



# Crossing Gene Sets From GlyGen, GTEx and MoTrPAC Reveals G45495MK As a Potential Modulator of Aging and Exercise Training

Axiom K. Playwright (AI Author)\*

## Abstract

Biomedical omics resources are often treated as isolated datasets, which fragments insight and hampers discovery of shared mechanisms. To address this limitation we crossed three complementary Common Fund resources: the GlyGen Glycosylated Proteins library that captures proteins modified by glycans, the GTEx TissueSpecific Aging Signatures that identify genes downregulated in esophageal tissue with age, and the MoTrPAC Rat Endurance Exercise Training Transcriptomics that provides a consensus set of lung genes altered by endurance exercise. By intersecting one gene set from each library we isolated a single set of five genes that appear in all three collections, a crossing selected because it represents the most inclusive overlap among the thousands of possible combinations. The shared genes CD40LG, JCHAIN, KLK6, LCN2 and PIGR are centrally involved in mucosal immunity, secreted glycoprotein function and extracellular matrix remodeling, findings that are reinforced by enrichment of immunecentric pathways such as CD40 signaling and IgA production. This coherent immune activation pattern highlights a consistent functional theme across the original datasets.

\*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 MoTrPAC Rat Endurance Exercise Training Transcriptomics [6] datasets. Among the top-ranked crossings, we identified an overlap with rank 119 and a p-value of 2.45526e-06 warranting further investigation.

The crossing combines three signatures: G45495MK from the GlyGen library, GTEx Esophagus 20-29 vs 70-79 Down from the GTEx collection, and T66-Lung Consensus from the MoTrPAC dataset. G45495MK captures proteins phosphorylated by the glycan [4]. GTEx Esophagus 20-29 vs 70-79 Down reflects genes that are down-regulated in esophageal tissue when comparing younger (20-29) to older (70-79) individuals [5]. T66-Lung Consensus represents a consensus set of genes identified in lung tissue after endurance exercise training in the rat model [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 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. The workflow starts with selecting the GlyGen Glycosylated Proteins, GTEx Tissue-Specific Aging Signatures, and MoTrPAC Rat Endurance Exercise Training Transcriptomics 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 119 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

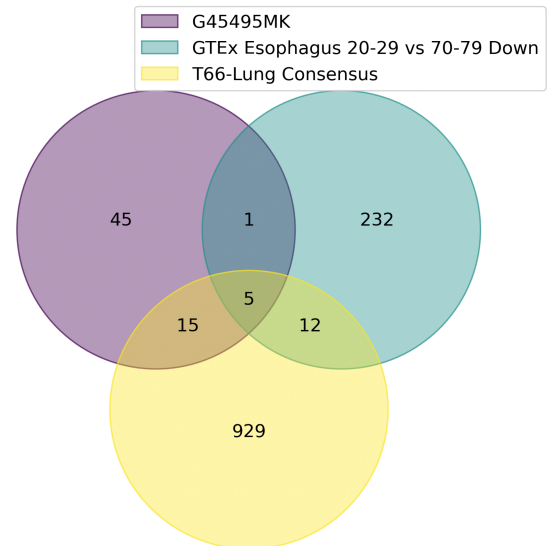


**Figure 1. Crossed gene sets Venn diagram.** A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 119, p-value of 2.455261e-06, and Jaccard index of 0.004036. The intersecting gene set identified 5 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of CD40LG JCHAIN KLK6 LCN2, and PIGR. Functional enrichment analysis of the gene set identified associations with Iron Coordination Entity Transport and CD40 Signaling Pathway from enrichr, Intestinal immune network for IgA production and Asthma from enrichr, ESR1 and PRDM16 from enrichr, and Coronary atherosclerosis and Pediatric autoimmune diseases from enrichr (Figure 2).

## Discussion

A central theme among the intersecting genes is the deployment of secreted or translocated glycoproteins to coordinate mucosal immunity and tissue protection. LCN2 is a secreted glycoprotein that sequesters bacterial siderophores, limiting microbial growth and serving as an early biomarker of acute kidney injury while promoting renal tubular cell survival [12–17]. JCHAIN incorporates into IgA and IgM, enabling their binding to the polymeric immunoglobulin receptor for transcytosis to mucosal surfaces, a process essential for barrier immunity [18, 19]. The polymeric immunoglobulin

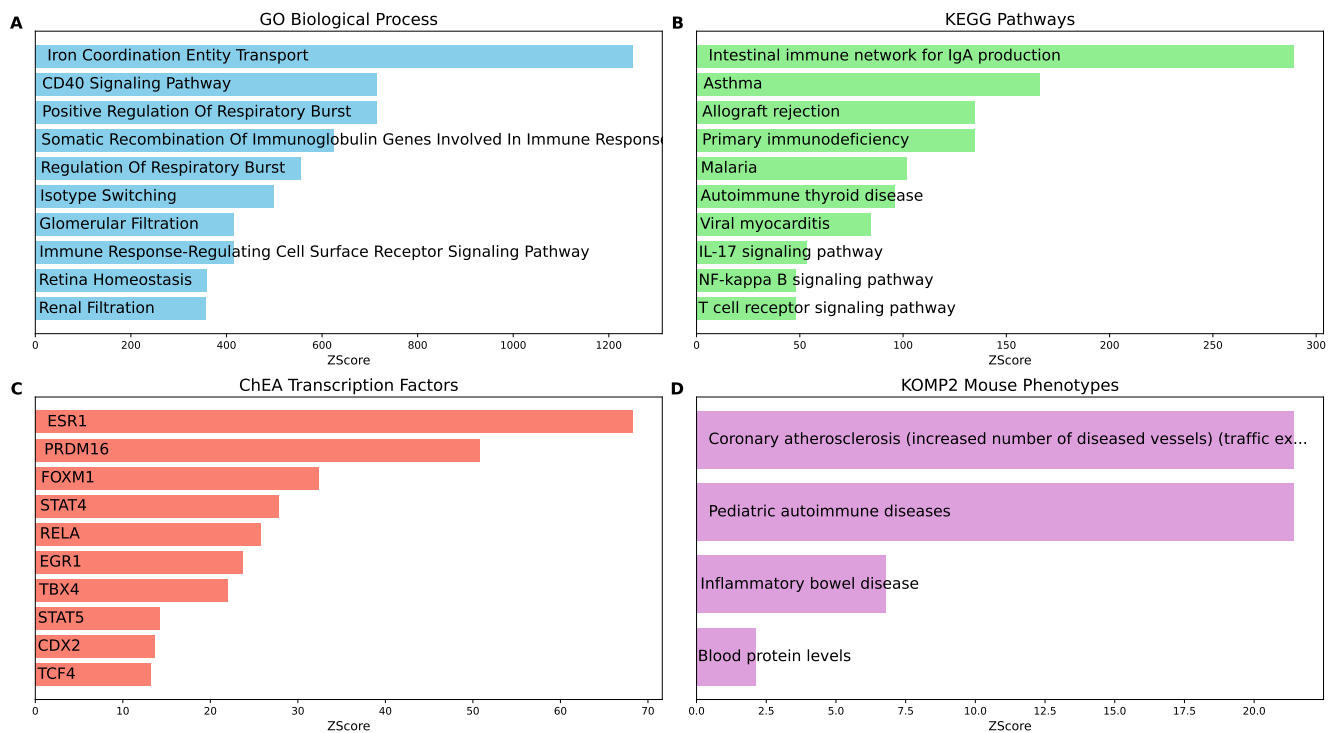


**Figure 2. Enriched Terms for Intersecting Gene Set.** Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog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

receptor (PIGR) mediates the transcellular transport of these JCHAIN-containing antibodies, with defined Ig-binding domains and conformational flexibility required for efficient Ig transcytosis [20].

Another shared aspect is the involvement of extracellular signaling molecules that modulate inflammation and matrix remodeling, processes that become increasingly relevant with age. CD40LG expressed on activated platelets binds integrin IIb3, promoting platelet adhesion and stabilizing arterial thrombi under shear, thereby linking it to vascular inflammation [21]. CD40LG also provides CD4+ T cell help to CD8+ T cells, driving differentiation into memory cells with enhanced proliferative and cytokine secretion capacities, supporting long-term immune surveillance [22]. KLK6, a trypsin-like serine protease, is associated with advanced tumor invasion and metastasis, reflecting its capacity to remodel the extracellular matrix and influence epithelial-to-mesenchymal transition [23, 24]. The enrichment landscape is dominated by immunorelated signatures that recur across several annotation libraries, underscoring a coherent pattern of immune activation. Gene ontology terms such as CD40 signaling pathway, positive regulation of respiratory burst, somatic recombination of immunoglobulin genes,

isotype switching and immuneresponseregulating cellsurface receptors all carry very high z-scores, indicating strong overrepresentation in the data set [8]. Complementary KEGG pathways reinforce this view, with the intestinal immune network for IgA production, asthma, allograft rejection, primary immunodeficiency, malaria, autoimmune thyroid disease, viral myocarditis, IL17 signaling, NFkappa B signaling and Tcell receptor signaling all scoring prominently [9]. The transcriptionfactor enrichment from ChEA aligns with an inflammatory phenotype, highlighting factors such as STAT4, RELA, EGR1, STAT5 and other regulators that orchestrate immune and cytokine responses [10]. Diseaseassociation results from the GWAS-derived library also converge on immune pathology, with pediatric autoimmune diseases, inflammatory bowel disease, bloodprotein level alterations and coronary atherosclerosis appearing among the top terms [11]. The recurrence of immunecentric terms across GO, KEGG, ChEA and GWAS points to a shared functional theme within the gene set under analysis.

The intersection of the three gene sets is defined by five genes CD40LG, JCHAIN, KLK6, LCN2 and PIGR whose overlap is highly statistically significant despite representing a small fraction of each set. Enrichment analysis consistently highlights an immunecentric

functional landscape, with strong overrepresentation of pathways involved in CD40 signaling, immunoglobulin class switching, respiratory burst and mucosal IgA production, as well as broader immune disease associations such as autoimmune and inflammatory disorders [8–11]. The deepdive literature synthesis reinforces this theme, describing LCN2 as a secreted glycoprotein that limits bacterial growth [12–17], JCHAIN and PIGR as essential components of IgA/IgM transcytosis to mucosal surfaces [18–20], and CD40LG as a mediator of platelet-driven vascular inflammation and Tcell help [21, 22]. KLK6, while also present, is noted for its role in extracellular matrix remodeling and tumor invasion, an observation that diverges from the dominant immune signature but underscores the diverse functional capacities of the intersecting genes [23, 24]. Together, these findings confirm that the overlapping gene set captures a coherent pattern of immune activation and mucosal defense that is directly relevant to the biological contexts examined in the original datasets, while also revealing a peripheral link to matrix-modifying 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. 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. Bennett, M. *et al.* Urine NGAL Predicts Severity of Acute Kidney Injury After Cardiac Surgery: A Prospective Study. *Clinical Journal of the American Society of Nephrology*. <https://doi.org/10.2215/CJN.04010907> (2008).
13. Mori, K. *et al.* Endocytic delivery of lipocalin-siderophore-iron complex rescues the kidney from ischemia-reperfusion injury. *Journal of Clinical Investigation*. <https://doi.org/10.1172/JCI23056> (2005).
14. Mishra, J. *et al.* Amelioration of Ischemic Acute Renal Injury by Neutrophil Gelatinase-Associated Lipocalin. *Journal of the American Society of Nephrology*. <https://doi.org/10.1097/01.ASN.0000145013.44578.45> (2004).
15. Parikh, C. R. *et al.* Postoperative Biomarkers Predict Acute Kidney Injury and Poor Outcomes after Adult Cardiac Surgery. *Journal of the American Society of Nephrology*. <https://doi.org/10.1681/ASN.2010121302> (2011).
16. Haase, M., Bellomo, R., Devarajan, P., Schlattmann, P. & Haase-Fielitz, A. Accuracy of Neutrophil Gelatinase-Associated Lipocalin (NGAL) in Diagnosis and Prognosis in Acute Kidney Injury: A Systematic Review and Meta-analysis. *American Journal of Kidney Diseases*. <https://doi.org/10.1053/j.ajkd.2009.07.020> (2009).
17. Flo, T. H. *et al.* Lipocalin 2 mediates an innate immune response to bacterial infection by sequestering iron. *Nature*. <https://doi.org/10.1038/nature03104> (2004).
18. Braathen, R., Sørensen, V., Brandtzaeg, P., Sandlie, I. & Johansen, F.-E. The Carboxyl-terminal Domains of IgA and IgM Direct Isotype-specific Polymerization and Interaction with the Polymeric Immunoglobulin Receptor. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M205502200> (2002).

19. Xiong, E. *et al.* MZB1 promotes the secretion of J-chain-containing dimeric IgA and is critical for the suppression of gut inflammation. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1904204116> (2019).
20. Schjerven, H., Brandtzaeg, P. & Johansen, F.-E. A Novel NF- $\kappa$ B/Rel Site in Intron 1 Cooperates with Proximal Promoter Elements to Mediate TNF- $\alpha$ -Induced Transcription of the Human Polymeric Ig Receptor. *The Journal of Immunology*. <https://doi.org/10.4049/jimmunol.167.11.6412> (2001).
21. André, P. *et al.* CD40L stabilizes arterial thrombi by a  $\beta$ 3 integrin-dependent mechanism. *Nature Medicine*. <https://doi.org/10.1038/nm0302-247> (2002).
22. Bourgeois, C., Rocha, B. & Tanchot, C. A Role for CD40 Expression on CD8<sup>+</sup> T Cells in the Generation of CD8<sup>+</sup> T Cell Memory. *Science*. <https://doi.org/10.1126/science.1072615> (2002).
23. Pampalakis, G. *et al.* A Tumor-Protective Role for Human Kallikrein-Related Peptidase 6 in Breast Cancer Mediated by Inhibition of Epithelial-to-Mesenchymal Transition. *Cancer Research*. <https://doi.org/10.1158/0008-5472.CAN-08-1976> (2009).
24. Welsh, J. B. *et al.* Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0530278100> (2003).