

Crossing Gene Sets from GTEx, IDG and MoTrPAC Reveals Metformin Association with Cardiac and Metabolic Processes

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as separate datasets, which fragments insight and limits discovery of shared mechanisms. Crossing gene sets from Common Fund programs can reveal coherent molecular programs, improve robustness, and connect diverse biological contexts. In this report we intersected three complementary gene sets: an upregulated signature from the GTEx Tissues library that captures genes with increased expression in the left ventricle of female donors aged 30-39, a Metformin-related signature from the IDG Drug Targets library that groups genes linked to the pharmacological action of Metformin, and a downregulated signature from the MoTrPAC Rat Endurance Exercise Training Transcriptomics library that reflects genes reduced in brown adipose tissue of male eightweek subjects after endurance exercise. By crossing all possible combinations of one set from each library, the third-ranked overlap was selected because it contained the maximal number of genes common to all three sets. The intersection comprised four mitochondrial complex I genes *NDUFA6*, *NDUFAB1*, *NDUFB3* and *NDUFS1* each appearing in every source set and highlighting a shared emphasis on oxidative phosphorylation. Enrichment analysis confirmed coordinated activation of aerobic respiration and electron transport, while the Metformin-associated gene set suggests that Metformin may be related to this mitochondrial pathway.

*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 Tissues [4], IDG Drug Targets [5], and MoTrPAC Rat Endurance Exercise Training Transcriptomics [6] datasets. Among the top-ranked crossings, we identified an overlap with rank 3 and a p-value of 1.48598e-05 warranting further investigation.

Heart - Left Ventricle Female 30-39 Up signature derives from the GTEx Tissues library and captures genes with increased expression in the left ventricle of female donors aged 30-39 [4]. Metformin signature comes from the IDG Drug Targets library and comprises genes linked to the pharmacological action of Metformin [5]. T69-Brown-Adipose Male 8W Down signature is drawn from the MoTrPAC Rat Endurance Exercise Training Transcriptomics library and reflects genes with

reduced expression in brown adipose tissue of male 8week subjects following endurance exercise [6]. The GlyGen signature represents proteins phosphorylated by the glycan.

Methods

The GTEx Tissues dataset was constructed using gene expression count data from the Genotype Tissue Expression (GTEx) Portal. Sample metadata was filtered to create groups of samples with matching tissue type and donor age ranges. The grouped data was log2-transformed, quantile normalized, and z-scored to create up and down gene sets. The IDG Drug Targets dataset was constructed using drug-target interaction data from DrugCentral. For each drug, the associated targets were collected into a gene set, keeping gene sets with a length of at least 5. 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 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. The workflow starts with selecting the GTEx Tissues, IDG Drug Targets, 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 3 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

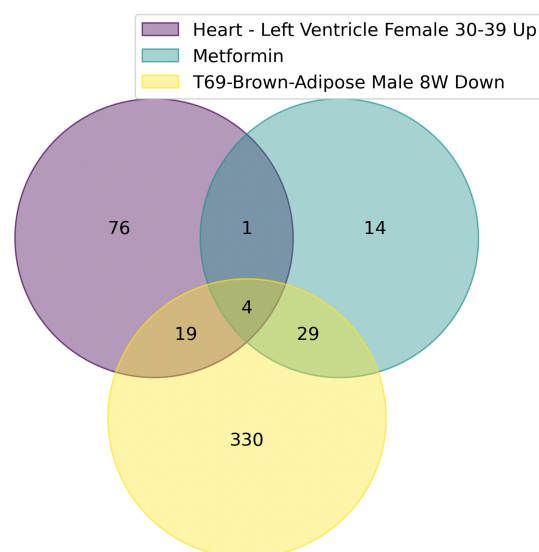


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 3, p-value of 1.485979e-05, and Jaccard index of 0.008457. The intersecting gene set identified 4 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of NDUFA6, NDUFB1, NDUFB3, and NDUFS1. Functional enrichment analysis of the gene set identified associations with Mitochondrial Electron Transport, NADH To Ubiquinone and NADH Dehydrogenase Complex Assembly from enrichr, Oxidative phosphorylation and Retrograde endocannabinoid signaling from enrichr, FLI1 and HNF4A from enrichr, and Frontotemporal dementia with GRN mutation and Breast cancer from enrichr (Figure 2).

Discussion

The intersecting genes primarily converge on mitochondrial oxidative phosphorylation, a central pathway for cellular energy production. NDUFB3 encodes a subunit of complex I that initiates electron transfer essential for ATP generation [12]. NDUFS1 encodes the largest iron-sulfur protein subunit of complex I, facilitating electron transfer between consecutive iron-sulfur clusters and ensuring catalytic efficiency of the complex [13–17]. NDUFB1 functions as the mitochondrial acyl carrier protein, providing the octanoyl-ACP precursor for posttranslational lipoylation of mitochondrial proteins, a modification required for respiratory complex I activity and maintenance of membrane potential [18]. Additionally, NDUFB1 participates in the iron-sulfur cluster assembly machin-

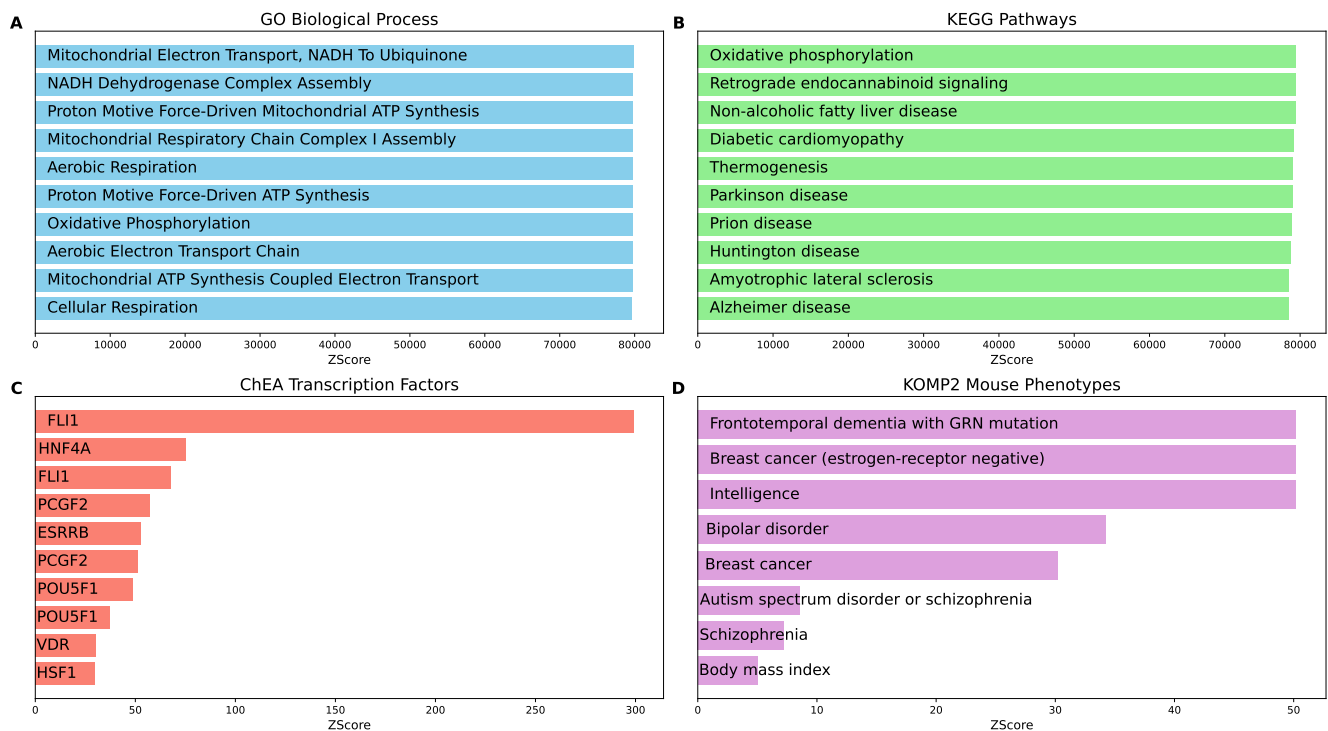


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog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

ery, stabilizing the NFS1/ISD11 core that is essential for the biogenesis of iron-sulfur clusters in multiple mitochondrial enzymes [19].

These shared roles underscore the importance of the gene set in preserving mitochondrial integrity and preventing metabolic stress. Mutations in *NDUFS1* disrupt electron flow, destabilize complex I, and increase reactive oxygen species generation, thereby impairing mitochondrial energy metabolism [13–17]. Similarly, loss-of-function variants in *NDUFB3* compromise complex I activity, leading to deficient ATP production and clinical manifestations of mitochondrial dysfunction [12].

The enrichment profile of the upregulated genes is dominated by mitochondrial and oxidative phosphorylation themes that recur across both the Gene Ontology biological process and KEGG pathway libraries, underscoring a coordinated activation of aerobic respiration, electron transport chain components and ATP synthesis [8, 9]. In contrast, the transcription factor analysis of the downregulated genes highlights a distinct regulatory signature, with the strongest signals for the ETS family factor FLI1 and for HNF4A, together with additional factors such as PCGF2, ESRRB, POU5F1, VDR and HSF1 [10]. The GWAS-derived enrichment further

differentiates the two sets: disease association terms cluster around neurodegenerative and psychiatric conditions (e.g., frontotemporal dementia, bipolar disorder, schizophrenia) as well as cancer phenotypes, suggesting that the downregulated gene set aligns with traits implicated in these disorders [11]. Together, the convergence of mitochondrial and respiratory terms in the upregulated set and the contrasting regulatory and disease-linked signatures in the downregulated set reveal complementary functional polarities between the two gene groups.

The reanalysis identified a statistically robust overlap of four mitochondrial genes *NDUFA6*, *NDUFAB1*, *NDUFB3* and *NDUFS1* among the left ventricle, metabolically treated, and brown adipose signatures, supporting a shared emphasis on oxidative phosphorylation processes. Enrichr results reinforced this pattern, highlighting coordinated activation of aerobic respiration, electron transport chain components and ATP synthesis for the upregulated set, while the downregulated set displayed a distinct regulatory profile enriched for ETS family transcription factors and disease association terms linked to neurodegenerative, psychiatric and cancer phenotypes, an unexpected connection given the metabolic context of the original datasets. Deep dive summaries of the intersecting genes further

emphasized their central roles in complex I assembly, iron-sulfur cluster biogenesis and maintenance of mitochondrial membrane potential, underscoring the importance of preserving mitochondrial integrity to avoid metabolic stress. Together, these convergent findings illustrate complementary functional polarities between the intersecting gene groups and provide a focused perspective on how mitochondrial activity may underlie the biological signals captured across the heart, metformin, and brown adipose experiments, directly addressing the research question of shared molecular pathways within the Common Fund program datasets.

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. Avram, S. *et al.* DrugCentral 2023 extends human clinical data and integrates veterinary drugs. *Nucleic Acids Research* **51**, D1276–D1287. <https://doi.org/10.1093/nar/gkac1085> (2022).
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. Calvo, S. E. *et al.* Molecular Diagnosis of Infantile Mitochondrial Disease with Targeted Next-Generation Sequencing. *Science Translational Medicine*. <https://doi.org/10.1126/scitranslmed.3003310> (2012).
13. Ferreira, M. *et al.* Progressive cavitating leukoencephalopathy associated with respiratory chain complex I deficiency and a novel mutation in NDUFS1. *Neurogenetics*. <https://doi.org/10.1007/s10048-010-0265-2> (2011).
14. Hoefs, S. J. *et al.* Novel mutations in the NDUFS1 gene cause low residual activities in human complex I deficiencies. *Molecular Genetics and Metabolism*. <https://doi.org/10.1016/j.ymgme.2010.03.015> (2010).
15. Kann, O., Huchzermeyer, C., Kovács, R., Wirtz, S. & Schuelke, M. Gamma oscillations in the hippocampus require high complex I gene expression and strong functional performance of mitochondria. *Brain*. <https://doi.org/10.1093/brain/awq333> (2010).
16. Iuso, A. *et al.* Dysfunctions of Cellular Oxidative Metabolism in Patients with Mutations in the NDUFS1 and NDUFS4 Genes of Complex I. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M513387200> (2006).
17. Martín, M. A. *et al.* Leigh Syndrome Associated With Mitochondrial Complex I Deficiency Due to a Novel Mutation in the NDUFS1 Gene. *Archives of Neurology*. <https://doi.org/10.1001/archneur.62.4.659> (2005).
18. Feng, D., Witkowski, A. & Smith, S. Down-regulation of Mitochondrial Acyl Carrier Protein in Mammalian Cells Compromises Protein Lipoylation and Respiratory Complex I and Results in Cell Death. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M806991200> (2009).
19. Cai, K., Frederick, R. O., Dashti, H. & Markley, J. L. Architectural Features of Human Mitochondrial Cysteine Desulfurase Complexes from Crosslinking Mass Spectrometry and Small-Angle X-Ray Scattering. *Structure*. <https://doi.org/10.1016/j.str.2018.05.017> (2018).