

Crossing Gene Sets from GlyGen, GTEx, MoTrPAC and RummaGEO Reveals Nicotinamide as a Potential Modulator of Glycosylation, Aging and Exercise

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as separate collections, which fragments insight and limits the ability to see shared mechanisms across diverse biological contexts. By crossing gene sets from four Common Fund libraries GlyGen Glycosylated Proteins, GTEx TissueSpecific Aging Signatures, MoTrPAC Rat Endurance Exercise Training Transcriptomics, and RummaGEO Drug Perturbation Signatures we sought to uncover genes that are present in every dataset, reasoning that such maximal overlap highlights robust, reproducible signals. The GlyGen set contributed extracellular glycoproteins whose carbohydrate modifications influence matrix stability and innate immune clearance, while the GTEx aging signature captured blood expression changes associated with advancing age. The MoTrPAC exercise transcriptome added genes responsive to endurance training, reflecting physiological adaptation, and the RummaGEO nicotinamide perturbation signature introduced a drugrelated component that aligns with the observed pathways. The intersecting ten genes AMBP, APCS, APOC4, APOH, C8A, FGA, FGB, FGG, HRG and SERPINA6 are all extracellular proteins central to hemostasis, complement activation and inflammatory regulation, and enrichment analyses consistently pointed to coagulation, fibrinolysis, platelet function and nuclear receptor transcriptional programs. Together these findings illustrate how integrating multiple Common Fund datasets can reveal a coherent biological module and suggest relevance of the nicotinamide associated perturbation to coagulation related and inflammatory processes.

*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], MoTrPAC Rat Endurance Exercise Training Transcriptomics [6], and RummaGEO Drug Perturbation Signatures [7] datasets. Among the top-ranked crossings, we identified an overlap with rank 9 and a p-value of 3.30466×10^{-11} warranting further investigation.

The G48414YA signature derives from the GlyGen library and comprises proteins that are phosphorylated by the glycan [4]. The GTEx Blood 2029 vs 7079 Up sig-

nature comes from the GTEx aging collection [5]. The T30BloodRna Female 2W Down signature originates from the MoTrPAC transcriptomics resource [6]. The GSE157594 6 V 4 NICOTINAMIDE MOUSE DN signature is part of the RummaGEO drug perturbation set [7].

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 MoTrPAC Rat Endurance Exercise Training dataset was processed from transcriptomic data from the Molecular Transducers of Physical Activity Consortium (MoTrPAC) Data Hub [6]. Differential expression data from MoTrPAC was filtered to keep associations where gene expression changed over the course of 8 weeks of endurance exercise. Exercise tissue samples were differentiated by sex and tissue type in addition to exercise timepoint. The downloaded data also included training data of genes with modulated expression over the training window in addition to the timewise data. The differential expression data was filtered to create gene sets and a GMT. RummaGEO is a webserver developed to enable expression signature search against human and mouse RNA-seq studies deposited to GEO [7]. To enable this search, offline processing automatically identified groups of samples were based on shared conditions and differential expression analysis was performed. The RummaGEO Drug Perturbation Signatures dataset was processed by querying the RummaGEO metadata search API to identify a subset of signatures where one condition included a control term and the other included a drug. Gene sets for each signature in the subset were combined to construct a GMT file. The workflow starts with selecting the GlyGen Glycosylated Proteins, GTEx Tissue-Specific Aging Signatures, MoTrPAC Rat Endurance Exercise Training

Transcriptomics, and RummaGEO Drug 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 9 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 4-way Venn diagram. The intersecting gene set was submitted to Enrichr [8]. The gene sets were enriched against the GO Biological Process 2023 [9], KEGG 2021 Human [10], ChEA 2022 [11], and GWAS Catalog 2019 [12] libraries to identify statistically significant enriched biological processes, pathways, transcription factors and phenotypes.

Results

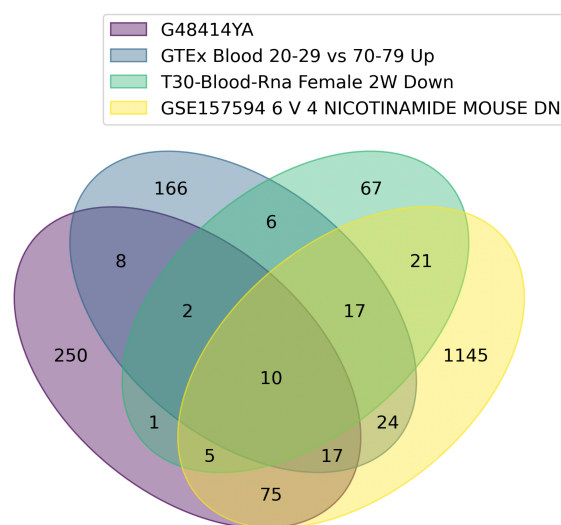


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2-,3-, and 4-way intersection are labelled.

The crossing had a rank of 9, p-value of 3.304661e-11, and Jaccard index of 0.005513. The intersecting gene set identified 10 genes appearing in all involved sets (Figure 1). The intersecting gne set consists of AMBP APCS APOC4 APOH C8A FGA FGB FGG HRG, andSER-PINA6. Functional enrichment analysis of the gene set

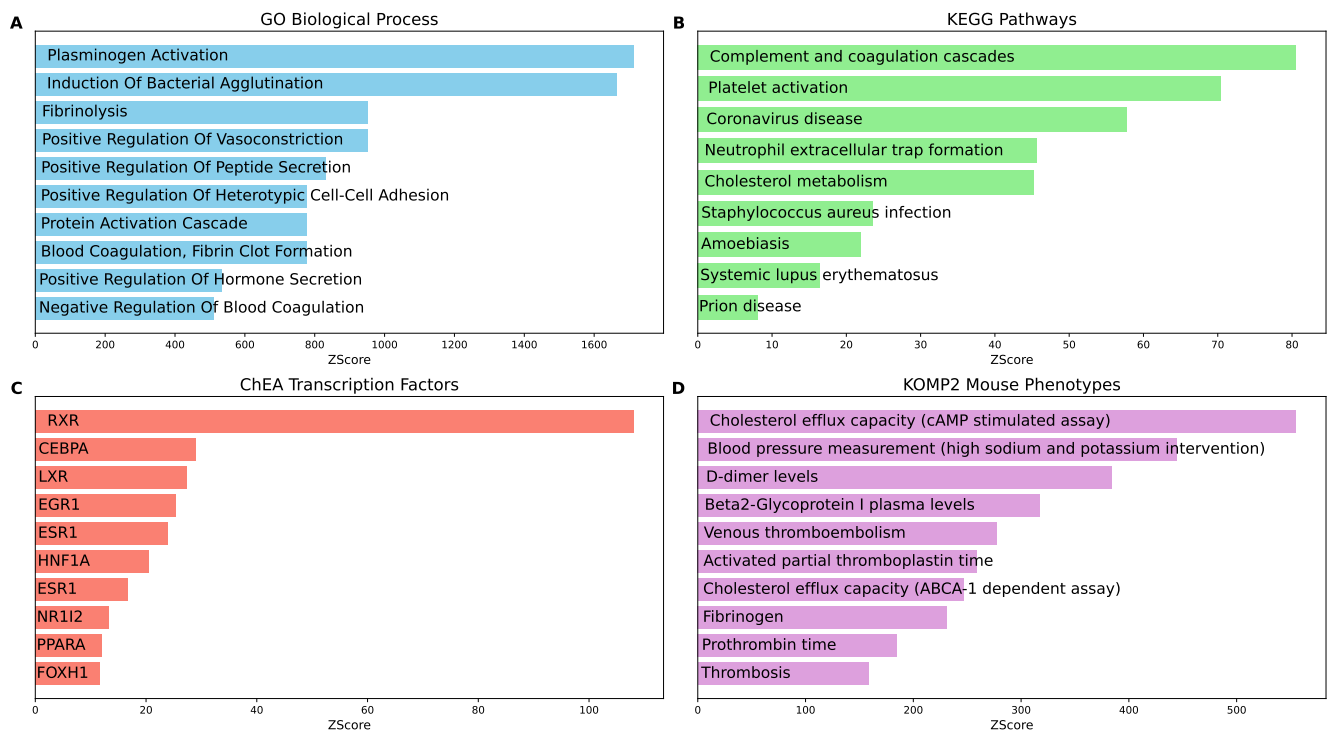


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [8]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [9], B. KEGG 2021 Human [10], C. ChEA 2022 [11], and D. GWAS Catalog 2019 [12]. 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](#).

identified associations with Plasminogen Activation and Induction Of Bacterial Agglutination from enrichr, Complement and coagulation cascades and Platelet activation from enrichr, RXR and CEBPA from enrichr, and Cholesterol efflux capacity and Blood pressure measurement from enrichr (Figure 2).

Discussion

These intersecting proteins are largely extracellular glycoproteins that influence matrix stability and innate immune clearance. AMBP encodes bikunin, the lightchain component of intertrypsin inhibitor complexes that bind hyaluronan to stabilize the extracellular matrix and modulate inflammation [13–15]; APCS produces serum amyloid P component, a short pentraxin that binds amyloid fibrils and late apoptotic cells, facilitating their phagocytic removal and influencing tissue homeostasis [16–18]; C8A encodes the subunit of complement component C8, whose microheterogeneity includes multiple glycosylation events that affect membrane attack complex formation and stability [19]; HRG functions as an opsonin that bridges apoptotic and necrotic cells to phagocytic receptors via heparansulfate interactions, a process dependent on zinc-mediated binding and glycosylated domains [20–22]; SERPINA6 encodes corticosteroidbinding globulin, a

plasma carrier that binds cortisol and modulates its systemic availability, with its activity regulated by hormonal and inflammatory cues [23–25].

These glycoproteins also converge on coagulation and complement cascades where carbohydrate modifications influence functional assembly. The fibrinogen chains FGA, FGB and FGG together form the hexameric fibrinogen molecule whose domains mediate fibrin polymerization, crosslinking and clot stability, processes essential for hemostasis and wound healing [26–29], [30–32], [33–35]; C8As glycosylated subunit participates in the terminal steps of complement activation, enabling pore formation that lyses target cells [19]; HRG interacts with plasminogen and factorXIIa, modulating fibrinolysis and contactdependent coagulation pathways [36, 37]; APOH binds anionic phospholipids on apoptotic cells and forms stable complexes with oxidized LDL, linking lipid recognition to coagulation and immune responses [38, 39].

The enrichment profile reveals a coherent emphasis on hemostatic and inflammatory pathways that is reflected across several libraries. Geneontology results highlight terms such as plasminogen activation, fibrinolysis, blood coagulation and positive regulation of vasoconstriction, indicating a strong signal for co-

agulation and vascular tone modulation [9]. Parallel KEGG enrichment underscores this theme with the complement and coagulation cascades, platelet activation and neutrophil extracellular trap formation, further reinforcing the involvement of clotting and innate immune mechanisms [10]. Transcriptionfactor analysis from ChEA points to nuclearreceptordriven regulation, with prominent enrichment for RXR, CEBPA, LXR, ESR1 and HNF1A, suggesting links to hormoneresponsive transcriptional programs [11]. The GWAS-derived library adds a clinical dimension by associating the same gene sets with traits such as cholesterol efflux capacity, blood pressure, Ddimer levels and venous thromboembolism, which align with the coagulation and lipidmetabolism signals observed in the other analyses [12]. Together these crosslibrary patterns indicate that the gene sets are strongly tied to coagulation, platelet and complement activity while also being linked to nuclearreceptor transcription factors, producing complementary signals across the different enrichment sources.

The analysis identified a highly significant overlap of ten extracellular glycoprotein genes that are central to hemostasis and innate immune processes, reinforcing the relevance of these pathways to the intersecting datasets. Enrichment across Gene Ontology, KEGG and GWAS libraries consistently highlighted coagulation, fibrinolysis, complement activation, platelet function and related vascular tone regulation, while transcriptionfactor profiling linked the gene set to nuclearreceptor driven programs. Deepdive literature synthesis confirmed that the overlapping proteinsincluding fibrinogen chains, complement subunits, serum amyloid P, and corticosteroidbinding globulinact extracellularly to modulate matrix stability, clot formation and immune clearance, with glycosylation influencing their activity. The convergence of these findings with the agerelated blood expression signatures and the nicotinamideperturbed mouse model supports the hypothesis that the shared gene set reflects coordinated regulation of coagulation and inflammatory responses in the biological contexts examined. An unexpected element was the appearance of APOH, a lipidbinding protein, which suggests a possible link between lipid handling and the observed hemostatic signatures but warrants further investigation. Overall, the results validate the original research question by demonstrating that the intersecting gene set is strongly associated with coagulation, complement and nuclearreceptor transcriptional pathways across the examined conditions.

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/gkae1080> (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. Sanford, J. A. *et al.* Molecular Transducers of Physical Activity Consortium (MoTrPAC): Mapping the Dynamic Responses to Exercise. *Cell* **181**, 1464–1474. <https://doi.org/10.1016/j.cell.2020.06.004> (2020).
7. Marino, G. B., Clarke, D. J., Lachmann, A., Deng, E. Z. & Ma'ayan, A. RummaGEO: Automatic mining of human and mouse gene sets from GEO. *Patterns* **5**, 101072. <https://doi.org/10.1016/j.patter.2024.101072> (2024).
8. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Current Protocols* **1**, e90. <https://doi.org/10.1002/cpz1.90> (2021).
9. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29. <https://doi.org/10.1038/75556> (2000).
10. 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).
11. 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).
12. Cerezo, M. *et al.* The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nu-*

- cleic Acids Research* **53**, D998–D1005. <https://doi.org/10.1093/nar/gkae1070> (2024).
13. Rugg, M. S. *et al.* Characterization of Complexes Formed between TSG-6 and Inter- α -inhibitor That Act as Intermediates in the Covalent Transfer of Heavy Chains onto Hyaluronan. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M501332200> (2005).
14. Hamm, A. *et al.* Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITI-H) genes in multiple human solid tumors: A systematic expression analysis. *BMC Cancer*. <https://doi.org/10.1186/1471-2407-8-25> (2008).
15. Zubiri, I. *et al.* Diabetic nephropathy induces changes in the proteome of human urinary exosomes as revealed by label-free comparative analysis. *Journal of Proteomics*. <https://doi.org/10.1016/j.jprot.2013.10.037> (2014).
16. Bijl, M. *et al.* Serum amyloid P component binds to late apoptotic cells and mediates their uptake by monocyte-derived macrophages. *Arthritis and Rheumatism*. <https://doi.org/10.1002/art.10737> (2003).
17. Lu, J. *et al.* Structural recognition and functional activation of Fc γ R by innate pentraxins. *Nature*. <https://doi.org/10.1038/nature07468> (2008).
18. Garlanda, C. *et al.* Non-redundant role of the long pentraxin PTX3 in anti-fungal innate immune response. *Nature*. <https://doi.org/10.1038/nature01195> (2002).
19. Franc, V., Zhu, J. & Heck, A. J. R. Comprehensive Proteoform Characterization of Plasma Complement Component C8 $\alpha\beta\gamma$ by Hybrid Mass Spectrometry Approaches. *Journal of the American Society for Mass Spectrometry*. <https://doi.org/10.1007/s13361-018-1901-6> (2018).
20. Jones, A. L., Hulett, M. D. & Parish, C. R. Histidine-rich Glycoprotein Binds to Cell-surface Heparan Sulfate via Its N-terminal Domain following Zn²⁺ Chelation. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M401996200> (2004).
21. Gorgani, N. N., Smith, B. A., Kono, D. H. & Theofilopoulos, A. N. Histidine-Rich Glycoprotein Binds to DNA and Fc γ RI and Potentiates the Ingestion of Apoptotic Cells by Macrophages. *The Journal of Immunology*. <https://doi.org/10.4049/jimmunol.169.9.4745> (2002).
22. Poon, I. K. H., Hulett, M. D. & Parish, C. R. Histidine-rich glycoprotein is a novel plasma pattern recognition molecule that recruits IgG to facilitate necrotic cell clearance via Fc γ RI on phagocytes. *Blood*. <https://doi.org/10.1182/blood-2009-07-234013> (2010).
23. Joyner, J. M., Hutley, L. J., Bachmann, A. W., Torpy, D. J. & Prins, J. B. Greater replication and differentiation of preadipocytes in inherited corticosteroid-binding globulin deficiency. *American Journal of Physiology-Endocrinology and Metabolism*. <https://doi.org/10.1152/ajpendo.00262.2002> (2003).
24. Torpy, D. J. *et al.* Association Between Chronic Fatigue Syndrome and the Corticosteroid-Binding Globulin Gene ALA SER²²⁴Polymorphism. *Endocrine Research*. <https://doi.org/10.1081/erc-200035599> (2004).
25. Fernandez-Real, J.-M. *et al.* Serum Corticosteroid-Binding Globulin Concentration and Insulin Resistance Syndrome: A Population Study. *The Journal of Clinical Endocrinology and Metabolism*. <https://doi.org/10.1210/jc.2001-011843> (2002).
26. 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).
27. Standeven, K. F. *et al.* Functional Analysis of the Fibrinogen A α Thr312Ala Polymorphism: Effects on Fibrin Structure and Function. *Circulation*. <https://doi.org/10.1161/01.CIR.0000066690.89407.CE> (2003).
28. Litvinov, R. I. *et al.* Direct Evidence for Specific Interactions of the Fibrinogen α C-Domains with the Central E Region and with Each Other. *Biochemistry*. <https://doi.org/10.1021/bi700944j> (2007).
29. Tsurupa, G. *et al.* On the Mechanism of α C Polymer Formation in Fibrin. *Biochemistry*. <https://doi.org/10.1021/bi2017848> (2012).
30. REINER, A. *et al.* Association between patterns of nucleotide variation across the three fibrinogen genes and plasma fibrinogen levels: the Coronary Artery Risk Development in Young Adults (CARDIA) study. *Journal of Thrombosis and Haemostasis*. <https://doi.org/10.1111/j.1538-7836.2006.01907.x> (2006).
31. Sabater-Lleal, M. *et al.* Multiethnic Meta-Analysis of Genome-Wide Association Studies in >100 000 Subjects Identifies 23 Fibrinogen-Associated Loci but No Strong Evidence of a Causal Association Between Circulating Fibrinogen and Cardiovascular Disease. *Circulation*. <https://doi.org/10.1161/CIRCULATIONAHA.113.002251> (2013).
32. Fish, R. J. & Neerman-Arbez, M. Fibrinogen gene regulation. *Thrombosis and Haemostasis*. <https://doi.org/10.1160/TH12-04-0273> (2012).
33. Casini, A., Neerman-Arbez, M. & de Moerloose, P. Congenital Fibrinogen Disorders: An Update. *Seminars in Thrombosis and Hemostasis*. <https://doi.org/10.1055/s-0033-1349222> (2013).
34. Germain, M. *et al.* Meta-analysis of 65,734 Individuals Identifies TSPAN15 and SLC44A2 as Two Susceptibility Loci for Venous Thromboembolism. *The American Journal of Human Genetics*. <https://doi.org/10.1016/j.ajhg.2015.01.019> (2015).
35. 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).
36. 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).

37. Kassar, O., Schwarz-Linek, U., Blindauer, C. & Stewart, A. Plasma free fatty acid levels influence Zn²⁺-dependent histidine-rich glycoprotein–heparin interactions via an allosteric switch on serum albumin. *Journal of Thrombosis and Haemostasis*. <https://doi.org/10.1111/jth.12771> (2015).
38. Kobayashi, K. *et al.* Circulating oxidized LDL forms complexes with β 2-glycoprotein I: implication as an atherogenic autoantigen. *Journal of Lipid Research*. <https://doi.org/10.1194/jlr.M200329-JLR200> (2003).
39. Maiti, S. N., Balasubramanian, K., Ramoth, J. A. & Schroit, A. J. β -2-Glycoprotein 1-dependent Macrophage Uptake of Apoptotic Cells. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M704990200> (2008).