

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

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as isolated collections, which fragments insight and limits the discovery of shared mechanisms across biological contexts. By crossing gene sets derived from multiple Common Fund programs, researchers can reveal coherent molecular programs that are robust across datasets and connect diverse perturbations to common physiological processes. In this study we combined four complementary resources: the GlyGen Glycosylated Proteins library, the GTEx TissueSpecific Aging signatures, the MoTrPAC Rat Endurance Exercise Training transcriptomics, and the RummaGEO Drug Perturbation signatures that include a sorafenib response profile. All possible onetoone combinations of one set from each library were evaluated and the crossing with the strongest joint inclusion of genes was selected for deeper analysis. The resulting fourway intersection contained eight genes: AMBP, APCS, APOH, C8A, FGA, FGB, SERPINA4 and SERPINA6 that appear in every contributing set, indicating maximal overlap among the examined libraries. Enrichment of these genes highlighted pathways of complement and coagulation, cholesterol metabolism, and regulation by hepatic nuclear receptors, suggesting a tightly linked network of extracellular plasma proteins that govern hemostasis, lipid handling and vascular homeostasis. Because sorafenib modulates the expression of this intersecting gene set, the drug emerges as a candidate for influencing these bloodborne processes and may hold therapeutic relevance for disorders of coagulation and inflammation.

*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 55 and a p-value of 1.54781e-09 warranting further investigation.

The G063560H signature derives from the GlyGen library and comprises proteins that are phosphorylated by the glycan [4]. The GTEx Blood 2029 vs 7079 Up signature originates from the GTEx TissueSpecific Aging collection and lists genes that show increased

expression in blood when comparing younger (2029) to older (7079) individuals [5]. The T30BloodRna Female 2W Down signature is part of the MoTrPAC Rat Endurance Exercise Training Transcriptomics resource and captures genes that are downregulated in female blood RNA after two weeks of the training protocol [6]. The GSE158458 2 V 1 SORAFENIB HUMAN DN signature comes from the RummaGEO Drug Perturbation dataset and reflects genes that are downregulated in human cells treated with sorafenib under 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 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 55 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

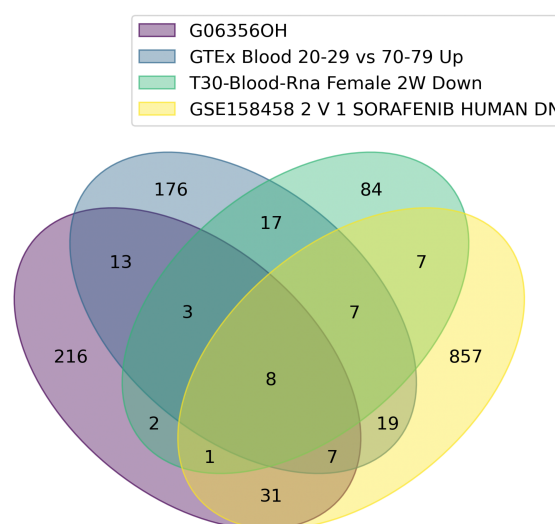


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2-,3-, and 4-way intersection are labelled.

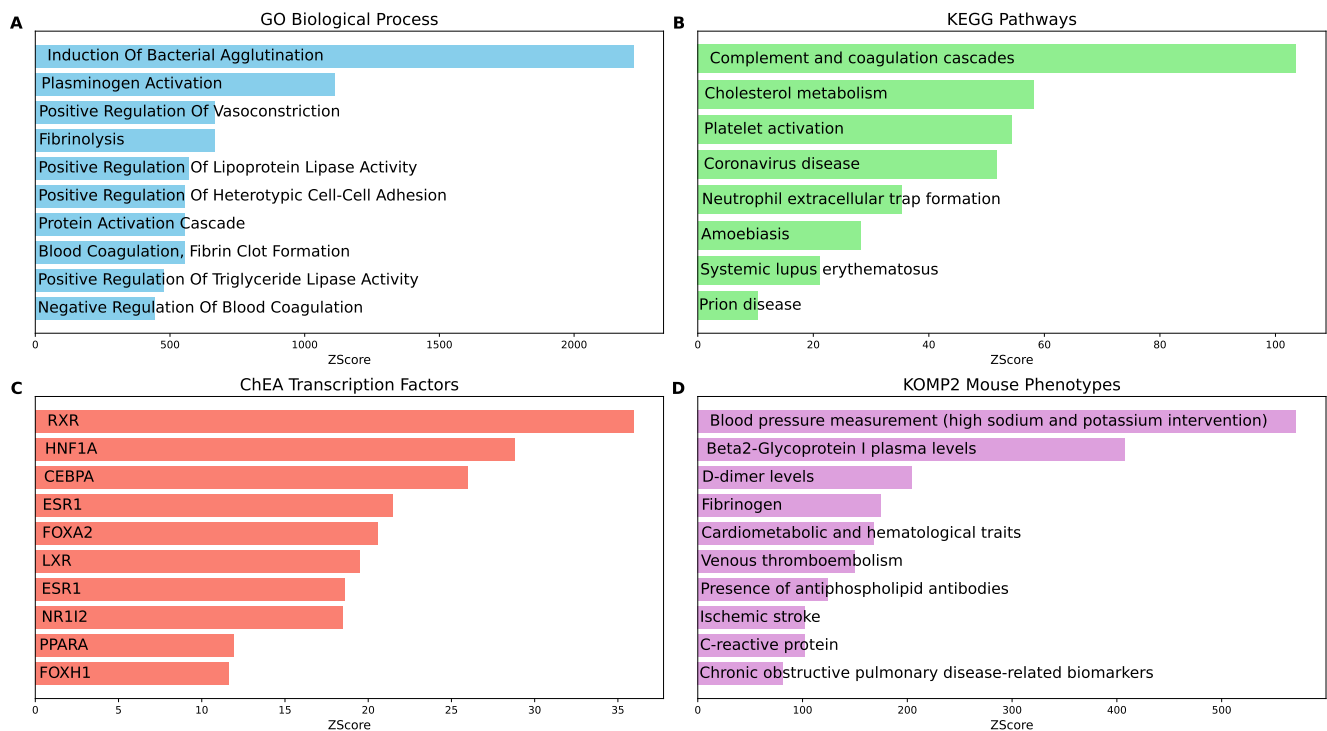


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [8]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [9], B. KEGG 2021 Human [10], C. ChEA 2022 [11], and D. GWAS Catalog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

The crossing had a rank of 55, p-value of 1.547811e-09, and Jaccard index of 0.005525. The intersecting gene set identified 8 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of AMBP, APCS, APOH, C8A, FGA, FGB, SERPINA4, and SERPINA6. 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, RXR and HNF1A from enrichr, and Blood pressure measurement and Beta2-Glycoprotein I plasma levels from enrichr (Figure 2).

Discussion

The shared biological functions of these proteins center on the organization and modification of extracellular and plasma glycoproteins that influence tissue integrity and inflammatory signaling. AMBP encodes bikunin, a structural component of intertrypsin inhibitor complexes that bind hyaluronan to stabilize the extracellular matrix and modulate inflammation and tissue remodeling [13–15]. APCS produces serum amyloid P component, a short pentraxin that binds amyloid fibrils and diverse molecular patterns, linking innate immune recognition to regulation of inflammatory responses [16–18]. The C8A subunit of the membrane

attack complex displays extensive microheterogeneity due to glycosylation, a modification that can affect MAC assembly and stability in immune defense [19]. FGA and FGB encode the α and β chains of fibrinogen, major plasma glycoproteins whose synthesis is regulated by inflammatory cues and whose structural domains govern fibrin polymerization and crosslinking, essential for clot formation and wound healing [20–22]. APOH, known as 2-glycoprotein I, is a multifunctional plasma glycoprotein that binds anionic phospholipids on apoptotic cells and oxidized LDL, contributing to clearance mechanisms and vascular homeostasis [23].

A second theme is the involvement of these genes in immune surveillance and vascular processes that shape phenotypic outcomes. APCS facilitates the scavenging of cellular debris and amyloid material, activities that are central to tissue homeostasis and have been linked to disease susceptibility [16, 18]. AMBP's role in extracellular matrix organization and inflammatory modulation positions it as a potential biomarker for renal pathology, reflecting relevance to tissue remodeling [13, 15]. C8A-mediated formation of the membrane attack complex enables cytolysis of altered host cells, a defense mechanism that can influence immune competence [19]. SERPINA4 and SERPINA6 encode serine protease inhibitors and corticosteroid-binding globulin,

respectively, both of which regulate vascular homeostasis, angiogenic signaling, and systemic glucocorticoid availability, thereby impacting inflammatory status and vascular health [24, 25].

The enrichment results display a consistent focus across libraries, with GO biological process terms highlighting fibrinolysis, blood coagulation, and positive regulation of vasoconstriction, while KEGG pathways emphasize complement and coagulation cascades, cholesterol metabolism, and platelet activation, indicating convergence on hemostatic and lipid handling functions [9, 10]. The ChEA transcription factor signatures identify nuclear receptor and hepatic regulators such as RXR, HNF1A, CEBPA, ESR1, FOXA2, LXR, NR1H2, PPARA and FOXH1, which are associated with the same biological themes observed in GO and KEGG [11].

The reanalysis confirms that the eight overlapping genes AMBP, APCS, APOH, C8A, FGA, FGB, SERPINA4 and SERPINA6 represent a tightly linked network of extracellular plasma proteins that collectively influence hemostasis, complement activity, lipid handling and vascular homeostasis. Enrichment of fibrinolysis, blood coagulation, platelet activation and complement cascade pathways, together with transcription factor signatures of hepatic nuclear receptors, reinforces the view that these genes converge on the regulation of bloodborne processes and systemic inflammatory signaling. Deepdive literature synthesis further emphasizes that the proteins encoded by these genes function as structural components of fibrinogen, regulators of the membrane attack complex, modulators of extracellular matrix stability and carriers of glucocorticoids, all of which shape immune surveillance and vascular phenotype. An unexpected observation is the prominence of SERPINA4 and SERPINA6, whose primary roles in serine protease inhibition and corticosteroid binding extend beyond classical coagulation functions, suggesting ancillary contributions to vascular and inflammatory homeostasis. Together, these convergent findings support the original hypothesis that the intersecting gene set underlies shared biological themes across the examined Common Fund datasets, linking age and sex-dependent blood expression changes, drug-induced transcriptional responses, and broader hemostatic and metabolic processes.

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. *Nucleic Acids Research* **53**, D998–D1005. <https://doi.org/10.1093/nar/gkae1070> (2024).
13. 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).

14. 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).
15. 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).
16. 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).
17. Pepys, M. B. *et al.* Targeted pharmacological depletion of serum amyloid P component for treatment of human amyloidosis. *Nature*. <https://doi.org/10.1038/417254a> (2002).
18. 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).
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. 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).
21. Fish, R. J. & Neerman-Arbez, M. Fibrinogen gene regulation. *Thrombosis and Haemostasis*. <https://doi.org/10.1160/TH12-04-0273> (2012).
22. 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).
23. 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).
24. 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).
25. Miao, R. Q., Agata, J., Chao, L. & Chao, J. Kallistatin is a new inhibitor of angiogenesis and tumor growth. *Blood*. <https://doi.org/10.1182/blood-2002-01-0185> (2002).