

Crossing Gene Sets From GlyGen, GTEx and LINCS Reveals POLE As a Potential Modulator of Aging and Gene Perturbation

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as isolated collections, limiting the ability to see shared molecular changes across different experimental contexts. By crossing gene sets from three Common Fund libraries—GlyGen glycosylated proteins, GTEx tissue specific aging signatures, and LINCS L1000 CRISPR knockout perturbations—we sought a robust overlap that captures elements present in every dataset and therefore reflects a maximally inclusive intersection among all possible combinations. The GlyGen set contributes proteins known to be modified by glycans, the GTEx set supplies genes that increase in expression in older blood donors, and the LINCS set provides genes reduced after POLE knockout in prostate cancer cells, together highlighting secreted plasma factors linked to systemic metabolism and tissue homeostasis. The analysis identified a single crossing containing twenty genes that appear in all three libraries, revealing a coordinated shift in circulating proteins during aging. Enrichment of this intersecting set points to negative regulation of hemostasis, complement and coagulation cascades, and cholesterol metabolism, and implicates transcriptional regulators such as CEBPA and ESR1. These results emphasize coordinated changes in coagulation and lipid handling pathways in the aged blood transcriptome.

*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 8 and a p-value of 3.80208e-24 warranting further investigation.

The crossing involves the signature G48414YA, which is the GlyGen signature comprising proteins phosphorylated by the glycan [4]; the signature GTEx Blood 20-29 vs 70-79 Up, which is the GTEx signature of genes showing increased expression in blood of individuals aged 20-29 compared with those aged 70-79 [5]; and the signature POLE::PC3::96H::KO::Cas9::XPR011::N09 down, which is the LINCS signature of genes reduced in expression following CRISPR-Cas9 knockout of POLE in PC3 cells [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, perturbation, 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 8 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

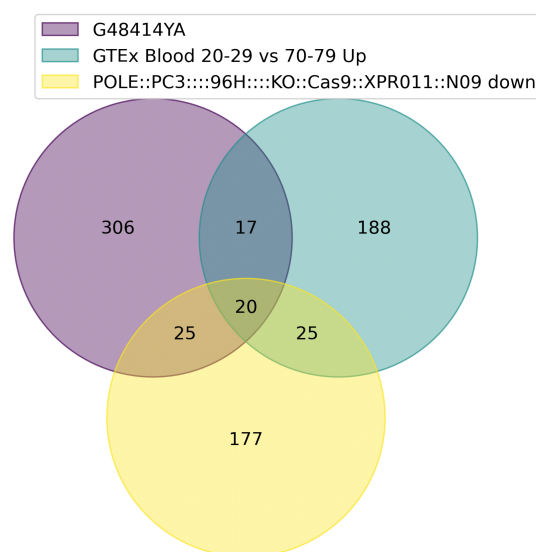


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 8, p-value of 3.802083×10^{-24} , and Jaccard index of 0.026385. The intersecting gene set identified 20 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of AFM AHSG ALB AMBP APCS APOB APOH AZGP1 C6 C9 CPB2 F9 FGG GC HRG KNG1 LBP PLG SERPINA7, and SERPIND1. Functional enrichment analysis of the gene set identified associations with Negative Regulation Of Hemostasis and Negative Regulation Of Blood Coagulation 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

These secreted glycoproteins function as transport and extracellular signaling mediators, linking systemic metabolism to tissue homeostasis. Afamin binds and shuttles hydrophobic molecules such as vitamin E throughout plasma and cerebrospinal fluid, while also acting as an extracellular chaperone that solubilizes acylmodified Wnt proteins, thereby influencing long-range signaling pathways [12–15]. In addition to its transport role, afamin levels correlate with metabolic alterations including insulin resistance and lipid dys-

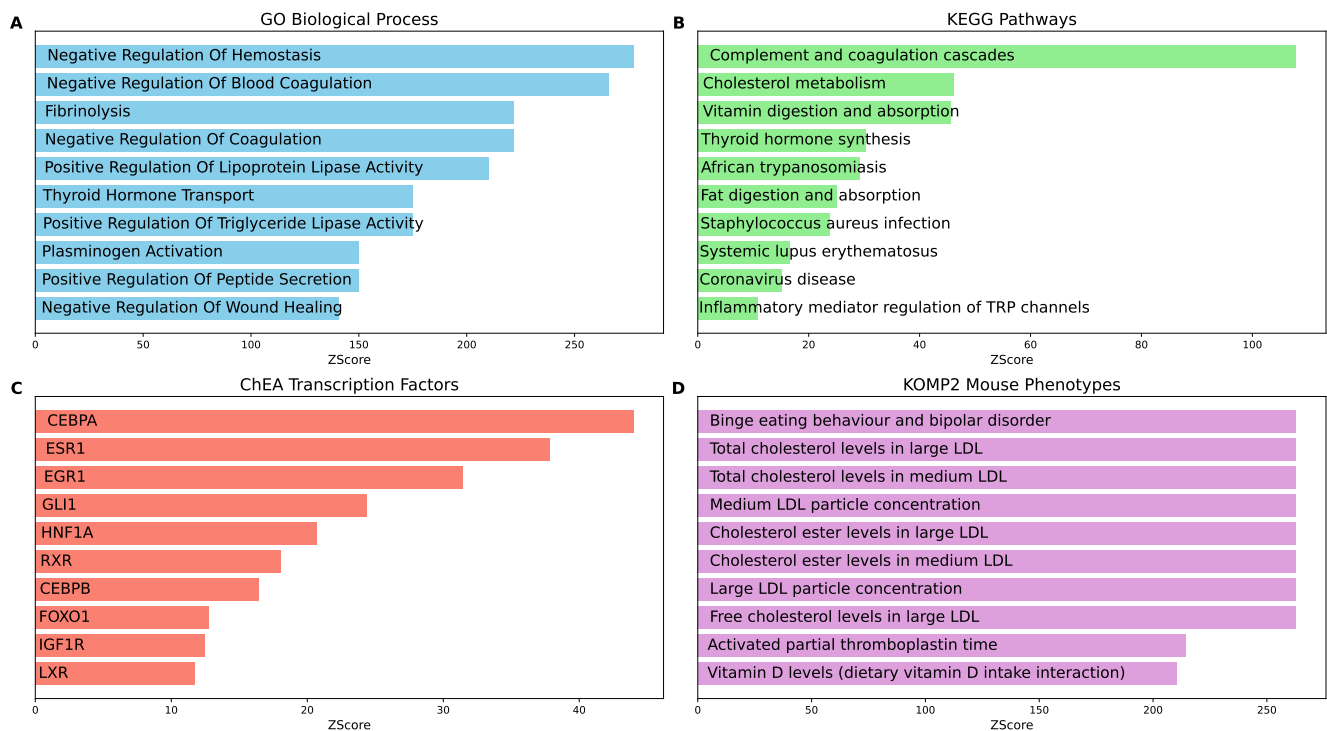


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog 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](#).

regulation, suggesting a broader involvement in age-related metabolic regulation [16]. Albumin contributes to antioxidant defenses by scavenging free radicals, a function that supports cellular protection against oxidative damage that accumulates with aging [17].

Collectively, these proteins regulate mineral and extracellular matrix dynamics, processes that are central to tissue integrity during aging. FetuinA (AHSG) forms soluble calciprotein particles that bind calciumphosphate clusters, preventing ectopic calcification and protecting vascular and softtissue compartments from pathological mineral deposition [18–21]. Bikunin, encoded by AMBP, serves as a structural component of intertrypsin inhibitor complexes that tether heavy chains to hyaluronan, stabilizing the extracellular matrix and modulating inflammatory remodeling responses [22–24]. Serum amyloidP component (APCS) binds apoptotic cells and amyloid fibrils, facilitating their clearance and linking innate immunity to the removal of damaged proteins that accumulate with age [25–27].

The enrichment profile of the upregulated gene set is dominated by terms related to blood clotting and lipid handling, with strong signals for negative regulation of hemostasis, fibrinolysis and other coagulation processes in the Gene Ontology Biological Process

analysis and a matching Complement and coagulation cascades term in the KEGG pathway results, indicating a coherent pattern of altered hemostatic control across functional annotation layers [8, 9]. In parallel, cholesterolcentric terms appear prominently in both the KEGG Cholesterol metabolism entry and the GWAS-derived library, which lists a suite of lipidrelated phenotypes such as total cholesterol levels in LDL subfractions and cholesterol ester concentrations, suggesting that the upregulated genes converge on lipid homeostasis pathways [9, 11]. The transcriptionfactor enrichment from ChEA points to regulators such as CEBPA, ESR1, EGR1 and HNF1A, providing a link between the observed pathway enrichments and upstream regulatory networks [10].

The reanalysis identified a highly significant overlap of twenty secreted plasma proteins between the G48414YA set, the GTEx blood agecomparison, and the POLEKO perturbation, underscoring a coordinated shift in circulatory factors during aging. Enrichment of these genes in Gene Ontology terms for negative regulation of hemostasis, fibrinolysis and the complement and coagulation cascades, together with KEGG pathways for cholesterol metabolism, reveals a dual alteration of hemostatic balance and lipid homeostasis in the aged blood transcriptome. Transcriptionfactor

signatures point to CEBPA, ESR1, EGR1 and HNF1A as upstream regulators, linking the observed functional changes to known metabolic and hormonal control networks. Deepdive summaries of individual proteins highlight roles in transport of hydrophobic molecules (AFM), antioxidant defense (ALB), inhibition of ectopic calcification (AHSG), extracellular matrix stabilization (AMBP) and clearance of apoptotic material (APCS), all of which contribute to systemic metabolic regulation and tissue integrity with age. The concurrent emergence of cholesterolcentric signals alongside coagulation pathways was somewhat unexpected, indicating that age-related remodeling of plasma composition involves both lipid handling and hemostatic control. Together, these findings reinforce the original research question by demonstrating that the intersecting gene sets converge on biologically coherent processes of blood clotting, lipid metabolism and extracellular matrix maintenance, providing a focused molecular portrait of age-associated changes in the circulatory system.

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. 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/gkae1070> (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. Kratzer, I. *et al.* Afamin is synthesized by cerebrovascular endothelial cells and mediates α -tocopherol transport across an *in vitro* model of the blood–brain barrier. *Journal of Neurochemistry*. <https://doi.org/10.1111/j.1471-4159.2008.05796.x> (2009).
15. Jerkovic, L. *et al.* Afamin Is a Novel Human Vitamin E-Binding Glycoprotein Characterization and In Vitro Expression. *Journal of Proteome Research*. <https://doi.org/10.1021/pr0500105> (2005).
16. Seeber, B., Morandell, E., Lunger, F., Wildt, L. & Dieplinger, H. Afamin serum concentrations are associated with insulin resistance and metabolic syndrome in polycystic ovary syndrome. *Reproductive Biology and Endocrinology*. <https://doi.org/10.1186/1477-7827-12-88> (2014).
17. Roche, M., Rondeau, P., Singh, N. R., Tarnus, E. & Bourdon, E. The antioxidant properties of serum albumin. *FEBS Letters*. <https://doi.org/10.1016/j.febslet.2008.04.057> (2008).
18. 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).
19. Jahnen-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).
20. Viegas, C. S. *et al.* Chronic Kidney Disease Circulating Calcioprotein Particles and Extracellular Vesicles Pro-

mote Vascular Calcification: A Role for GRP (Gla-Rich Protein). *Arteriosclerosis, Thrombosis, and Vascular Biology*. <https://doi.org/10.1161/ATVBAHA.117.310578> (2018).

21. 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).
22. 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).
23. 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).
24. 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).
25. 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).
26. 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).
27. 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).