

Crossing Gene Sets from GTEx, MoTrPAC and RummaGEO Reveals Azacitidine as a Potential Modulator of Brain Aging and Exercise Metabolism

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as isolated collections, limiting the ability to see shared biological signals across studies. By crossing gene set libraries derived from the Common Fund Data Ecosystem, we can combine distinct experimental contexts and reveal robust molecular programs that appear in multiple conditions. In this work we selected three resources: the GTEx tissuespecific aging signatures that capture genes downregulated in the aging brain, the MoTrPAC rat endurance exercise transcriptomics that identify genes upregulated in brown adipose tissue after training, and the RummaGEO drug perturbation signatures that reflect genes induced by the azacitidine treatment. The three sets were intersected to find genes present in all of them, representing a maximal shared component among the many possible overlaps. The resulting nineteengene intersection highlighted a coherent network of immune receptor regulators, including PTPRC and CCR5, whose posttranslational modifications control signaling thresholds. Enrichment analysis linked these genes to immune functions such as superoxide production, myeloid cell immunity, antigen receptor signaling, and also to bone remodeling pathways. The inclusion of azacitidine responsive genes underscores the relevance of this drugperturbation signature within the shared immune and remodeling axis.

*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 GTEx Tissue-Specific Aging Signatures [4], MoTrPAC Rat Endurance Exercise Training Transcriptomics [5], and RummaGEO Drug Perturbation Signatures [6] datasets. Among the top-ranked crossings, we identified an overlap with rank 24 and a p-value of 4.34592e-17 warranting further investigation.

The GTEx Brain 2029 vs 3039 Down signature derives from the GTEx TissueSpecific Aging Signatures library, representing a set of genes identified in that aging-focused analysis [4]; the T69BrownAdipose Male 8W Up signature originates from the MoTrPAC Rat Endurance Exercise Training Transcriptomics collection, reflecting genes highlighted in that exerciserelated dataset [5]; and the GSE176008 0 V 1 AZACITIDINE HUMAN UP signature comes from the RummaGEO Drug Perturbation Signatures resource, indicating genes emphasized in

that drugresponse study [6].

Methods

The GTEx Tissue-Specific Aging dataset was constructed using gene expression count data from the Genotype Tissue Expression (GTEx) Portal [4]. 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 [5]. 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 time-point. 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 [6]. 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 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 24 was selected for further investigation, and the inter-

secting 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

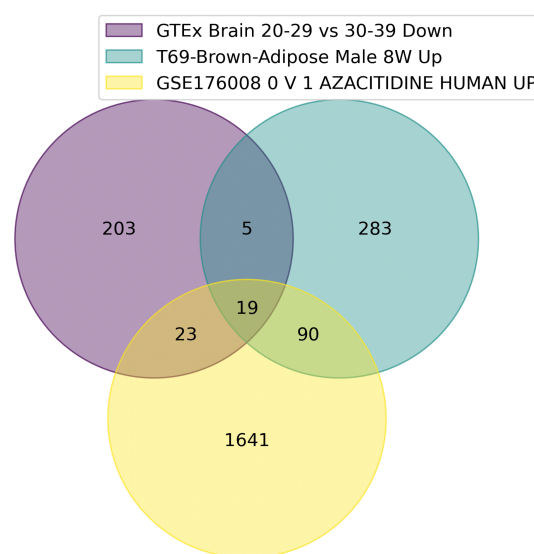


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 24, p-value of 4.345922e-17, and Jaccard index of 0.008392. The intersecting gene set identified 19 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of APBB1IP ARHGAP9 CCR5 CLEC7A CSF1R CYTH4 DAPP1 FCGR1A HCK IRF8 ITGAM ITGAX MYO1G PARVG PIK3R5 PTPRC TLR7 TYROBP, and WDFY4. Functional enrichment analysis of the gene set identified associations with Positive Regulation Of Superoxide Anion Generation and Regulation Of Superoxide Anion Generation from enrichr, Acute myeloid leukemia and Fc gamma R-mediated phagocytosis from enrichr, CIITA and SOX2 from enrichr, and Survival in colon cancer and Systemic sclerosis from enrichr (Figure 2).

Discussion

The shared biological theme among these genes is the regulation of immune receptor function through posttranslational modifications that influence cell sig-

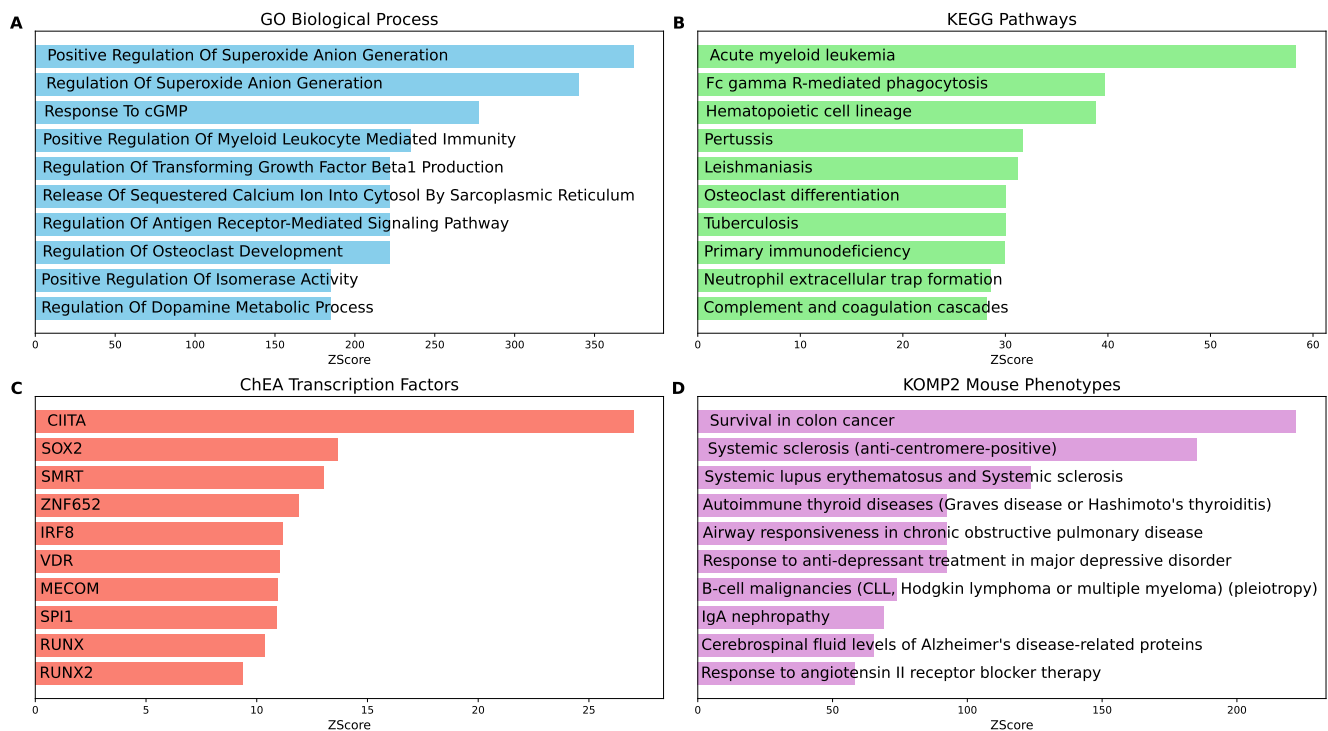


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog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

naling. One key example is the CD45 phosphatase encoded by PTPRC, whose activity can be altered by changes in its glycosylation state, thereby modulating signaling thresholds in lymphocytes and impacting disease processes [12–21]. Another illustrative case is the chemokine receptor CCR5, where tyrosine sulfation of its Nterminal region is essential for highaffinity interactions with viral ligands and therapeutic antagonists, underscoring the broader importance of receptor modifications in immune responses [22].

The enrichment profile of the upregulated genes reveals a striking convergence on immunerelated and boneremodeling processes across several annotation libraries. In the Gene Ontology collection, the most prominent terms involve regulation of superoxide anion generation, positive regulation of myeloid leukocyte mediated immunity, transforming growth factor-beta signaling, calcium ion release into the cytosol, antigenreceptormediated signaling, osteoclast development, and dopamine metabolism, indicating coordinated activation of oxidative, inflammatory, and developmental pathways [8]. Consistent themes emerge in the KEGG analysis, where pathways such as acute myeloid leukemia, FcgammaRmediated phagocytosis, hematopoietic cell lineage, osteoclast differentiation, and a suite of infectious diseaserelated cascades (per-

tussis, leishmaniasis, tuberculosis, neutrophil extracellular trap formation, complement and coagulation cascades) are strongly enriched, reinforcing the link to hematopoietic and innate immune functions [9]. Transcriptionfactor enrichment (ChEA) highlights master regulators of immune cell identity and differentiation, including CIITA, IRF8, VDR, SPI1, and members of the RUNX family, together with pluripotencyassociated factor SOX2, suggesting that the upregulated cohort is driven by a network of factors that orchestrate antigen presentation, myeloid lineage commitment, and bonerelated gene programs [10]. The GWASderived (KomP) results further connect the gene set to autoimmune and inflammatory disease phenotypes, with top associations to systemic sclerosis, systemic lupus erythematosus, autoimmune thyroid disease, chronic obstructive pulmonary disease responsiveness, majordepressivedisorder treatment response, Bcell malignancies, IgA nephropathy, Alzheimers disease biomarkers, and angiotensinIIreceptorblocker response, underscoring a broad relevance to disorders of immune dysregulation and tissue remodeling [11].

The crossing of the three gene sets yielded a highly significant overlap of nineteen genes ($p=4.310$) despite a modest Jaccard similarity, indicating that a focused subset of transcripts is shared among agere-

lated brain downregulation, early stage brown adipose upregulation, and AZACITIDINE-induced human upregulation. Enrichment analysis of these intersecting genes consistently highlighted immune-related functions, including regulation of superoxide production, myeloid leukocyte immunity, antigen receptor signaling, and pathways such as acute myeloid leukemia, Fc-mediated phagocytosis, hematopoietic cell lineage, and osteoclast differentiation. Transcription factor profiling pointed to master regulators of immune cell identity (CIITA, IRF8, SPI1, RUNX family) and the pluripotency factor SOX2, while GWAS links connected the gene set to a broad spectrum of autoimmune and inflammatory conditions. The deep dive perturbation data underscored the importance of posttranslational modifications in modulating immune receptor activity, exemplified by CD45 (PTPRC) glycosylation and CCR5 tyrosine sulfation, which influence signaling thresholds and ligand interactions. An unexpected theme emerging from the enrichment is the prominence of bone remodeling processes, such as osteoclast development and dopamine metabolism, suggesting that the shared gene signature may bridge immune regulation with tissue remodeling pathways. Together, these findings reinforce the relevance of the intersecting genes to the immune and remodeling axes interrogated in the original research question, while also revealing unanticipated connections that merit further investigation.

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. 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).
5. 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).
6. 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).
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. Clark, M. C. *et al.* Galectin-3 binds to CD45 on diffuse large B-cell lymphoma cells to regulate susceptibility to cell death. *Blood*. <https://doi.org/10.1182/blood-2012-06-438234> (2012).
13. Ruiz-Perez, F. *et al.* Serine protease autotransporters from *Shigella flexneri* and pathogenic *Escherichia coli* target a broad range of leukocyte glycoproteins. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1101006108> (2011).
14. Guo, H., Cruz-Munoz, M.-E., Wu, N., Robbins, M. & Vellette, A. Immune Cell Inhibition by SLAMF7 Is Mediated by a Mechanism Requiring Src Kinases, CD45, and SHIP-1 That Is Defective in Multiple Myeloma Cells. *Molecular and Cellular Biology*. <https://doi.org/10.1128/MCB.01107-14> (2015).
15. Fulcher, J. A. *et al.* Galectin-1 Co-clusters CD43/CD45 on Dendritic Cells and Induces Cell Activation and Migration through Syk and Protein Kinase C Signaling. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M109.037507> (2009).
16. Zhang, M. Cell surface sialic acids do not affect primary CD22 interactions with CD45 and surface IgM nor the rate of constitutive CD22 endocytosis. *Glycobiology*. <https://doi.org/10.1093/glycob/cwh126> (2004).
17. Zhou, C., Saxon, A. & Zhang, K. Human Activation-Induced Cytidine Deaminase Is Induced by IL-4 and

Negatively Regulated by CD45: Implication of CD45 as a Janus Kinase Phosphatase in Antibody Diversification. *The Journal of Immunology*. <https://doi.org/10.4049/jimmunol.170.4.1887> (2003).

18. Fernandes, R. A. *et al.* Immune receptor inhibition through enforced phosphatase recruitment. *Nature*. <https://doi.org/10.1038/s41586-020-2851-2> (2020).
19. Tromp, A. T. *et al.* Human CD45 is an F-component-specific receptor for the staphylococcal toxin Pantón-Valentine leukocidin. *Nature Microbiology*. <https://doi.org/10.1038/s41564-018-0159-x> (2018).
20. Vogel, A., Strassburg, C. P. & Manns, M. P. 77 C/G mutation in the tyrosine phosphatase CD45 gene and autoimmune hepatitis: evidence for a genetic link. *Genes amp; Immunity*. <https://doi.org/10.1038/sj.gene.6363918> (2003).
21. Ishikawa, H. *et al.* Requirements of src family kinase activity associated with CD45 for myeloma cell proliferation by interleukin-6. *Blood*. <https://doi.org/10.1182/blood.v99.6.2172> (2002).
22. Huang, C.-c. *et al.* Structures of the CCR5 N Terminus and of a Tyrosine-Sulfated Antibody with HIV-1 gp120 and CD4. *Science*. <https://doi.org/10.1126/science.1145373> (2007).