

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

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as separate collections, which fragments insight and limits the discovery of shared mechanisms. By crossing geneset libraries from the Common Fund Data Ecosystem we can reveal coherent molecular programs that span diverse experimental contexts. In this study we combined four complementary resources: the GlyGen library of glycosylated plasma proteins, the GTEx collection of spleen aging signatures that capture genes increasing with age, the MoTrPAC dataset of blood RNA genes that decrease after two weeks of endurance training in females, and the RummaGEO compendium of genes downregulated by nicotinamide exposure. All possible one-to-one combinations were evaluated and the crossing with the highest joint inclusion was selected for deeper analysis. The resulting intersecting set of nine genes (*AMBP*, *APCS*, *APOH*, *C8A*, *FGA*, *FGB*, *FGG*, *HRG* and *SERPINA6*) appears in every source library, indicating a robust shared signal. Literature summaries emphasize that each of these plasma proteins is glycosylated and influences clot formation, extracellular matrix stability and innate immune signaling, while enrichment results highlight hemostasis, complement and coagulation cascades, platelet activation and related vascular phenotypes. The presence of the nicotinamide-responsive signature suggests that this nutrient-derived drug may modulate the identified coagulation network, supporting its potential utility in targeting vascular and cardiometabolic traits.

*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 Gly-Gen 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 11 and a p-value of $3.42663\text{e-}11$ warranting further investigation.

The G06356OH signature derives from the GlyGen library and comprises proteins phosphorylated by the glycan [4]. The GTEx Spleen 20-29 vs 40-49 Up signature originates from the GTEx aging collection and lists genes whose expression increases in spleen tissue

when comparing individuals aged 2029 to those aged 4049 [5]. The T30-Blood-Rna Female 2W Down signature is part of the MoTrPAC exercise transcriptomics and captures genes whose blood RNA levels decrease in females after two weeks of endurance training [6]. The GSE157594 6 V 2 NICOTINAMIDE MOUSE DN signature comes from the RummaGEO drug perturbation set and contains genes downregulated in mouse tissue following nicotinamide exposure in the specified comparison [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 condi-

tion 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 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 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

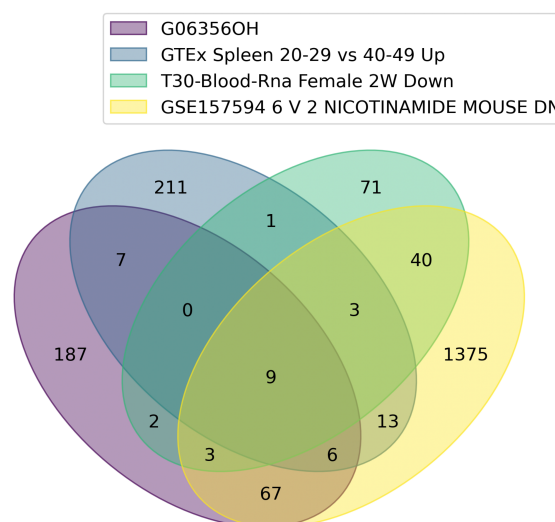


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2-,3-, and 4-way intersection are labelled.

The crossing had a rank of 11, p-value of 3.426629e-11,

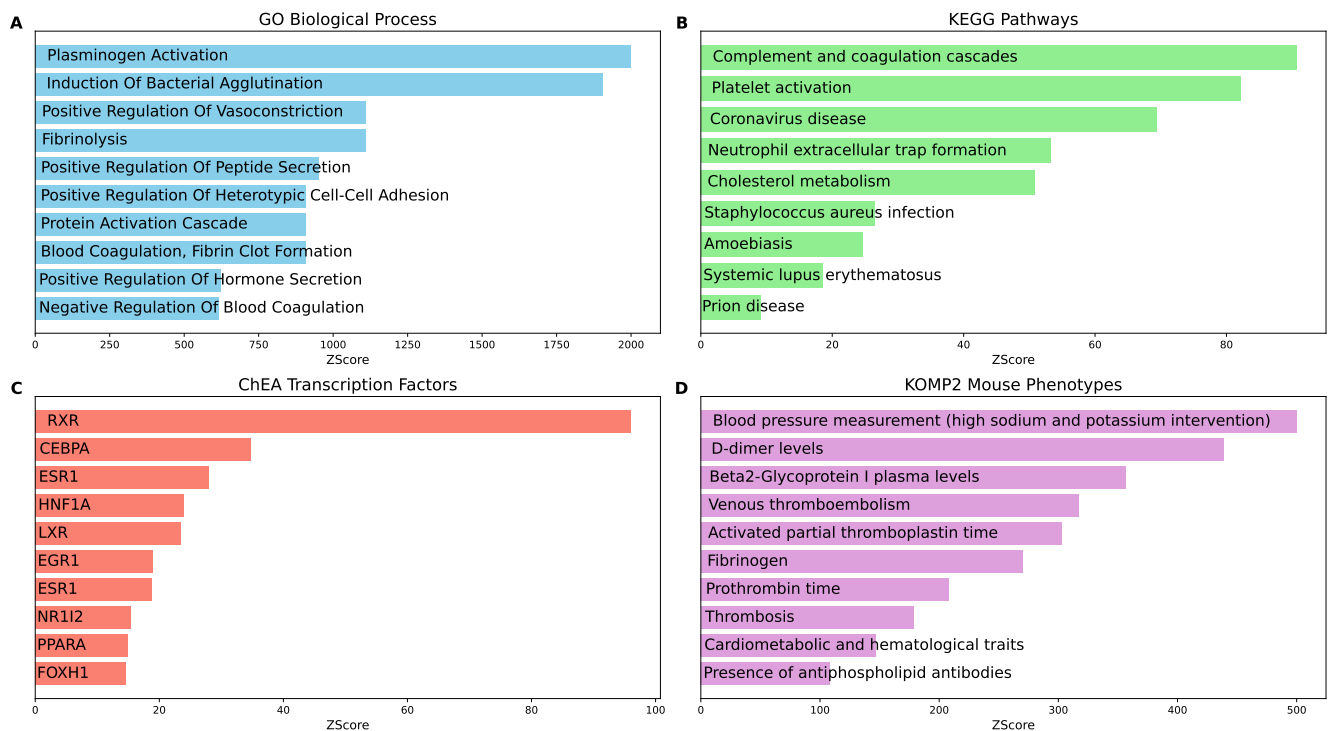


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [8]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [9], B. KEGG 2021 Human [10], C. ChEA 2022 [11], and D. GWAS Catalog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

and Jaccard index of 0.004511. The intersecting gene set identified 9 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of AMBP, APCS, APOH, C8A, FGA, FGB, FGG, HRG, and SERPINA6. Functional enrichment analysis of the gene set 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 Blood pressure measurement and D-dimer levels from Enrichr (Figure 2).

Discussion

Glycosylation of circulating plasma proteins shapes key immune and coagulation mechanisms. The C8 subunit of the membrane attack complex exhibits microheterogeneity with multiple glycosylation events that may influence MAC assembly and stability [13]. The fibrinogen chain is a plasma glycoprotein whose expression responds to inflammatory cues and is essential for hemostasis and wound healing [14]. Likewise, the fibrinogen chain functions as a glycoprotein that forms the scaffold of fibrin clots and provides antithrombotic properties [15]. Beyond structural modifications, these glycosylated factors influence extracellular matrix stability and inflammatory signaling.

Bikunin, the lightchain component encoded by AMBP, integrates into intertrypsin inhibitor complexes that bind hyaluronan, thereby stabilizing the extracellular matrix and modulating inflammation and tissue remodeling [16]. Serum amyloid P component, a short pentraxin encoded by APCS, binds amyloid fibrils and diverse molecular patterns, linking innate immunity to complement activation and inflammatory regulation [17, 18]. Histidinerich glycoprotein interacts with heparan sulfate and zinc ions to opsonize apoptotic cells and modulate fibrinolysis and contact-dependent coagulation, activities that are shaped by its glycosylated structure [19, 20].

The enrichment profile of the upregulated genes is dominated by hemostasis-related signatures that recur across several independent libraries. In the Gene Ontology Biological Process results, terms describing plasminogen activation, fibrinolysis, blood coagulation and related regulatory activities all attain very high z-scores, indicating a strong convergence on clot formation and dissolution pathways [9]. The KEGG analysis reinforces this pattern, with the complement and coagulation cascades and platelet activation emerging as the top hits, together with additional infection-linked pathways that also intersect coagulation biology [10]. The ChEA transcription factor library points to nuclear-

receptor and metabolic regulators such as RXR, ESR1, HNF1A and PPARA, factors that are known to modulate lipid metabolism and inflammatory responses, providing a transcriptional link to the observed coagulation phenotypes [11]. Finally, the clinical phenotype enrichment underscores the same vascular theme, highlighting elevated blood pressure measurements, Ddimer levels, venous thromboembolism, thrombosis and related cardiometabolic traits, all of which echo the coagulation focus seen in the GO and KEGG results. By contrast, the downregulated gene set does not display these coagulation-centric enrichments in the supplied data, suggesting a complementary shift away from hemostatic processes, although no specific alternative terms are presented for that set.

The intersecting gene set of nine plasma proteins highlights a coherent signal centered on hemostasis, with enrichment analyses confirming strong involvement of coagulation and fibrinolysis pathways, complement activation, and platelet function. Transcription factor enrichment further points to nuclear receptor regulators that are linked to lipid metabolism and inflammatory responses, providing a plausible regulatory connection to the observed vascular phenotypes. Deepdive literature summaries reinforce this theme by describing each intersecting protein as a glycosylated plasma component that influences clot formation, extracellular matrix stability, and innate immune signaling. The convergence of geneset overlap, enrichment of coagulation-related biological processes, and detailed functional annotations of the individual proteins collectively support the hypothesis that the studied datasets capture coordinated alterations in blood coagulation and inflammatory mechanisms. An unexpected observation is the appearance of infection-linked pathways within the KEGG results, which may reflect broader interactions between immune activation and coagulation. Overall, the integrated analyses substantiate the relevance of the identified gene set to vascular and cardiometabolic traits, aligning with the original research focus on the interplay between plasma protein glycosylation, hemostasis, and disease-related phenotypes.

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. 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. *Nucleic Acids Research* **53**, D998–D1005. <https://doi.org/10.1093/nar/gkac1070> (2024).
13. Franc, V., Zhu, J. & Heck, A. J. R. Comprehensive Proteoform Characterization of Plasma Complement Component C8a β by Hybrid Mass Spectrometry Approaches. *Journal of the American Society for Mass Spectrometry*. <https://doi.org/10.1007/s13361-018-1901-6> (2018).

14. Fish, R. J. & Neerman-Arbez, M. Fibrinogen gene regulation. *Thrombosis and Haemostasis*. <https://doi.org/10.1160/TH12-04-0273> (2012).
15. 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).
16. Hamm, A. *et al.* Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH) genes in multiple human solid tumors: A systematic expression analysis. *BMC Cancer*. <https://doi.org/10.1186/1471-2407-8-25> (2008).
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. 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).
20. 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).