponding to the genes that have healthy-specific variants are colored in blue **A**: A spring embedding visualization of the subnetwork. **B**: The direct neighborhood of SLC25A10 in the subnetwork.

<https://doi.org/10.1371/journal.pone.0284084.g003>

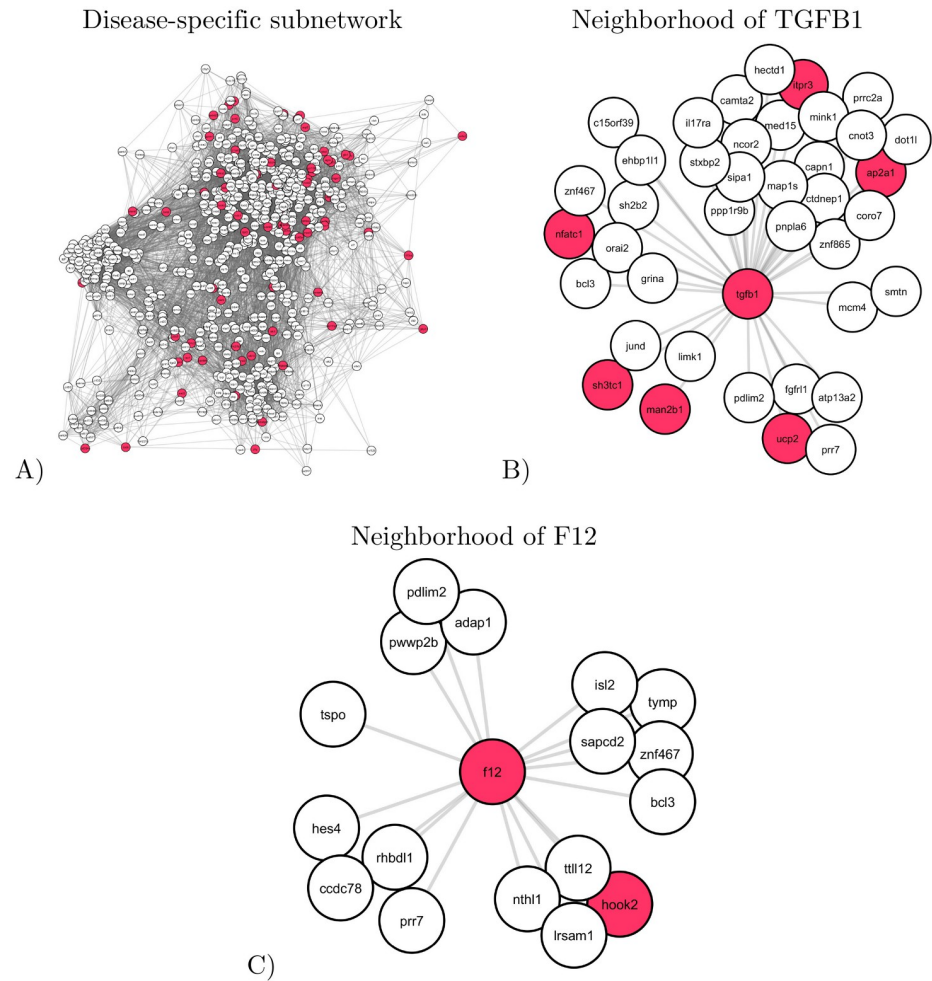
**Disease-specific subnetwork.** The cluster containing the largest percentage of genes with disease-specific variants corresponds to the subnetwork containing 695 nodes (genes) and 11,862 edges. Illustrated in Fig 4, panel A, this subnetwork contains 9.93% of genes with disease-specific variants (genes mutated in all diseased subjects, but not mutated in the healthy one, colored in red) and does not contain any gene with healthy-specific variant (genes that are mutated only in the healthy subject).

Functionally, the disease-specific subnetwork is significantly enriched in “Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal” (R-HSA-141444) Reactome pathway. MAD2 is known to be involved in platelet production [33]; thus, a dysregulation of MAD2 may result in inappropriate platelet adhesion/activation and thrombosis [33]. The disease-specific subnetwork is also significantly enriched in “attachment of mitotic spindle microtubules to kinetochore” (GO:0051315), “metaphase plate congression” (GO:0051310) and “chromosome segregation” (GO:0007059) GO Biological Processes, all being part of the “cell cycle process” (GO:0022402). Wang *et al.* [34] showed that platelets could inhibit proliferation mainly through the arrest of the cell cycle and inhibition of DNA synthesis. Thus, the Reactome and GO enrichments detailed above may result from platelet alterations.

Moreover, the disease-specific subnetwork contains two thrombophilia-related genes present in the DisGeNet database, TGFB1 (transforming growth factor beta-1, TGF $\beta$ -1) and F12 (coagulation factor XII), which are both disease-specific variants (Fig 4, panels B and C). TGFB1 gene regulates proliferation, maturation, differentiation, motility and apoptosis of cells [35]. It is also involved in the formation of blood vessels, wound healing, development of muscle tissue and body fat, inflammatory processes in the immune system and prevention of tumor growth [36, 37]. There are studies associating TGFB1 variants with heart disease, hypertension, myocardial infarction, and coronary artery disease [36, 38]. Though there is no direct interaction of proteins TGFB1 and thrombin, there are studies connecting their functions in immunity and hemostasis [39, 40]. Since all subjects in this study are heterozygous carriers of the Prothrombin Belgrade variant, it is possible that the TGFB1 variant could have an additive contribution to the clinical outcome of the subjects and lead to severe clinical manifestations.

F12 gene (coagulation factor XII) is essential for normal blood clotting [41]. An inactive form of F12 circulates in the bloodstream until it is activated. Upon activation, F12 interacts with the coagulation factor XI and with a plasma protein, prekallikrein. This process leads to





**Fig 4. An illustration of the disease-specific subnetwork.** Nodes corresponding to the genes that have disease-specific variants are colored in red. **A:** A spring embedding visualization of the subnetwork. **B and C:** The direct neighborhoods of TGFB1 (panel B) and F12 (panel C) in the subnetwork.

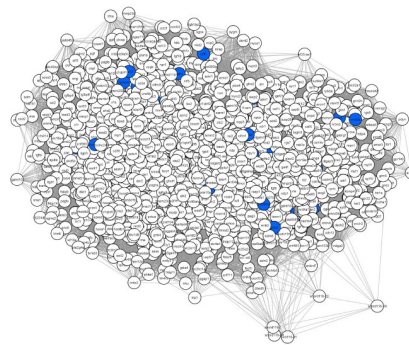
<https://doi.org/10.1371/journal.pone.0284084.g004>

increased permeability of blood vessel walls and inflammation [41]. In our study, we detect several different variants in F12 gene for symptomatic subjects. However, their relevance for the associated molecular mechanism of the Prothrombin Belgrade variant needs to be further investigated.

Among the direct neighbors of TGFB1 and F12 genes within the disease-specific subnetwork, we find disease-specific variants that are related to thrombophilia. For instance, among the neighbors of TGFB1, UCP2 gene (mitochondrial uncoupling protein-2) is mutated in diseased subjects. UCP2 is known to be associated with venous thrombosis [42].

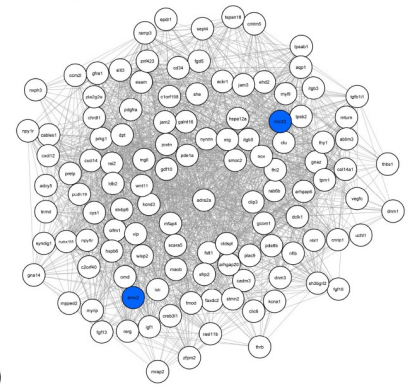
**ADRA2A and TBXA2R subnetwork.** The cluster containing ADRA2A and TBXA2R corresponds to the subnetwork of 1,591 nodes (genes) and 82,871 edges. Illustrated in Fig 5, panel A, the subnetwork contains 13 genes with healthy-specific variants (blue nodes in Fig 5). Functionally, the ADRA2A and TBXA2R subnetwork is significantly enriched in "Signaling by PDGF" (R-HSA-186797) Reactome pathway, namely the Platelet-derived Growth Factor (PDGF) signaling network. PDGF is a potent stimulator of growth and motility of connective tissue cells whose increase in protein expression has been associated with acute and chronic

Subnetwork containing ADRA2A and TBXA2R



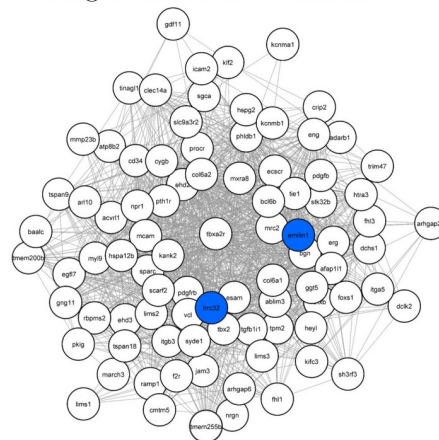
A)

Neighborhood of ADRA2A



B)

Neighborhood of TBXA2R



C)

**Fig 5. An illustration of the subnetwork containing ADRA2A and TBXA2R.** Nodes corresponding to the genes that have healthy-specific variants are colored in blue. **A:** A spring embedding visualization of the subnetwork. **B and C:** The direct neighborhoods of ADRA2A (panel B) and TBXA2R (panel C) in the subnetwork.

<https://doi.org/10.1371/journal.pone.0284084.g005>

venous thrombosis [43]. Moreover, the ADRA2A and TBXA2R subnetwork is significantly enriched in “angiogenesis” (GO:0001525), “nervous system development” (GO:0007399) and “vasculogenesis” (GO:0001570) GO Biological Processes, all related to thrombophilia, or vascular disease.

ADRA2A and TBXA2R are thrombophilia-related genes mutated in the healthy subject. ADRA2A encodes for alpha-2-adrenergic receptors, expressed at the platelet surface and coupled with the G protein. The activation of these receptors leads to platelet activation and aggregation, which are essential components in thrombus formation [44]. This indicates the pivotal role of alpha-2-adrenergic receptors in hemostasis and suggests that variants of ADRA2A detected in the healthy subject could interact with the Prothrombin Belgrade variant. However, functional studies with mutated alpha-2-adrenergic receptors should be conducted to investigate whether those variants could represent gene modifiers responsible for the health status of the Prothrombin Belgrade variant carrier. TBXA2R encodes for the thromboxane A2 receptor, a member of the G protein-coupled receptor family and one of the most important mediators in the process of hemostasis. Thromboxane binding leads to receptor activation and induces platelet aggregation and activation. The thromboxane A2 receptor has two isoforms in

humans, produced by alternative splicing of the primary transcript, which differs in the length of the cytoplasmic C-terminal end (15 amino acids for  $\alpha$  and 79 amino acids for  $\beta$  isoform). The cytoplasmic tail is important for proper coupling of G protein and signal transduction, thus indicating that these two isoforms might have differential roles in the etiology of human diseases [45, 46]. In our study, the healthy subject has been shown to have variants in the TBXA2R gene, located in the cytoplasmic region of the  $\beta$  isoform of the thromboxane A2 receptor. So far, reported variants in this region are associated with the loss of receptor function and impaired thromboxane A2 agonist-induced platelet aggregation, resulting in bleeding disorders [45, 46]. We hypothesize that this variant could inhibit signal transduction, leading to decreased platelet aggregation, thus protecting the healthy subject from the occurrence of thrombosis. The effect of this variant on platelet function and its interaction with the Prothrombin Belgrade mechanism should be investigated in future studies.

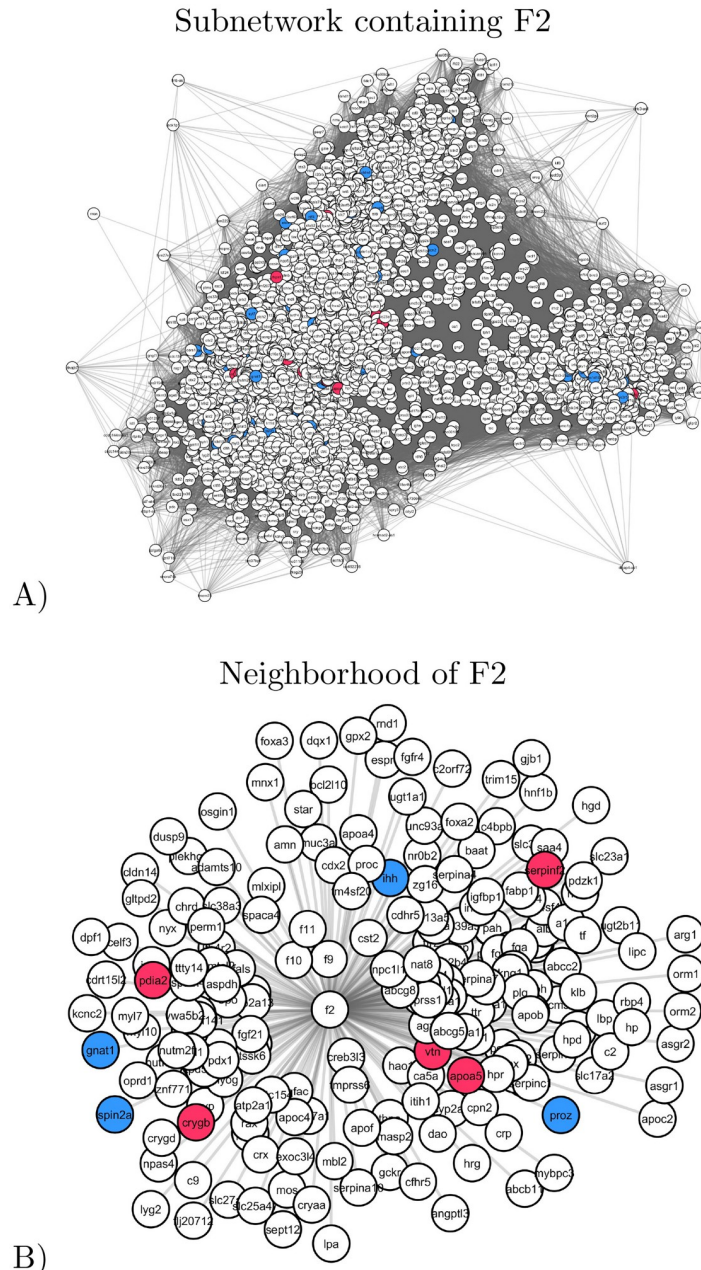
For these reasons, the importance of analyzing ADRA2A and TBXA2R subnetwork lies in the fact that their variants could protect the healthy carrier from thrombosis disorders. Within this subnetwork, ADRA2A has 117 neighbors (Fig 5, panel B), out of which two are mutated only in the healthy subject: LRRC32 and EMX2. Interestingly, LRRC32 is also a neighbor of TBXA2R (Fig 5, panel C), which has 88 neighbors in total. LRRC32 is a key regulator of transforming growth factor beta, inducing a latent state of TGF- $\beta$  in extracellular space [47]. This finding highlights the importance of TGFBI, as already mentioned in Section Disease-specific subnetwork.

**F2 subnetwork.** The subnetwork, containing F2, corresponds to a large network of 2,898 nodes and 178,748 edges. Illustrated in Fig 6, panel A, the F2 subnetwork contains 1.66% of genes with healthy-specific variants (genes mutated only in the healthy subject, in blue) and 0.76% of disease-specific variants (genes mutated in all diseased subjects but not mutated in the healthy one, colored in red).

Due to the large size of this subnetwork, functional enrichment analysis is not relevant. Instead, we focus on the thrombophilia-related genes. This subnetwork contains 23 thrombophilia-related genes, including many genes related to the activation of thrombin (F2, F5, F7, F9 and F11). Their variants are presented in S1 Table in S1 File. Within this subnetwork, F2 has 211 neighbors (Fig 6, panel B), out of which 5 have disease-specific variants, APOA5, CRYGB, PDIA2, SERPINF2 and VTN, and 4 have healthy-specific variants, GNAT1, IHH, PROZ and SPIN2A. While these genes are not related to thrombophilia in the DisGeNet database, we found literature evidence that they may be relevant to thrombophilia. For instance, APOA5 (apolipoprotein A5) is an important determinant of plasma triglyceride levels, which are associated with increased thrombosis risk [48]. CRYGB has been identified as a differentially expressed gene in Coronary Heart Disease [49]. PDIA2 is involved in the migration of human vascular smooth muscle cells [50]. SERPINF2 is an obvious candidate for venous thrombosis as it codes for a serpin protease inhibitor that acts as an inhibitor of plasmin [51]. Promoter variants of VTN are associated with vascular diseases [52]. The deletion of GNAT1 inhibits the development of retinal vascular pathology in early diabetic retinopathy [53]. IHH is involved in the Hedgehog (HH) pathway, whose functions include maintenance of the endothelial compartment and orchestrating revascularization upon vessel occlusion-induced ischemia [54]. PROZ (Protein Z) is associated with venous thrombosis [55].

## Discussion and conclusion

We proposed a data-integration method based on SONMTF to uncover molecular mechanisms behind a rare subtype of hereditary thrombophilia. Our integrative framework can cope with the lack of germline variant data, revealing interesting gene clusters. The results show that



**Fig 6. An illustration of the subnetwork containing F2, F5, F7, F9 and F11.** Nodes corresponding to the genes that have disease-specific variants are colored in red and nodes corresponding to the genes that have healthy-specific variants are colored in blue. **A:** A spring embedding visualization of the subnetwork. **B:** The direct neighborhood of F2 in the subnetwork.

<https://doi.org/10.1371/journal.pone.0284084.g006>

we can overcome the scarcity of samples by integrating omics network data with germline variant data. In other words, we can apply our integrative framework to any type of rare disease. Moreover, we applied our method to all coding variants present in our datasets without any filter, allowing for data-driven suggestions of candidate disease-related genes for the selected disease.



A possible limitation of this approach is that it requires phenotype information for each germline variant sample. As detailed in [S1 File](#), our method is not completely unsupervised since it uses the observed phenotypes to constrain the grouping of the subjects. However, this is a small drawback compared to the benefit of uncovering molecular insights into rare diseases. Another potential limitation of our study is the construction of the co-expression network used as input into our NMTF-based data-integration model. In our study, we followed the approach from Malod-Dognin *et al.* [13], which yielded good results for the study of cancer. However, co-expression network construction and thresholding are still open questions with no gold standard solutions [56].

A possible future development is the incorporation of other prior information into the model, e.g., the zygosity of subjects or the type of coding variants. Another possible future development could be to assess how different co-expression measurements and thresholding strategies affect the ability of the framework to identify additional disease-related gene variants.

The analyses performed on the healthy-specific and disease-specific subnetworks obtained by our SONMF integrative framework show that several candidate disease-related genes for antithrombin resistance need further investigation. For instance, in the healthy-specific subnetwork, CD320, DHCR7, EPN1, FN3KRP, GCSH, MPST, RTEL1 and SLC27A4 may help to protect the healthy carrier from disease-specific variants. They are all connected to SLC25A10, the only thrombophilia-annotated gene in the subnetwork and two of them are validated in literature to have a role in thrombosis. The analysis of the disease-specific subnetwork revealed that the UCP2 gene might have an important role in thrombophilia as it is connected to TGFB1, a well-known thrombophilia gene present in the DisGeNet database. Moreover, the analysis of the F2 subnetwork revealed that APOA5, CRYGB, PDIA2, SERPINF2, VTN, GNAT1, IHH, PROZ and SPIN2A are mutated and directly connected to F2. We found literature evidence of the involvement of these genes in blood and vascular disorders; thus, they should be further investigated for their role in thrombophilia. Out of the 20 candidate genes that we newly proposed for thrombophilia, 17 of them have already been annotated for other diseases (detailed in [S1 File](#), Section 4). Among them, MPST, SERPINF2, VTN and PROZ are annotated for bleeding disorders, providing interesting suggestions for candidate disease-related genes.

Our results are in concordance with the clinical study of Miljic *et al.* [57], which implies that platelets could participate in the mechanism of the Prothrombin Belgrade variant. They suggest that due to the antithrombin resistance, mutated thrombin is poorly inactivated, resulting in a prolonged period of platelet activation [57]. We hypothesize that variants in ADRA2A and TBXA2R genes might be associated with decreased platelet activation, compensating for the prolonged platelet activation caused by impaired thrombin inhibition in prothrombin Belgrade mutation carriers. This suggests that newly detected variants in these two genes might have a protective effect and represent gene modifiers in the Prothrombin Belgrade mechanism. However, further functional studies should be conducted to investigate their role in the Prothrombin Belgrade variant mechanism.

## Supporting information

**S1 File. Supplementary materials, figures and tables.**  
(PDF)

## Acknowledgments

The authors would like to express their deepest gratitude to Dr. Radoje Drmanac from Complete Genomics, for the assistance with sample sequencing and continuous support during this

study. Also, the authors thank Dr. Mirjana Kovač and Dr. Predrag Miljić for their great contribution in selection of study participants.

## Author Contributions

**Conceptualization:** Branko Tomić.

**Data curation:** Maja Gvozdenov, Branko Tomić, Sofija Dunjić Manevski, Valentina Djordjević.

**Funding acquisition:** Branko Tomić, Nataša Pržulj.

**Investigation:** Valentina Djordjević.

**Methodology:** Noël Malod-Dognin, Gaia Ceddia, Nataša Pržulj.

**Supervision:** Noël Malod-Dognin, Nataša Pržulj.

**Validation:** Noël Malod-Dognin, Gaia Ceddia.

**Visualization:** Gaia Ceddia.

**Writing – original draft:** Gaia Ceddia.

**Writing – review & editing:** Noël Malod-Dognin, Gaia Ceddia, Branko Tomić, Valentina Djordjević, Nataša Pržulj.

## References

1. Rosendaal FR. Venous thrombosis: the role of genes, environment, and behavior. *ASH Education Program Book*. 2005; 2005(1):1–12.
2. Franco RF, Reitsma PH. Genetic risk factors of venous thrombosis. *Human Genetics*. 2001; 109(4):369–384. <https://doi.org/10.1007/s004390100593> PMID: 11702218
3. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM, et al. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. *The American Journal of Human Genetics*. 2000; 67(6):1452–1459.
4. Raskob GE, Angchaisuksiri P, Blanco AN, Buller H, Gallus A, Hunt BJ, et al. Thrombosis: a major contributor to global disease burden. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2014; 34(11):2363–2371. <https://doi.org/10.1161/ATVBAHA.114.304488> PMID: 25304324
5. Miyawaki Y, Suzuki A, Fujita J, Maki A, Okuyama E, Murata M, et al. Thrombosis from a prothrombin mutation conveying antithrombin resistance. *New England Journal of Medicine*. 2012; 366(25):2390–2396. <https://doi.org/10.1056/NEJMoa1201994> PMID: 22716977
6. Djordjevic V, Kovac M, Miljic P, Murata M, Takagi A, Pruner I, et al. A novel prothrombin mutation in two families with prominent thrombophilia—the first cases of antithrombin resistance in a Caucasian population. *Journal of Thrombosis and Haemostasis*. 2013; 11(10):1936–1939. <https://doi.org/10.1111/jth.12367> PMID: 23927452
7. Frayling T. Genome-wide association studies: the good, the bad and the ugly. *Clinical Medicine*. 2014; 14(4):428. <https://doi.org/10.7861/clinmedicine.14-4-428> PMID: 25099848
8. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*. 2019; 20(8):467–484. <https://doi.org/10.1038/s41576-019-0127-1> PMID: 31068683
9. Lin X, Boutros PC. Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*. 2020; 21(1):1–10. <https://doi.org/10.1186/s12859-019-3312-5> PMID: 31906867
10. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*. 2019; 50:71–91. <https://doi.org/10.1016/j.inffus.2018.09.012> PMID: 30467459
11. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Research*. 2018; 47(D1):D529–D541. <https://doi.org/10.1093/nar/gky1079>



12. Obayashi T, Kagaya Y, Aoki Y, Tadaka S, Kinoshita K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Research*. 2018; 47(D1):D55–D62. <https://doi.org/10.1093/nar/gky1155>
13. Malod-Dognin N, Petschnigg J, Windels SF, Povh J, Hemingway H, Ketteler R, et al. Towards a data-integrated cell. *Nature Communications*. 2019; 10(1):805. <https://doi.org/10.1038/s41467-019-08797-8> PMID: 30778056
14. Rauscher B, Heigwer F, Henkel L, Hielscher T, Voloshanenko O, Boutros M. Toward an integrated map of genetic interactions in cancer cells. *Molecular Systems Biology*. 2018; 14(2). <https://doi.org/10.15252/msb.20177656> PMID: 29467179
15. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*. 2014; 42(D1):D756–D763. <https://doi.org/10.1093/nar/gkt1114> PMID: 24259432
16. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*. 2016; p. gkw943. <https://doi.org/10.1093/nar/gkw943> PMID: 27924018
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*. 2000; 25(1):25.
18. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*. 2017; 46(D1):D649–D655. <https://doi.org/10.1093/nar/gkx1132>
19. Ding C, Li T, Peng W, Park H. Orthogonal nonnegative matrix tri-factorizations for clustering. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2006. p. 126–135.
20. Peng W, Li L, Dai W, Du J, Lan W. Predicting protein functions through non-negative matrix factorization regularized by protein-protein interaction network and gene functional information. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE; 2019. p. 86–89.
21. Tang X, Cai L, Meng Y, Xu J, Lu C, Yang J. Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Frontiers in Immunology*. 2021; 11:3824. <https://doi.org/10.3389/fimmu.2020.603615> PMID: 33584672
22. Guo X, Xiao H, Guo S, Dong L, Chen J. Identification of breast cancer mechanism based on weighted gene coexpression network analysis. *Cancer Gene Therapy*. 2017; 24(8):333–341. <https://doi.org/10.1038/cgt.2017.23> PMID: 28799567
23. Wu P, An M, Zou HR, Zhong CY, Wang W, Wu CP. A robust semi-supervised NMF model for single cell RNA-seq data. *PeerJ*. 2020; 8:e10091. <https://doi.org/10.7717/peerj.10091> PMID: 33088619
24. Gligorijević V, Malod-Dognin N, Pržulj N. Patient-specific data fusion for cancer stratification and personalised treatment. In: *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific; 2016. p. 321–332.
25. Wang H, Nie F, Huang H, Ding C. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. In: *2011 IEEE 11th International Conference on Data Mining*. IEEE; 2011. p. 774–783.
26. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*. 2004; 101(12):4164–4169. <https://doi.org/10.1073/pnas.0308531101> PMID: 15016911
27. Rice JA. *Mathematical statistics and data analysis*. Cengage Learning; 2006.
28. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995; 57(1):289–300.
29. Capriotti E, Fariselli P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Research*. 2017; 45(W1):W247–W252. <https://doi.org/10.1093/nar/gkx369> PMID: 28482034
30. Senis Y. Protein-tyrosine phosphatases: a new frontier in platelet signal transduction. *Journal of Thrombosis and Haemostasis*. 2013; 11(10):1800–1813. <https://doi.org/10.1111/jth.12359> PMID: 24015866
31. Cai Y, Zeng C, Su Q, Zhou J, Li P, Dai M, et al. Association of RTEL1 gene polymorphisms with stroke risk in a Chinese Han population. *Oncotarget*. 2017; 8(70):114995. <https://doi.org/10.18632/oncotarget.22980> PMID: 29383136
32. Hoss GRW, Poloni S, Blom HJ, Schwartz IVD. Three Main Causes of Homocystinuria: CBS, cblC and MTHFR Deficiency. What do they Have in Common? *Journal of Inborn Errors of Metabolism and Screening*. 2019; 7. <https://doi.org/10.1590/2326-4594-jiems-2019-0007>

33. Ye B, Li C, Yang Z, Wang Y, Hao J, Wang L, et al. Cytosolic carboxypeptidase CCP6 is required for megakaryopoiesis by modulating Mad2 polyglutamylation. *Journal of Experimental Medicine*. 2014; 211(12):2439–2454. <https://doi.org/10.1084/jem.201411123> PMID: 25332286
34. Wang Y, Zhang H. Platelet-induced inhibition of tumor cell growth. *Thrombosis Research*. 2008; 123(2):324–330. <https://doi.org/10.1016/j.thromres.2008.06.021> PMID: 18694587
35. Yu Y, Feng XH. TGF- $\beta$  signaling in cell fate control and cancer. *Current Opinion in Cell Biology*. 2019; 61:56–63. <https://doi.org/10.1016/j.ceb.2019.07.007> PMID: 31382143
36. Caja L, Dituri F, Mancarella S, Caballero-Diaz D, Moustakas A, Giannelli G, et al. TGF- $\beta$  and the Tissue Microenvironment: Relevance in Fibrosis and Cancer. *International Journal of Molecular Sciences*. 2018; 19(5):1294. <https://doi.org/10.3390/ijms19051294> PMID: 29701666
37. Lodyga M, Hinz B. TGF- $\beta$ 1—A truly transforming growth factor in fibrosis and immunity. In: *Seminars in Cell & Developmental Biology*. vol. 101. Elsevier; 2020. p. 123–139.
38. Chen Y, Dawes PT, Packham JC, Matthey DL. Interaction between smoking and functional polymorphism in the TGFB1 gene is associated with ischaemic heart disease and myocardial infarction in patients with rheumatoid arthritis: a cross-sectional study. *Arthritis Research & Therapy*. 2012; 14(2):1–11. <https://doi.org/10.1186/ar3804> PMID: 22513132
39. Bae JS, Kim IS, Rezaie AR. Thrombin down-regulates the TGF- $\beta$ -mediated synthesis of collagen and fibronectin by human proximal tubule epithelial cells through the EPCR-dependent activation of PAR-1. *Journal of Cellular Physiology*. 2010; 225(1):233–239. <https://doi.org/10.1002/jcp.22249> PMID: 20506163
40. Gunaje JJ, Bhat GJ. Distinct Mechanisms of Inhibition of Interleukin-6-Induced Stat3 Signaling by TGF- $\beta$  and  $\alpha$ -Thrombin in CCL39 Cells. *Molecular Cell Biology Research Communications*. 2000; 4(3):151–157. <https://doi.org/10.1006/mcbr.2001.0272> PMID: 11281729
41. Didiasova M, Wujak L, Schaefer L, Wygrecka M. Factor XII in coagulation, inflammation and beyond. *Cellular Signalling*. 2018; 51:257–265. <https://doi.org/10.1016/j.cellsig.2018.08.006> PMID: 30118759
42. Heil SG, Vermeulen SH, Van der Rijt-Pisa BJ, den Heijer M, Blom HJ. Role for mitochondrial uncoupling protein-2 (UCP2) in hyperhomocysteinemia and venous thrombosis risk? *Clinical Chemistry and Laboratory Medicine*. 2008; 46(5):655–659. <https://doi.org/10.1515/CCLM.2008.134> PMID: 18839467
43. Alhabibi AM, Eldewi DM, Wahab MAA, Farouk N, El-Hagrasy HA, Saleh OI. Platelet-derived growth factor-beta as a new marker of deep venous thrombosis. *Journal of Research in Medical Sciences: the Official Journal of Isfahan University of Medical Sciences*. 2019; 24. [https://doi.org/10.4103/jrms.JRMS\\_965\\_18](https://doi.org/10.4103/jrms.JRMS_965_18) PMID: 31160915
44. Ho MK, Wong YH. G z signaling: Emerging divergence from G i signaling. *Oncogene*. 2001; 20(13):1615–1625. <https://doi.org/10.1038/sj.onc.1204190> PMID: 11313909
45. Kinsella BT. Thromboxane A2 signalling in humans: a ‘Tail’ of two receptors. *Biochemical Society Transactions*. 2001; 29(6):641–654. <https://doi.org/10.1042/bst0290641> PMID: 11709048
46. Powell KL, Stevens V, Upton DH, McCracken SA, Simpson AM, Cheng Y, et al. Role for the thromboxane A2 receptor  $\beta$ -isoform in the pathogenesis of intrauterine growth restriction. *Scientific Reports*. 2016; 6(1):1–15. <https://doi.org/10.1038/srep28811>
47. Tran DQ, Andersson J, Wang R, Ramsey H, Unutmaz D, Shevach EM. GARP (LRRC32) is essential for the surface expression of latent TGF- $\beta$  on platelets and activated FOXP3+ regulatory T cells. *Proceedings of the National Academy of Sciences*. 2009; 106(32):13445–13450. <https://doi.org/10.1073/pnas.0901944106> PMID: 19651619
48. Morelli VM, Lijfering WM, Bos MH, Rosendaal FR, Cannegieter SC. Lipid levels and risk of venous thrombosis: results from the MEGA-study. *European Journal of Epidemiology*. 2017; 32(8):669–681. <https://doi.org/10.1007/s10654-017-0251-1> PMID: 28540474
49. Joehanes R, Ying S, Huan T, Johnson AD, Raghavachari N, Wang R, et al. Gene expression signatures of coronary heart disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2013; 33(6):1418–1426. <https://doi.org/10.1161/ATVBAHA.112.301169> PMID: 23539218
50. Peña E, Arderiu G, Badimon L. Protein disulphide-isomerase A2 regulated intracellular tissue factor mobilisation in migrating human vascular smooth muscle cells. *Thrombosis and Haemostasis*. 2015; 113(04):891–902. <https://doi.org/10.1160/TH14-09-0776> PMID: 25631539
51. Germain M, Saut N, Grelliche N, Dina C, Lambert JC, Perret C, et al. Genetics of venous thrombosis: insights from a new genome wide association study. *PloS One*. 2011; 6(9):e25581. <https://doi.org/10.1371/journal.pone.0025581> PMID: 21980494
52. Wang Y, Xu J, Chen J, Fan X, Zhang Y, Yu W, et al. Promoter variants of VTN are associated with vascular disease. *International Journal of Cardiology*. 2013; 168(1):163–168. <https://doi.org/10.1016/j.ijcard.2012.09.100> PMID: 23041018

53. Liu H, Tang J, Du Y, Saadane A, Samuels I, Veenstra A, et al. Transducin1, phototransduction and the development of early diabetic retinopathy. *Investigative Ophthalmology & Visual Science*. 2019; 60(5):1538–1546. <https://doi.org/10.1167/iovs.18-26433> PMID: 30994864
54. Queiroz K, Bijlsma MF, Tio RA, Zeebregts CJ, Dunaeva M, Ferreira CV, et al. Dichotomy in Hedgehog signaling between human healthy vessel and atherosclerotic plaques. *Molecular Medicine*. 2012; 18(7):1122–1127. <https://doi.org/10.2119/molmed.2011.00250> PMID: 22371306
55. Le Cam-Duchez V, Bagan-Triquet A, Barbay V, Mihout B, Borg J. The G79A polymorphism of protein Z gene is an independent risk factor for cerebral venous thrombosis. *Journal of Neurology*. 2008; 255(10):1521–1525. <https://doi.org/10.1007/s00415-008-0958-8> PMID: 18677630
56. Burns JJ, Shealy BT, Greer MS, Hadish JA, McGowan MT, Biggs T, et al. Addressing noise in co-expression network construction. *Briefings in Bioinformatics*. 2022; 23(1):bbab495. <https://doi.org/10.1093/bib/bbab495> PMID: 34850822
57. Miljic P, Gvozdenov M, Takagi Y, Takagi A, Pruner I, Dragojevic M, et al. Clinical and biochemical characterization of the prothrombin Belgrade mutation in a large Serbian pedigree: new insights into the anti-thrombin resistance mechanism. *Journal of Thrombosis and Haemostasis*. 2017; 15(4):670–677. <https://doi.org/10.1111/jth.13618> PMID: 28075532