

Crossing Gene Sets from GlyGen, GTEx and LINCS Reveals SEPHS2 as a Potential Modulator of Aging and Glycosylation

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as isolated datasets, which fragments insight and hampers discovery of shared biological themes. Crossing geneset libraries from Common Fund programs can unify disparate evidence, increase robustness, and reveal coordinated molecular programs. In this report we combined three resources GlyGen Glycosylated Proteins, GTEx Tissue Specific Aging Signatures, and LINCS L1000 CRISPR KO Perturbation Signatures by extracting all possible one-to-one intersections and selecting the top crossing based on the most conservative Fisher exact test p-value and highest Jaccard inclusion. The chosen crossing, ranked eleventh among 5,000 candidates, yielded a 19-gene set that appears in every constituent library, reflecting maximal shared representation across the datasets. GlyGen contributes a panel of secreted glycoproteins that bind hydrophobic ligands and modulate extracellular signaling, GTEx supplies age-associated blood expression changes, and LINCS adds genes whose expression declines after SEPHS2 knockout, together highlighting a coherent secreted protein axis. Enrichment analysis identified hemostasis, coagulation, fibrinolysis, and cholesterol metabolism as dominant biological themes, with transcription factor signatures for CEBPA and ESR1 linking metabolic and endocrine regulation. The convergence of these functional annotations underscores a coordinated age-related modulation of blood clotting and lipid handling, illustrating the power of multilibrary crossing to pinpoint biologically meaningful intersections.

*The Ma'ayan Laboratory, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Introduction

Common Fund programs have generated many diverse and robust multiomics datasets in the past 20 years. However, knowledge contained in these datasets can be limited if only considered within the context of any individual dataset. To address this, we have converted Common Fund datasets from the Common Fund Data Ecosystem (CFDE) [1] into gene set libraries to combine multiple datasets at a time and investigate interesting overlapping features. Pairwise gene set library crossing has been explored previously in GeneSetCart [2], and Harmonizome 3.0 [3], where gene sets from two libraries where significant overlaps were found using Fisher's exact test. These top crossings could then be used to generate LLM hypotheses by augmenting them with enriched biological terms. Here we extend the strategy by combining more libraries at once. Each

combination of three gene sets (one from each library) is analyzed to identify significant pairwise intersections and a robust 3-way intersecting gene set.

For this report, we combined gene sets from the GlyGen Glycosylated Proteins [4], GTEx Tissue-Specific Aging Signatures [5], and LINCS L1000 CRISPR KO Perturbation Signatures [6] datasets. Among the top-ranked crossings, we identified an overlap with rank 11 and a p-value of 1.93432e-23 warranting further investigation.

G48414YA is a GlyGen signature comprising proteins phosphorylated by the glycan [4]. GTEx Blood 20-29 vs 70-79 Up is a GTEx signature that lists genes showing increased expression in blood when comparing individuals aged 20-29 to those aged 70-79 [5]. SEPHS2::A375:::96H:::KO::Cas9::XPR036::112 down

is a LINCS L1000 CRISPR KO signature representing genes whose expression is reduced after CRISPRCas9 knockout of SEPHS2 in the A375 cell line measured at 96hours in experiment XPR036 replicate I12 [6].

Methods

The GlyGen Glycosylated Proteins dataset was processed from human proteoform glycosylation site citation datasets from GlyGen [4]. Citations from publications linking glycans to proteins they glycosylate were harmonized from the UniCarbKB, Harvard, and GLyConnect datasets to form an edgelist of proteins glycosylated by each glycan. Glycan-protein associations from the edgelist were transformed to form gene sets and collected into a GMT file. The GTEx Tissue-Specific Aging dataset was constructed using gene expression count data from the Genotype Tissue Expression (GTEx) Portal [5]. Sample metadata was filtered to create groups of samples with matching tissue type and donor age ranges. Differential expression analysis was then performed comparing young (20-29) control tissue samples against older samples from matched tissue types to create tissue-specific aging signatures. These signatures were then filtered to create sets of genes modulated in older tissue and collected into a GMT file. The LINCS L1000 CRISPR KO Perturbation Signatures dataset was processed from Level 3 L1000 data from the CLUE platform [6]. Replicate perturbation profiles were grouped by matching timepoint, perturbagen, detection plate, and well ID metadata. Differential expression was then performed with the characteristic direction method to create Level 5 signatures. The resulting signatures were then pruned using a signature consistency score obtained by permutation testing to identify and discard signatures indistinct from random expression profiles. Remaining signatures were then collected into a GMT file. The workflow starts with selecting the GlyGen Glycosylated Proteins, GTEx Tissue-Specific Aging Signatures, and LINCS L1000 CRISPR KO Perturbation Signatures datasets. Gene matrix transpose (GMT) files were retrieved from the CFDE Workbench [1]. GMTs were crossed to analyze all possible combinations of one gene set from each library. For each combination, the p-value was computed by computing the hypergeometric test of each gene set against the intersection of the remaining sets, using the most conservative p-value as the value for the crossing. The top 5,000 combinations were kept, ranked by ascending (most conservative) p-value and descending Jaccard index. The crossing with rank 11 was selected for further investigation, and the intersecting gene set was extracted. The terms involved in the crossing were extracted from their respective gene

set library and used to create a 3-way Venn diagram. The intersecting gene set was submitted to Enrichr [7]. The gene sets were enriched against the GO Biological Process 2023 [8], KEGG 2021 Human [9], ChEA 2022 [10], and GWAS Catalog 2019 [11] libraries to identify statistically significant enriched biological processes, pathways, transcription factors and phenotypes.

Results

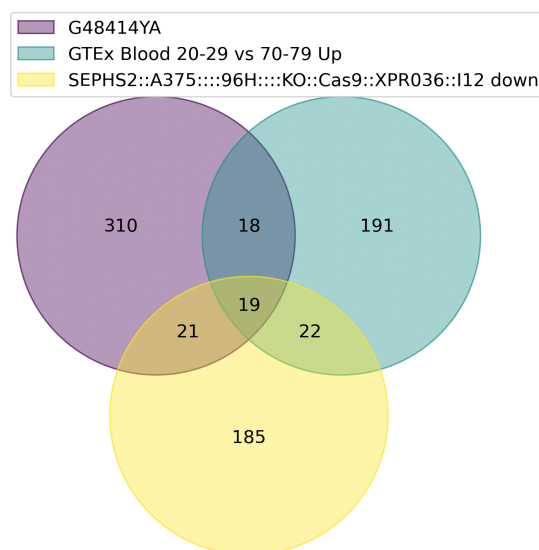


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 11, p-value of 1.934317e-23, and Jaccard index of 0.024804. The intersecting gene set identified 19 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of AFM AHSB ALB AMBP APOB APOH AZGP1 C9 CPB2 F9 FGA FGB FGG GC HRG LBP PLG SERPINA3, and SERPINA7. Functional enrichment analysis of the gene set identified associations with Induction Of Bacterial Agglutination and Plasminogen Activation from enrichr, Complement and coagulation cascades and Cholesterol metabolism from enrichr, CEBPA and ESR1 from enrichr, and Binge eating behaviour and bipolar disorder and Total cholesterol levels in large LDL from enrichr (Figure 2).

Discussion

The shared set is dominated by secreted glycoproteins that act as carriers for hydrophobic or vitamin-derived ligands and thereby influence extracellular signaling. Afamin is a liver-derived glycoprotein that binds vitamin E and other hydrophobic molecules and addition-

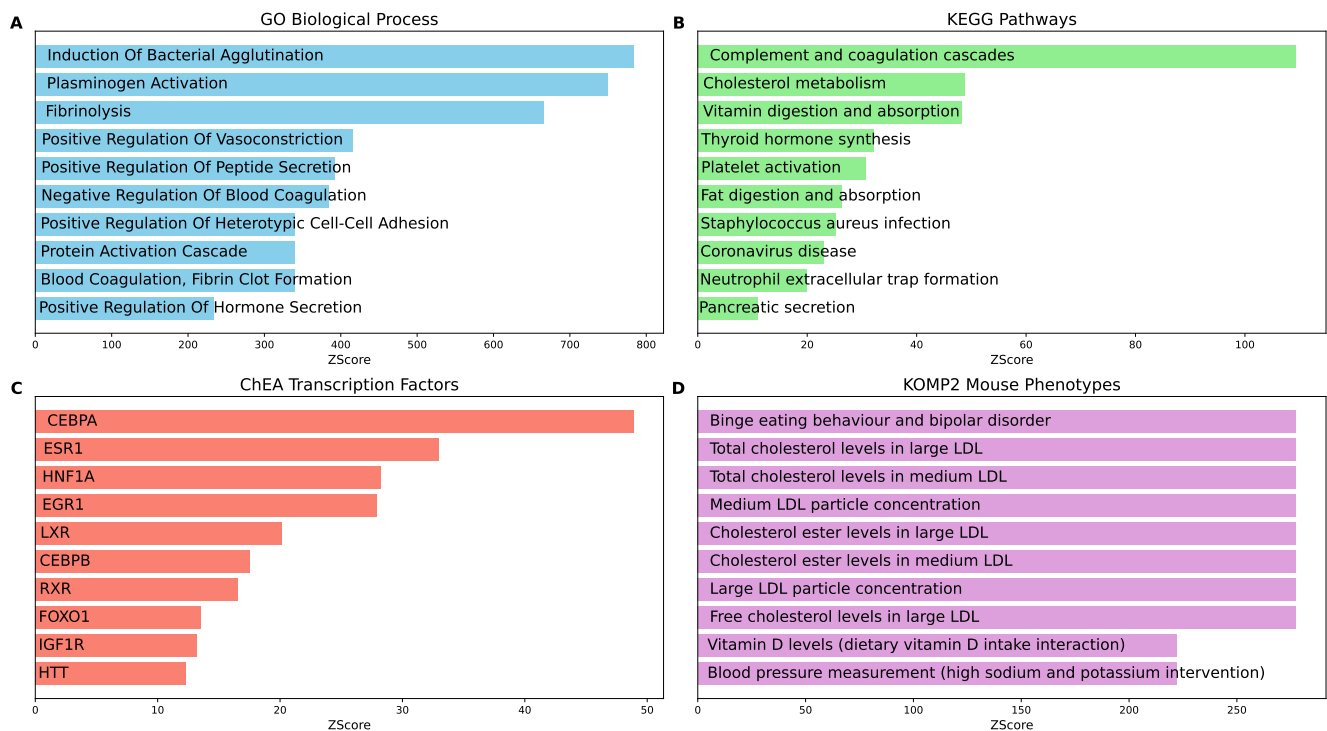


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog 2019 [11]. Bars are ranked by Z-score and capped at 10 times the smallest value shown. Color indicates the source library. The full enrichment results are available to view at [Enrichr](#).

ally functions as an extracellular chaperone that solubilizes acylated Wnt proteins, suggesting a role in long-range signaling pathways [12, 13]. Alpha2Heremans-Schmid glycoprotein (fetuina) is another liver-secreted protein that prevents ectopic mineral deposition by forming soluble calciprotein particles that stabilize calcium-phosphate complexes, thereby protecting soft tissues from pathological calcification [14–17]. The vitamin D-binding protein encoded by GC circulates as a glycoprotein that binds and transports 25-hydroxyvitamin D, controlling its half-life and tissue availability and linking vitamin D status to bone health and immune function [18–20]. Lipopolysaccharide-binding protein is an acute-phase glycoprotein that captures bacterial lipopolysaccharide and presents it to CD14/TLR complexes, amplifying innate immune responses and serving as a biomarker of microbial exposure and systemic inflammation [21–23].

A second common theme is the involvement of these glycoproteins in coagulation, fibrinolysis and extracellular matrix remodeling, processes that are known to change with age. Plasma carboxypeptidase B2 (TAFI) regulates clot stability by removing C-terminal lysines from fibrin degradation products, thereby reducing plasminogen activation, and it also inactivates inflammatory mediators such as C3a and C5a, linking

hemostasis to inflammation [24–27]. Plasminogen is a circulating glycoprotein zymogen whose activation generates plasmin for fibrin dissolution; glycation of its lysine-binding sites in diabetes impairs this conversion, illustrating how posttranslational modification can modulate fibrinolytic capacity [28–31]. Histidinerich glycoprotein binds plasminogen and factor XIIIa, influencing both fibrinolysis and contact-dependent coagulation, and its interactions are modulated by zinc and fatty acids, highlighting a multifaceted regulatory role in vascular homeostasis [32–34]. The three fibrinogen chains (FGA, FGB, FGG) are glycoprotein subunits that assemble into the fibrinogen hexamer, providing the structural scaffold for clot formation; mutations in any of these chains disrupt polymerization or crosslinking, leading to quantitative or qualitative fibrinogen deficiencies that affect clot stability [35–37].

The enrichment profile for the regulated gene sets shows a coherent emphasis on hemostasis and lipid-related biology that is reflected across several independent libraries. Geneontology results highlight multiple processes tied to blood clotting, fibrinolysis and vasoconstriction, while the KEGG analysis independently identifies the complement and coagulation cascades together with platelet activation, reinforcing a consistent signal of coagulation pathway involvement.

Parallel to these hemostatic themes, both the KEGG results report cholesterol metabolism and thyroid hormone synthesis, and the phenotype-focused data emphasize total cholesterol levels, LDL particle concentrations and cholesterol ester content, indicating regulation of lipid handling. The transcription factor enrichment (ChEA) points to regulators such as CEBPA, ESR1 and HNF1A, which are associated with metabolic and endocrine programs, linking to the observed cholesterol and hormone-related enrichments. In contrast, the KEGG set also contains infection-related entries (Staphylococcus aureus infection, Coronavirus disease, neutrophil extracellular trap formation) that are not echoed in the GO or ChEA outputs, indicating a more specific component of the response that is not shared across the broader functional annotations. Together, the overlapping coagulation and cholesterol themes across GO, KEGG and the phenotype data suggest a unified biological axis, whereas the infection-specific KEGG terms represent a distinct, library-unique facet of the geneset activity. [8], [9], [10]

The reanalysis highlights a tightly focused set of nineteen secreted glycoproteins that bridge the three queried gene sets, a convergence that is statistically robust and points to coordinated changes in blood biology. Enrichment of hemostasis-related processes across Gene Ontology, KEGG and phenotype libraries confirms that coagulation, fibrinolysis and platelet activation are central themes, while parallel enrichment of cholesterol metabolism, thyroid hormone synthesis and related phenotypes underscores a linked lipid-handling axis. Deepdive literature synthesis reinforces these themes by describing the proteins as carriers of hydrophobic ligands, modulators of vitamin transport, and regulators of clot stability and extracellular matrix remodeling, with specific examples such as afamin, fetuinA, vitaminD binding protein, lipopolysaccharide-binding protein, carboxypeptidaseB2, plasminogen, histidine-rich glycoprotein and the fibrinogen chains. The transcription factor enrichment for CEBPA, ESR1 and HNF1A aligns with the metabolic and endocrine signals reflected in the cholesterol-related findings. An unexpected facet emerges from KEGG, which also flags infection-related pathways such as Staphylococcus aureus infection and coronavirus disease; these terms are not mirrored in the GO or transcription factor results and may represent a distinct, library-specific signal. Together, the overlapping gene set, the functional enrichments and the literature-based characterizations converge on a unified biological axis of age-associated modulation of coagulation and lipid metabolism in blood, directly addressing the original research question while acknowledging the singular infection-related

signal as an outlier.

Acknowledgements

This project was funded by the NIH grant OT2OD036435. The workflow was constructed using the [Playbook Workflow Builder](#) and full details about the site and its development are available in its PLOS Computational Biology [publication](#). The OpenAI gpt-oss-120b LLM was used to assist in writing the article.

References

1. Evangelista, J. E. *et al.* The CFDE Workbench: Integrating Metadata and Processed Data from Common Fund Programs. *Journal of Molecular Biology*, 169631. <https://doi.org/10.1016/j.jmb.2026.169631> (2026).
2. Marino, G. B., Olaiya, S., Evangelista, J. E., Clarke, D. J. B. & Ma'ayan, A. GeneSetCart: assembling, augmenting, combining, visualizing, and analyzing gene sets. *GigaScience* **14**. <https://doi.org/10.1093/gigascience/giaf025> (2025).
3. Diamant, I., Clarke, D. J. B., Evangelista, J. E., Lingam, N. & Ma'ayan, A. Harmonizome 3.0: integrated knowledge about genes and proteins from diverse multi-omics resources. *Nucleic Acids Research* **53**, D1016–D1028. <https://doi.org/10.1093/nar/gkac1080> (2024).
4. York, W. S. *et al.* GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology* **30**, 72–73. <https://doi.org/10.1093/glycob/cwz080> (2019).
5. And Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330. <https://doi.org/10.1126/science.aaz1776> (2020).
6. Evangelista, J. E. *et al.* SigCom LINCS: data and meta-data search engine for a million gene expression signatures. *Nucleic Acids Research* **50**, W697–W709. <https://doi.org/10.1093/nar/gkac328> (2022).
7. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Current Protocols* **1**, e90. <https://doi.org/10.1002/cpz1.90> (2021).
8. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29. <https://doi.org/10.1038/75556> (2000).
9. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research* **51**, D587–D592. <https://doi.org/10.1093/nar/gkac963> (2022).
10. Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Research* **47**, W212–W224. <https://doi.org/10.1093/nar/gkz446> (2019).
11. Cerezo, M. *et al.* The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research* **53**, D998–D1005. <https://doi.org/10.1093/nar/gkac1070> (2024).

12. Naschberger, A. *et al.* Structural Evidence for a Role of the Multi-functional Human Glycoprotein Afamin in Wnt Transport. *Structure*. <https://doi.org/10.1016/j.str.2017.10.006> (2017).
13. Voegelé, A. F. *et al.* Characterization of the Vitamin E-Binding Properties of Human Plasma Afamin. *Biochemistry*. <https://doi.org/10.1021/bi026513v> (2002).
14. Viegas, C. S. *et al.* Chronic Kidney Disease Circulating Calciprotein Particles and Extracellular Vesicles Promote Vascular Calcification: A Role for GRP (Gla-Rich Protein). *Arteriosclerosis, Thrombosis, and Vascular Biology*. <https://doi.org/10.1161/ATVBAHA.117.310578> (2018).
15. Schäfer, C. *et al.* The serum protein α 2-Heremans-Schmid glycoprotein/fetuin-A is a systemically acting inhibitor of ectopic calcification. *Journal of Clinical Investigation*. <https://doi.org/10.1172/JCI17202> (2003).
16. Jähnen-Dechent, W., Heiss, A., Schäfer, C. & Ketteler, M. Fetuin-A Regulation of Calcified Matrix Metabolism. *Circulation Research*. <https://doi.org/10.1161/CIRCRESAHA.110.234260> (2011).
17. Heiss, A. *et al.* Structural Basis of Calcification Inhibition by α 2-HS Glycoprotein/Fetuin-A. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M210868200> (2003).
18. Wang, T. J. *et al.* Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *The Lancet*. [https://doi.org/10.1016/S0140-6736\(10\)60588-0](https://doi.org/10.1016/S0140-6736(10)60588-0) (2010).
19. Ahn, J. *et al.* Genome-wide association study of circulating vitamin D levels. *Human Molecular Genetics*. <https://doi.org/10.1093/hmg/ddq155> (2010).
20. Engelman, C. D. *et al.* Genetic and Environmental Determinants of 25-Hydroxyvitamin D and 1,25-Dihydroxyvitamin D Levels in Hispanic and African Americans. *The Journal of Clinical Endocrinology and Metabolism*. <https://doi.org/10.1210/jc.2007-2702> (2008).
21. Schröder, N. W. *et al.* Lipoteichoic Acid (LTA) of *Streptococcus pneumoniae* and *Staphylococcus aureus* Activates Immune Cells via Toll-like Receptor (TLR)-2, Lipopolysaccharide-binding Protein (LBP), and CD14, whereas TLR-4 and MD-2 Are Not Involved. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M212829200> (2003).
22. Ranao, D. R. E., Kelley, S. L. & Tapping, R. I. Human Lipopolysaccharide-binding Protein (LBP) and CD14 Independently Deliver Triacylated Lipoproteins to Toll-like Receptor 1 (TLR1) and TLR2 and Enhance Formation of the Ternary Signaling Complex. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M113.453266> (2013).
23. Jack, R. S. *et al.* Lipopolysaccharide-binding protein is required to combat a murine Gram-negative bacterial infection. *Nature*. <https://doi.org/10.1038/39622> (1997).
24. VAN DE WOUWER, M. *et al.* The lectin-like domain of thrombomodulin interferes with complement activation and protects against arthritis. *Journal of Thrombosis and Haemostasis*. <https://doi.org/10.1111/j.1538-7836.2006.02033.x> (2006).
25. Du, X.-Y. *et al.* Regulation of Chemerin Bioactivity by Plasma Carboxypeptidase N, Carboxypeptidase B (Activated Thrombin-activable Fibrinolysis Inhibitor), and Platelets. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M805000200> (2009).
26. Mosnier, L. O., Buijtenhuijs, P., Marx, P. F., Meijers, J. C. M. & Bouma, B. N. Identification of thrombin activatable fibrinolysis inhibitor (TAFI) in human platelets. *Blood*. <https://doi.org/10.1182/blood-2002-09-2944> (2003).
27. Myles, T. *et al.* Thrombin Activatable Fibrinolysis Inhibitor, a Potential Regulator of Vascular Inflammation. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M306977200> (2003).
28. Tersteeg, C. *et al.* Plasmin Cleavage of von Willebrand Factor as an Emergency Bypass for ADAMTS13 Deficiency in Thrombotic Microangiopathy. *Circulation*. <https://doi.org/10.1161/CIRCULATIONAHA.113.006727> (2014).
29. Law, R. H. *et al.* The X-ray Crystal Structure of Full-Length Human Plasminogen. *Cell Reports*. <https://doi.org/10.1016/j.celrep.2012.02.012> (2012).
30. Whyte, C. S. *et al.* Plasminogen associates with phosphatidylserine-exposing platelets and contributes to thrombus lysis under flow. *Blood*. <https://doi.org/10.1182/blood-2014-09-599480> (2015).
31. Ajjan, R. A. *et al.* Diabetes is associated with posttranslational modifications in plasminogen resulting in reduced plasmin generation and enzyme-specific activity. *Blood*. <https://doi.org/10.1182/blood-2013-04-494641> (2013).
32. Jones, A. L., Hulett, M. D., Altin, J. G., Hogg, P. & Parish, C. R. Plasminogen Is Tethered with High Affinity to the Cell Surface by the Plasma Protein, Histidine-rich Glycoprotein. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M406027200> (2004).
33. Rolny, C. *et al.* HRG Inhibits Tumor Growth and Metastasis by Inducing Macrophage Polarization and Vessel Normalization through Downregulation of PlGF. *Cancer Cell*. <https://doi.org/10.1016/j.ccr.2010.11.009> (2011).
34. TSUCHIDA-STRAETEN, N. *et al.* Enhanced blood coagulation and fibrinolysis in mice lacking histidine-rich glycoprotein (HRG). *Journal of Thrombosis and Haemostasis*. <https://doi.org/10.1111/j.1538-7836.2005.01238.x> (2005).
35. Fish, R. J. & Neerman-Arbez, M. Fibrinogen gene regulation. *Thrombosis and Haemostasis*. <https://doi.org/10.1160/TH12-04-0273> (2012).
36. Attanasio, C., David, A. & Neerman-Arbez, M. Outcome of donor splice site mutations accounting for congenital afibrinogenemia reflects order of intron removal in the fibrinogen alpha gene (FGA). *Blood*. <https://doi.org/10.1182/blood-2002-03-0853> (2003).
37. Uitte de Willige, S. *et al.* Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen γ' levels.

Blood. <https://doi.org/10.1182/blood-2005-05-2180> (2005).