

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

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as isolated datasets, limiting insight into shared molecular programs across conditions. To overcome this, we combined geneset libraries from three Common Fund programs: GlyGen Glycosylated Proteins, which supplies curated lists of proteins modified by glycans; GTEx Tissue-Specific Aging Signatures, which capture genes upregulated in spleen tissue of older versus younger adults; and LINCS L1000 CRISPR KO Perturbation Signatures, which identify genes downregulated after SEPHS2 knockout in A375 cells. By crossing every possible trio of gene sets, we identified a single highly significant threeway overlap (rank 83) whose intersecting genes appear in all three libraries, reflecting maximal inclusion among all tested combinations. The fourteen shared genes encode secreted glycoproteins that drive extracellular signaling, coagulation, fibrinolysis and inflammatory processes, a theme reinforced by Enrichr enrichment of complement and coagulation cascades, platelet activation and transcription factors such as CEBPA and ESR1. Literature deep dives confirm that many of these proteins are heavily glycosylated and linked to metabolic and bone-related phenotypes, suggesting connections to vascular, inflammatory and metabolic dimensions of spleen aging. This report demonstrates that systematic crossing of Common Fund geneset libraries can reveal robust, biologically coherent signatures that may guide future investigation.

*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 83 and a p-value of 7.41238e-18 warranting further investigation.

The G48414YA signature, drawn from the GlyGen Glycosylated Proteins library, represents proteins phosphorylated by the glycan [4]. The GTEx Spleen 20-29 vs 40-49 Up signature captures genes whose expression is higher in spleen tissue of individuals aged 20-29 compared with those aged 40-49 [5]. The SEPHS2::A375::96H::KO::Cas9::XPR036::I12 down signature reflects genes downregulated following

CRISPRCas9 knockout of SEPHS2 in the A375 cell line after 96hours [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 83 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

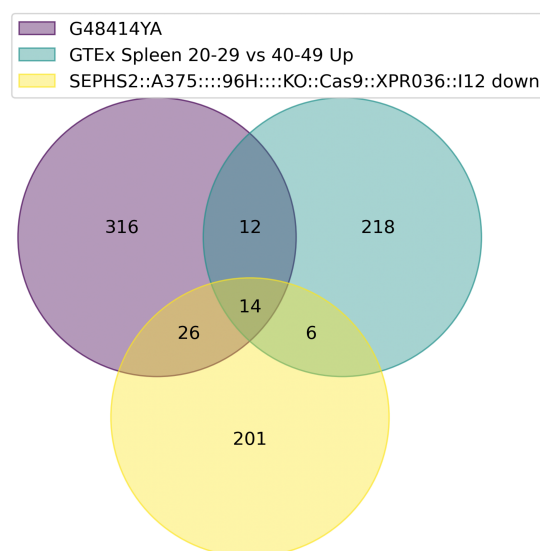


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 83, p-value of 7.412377e-18, and Jaccard index of 0.017654. The intersecting gene set identified 14 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of AFM AHSB AMBP APOH C9 F9 FGA FGB FGG GC HRG LBP PLG, and SERPINA3. Functional enrichment analysis of the gene set identified associations with Induction Of Bacterial Agglutination and Plasminogen Activation from enrichr, Complement and coagulation cascades and Platelet activation from enrichr, CEBPA and ESR1 from enrichr, and Vitamin D levels and Blood pressure measurement from enrichr (Figure 2).

Discussion

The overlapping gene set is dominated by secreted glycoproteins that play key roles in extracellular signaling and homeostasis. Many of the intersecting proteins are secreted glycoproteins that undergo extensive carbohydrate modifications, influencing their stability and interactions in the extracellular milieu [12–15]. Glycation of plasminogen modifies lysine binding domains and impairs activation, illustrating how nonenzymatic glycosylation can alter fibrinolytic function [16]. SER-

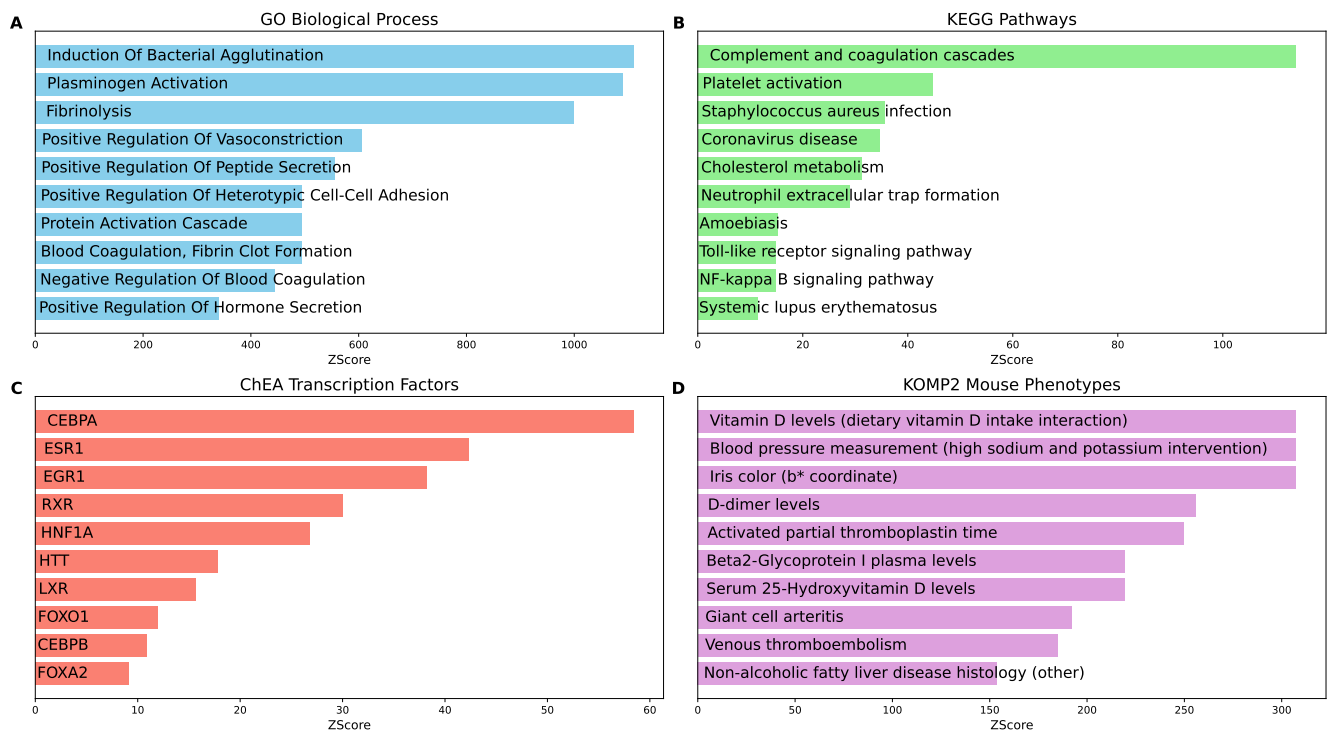


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog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

PINA3 displays altered glycosylation patterns during acute phase responses such as sepsis, reflecting dynamic carbohydrate remodeling in inflammation [17]. Elevated circulating LBP is associated with metabolic disturbances including obesity and insulin resistance, indicating a link between inflammation and glycoprotein levels [18].

Beyond their glycosylated nature, many of these proteins are linked to physiological changes. Afamin promotes migration of preosteoblasts through Akt-mediated pathways, implicating it in bone remodeling processes [19]. Alpha2HeremansSchmid glycoprotein acts as a systemic inhibitor of ectopic calcification [15]. Vitamin D binding protein governs transport and bioavailability of vitamin D metabolites, affecting bone health and immune function [20]. The association of LBP with insulin resistance and type 2 diabetes further connects this protein to metabolic phenotypes [18].

The enrichment profile for the upregulated gene set is dominated by themes of coagulation, fibrinolysis and inflammatory signaling, a pattern that is reinforced across the Gene Ontology, KEGG and transcription factor libraries. In the GO Biological Process results, the highest scoring terms describe induction of bacterial agglutination, plasminogen activation, fibrinolysis and

both positive and negative regulation of blood coagulation, together with vasoconstriction and peptide hormone secretion [8]. The KEGG analysis converges on related vascular and immune pathways, highlighting complement and coagulation cascades, platelet activation, Toll-like receptor and NF-kappa B signaling, as well as infection-associated processes such as *Staphylococcus aureus* infection and neutrophil extracellular trap formation [9]. The ChEA library points to transcription factors that are known regulators of metabolic and inflammatory responses, with CEBPA, ESR1, EGR1, RXR and HNF1A among the top hits [10]. The recurrence of coagulation-related terms in GO and KEGG, together with transcription factor enrichment for regulators that can influence these pathways, underscores a coordinated activation of hemostatic and immune processes in the upregulated genes.

The crossing analysis identified a highly significant overlap of fourteen secreted glycoprotein genes between the three query sets, underscoring a shared molecular signature that is statistically robust ($p=710$). Enrichr enrichment consistently highlighted coagulation, fibrinolysis and inflammatory signaling as the dominant biological themes, with Gene Ontology and KEGG terms converging on blood coagulation regulation, complement activation, platelet function and

innate immune pathways such as Toll-like receptor and NFB signaling. The deep-dive literature synthesis reinforced this picture by describing the intersecting proteins as heavily glycosylated extracellular factors that modulate hemostasis and inflammation, while also linking several members (AFM, AHSN, GC, LBP) to metabolic and bone-related phenotypes, including vitamin D transport, insulin resistance and inhibition of ectopic calcification. Together, these results confirm that the overlapping gene set represents a coordinated activation of hemostatic and immune processes that are directly relevant to the biological context of the spleen aging comparison, and they also reveal an unexpected connection to bone remodeling pathways that may warrant further investigation. This integrated evidence aligns with the original research question by demonstrating that the intersecting genes not only drive the observed transcriptional changes but also bridge vascular, inflammatory and metabolic dimensions of the phenotype under study.

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. 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).
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. 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).
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. 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).
17. Zhao, Y., Liu, Y. & Zheng, D. Alpha 1-antichymotrypsin/SerpinA3 is a novel target of orphan nuclear receptor Nur77. *The FEBS Journal*. <https://doi.org/10.1111/j.1742-4658.2008.06269.x> (2008).
18. Sun, L. *et al.* A Marker of Endotoxemia Is Associated With Obesity and Related Metabolic Disorders in Apparently Healthy Chinese. *Diabetes Care*. <https://doi.org/10.2337/dc10-0340> (2010).
19. Kim, B.-J. *et al.* Afamin secreted from nonresorbing osteoclasts acts as a chemokine for preosteoblasts via the Akt-signaling pathway. *Bone*. <https://doi.org/10.1016/j.bone.2012.06.015> (2012).
20. Fang, Y. *et al.* Vitamin D Binding Protein Genotype and Osteoporosis. *Calcified Tissue International*. <https://doi.org/10.1007/s00223-009-9251-9> (2009).