

Crossing Gene Sets from LINCS, Metabolomics and MoTrPAC Reveals Oxoglutaric Acid as a Potential Modulator of Energy Metabolism and Exercise Adaptation

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are often treated as separate datasets, which fragments insight and limits the discovery of shared molecular programs. To overcome this limitation we combined gene sets from three Common Fund libraries: a CRISPR knockout signature from the LINCS L1000 Perturbation collection that records the transcriptional response to loss of the kinase FAM20C in HT29 cells, an enzyme set linked to the metabolite oxoglutaric acid from the Metabolomics EnzymeMetabolite Associations, and a downregulated gene set from brown adipose tissue after eight weeks of endurance exercise in the MoTrPAC rat transcriptomics archive. These resources were accessed through the CFDE Workbench, crossed computationally, and the most inclusive threeway overlap was selected for further study because it contained the greatest number of genes present in all three libraries. The intersecting genes DLSD, DLST, IDH2 and IDH3A are core components of the tricarboxylic acid cycle and were shown by Enrichr analysis to be strongly associated with metabolic processes, specific transcription factor regulators and disease-related phenotypes. Deepdive literature synthesis highlighted their essential roles in mitochondrial energy production, redox balance and links to cell death pathways. Together the findings illustrate how integrating disparate Common Fund datasets can reveal a coherent metabolic module that is transcriptionally coordinated across perturbation, metabolite association and exercise contexts.

*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 36 and a p-value of 5.45910e-07 warranting further investigation.

The first signature, FAM20C::HT29:::96H:::KO::Cas9::XPR016::L12 up, is a CRISPR knockout perturbation from the LINCS library that captures the transcriptional response of HT29 cells 96 hours after loss of FAM20C, with the listed genes showing increased expression [4]. The second signature, Oxoglutaric Acid, comes from the

Metabolomics library and represents the set of enzymes associated with the metabolite oxoglutaric acid [5]. The third signature, T69BrownAdipose Male 8W Down, originates in the MoTrPAC transcriptomics collection and denotes the genes that are downregulated in brown adipose tissue of male subjects after eight weeks of endurance exercise [6]. The GlyGen signature, not listed among the top three, denotes proteins that are phosphorylated by the glycan.

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, perturbation, 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 pos-

sible 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 36 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

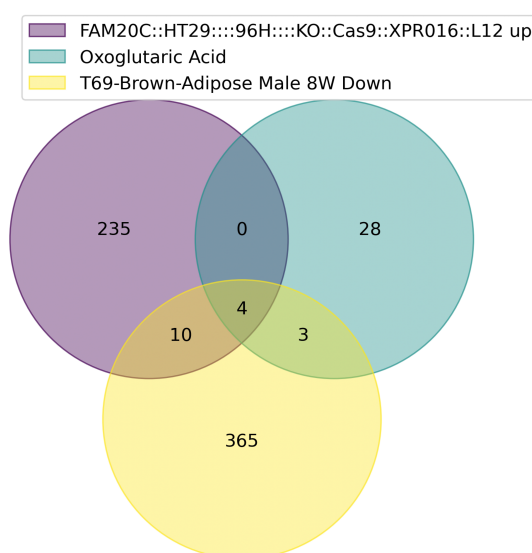


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 36, p-value of 5.459101e-07, and Jaccard index of 0.006202. The intersecting gene set identified 4 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of DLD, DLST, IDH2, and IDH3A. Functional enrichment analysis of the gene set identified associations with Tricarboxylic Acid Metabolic Process and acetyl-CoA Biosynthetic Process From Pyruvate from enrichr, Citrate cycle and Tryptophan metabolism from enrichr, E4F1 and HOXA2 from enrichr, and Mood instability and Non-small cell lung

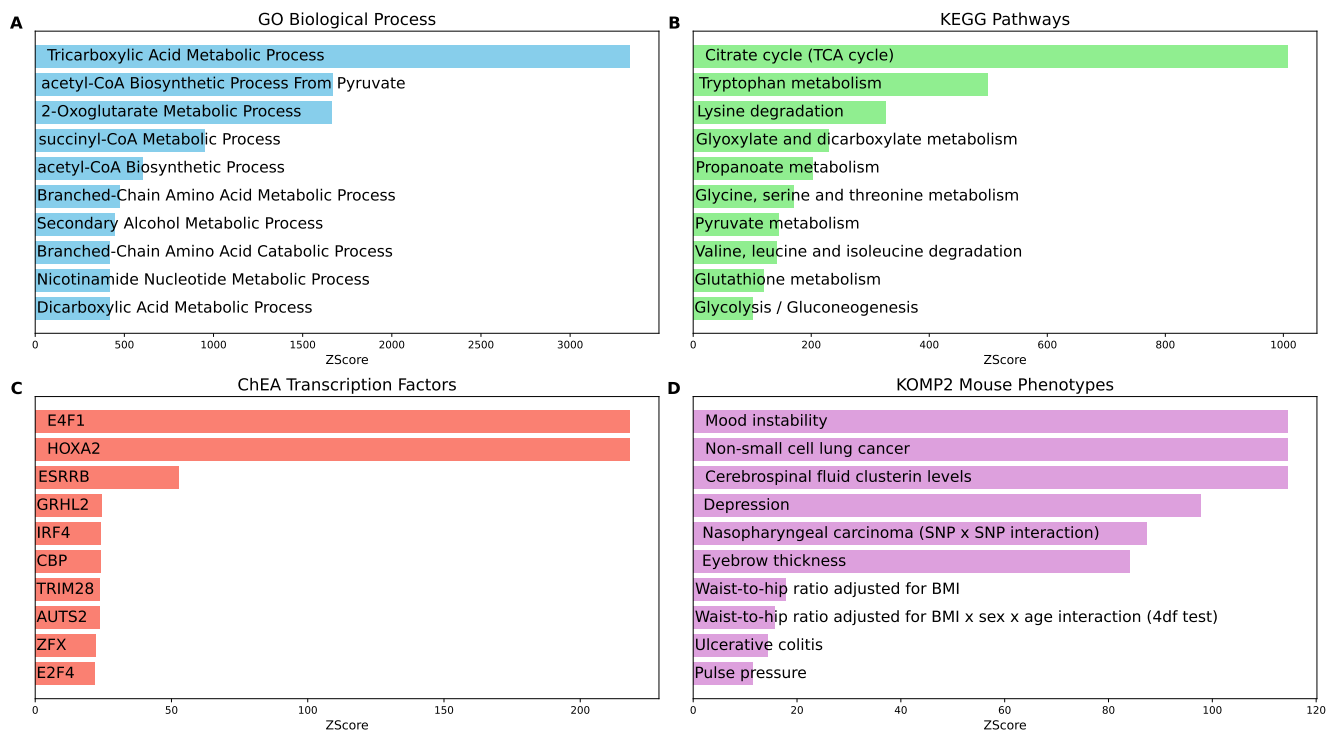


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog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

cancer from enrichr (Figure 2).

Discussion

These mitochondrial enzymes collectively sustain cellular energy production and redox balance. DLD catalyzes the oxidation of dihydrolipoamide and its partial deficiency disrupts mitochondrial dehydrogenase activities, predisposing neural and muscular tissues to heightened vulnerability against neurotoxins [12, 13]; DLST functions as the E2 component of the ketoglutarate dehydrogenase complex, and reduced DLST levels exacerbate oxidative stress and neuronal vulnerability in neurodegenerative disease models [14, 15]; IDH2 generates NADPH within mitochondria, a cofactor essential for maintaining redox homeostasis under metabolic stress [16]; IDH3A contributes NADH for ATP synthesis, linking its activity to overall energy homeostasis in the tricarboxylic acid cycle [17].

Alterations in the activity of these enzymes intersect with cell death and metabolic pathways. DLD modulates mitochondrial reactive oxygen species and iron levels under cystine deprivation, thereby contributing to ferroptotic cell death [18]; DLST overexpression correlates with aggressive neuroblastoma while its down-regulation leads to oxidative stress in neurodegenera-

tive contexts, highlighting its role in both tumorigenesis and neuronal decline [15, 19]; IDH2 supports reductive carboxylation of glutamine-derived ketoglutarate during hypoxia, sustaining lipid synthesis when oxygen is limited and influencing metabolic adaptations linked to cellular stress [16]; IDH3A overexpression reduces ketoglutarate levels, stabilizing HIF1 and promoting angiogenesis [20].

The enrichment profile for the upregulated gene set shows a striking convergence of metabolic themes across the Gene Ontology Biological Process and KEGG libraries, with terms describing the tricarboxylic acid cycle, acetylCoA biosynthesis, branchedchain aminoacid metabolism and related dicarboxylicacid and pyruvate processes all receiving very high zscores. This consistency reinforces a central role for core carbonfuel and aminoacid catabolic pathways in the upregulated cohort. In the transcriptionfactor library, the top hits include E4F1, HOXA2 and several other regulators, suggesting that the same gene set is also targeted by factors that have been linked to metabolic regulation. The phenotype/GWAS library contributes a distinct but complementary set of terms, such as mood instability, nonsmall cell lung cancer and ulcerative colitis, indicating that the metabolic signature may intersect with diseaserelated traits. Together, the

overlapping metabolic terms across GO and KEGG, the transcriptionfactor enrichment in ChEA and the diseaseoriented terms in GWAS point to a coherent pattern in which upregulated genes are coordinated around energyproducing pathways while also being associated with clinical phenotypes, highlighting a complementary relationship between molecular function and phenotypic outcome [8–11].

The crossing analysis identified a highly significant overlap of four mitochondrial enzymesDLSD, DLST, IDH2 and IDH3A between the three query sets, despite the modest Jaccard similarity, underscoring a focused convergence on core tricarboxylic acid cycle components. Deepdives of these genes consistently highlighted their central roles in sustaining cellular energy production, redox balance and mitochondrial function, while also revealing links to celldeath pathways and diseaseassociated stress responses. Enrichr enrichment reinforced this picture, showing striking alignment of the upregulated gene set with metabolic processes in Gene Ontology and KEGG, including tricarboxylic acid cycle, acetylCoA biosynthesis and branchedchain aminoacid catabolism, and identifying transcriptionfactor regulators such as E4F1 and HOXA2 that are tied to metabolic control. Complementary diseaseoriented terms from the GWAS library, notably mood instability, nonsmall cell lung cancer and ulcerative colitis, suggest that the metabolic signature extends into clinically relevant phenotypes, an observation that is somewhat unexpected given the primarily biochemical focus of the gene set. Together, these results support the overarching research question by demonstrating that the intersecting genes form a coherent metabolic module that is both transcriptionally coordinated and associated with diverse disease traits, linking molecular energy pathways to broader phenotypic outcomes.

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. Klivenyi, P. *et al.* Mice deficient in dihydrolipoamide dehydrogenase show increased vulnerability to MPTP, malonate and 3-nitropropionic acid neurotoxicity. *Journal of Neurochemistry*. <https://doi.org/10.1046/j.1471-4159.2003.02263.x> (2004).
13. Brautigam, C. A., Chuang, J. L., Tomchick, D. R., Machius, M. & Chuang, D. T. Crystal Structure of Human Dihydrolipoamide Dehydrogenase: NAD⁺/NADH Binding and the Structural Basis of Disease-causing Mutations. *Journal of Molecular Biology*. <https://doi.org/10.1016/j.jmb.2005.05.014> (2005).
14. Yang, L. *et al.* Mice deficient in dihydrolipoyl succinyl transferase show increased vulnerability to mitochondrial toxins. *Neurobiology of Disease*. <https://doi.org/10.1016/j.nbd.2009.07.023> (2009).
15. Shi, Q., Chen, H.-L., Xu, H. & Gibson, G. E. Reduction in the E2k Subunit of the α -Ketoglutarate Dehydrogenase

Complex Has Effects Independent of Complex Activity. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M409064200> (2005).

16. Wise, D. R. *et al.* Hypoxia promotes isocitrate dehydrogenase-dependent carboxylation of α -ketoglutarate to citrate to support cell growth and viability. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1117773108> (2011).
17. Findlay, A. S. *et al.* Mouse Idh3a mutations cause retinal degeneration and reduced mitochondrial function. *Disease Models and Mechanisms*. <https://doi.org/10.1242/dmm.036426> (2018).
18. Shin, D., Lee, J., You, J. H., Kim, D. & Roh, J.-L. Dihydrolipoamide dehydrogenase regulates cystine deprivation-induced ferroptosis in head and neck cancer. *Redox Biology*. <https://doi.org/10.1016/j.redox.2019.101418> (2020).
19. Anderson, N. M. *et al.* Metabolic Enzyme DLST Promotes Tumor Aggression and Reveals a Vulnerability to OXPHOS Inhibition in High-Risk Neuroblastoma. *Cancer Research*. <https://doi.org/10.1158/0008-5472.CAN-20-2153> (2021).
20. Zeng, L. *et al.* Aberrant IDH3a expression promotes malignant tumor growth by inducing HIF-1-mediated metabolic reprogramming and angiogenesis. *Oncogene*. <https://doi.org/10.1038/onc.2014.411> (2014).