

Crossing Gene Sets from LINCS, Metabolomics and MoTrPAC Reveals FAD as a Potential Modulator of Gene Regulation and Brown Adipose Tissue Remodeling

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as isolated datasets, which fragments insight and limits the discovery of shared molecular programs. Combining gene sets from multiple Common Fund programs can reveal robust, reproducible signatures that connect diverse perturbations to common biology. In this study we crossed three curated libraries retrieved from the Common Fund Data Ecosystem: the LINCS L1000 CRISPR knockout signature that captures genes upregulated when a target gene is deleted, the Metabolomics enzymemetabolite association set that defines enzymes linked to the cofactor flavin adenine dinucleotide, and the MoTrPAC rat endurance exercise transcriptome that lists genes downregulated in brown adipose tissue after eight weeks of training. All possible triples of one set from each library were evaluated for statistical overlap and the ninth ranked combination, which maximized the number of genes shared across all three sets, was selected for detailed analysis. The intersecting set comprises the mitochondrial dehydrogenases ACADM, ACADVL, GCDH and IVD, enzymes that initiate fattyacid betaoxidation and branchedchain aminoacid catabolism. Enrichment profiling highlighted coordinated fattyacid oxidation, PPAR signaling and related metabolic pathways, while transcriptionfactor analysis identified regulators such as E2F1, RUNX2 and PPARG. Together these results demonstrate that the shared dehydrogenases form a core metabolic signature linking CRISPR perturbation, cofactor biology and exerciseinduced adipose remodeling, illustrating the power of multilibrary crossing to uncover biologically meaningful overlaps.

*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 LINCS L1000 CRISPR KO Perturbation Signatures [4], Metabolomics Metabolite Enzyme Associations [5], and MoTrPAC Rat Endurance Exercise Training Transcriptomics [6] datasets. Among the top-ranked crossings, we identified an overlap with rank 9 and a p-value of 8.95033e-08 warranting further investigation.

The three signatures comprising the topranked crossing are the URB2::BICR6:::96H:::KO::Cas9::XPR027::G11 up signature, the FAD signature, and the T69BrownAdipose Male 8W Down signature. The URB2::BICR6:::96H:::KO::Cas9

up signature originates from the LINCS L1000 CRISPR KO Perturbation library and reflects a CRISPR knock-out experiment in which loss of the target gene leads to upregulation of the G11 transcript [4]. The FAD signature is drawn from the Metabolomics Metabolite Enzyme Associations library and denotes the set of enzymes that are associated with the cofactor flavin adenine dinucleotide (FAD) [5]. The T69BrownAdipose Male 8W Down signature comes from the MoTrPAC Rat Endurance Exercise Training Transcriptomics dataset and captures genes that are downregulated in brown adipose tissue of male rats after eight weeks of endurance training [6]. The GlyGen signature, which represents proteins phosphorylated by the glycan, is also part of the broader analysis context.

Methods

The LINCS L1000 CRISPR KO Perturbation Signatures dataset was processed from Level 3 L1000 data from the CLUE platform [4]. 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 Metabolomics Enzyme-Metabolite Associations dataset was processed from associations between metabolites and related enzymes [5]. The Metabolomics Workbench metabolite dataset was created by integrating structures and annotations from various open resources. The Human Metabolome Gene/Protein (MGP) database containing information about metabolites and linked genes/proteins was used to create an edgelist of metabolite-enzyme associations for metabolites with at least one public study deposited to the Metabolomics Workbench. The enzyme-metabolite edgelist was harmonized to create gene sets which were collected into a GMT. 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 LINCS L1000 CRISPR KO Perturbation Signatures, Metabolomics Metabolite Enzyme Associations, 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 9 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

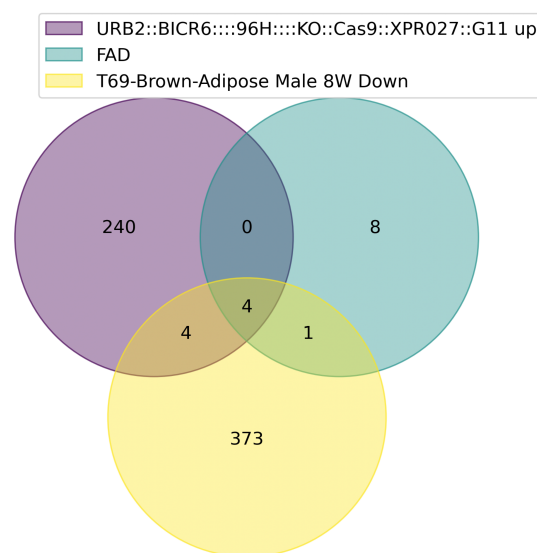


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 9, p-value of 8.950326e-08, and Jaccard index of 0.006349. The intersecting gene set identified 4 genes appearing in all involved sets (Figure 1). The intersecting gne set consists of ACADM

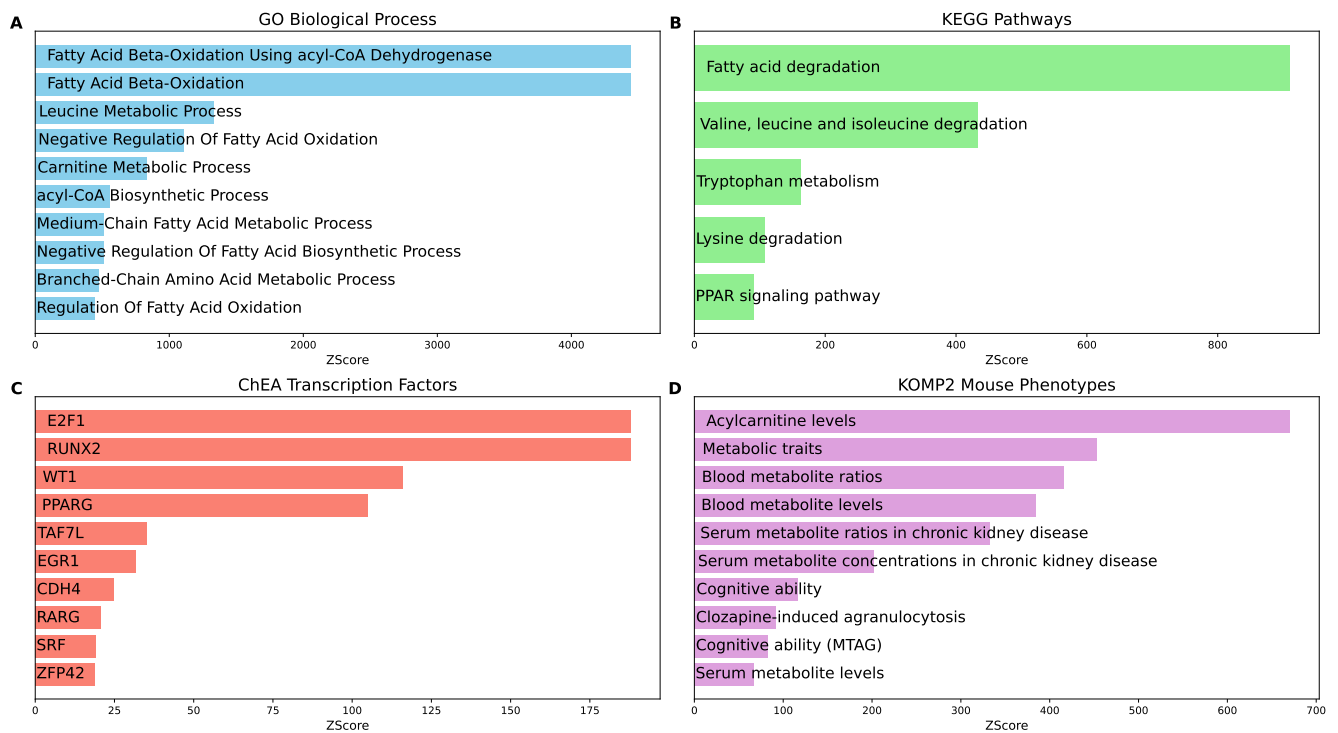


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog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

ACADVL GCDH, and IVD. Functional enrichment analysis of the gene set identified associations with Fatty Acid Beta-Oxidation Using acyl-CoA Dehydrogenase and Fatty Acid Beta-Oxidation from enrichr, Fatty acid degradation and Valine, leucine and isoleucine degradation from enrichr, E2F1 and RUNX2 from enrichr, and Acylcarnitine levels and Metabolic traits from enrichr (Figure 2).

Discussion

The intersecting gene set comprising ACADM, ACADVL, GCDH and IVD encodes mitochondrial dehydrogenases that initiate key catabolic pathways essential for cellular energy production [12–15]. These enzymes each catalyze the first dehydrogenation step in the mitochondrial oxidation of fatty acids or the oxidative degradation of branchedchain amino acids, thereby driving the conversion of substrateCoA derivatives into downstream metabolites [16–19]. Their activity sustains lipid and aminoacid homeostasis by preventing the buildup of upstream intermediates that can become toxic, as demonstrated by the role of ACADM in maintaining lipid balance, GCDH in averting neurotoxic metabolite accumulation, and IVD in regulating branchedchain aminoacid catabolism [18–20]. Loss or downregulation of these dehydrogenases is linked to

metabolic disturbances such as intracellular lipid accumulation, neurodegeneration, and diseaseassociated phenotypes, highlighting their importance for overall metabolic health [16, 20, 21].

The enrichment profile for the upregulated genes is dominated by metabolic processes that recur across three independent libraries. In the Gene Ontology Biological Process analysis the most extreme zscores are assigned to fattyacid oxidation, negative regulation of fattyacid oxidation and related carnitine and acylCoA biosynthetic pathways, indicating a strong bias toward lipid catabolism [8]. A parallel signal emerges from the KEGG results, where fattyacid degradation and the PPAR signaling pathway appear alongside branchedchain aminoacid catabolism (valine, leucine and isoleucine degradation) and tryptophan metabolism, reinforcing the notion that the upregulated set is enriched for coordinated fattyacid and aminoacid metabolic pathways [9]. The GWASstyle library further supports this metabolic theme, with top terms describing acylcarnitine levels, blood metabolite ratios and concentrations, and broader metabolictrait associations, all of which align with the lipidoxidation signatures from GO and KEGG [11]. In contrast, the transcriptionfactor enrichment (ChEA) highlights a distinct regulatory landscape. The most prominent

factors include E2F1, RUNX2 and PPARG, together with WT1, TAF7L, EGR1 and others, suggesting that the downregulated genes are preferentially bound by transcription factors associated with various regulatory functions [10]. Notably, PPARG appears both as a transcription factor hit and as a component of the PPAR signaling pathway in KEGG, providing a point of convergence between the metabolic and regulatory signatures [9, 10].

The crossing analysis identified a highly significant overlap of four mitochondrial dehydrogenases ACADM, ACADVL, GCDH and IVD between the URB2up, FAD and T69BrownAdipose male 8week down gene sets, confirming that these enzymes serve as a shared molecular signature across the disparate perturbations. Enrichr profiling of the upregulated genes reinforced this link, revealing a coordinated enrichment for fatty acid oxidation, PPAR signaling and branched-chain amino acid catabolism, all hallmarks of lipid and amino acid metabolism in brown adipose tissue. The transcription factor enrichment for the downregulated genes highlighted regulators such as E2F1, RUNX2 and PPARG, with PPARG also appearing in the metabolic KEGG pathway, suggesting a convergence of regulatory and metabolic signals. Deepdive literature synthesis further underscored the central role of the intersecting dehydrogenases in initiating mitochondrial catabolism of fatty acids and branched chain amino acids, and indicated that their loss can provoke lipid accumulation and metabolic dysfunction. The only unexpected element was the prominence of transcription factor signatures that did not directly map onto the metabolic terms, hinting at additional layers of regulation. Together, these results support the original hypothesis that the intersecting gene set drives key metabolic pathways in brown adipose tissue and that perturbations affecting these enzymes are likely to influence the metabolic phenotype under 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. 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).
5. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research* **44**, D463–D470. <https://doi.org/10.1093/nar/gkv1042> (2015).
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. Huss, J. M., Kopp, R. P. & Kelly, D. P. Peroxisome Proliferator-activated Receptor Coactivator-1 α (PGC-1 α) Coactivates the Cardiac-enriched Nuclear Receptors Estrogen-related Receptor- α and - γ . *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M206324200> (2002).
13. Sauer, S. W. *et al.* Bioenergetics in Glutaryl-Coenzyme A Dehydrogenase Deficiency. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M502845200> (2005).
14. Benatti, R. O. *et al.* Maternal high-fat diet consumption modulates hepatic lipid metabolism and microRNA-122 (<i>miR-122</i>) and microRNA-370 (<i>miR-370</i>) expression in offspring. *British Journal of Nutrition*. <https://doi.org/10.1017/S0007114514000579> (2014).

15. Mohsen, A.-W. A. & Vockley, J. Kinetic and spectral properties of isovaleryl-CoA dehydrogenase and interaction with ligands. *Biochimie*. <https://doi.org/10.1016/j.biochi.2014.11.007> (2015).
16. Choi, J. *et al.* Potential roles of PINK1 for increased PGC-1 α -mediated mitochondrial fatty acid oxidation and their associations with Alzheimer disease and diabetes. *Mitochondrion*. <https://doi.org/10.1016/j.mito.2014.09.005> (2014).
17. Külkens, S. *et al.* Late-onset neurologic disease in glutaryl-CoA dehydrogenase deficiency. *Neurology*. <https://doi.org/10.1212/01.WNL.0000167428.12417.B2> (2005).
18. Illig, T. *et al.* A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics*. <https://doi.org/10.1038/ng.507> (2009).
19. Luís, P. B. *et al.* Role of Isovaleryl-CoA Dehydrogenase and Short Branched-Chain Acyl-CoA Dehydrogenase in the Metabolism of Valproic Acid: Implications for the Branched-Chain Amino Acid Oxidation Pathway. *Drug Metabolism and Disposition*. <https://doi.org/10.1124/dmd.110.037606> (2011).
20. Radha Rama Devi, A. *et al.* Spectrum of mutations in Glutaryl-CoA dehydrogenase gene in glutaric aciduria type I – Study from South India. *Brain and Development*. <https://doi.org/10.1016/j.braindev.2015.05.013> (2016).
21. Ryu, M.-J., Lee, C., Kim, J., Shin, H.-S. & Yu, M.-H. Proteomic analysis of stargazer mutant mouse neuronal proteins involved in absence seizure. *Journal of Neurochemistry*. <https://doi.org/10.1111/j.1471-4159.2007.05100.x> (2007).