A Systems-Level Study of Transcriptional Regulation of

*C. elegans* Metabolism

A Dissertation Presented

By

Shivani Nanda

Submitted to the Faculty of the

Morningside Graduate School of Biomedical Sciences at

UMass Chan Medical School

In partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

July 14, 2023

SYSTEMS BIOLOGY
A Systems-Level Study of Transcriptional Regulation of

*C. elegans* Metabolism

A Dissertation Presented

By

Shivani Nanda

This work was undertaken in the Morningside Graduate School of Biomedical Sciences

Interdisciplinary Graduate Program
Under the mentorship of

-----------------------------------------------------------------------------------------
A.J. Marian Walhout, Ph.D.; Thesis Advisor
-----------------------------------------------------------------------------------------
Lutfu Safak Yilmaz, Ph.D.; Thesis Co-Advisor
-----------------------------------------------------------------------------------------
Athma Pai, Ph.D.; Member of Committee
-----------------------------------------------------------------------------------------
Manuel Garber, Ph.D.; Member of Committee
-----------------------------------------------------------------------------------------
Michael Lee, Ph.D.; Member of Committee
-----------------------------------------------------------------------------------------
John Murray, Ph.D.; External Member of Committee
-----------------------------------------------------------------------------------------
Zhiping Weng, Ph.D.; Chair of Committee
-----------------------------------------------------------------------------------------
Mary Ellen Lane, Ph.D.
Dean of the Morningside Graduate School of Biomedical Sciences

July 14, 2023
DEDICATION

To my Mumma, who is my guide, my pillar, my friend, my unwavering confidante, and the air that I breathe, and to my Papa, who dreamed big for me before I could and who is my celestial advocate, forever watching me from the heavens above.
ACKNOWLEDGEMENTS

My Ph.D. journey has been enriched by the support of many individuals, to whom I owe immense gratitude and appreciation. First and foremost, I want to acknowledge my esteemed thesis advisor Dr. Marian Walhout, who opened the door for me to do research in her lab, even though I had no experience in systems biology or metabolism. She has an amazing way of breaking down complex concepts and making science seem less intimidating. It became quickly apparent during my rotation that her mentorship would be indispensable to my upcoming doctorate years. True to her word, she has been both my biggest cheerleader and toughest critic. With her help, I transitioned from a student into a scientist. She guided me on how to “ask the right questions,” foster critical thinking, and communicate science in an effective manner. Her careful review of each draft of my paper, and her meticulous examination of every figure, ensured that I delivered my best work. She has been there for me during the good times and the bad, and without her as my advisor, my Ph.D. journey would have been incomplete.

Next, I want to thank my dear thesis co-advisor Dr. Safak Yilmaz, who is the coolest mentor and friend one can have. His technical and mathematical knowledge is extraordinary. It’s been an honor working alongside him in the same office, that also gave me the convenience of throwing questions at him anytime, bouncing ideas off him, or venting my frustrations, and he would amazingly stop what he was doing to give me his full attention. I have learnt so much from him, and thanks to his guidance, I have become a better coder and statistician. The best thing about him is that no matter how his day is going, he always seems happy and has a smile on his face. Even though we have found that metabolism is extensively activated, or repressed by transcription, my Ph.D. journey has only ever been extensively activated by Marian and Safak. I am immensely lucky to be mentored by both of them.

I have received great advice and feedback on my research from my TRAC and QE committee including Dr. Zhiping Weng, Dr. Michael Lee, Dr. Manuel Garber, Dr. Athma Pai, Dr. Robert Brewster, Dr. Kate Fitzgerald, and Dr. Cole Haynes. I extend my gratitude to all of them. Special thanks to my external examiner Dr. John Murray, for flying all the way from Pennsylvania, for my defense. I would like to thank Dr. Job Dekker for asking the most insightful questions during my seminars, and for being a motivating force in the department.
I want to specially acknowledge Caryn Navaro, for her valuable feedback on my paper, and Lynn Jaime for printing out my paper figures in her office so that Marian received them first thing in the morning. My special shoutout to my Walhout labmates who are all amazing scientists and human beings and have provided their constructive feedback on my project during lab meetings. I would begin by thanking our lab alumni Cedric Diot and Auri Garcia-Gonzalez, for reading my QE proposal thoroughly and teaching me a thing or two about writing. Brent Horowitz is an excellent friend and one of the kindest people. Amongst countless other things, I want to especially thank him for tracking every lab birthday and making me try out climbing. I want to thank Amy Holdorf for sharing her insights on making publishing-quality figures; Xuhang “Hang” Li for many engaging discussions, and for setting benchmark for hard work and dedication in the lab; Gabrielle “Gabby” Giese for wearing multiple hats in the lab and for always updating the “burn book”; Yong-Uk Lee for many coffee trips to Dunkins, that one climbing sesh and never-boring conversations in between; Olga Ponomarova for being her pleasant and motivating self during the morning writing sessions; Hefei “Fei” Zhang for showing how good science can also happen while being calm and zen; Zeynep Mirza for being effortlessly smart, kind and brilliant; Daniel Richards for corroborating my beliefs on ‘intermittent fasting’, Nana for sharing the list of transporters, and lastly Sushila Bhattacharya for her clear heart and fulfilling my craving to talk in Hindi whenever I wanted. I have been fortunate to forge instant connections and friendships with Brittany Rosener, Sunil Guharajan, and Snehal Sambare from the department. Our interactions, though limited, have been meaningful and truly enriching.

I would like to thank my mentors prior to joining UMass who steered me in the right direction and inculcated the potential of a researcher in me, including Dr. U Deva Priyakumar (IIIT Hyderabad), Dr. Smita Agrawal (Strand Life Sciences), and Dr. Yasha Hasija (Delhi Technological University). Furthermore, I am thankful to Dr. Vikas Dhingra from Takeda, who is not only my uncle but also a consistent source of inspiration. As the pioneer in our family to venture into research, his journey has been a beacon for me. His informal yet impactful mentorship over the years has been instrumental in shaping my research path.

I have found amazing friends at UMass Medical School. Christine Carbone (my ride-or-die/ bestie), Megan Honeywell (the sweetest and most creative soul), Karthik Ramesh (someone who would give the shirt off his back for his friends),
Debanjan Goswamy (extremely genuine and funny), Havisha Honwad (multitalented and most entertaining), Shannon Bailey (coolest cat who shares my love for trash shows) and Eleni Jaecklein (wittiest of all), are all like family to me, and I have shared profound memories with each one of them. I would also like to thank my Ph.D. cohort for giving me the privilege of knowing and being friends with a very smart group of people especially Michelle Mosqueda, Mihir Doshi, and Mike Feyder. My family-like-friends at “Cirrus” provided me with a sense of Indian community and comfort that is irreplaceable, especially Niki, Mitul, Pallavi, Ashish, Honey, Sumit, Nidhi, and Aditya. My heartfelt appreciation extends to my lifelong friends, Pooja Sharma and Ridhima Dhamija, who hail from my hometown. Their friendship, akin to sisterhood, has been a source of steadfast support. I met some of the most dependable people at IIIT, whom I call friends for lifetime: Pushpa Itagi, Parshvika Sharma, Sankalp Khare, Neha Prasad, and Gayathri Lakshmi.

Finally, I can’t end this without a massive thank you to my family. Their support has been my rock, and I honestly would not have gotten this far without them. My mumma, Poonam Arya Nanda, a retired mathematics lecturer, is the one who sparked my love for science and math when I was a kid. She has always put her kids first and taught me what it means to be a strong woman. Whenever I have felt like the ground was slipping away beneath me, she was there, motivating me every step of the way. My papa, the late Sushil Kumar, was a powerful force in my life, and provided me with great education and best resources to succeed. He taught me how to be ambitious and always encouraged me to dream big. My younger brother, Rahul Nanda, may be younger, but he displays wisdom beyond his years. His advice has always provided me with valuable perspectives and insights. At my lowest points, it was his encouragement that kept me moving forward. I thank my grandparents KK Arya, the late Sudershan Arya, the late Vishan Devi and the late KC Nandwani for teaching the power of resilience and patience. All my uncles and aunts (Pankaj Arya, Priti Arya, Parveen Arya, and Mukta Dhingra), and cousins (Divyanshu, Atul, Anushka, Mrigank, and Ishani) have been supportive and always brought me so much joy. And of course, I thank my partner and husband, Vinay Garg. He has been with me through every phase of this Ph.D. journey over the last four years. He taught me how to organize and write my code like a “software developer” and not “bioinformatician”, though I am not quite sure I have nailed it yet. To all of them, I cannot say thank you enough.
ABSTRACT

Metabolism is controlled to ensure organismal development and homeostasis. Several mechanisms regulate metabolism, including allosteric control and transcriptional regulation of metabolic enzymes and transporters. Until now, metabolism regulation has mostly been described for individual genes and pathways, and the extent of transcriptional regulation of the entire metabolic network largely remains largely unknown.

In the first part of this thesis, I used variation in mRNA levels as a proxy for transcriptional regulation of metabolism and found that three-quarters of all metabolic genes are transcriptionally regulated in the nematode *Caenorhabditis elegans*. I also found that many annotated metabolic pathways are coexpressed. With the aid of gene expression data and the iCEL1314 metabolic network model, I defined coregulated sub-pathways in an unbiased manner. By utilizing a large gene expression compendium, I determined the conditions under which sub-pathways exhibit strong coexpression. Additionally, we developed 'WormClust', a web application that facilitates a gene-by-gene query of genes to view their association with metabolic (sub)-pathways.

In the second part of this thesis, I addressed the "missing" annotations in *C. elegans* metabolism. Through the analysis of various datasets spanning space, time, and conditions, I associated orphan metabolic genes, transporters,
transcription factors, and RNA-binding proteins with the existing metabolic network.

Overall, this study illuminates the ubiquity of transcriptional regulation of metabolism and provides a blueprint for similar studies in other organisms, including humans.
TABLE OF CONTENTS

DEDICATION .................................................................................................................................................. iii
ACKNOWLEDGEMENTS .............................................................................................................................. iv
ABSTRACT ................................................................................................................................................... vii
TABLE OF CONTENTS .............................................................................................................................. ix
LIST OF FIGURES ....................................................................................................................................... xiii
LIST OF TABLES .......................................................................................................................................... xv
LIST OF FILES ........................................................................................................................................... xvi
LIST OF COPYRIGHTED MATERIAL ....................................................................................................... xvii

CHAPTER I : INTRODUCTION ...................................................................................................................... 1

Preface .......................................................................................................................................................... 1

Introduction to metabolism .......................................................................................................................... 1
  Metabolic homeostasis ............................................................................................................................... 2
  Metabolic network model .......................................................................................................................... 4
    Genome-scale network reconstruction .................................................................................................... 5
    Flux balance analysis (FBA) .................................................................................................................... 7
    Limitations of FBA ................................................................................................................................ 9

Regulation of metabolism .......................................................................................................................... 10
  Allosteric regulation of metabolism ........................................................................................................ 11
  Post-translational modifications in metabolism ....................................................................................... 12
  Post-transcriptional regulation of metabolism ........................................................................................ 14
    Role of RNA binding proteins (RBPs) in metabolism ........................................................................... 14
  Transcriptional regulation of metabolism ................................................................................................ 16
    Transcriptional initiation, elongation, and termination ......................................................................... 18
    Role of transcription factors (TFs) in regulating metabolism ................................................................. 19
    Metabolic rewiring ................................................................................................................................ 20
    Cost versus benefits ............................................................................................................................... 21

Studies illustrating transcriptional regulation of metabolic processes... .................................................. 24
  Studies in unicellular organisms .............................................................................................................. 24
  Studies in multicellular organisms ......................................................................................................... 26

Caenorhabditis elegans as a model organism ............................................................................................ 28
  Advantages ............................................................................................................................................... 29
    iCEL1314 - a C. elegans metabolic network model ............................................................................. 30
  Examples of transcriptional regulation of metabolism in C. elegans ..................................................... 31
Knowledge gaps and thesis objectives ........................................... 38

CHAPTER II: TRANSCRIPTIONAL REGULATION OF METABOLISM AT A SYSTEMS-LEVEL IN C. ELEGANS ........................................... 44

Preface .......................................................................................... 44

Background .................................................................................... 44

Results .......................................................................................... 48

Methods ......................................................................................... 93
Gene-centric coexpression with metabolic (sub)-pathways by WormClust. 106
Data availability ............................................................................................................. 106

CHAPTER III: SYSTEMATICALLY DECIPHERING THE “MISSING” KNOWLEDGE IN C. ELEGANS METABOLISM .................................................................................. 108

Background .......................................................................................................................... 108
Transporters .......................................................................................................................... 110
Orphan metabolic genes ........................................................................................................ 111
TFs ...................................................................................................................................... 112
RBP ................................................................................................................................... 113

Results .................................................................................................................................. 115
60% of protein-coding C. elegans genes are poorly annotated ................................................. 115
Majority of C. elegans genes are associated with metabolic pathways .................................... 117
RBP ...................................................................................................................................... 119
Transporters are highly enriched in coexpression with energy-related genes ....................... 121
Transporters associated with metabolism .............................................................................. 124
Vacular ATPases facilitates the transport of multiple metabolites ............................................. 126
Many transporters are coexpressed with peroxisomal FA degradation .................................. 131
Transporters enriched to aminoacyl-tRNA biosynthesis are mainly involved in transporting amino acids or tRNA import .......................................................... 132
Orphan genes associated with metabolism ............................................................................ 133
Orphan genes associated with vacular ATPases ..................................................................... 138
Orphan genes associated with peroxisomal FA degradation .................................................. 140
Orphan genes associated to electron transport chain ............................................................. 143
Orphan genes associated to Glycolysis/ Gluconeogenesis, canonical propionate degradation pathway, and glyoxylate/decarboxylate metabolism ................................................................. 147
TFs associated with metabolism ............................................................................................. 148
NHRs associated with metabolism .......................................................................................... 151
Other TFs associated with metabolism .................................................................................... 153
RBP ..................................................................................................................................... 158
RBP ...................................................................................................................................... 160
RBP ...................................................................................................................................... 163
RBP ...................................................................................................................................... 166

Methods .................................................................................................................................. 175
Datasets .................................................................................................................................. 175
Scoring and Clustering of Gene-Pathway Enrichment Associations ........................................ 175
Building the network graph of gene-pathway associations ...................................................... 176

CHAPTER IV: DISCUSSION..................................................................................................... 177

Part I: Understanding the transcriptional regulation of metabolism at a systems-level in C. elegans ................................................................................................. 177
Transcriptional regulation occurs extensively at the level of genes, reactions and pathways ................................................................................................................................. 177
Caveats and future directions ................................................................. 181
Coexpression of metabolic genes functioning together- an evolutionarily
conserved principle .............................................................................. 181
Caveats of this principle and future directions ........................................ 181
Metabolic wiring and rewiring: the role of transcriptional regulation ....... 183
Transcriptional and post-transcriptional regulation ............................... 185
Future directions to segregate transcriptional and post-transcriptional
regulation .............................................................................................. 186
Coflux fine-tunes a simple guilt-by-association approach ....................... 189
Regulation of metabolic pathways influenced by control point(s) .......... 190
Future direction to determine control points of metabolic pathways ..... 191

Part II: Systematically deciphering the “missing” knowledge in C.
elegans metabolism ............................................................................... 192
Most C. elegans genes are associated with metabolism ....................... 193
Deriving gene-pathway associations: A scoring system approach ........ 194
Many C. elegans genes are associated with peroxisomal FA degradation.. 195
Ambiguous associations between paralogous genes and metabolic reactions
.............................................................................................................. 197
Future directions to resolve OR genes ................................................. 199

Proposed OR gene resolution algorithm ............................................. 200
Not all TFs are coexpressed with their regulated genes ....................... 202
Future directions to improve TF-metabolic pathway associations .......... 202

Experimental validation of predicted gene-metabolic pathway
associations ............................................................................................ 204

Future computational approaches to further explore the systems-level
view of metabolism and gene regulation ............................................. 206
Integration of regulatory information with metabolic network model-
Regulatory FBA (rFBA) ........................................................................ 206
Limitations of rFBA ............................................................................. 207
Machine-learning approaches to detect novel metabolic pathways ........ 208

Applicability to other organisms, including humans ......................... 210

BIBLIOGRAPHY .................................................................................. 212
LIST OF FIGURES

Figure 1.1: Schematic diagram to show metabolism, maintenance of homeostasis and consequence of failure to maintain homeostasis in the cell. .......................... 3

Figure 1.2: Range of mechanisms employed by cell to regulate metabolism that modulate metabolic flux over different timescales. ........................................ 10

Figure 1.3: Interaction between transcription and metabolism .................................. 17

Figure 1.4: Metabolic rewiring- Transcriptional rerouting that occurs in response to a changed condition. ............................................................... 21

Figure 1.5: Activation of the propionate degradation shunt pathway is controlled by a persistence detector. ................................................................. 34

Figure 1.6: Schematic diagram of B12-mechanism II .............................................. 36

Figure 1.7: Diagram showing the major objective of study: ‘What is the overall extent of transcriptional regulation of metabolism in C. elegans?’ .................. 42

Figure 2.1: Transcriptional Regulation of Metabolism during Development........... 50

Figure 2.2: Transcriptional Regulation of Metabolic Genes across Different Tissues ................................................................. 53

Figure 2.3: Transcriptional Regulation of Metabolism Across Development using alternative criterion ............................................................... 55

Figure 2.4: Analysis of Metabolic Gene Expression during Development, in Different Tissues and in a Gene Expression Compendium ................................. 56

Figure 2.5: Pathway and Phenotypic Enrichment Analysis of Metabolic Genes Based on Expression Variation ........................................................... 59

Figure 2.6: Transcriptional Regulation of Metabolic Genes across a Compendium ....................................................................................... 61

Figure 2.7: Supervised approach to investigate coexpression of metabolic pathways ...................................................................................... 66

Figure 2.8: Reaction-level analysis of metabolic pathways ...................................... 68

Figure 2.9: A semi-supervised approach to extract coexpressed and flux-dependent metabolic genes ........................................................................ 71

Figure 2.10: Mean Silhouette Score (MSS) to Evaluate Cluster Quality of Stringent and Relaxed Clusters ............................................................... 73

Figure 2.11: A Semi-supervised approach defines metabolic pathway boundaries ..................................................................................... 76
Figure 2.12: Semi-supervised clustering of product matrix using relaxed parameters................................................................. 79
Figure 2.13: Extraction of metabolic communities ................................................................. 82
Figure 2.14: Condition analysis of metabolic gene coexpression................................. 85
Figure 2.15: WormClust: A web application that enables querying of genes to identify coexpression with metabolic (sub)-pathways................................. 90
Figure 2.16: Enrichment of TFs to (Sub-)Pathways ................................................................. 92
Figure 3.1: Possible “missing” knowledge to C. elegans metabolism: .................. 109
Figure 3.2: Pie chart depicting percentage of C. elegans genes with different levels of UniProt Annotation Scores(left); Bar graph showing percentage of different categories of genes with different annotation scores (right). ...................... 115
Figure 3.3: Pie charts depicting percentage of iCEL1314 genes, orphan genes, RBPs, TFs, transporters, rest of non-metabolic genes and all C. elegans genes significantly associated (FDR<0.05, NES>=2) with metabolic pathways across time, space and conditions. The green part depicts the significant association, whereas the red portion indicates non-significant association.................. 119
Figure 3.4: Significant enrichment of C. elegans orphan genes, RBPs, TFs and transporters with metabolic pathways.................................................................................................................................................. 122
Figure 3.5: Significant enrichment of orphan genes, RBPs, TFs and transporters to metabolic communities 1, 2 and 3 (Figure 2.13A)................................................................. 123
Figure 3.6: Heatmaps showing clustering of transporter genes significantly enriched (FDR ≤ 0.05 and NES ≥ 2) to the metabolic pathways, based on score, before and after filtering for score>1 criteria................................................................. 125
Figure 3.7: Transporters significantly associated to metabolic pathways .......... 131
Figure 3.8: Network graph showing significant orphan gene-pathway associations with an score of more than 1........................................................................................................... 134
Figure 3.9: Clustered heatmap showing the score of orphan genes based on their significant enrichment to coexpression, to metabolic pathways. Scores less than 2 have been masked........................................................................................................ 135
Figure 3.10: Orphan genes significantly associated with metabolic pathways .... 139
Figure 3.11: Significant TF-Pathway associations ................................................................. 150
Figure 3.12: Significant NHR-metabolic pathway associations with a score of 2 or more................................................................................................................ 152
Figure 3.13: Significant associations of non-NHR TFs with metabolic pathways ................................................................................................................................. 156
Figure 3.14: Clustered heatmap of RBP associations with metabolic pathways based on FDR, for all scores

Figure 3.15: Network graph showing significant RBP-metabolic pathway associations with a score of more than 1, while excluding ribosomal proteins.

Figure 3.16: Significant associations of RBPs with growth

Figure 3.17: Significant associations of RBPs with energy

Figure 3.18: Significant associations of RBPs with pathways other than energy or growth-related

Figure 4.1: Summary of systems-level study of transcriptional regulation of metabolism

LIST OF TABLES

Table 2.1: Tools Table

Table 3.1: High-scoring transporters significantly enriched to metabolic pathways with score of 3 or 4

Table 3.2: High-scoring orphan metabolic genes significantly associated with metabolic pathways (Score of 3 or 4 only)

Table 3.3: High-scoring TFs significantly enriched to metabolic pathways with a score of 3 or 4

Table 3.4: High-scoring RBPs significantly associated with metabolic pathways with a score of 3 or 4
LIST OF FILES

File 2.1: Description of Metabolic genes distinguished into classes; A, B, C and D

File 2.2: Variations scores (VS) of metabolic and non-metabolic genes across development

File 2.3: Coefficient of variation (CV) of metabolic and non-metabolic genes across tissues

File 2.4: Metabolic genes divided into quadrants based on development VS and tissue CV values

File 2.5: List of datasets used in gene expression compendium

File 2.6: CV of metabolic genes across 177 datasets

File 2.7: Self-enrichment of coexpression on the WormPaths pathways

File 2.8: iCEL1314 model, including the pre-defined redundant reaction list used for flux calculation

File 2.9: Stringent clusters (minClusterSize=3, deepSplit=2) selected on the basis of mean silhouette scores (MSS)

File 2.10: Relaxed clusters (minClusterSize=6, deepSplit=3 selected on the basis of mean silhouette scores (MSS)

File 2.11: Communities B, C, and D based on product matrix clustered heatmap

File 3.1: Scores, ranging from 1 to 4 of transporters significantly enriched to metabolic pathways

File 3.2: Scores, ranging from 1 to 4 of orphan genes significantly enriched to metabolic pathways

File 3.3: Scores, ranging from 1 to 4 of TFs significantly enriched to metabolic pathways

File 3.4: Scores, ranging from 1 to 4 of RBPs significantly enriched to metabolic pathways


CHAPTER I : INTRODUCTION

Preface

A part of this introductory chapter has been adapted from a perspective written by Gabrielle Giese, me (Shivani Nanda), Amy Holdorf and Marian Walhout titled “Transcriptional regulation of metabolic flux: A Caenorhabditis elegans perspective”, published in the journal Current Opinion in Systems Biology in 2019.

Introduction to metabolism

Metabolism constitutes all biochemical processes in living organisms that produce energy for vital processes and for synthesizing new organic material. These processes allow organisms to grow, reproduce, maintain their structures, and respond to their environments. Metabolism can be broadly divided into two categories: anabolism and catabolism (Figure 1.1). Anabolism is the set of metabolic processes that involve the synthesis of complex molecules from simpler ones. These processes typically require energy provided by energy-rich molecules such as ATP (adenosine triphosphate). They are responsible for growth and cellular repair, as well as the production of essential biomolecules, such as proteins, nucleic acids, and lipids. Catabolism, on the other hand, is the set of metabolic processes that break down complex molecules into simpler ones,
releasing energy in the process. This energy can be used to fuel anabolic reactions or other cellular processes. Examples of catabolic processes include the breakdown of carbohydrates, lipids, and proteins for energy production. It takes the perfect coordination of these processes to maintain metabolic homeostasis in living organisms.

**Metabolic homeostasis**

Metabolic homeostasis refers to the dynamic balance that the body maintains to keep its internal environment steady and constant, even during fluctuating dietary and environmental conditions. This equilibrium involves many biochemical reactions that occur simultaneously and continuously in the body’s cells, aimed at maintaining vital physiological processes. For example, after a meal, the body breaks down the consumed carbohydrates into glucose. This increase in blood glucose triggers the pancreas to release insulin, a hormone that signals cells to take in glucose and use it for energy or store it for later use. This process helps maintain blood glucose levels within a normal range, thus illustrating metabolic homeostasis.

In humans, failure to maintain homeostasis can lead to a variety of metabolic disorders such as diabetes, inborn errors in human metabolism, obesity, and hypertension\(^1,2\) (Figure 1.1). The metabolic homeostasis of glucose is disrupted in both type 1 and type 2 diabetes, leading to high blood sugar levels\(^3\text{–}^7\).
Altered metabolism also plays a key role in cancer where the metabolic network is rewired to increase cell proliferation and angiogenesis\textsuperscript{8,9}.

*Figure 1.1:* Schematic diagram to show metabolism, maintenance of homeostasis and consequence of failure to maintain homeostasis in the cell.

Through careful regulation, metabolic pathways maintain energy and nutrient balance in the cell, and by extension, the whole organism. The interconnections between these pathways, form a network that allows the flow and transformation of energy and matter within the cell, collectively referred to as the metabolic network. Each metabolite can be a part of multiple pathways, and changes in one pathway can influence others, making metabolic network a complex and dynamic system.

Understanding the metabolic network and its regulation is crucial for maintaining homeostasis. A metabolic network can be computationally simulated...
in the form of metabolic network model, for providing a systematic framework to understand and analyze the flow of metabolites through the biochemical reactions in a cell.

**Metabolic network model**

A metabolic network model is the mathematical representations of the metabolic reactions and pathways within an organism. Metabolic network models are widely used in systems biology, biotechnology, and medicine for various purposes, such as predicting phenotypes, identifying potential drug targets, optimizing metabolic engineering strategies and the investigation of the underlying molecular mechanisms of various biological processes. There are different types of metabolic network models, including genome-scale metabolic models (GEM), and kinetic models.

A GEM links metabolites as nodes via enzymatic reactions as edges. The chemical reactions in this network have been elucidated one reaction at a time, through years of biochemical and genetic studies. Subsequently, specific enzymes have been linked to each metabolic reaction, and the corresponding genes for these enzymes have been identified. However, the experimental connections of genes and proteins to reactions have been done for only a limited number of model organisms, and such connections are extrapolated to other organisms by homology-based computational annotations to generate what is
referred to as ‘genome-scale network reconstructions’\textsuperscript{20–22}. The first GEM was created for \textit{Haemophilus influenza} and appeared shortly after its genome was first sequenced\textsuperscript{23}, and GEMs have now grown to the level where they enable predictive biology\textsuperscript{24,25}. There are established protocols for genome-scale network reconstructions\textsuperscript{26,27}, which can also be partially automated\textsuperscript{28,29}.

\textbf{Genome-scale network reconstruction}

Genome-scale network reconstruction is the process of systematically assembling a comprehensive representation of the metabolic, signaling, or gene regulatory network of an organism based on its genome sequence, annotation, and available experimental data. The most common type of genome-scale network reconstruction is for GEMs. The reconstruction process involves the following steps:

- Genome annotation: The starting point for network reconstruction is a well-annotated genome, that involves identifying genes, their functions, and the corresponding enzymes or other proteins they encode.
- Collection of biochemical reactions: Biochemical reactions catalyzed by the enzymes encoded in the genome are collected from databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG)\textsuperscript{30}, MetaCyc\textsuperscript{31,32}, or BiGG\textsuperscript{33}, and from the literature.
• Determination of gene-protein-reaction (GPR) associations: Each biochemical reaction is associated with the corresponding enzyme(s) and the gene(s) encoding the enzyme(s).

• Metabolite information: The metabolites participating in the biochemical reactions are identified and annotated with their chemical structures and properties.

• Biomass composition: A biomass equation representing the cellular composition (e.g., proteins, lipids, nucleic acids, and other cellular components) is included in the model to simulate growth and growth-associated energy requirements.

• Transport and exchange reactions: Reactions representing the transport of metabolites across cellular compartments (e.g., cytoplasm, mitochondria) and the exchange of metabolites with the extracellular environment are included in the model.

• Model validation: The reconstructed network is tested for its ability to predict growth phenotypes, gene essentiality, or metabolic capabilities under various environmental conditions and genetic perturbations. The model is iteratively refined based on the agreement between the predicted and experimental phenotypes.

• Model curation: The model is curated to ensure its quality and consistency, including the removal of dead-end metabolites, gap-filling of missing
reactions, and reconciliation of discrepancies in gene annotations or reaction stoichiometry.

The resulting GEM can be used for flux balance analysis (FBA), which is an efficient constraint-based modeling approach. Genome-scale network reconstructions have been developed for a wide range of organisms, including bacteria, yeast, plants, and animals (including humans), and have found numerous applications in systems biology, biotechnology, and medicine.

Flux balance analysis (FBA)

FBA is the computational method used to derive information from GEMs. FBA calculates the flux, or rates of the reactions, and hence the flow of metabolites through these metabolic networks. The objective of FBA is to predict the optimal flux distribution within a metabolic network, usually to maximize the growth rate or production of a specific metabolite. Here are the different steps involved in FBA:

- Reconstruction of the metabolic network: Described in previous section.
- Building the stoichiometric matrix: The stoichiometric matrix (S) represents the stoichiometry of the metabolic network. Each row in the matrix corresponds to a metabolite and each column corresponds to a reaction. The entries in the matrix represent the stoichiometric coefficients of the metabolites in the reactions.
• Applying mass balance constraints: The next step involves applying mass balance constraints to the system. This is based on the assumption that the system is in a steady state, meaning that the concentration of each metabolite does not change over time. In terms of the stoichiometric matrix, this assumption is represented as $Sv = 0$, where $v$ is the vector of reaction fluxes.

• Capacity constraints definition: Constraints defined for each reaction in the network usually include lower and upper bounds on the fluxes, which represent the minimum and maximum rates at which reactions can occur. These can be based on experimental data or literature values.

• Objective function definition: An objective function is defined to represent the biological goal of the cell, such as maximizing biomass or energy production, or the production of a specific metabolite. The objective function is a linear combination of the fluxes of certain reactions, typically involving only one reaction.

• Linear programming problem (LPP) formulation: The mass balance equations with the stoichiometric matrix, constraints, and objective function are combined to formulate a linear programming problem.

• Solving the LPP: The LPP is solved using computational algorithms, such as the simplex algorithm or interior-point methods. The solution provides
the optimal flux distribution in the metabolic network that satisfies the constraints and maximizes (or minimizes) the objective function.

- Analysis and interpretation of results: The optimal flux distributions with different constraints and objectives, are analyzed to gain insights into the metabolic capabilities of the organism under different conditions.

**Limitations of FBA**

FBA is a powerful tool for understanding cellular metabolism, but it has limitations\textsuperscript{15,34}. The method assumes a metabolic steady state, where the concentration of metabolites inside the cell does not change over time. However, this is not always the case, particularly under dynamic environmental conditions or during different phases of cell growth. FBA does not consider kinetic parameters, which are often critical for understanding metabolic behavior. Thus, it cannot predict the dynamic response of the system to changes over time. While FBA generally respects the directionality of reactions based on thermodynamics, it does not fully integrate thermodynamic constraints, which can lead to predictions that are thermodynamically infeasible. Despite these limitations, FBA is a valuable tool for analyzing metabolic networks due to its simplicity, scalability to large networks, and ability to make useful predictions without detailed kinetic information. Consequently, the approach has proven to be valuable for studying metabolism and guiding metabolic engineering efforts\textsuperscript{15}.
Regulation of metabolism

Organisms constantly face fluctuating environmental and dietary conditions, necessitating the regulation of metabolism to maintain homeostasis. Cells employ a diverse set of mechanisms for this metabolic regulation, encompassing transcriptional, post-transcriptional, and allosteric regulations, as well as post-translational modifications. These different mechanisms allow the modulation of metabolic flux over various timescales to meet the diverse and dynamic metabolic needs of a cell, which span from immediate requirements in order of seconds to those that persist over for extended periods (Figure 1.2).

Figure 1.2: Range of mechanisms employed by cell to regulate metabolism that modulate metabolic flux over different timescales.
Allosteric regulation of metabolism

Allostery has long been appreciated as a major regulatory mechanism of metabolism where metabolic enzyme activity is modulated by physical interactions with metabolites. It is a fast-acting mechanism where metabolites directly modulate enzyme activity by causing conformational changes in the enzyme (Figures 1.2 and 1.3). For instance, the enzyme phosphofructokinase (PFK) is a central enzyme in the glycolytic pathway, a metabolic process where glucose is broken down to produce energy in the form of ATP. Specifically, PFK catalyzes the conversion of fructose 6-phosphate to fructose 1,6-biphosphate, a key regulatory step in glycolysis. The activity of PFK is subject to intricate allosteric regulation to ensure the optimal balance of energy production in the cell. Under conditions where the cell's energy supply, represented by ATP levels, is high, ATP molecules bind to an allosteric site on the PFK enzyme separate from its active site. This binding induces a conformational change in PFK that decreases its affinity for fructose 6-phosphate, effectively reducing its enzymatic activity. This inhibition acts as a feedback mechanism, slowing down glycolysis when the cell's energy needs are satisfied, thereby preventing excessive production of ATP. However, this inhibitory effect is not permanent. When the ATP levels in the cell drop and it is converted to AMP, a marker of low energy, the situation changes. AMP competes with ATP for the same allosteric site on PFK but has a different effect: it increases PFK's affinity for fructose 6-phosphate, thereby enhancing the enzyme's
activity. This effectively speeds up glycolysis, allowing more glucose to be broken down and more ATP to be produced to meet the cell’s increased energy demand\textsuperscript{41,42}. Likewise, fructose-1,6-biphosphatase (FBPase-1), an enzyme participating in gluconeogenesis, mediates the reverse reaction of PFK, converting fructose-1,6-biphosphate back to fructose-6-phosphate. Its activity is also governed allosterically: AMP (indicating low energy levels) suppresses its function, while citrate (indicating high energy levels) amplifies it\textsuperscript{43}.

During TCA cycle, isocitrate dehydrogenase, the enzyme that catalyzes the oxidative decarboxylation of isocitrate to alpha-ketoglutarate is allosterically activated by ADP and inhibited by ATP and NADH\textsuperscript{44–46}. The story is similar in glycogen metabolism. Here, glycogen phosphorylase, regulated allosterically, is spurred into action by AMP, and suppressed by ATP and glucose-6-phosphate (high energy levels)\textsuperscript{47–49}. Conversely, glycogen synthase, another player in glycogen metabolism, gets allosterically activated by glucose-6-phosphate and restrained by AMP\textsuperscript{50}. These allosteric mechanisms ensure metabolic homeostasis, by meeting the cell's immediate needs during fluctuating conditions.

**Post-translational modifications in metabolism**

Like allosteric regulation, post-translational modifications regulate metabolic enzymes by altering their activity, stability, or interaction with other proteins. They also allow cells to rapidly adjust metabolic fluxes in response to changing cellular
conditions, nutrient availability and signaling events. Some common types of post-translational modifications include phosphorylation, acetylation, and ubiquitination. For instance, consider glycogenolysis, where glycogen phosphorylase, the enzyme that breaks down glycogen into glucose-1-phosphate, undergoes phosphorylation. This modification, facilitated by phosphorylase kinase (which in turn is activated by protein kinase A), enhances the enzyme's activity, thus promoting glycogen breakdown. Whereas, in glycogenesis, phosphorylation of glycogen synthase, the enzyme that catalyzes synthesis of glycogen from glucose-6-phosphate, inhibits its activity, reducing glycogen synthesis\textsuperscript{51–53}.

The role of acetylation in metabolic enzyme regulation is also significant. For example, the acetylation of a particular lysine residue effectively deactivates acetyl-CoA synthetase, an enzyme involved in the formation of acetyl-CoA from acetate and CoA. However, its activity can be reinstated via deacetylation by enzymes known as sirtuins\textsuperscript{54–56}. PTMs also have a substantial role in cholesterol synthesis, particularly through the regulation of HMG-CoA reductase via ubiquitination. High cellular cholesterol levels trigger the ubiquitination of this enzyme, marking it for degradation by the proteasome and consequently reducing cholesterol synthesis\textsuperscript{57}. These instances show diverse ways in which PTMs regulate metabolic enzymes.
Post-transcriptional regulation of metabolism

Post-transcriptional regulation of metabolism refers to the control of metabolic processes that occurs after the transcription of genes into messenger RNA (mRNA). It involves acting on mRNA molecules and their translation into proteins, which are responsible for carrying out metabolic reactions. Several mechanisms contribute to post-transcriptional regulation of metabolism, including mRNA stability and decay, alternative splicing, and translation regulation. Post-transcriptional regulation by microRNAs (miRNA) and RNA-binding proteins (RBPs) is important for many cellular functions.

Role of RNA binding proteins (RBPs) in metabolism

RBPs are proteins that specifically interact with RNA molecules through one or more RNA-binding domains. They modulate the stability, processing, localization, and translation of RNA molecules. RBPs can either stabilize or destabilize RNA molecules, recruit other factors involved in RNA metabolism, or directly modulate the translation of target mRNAs. They also play more diverse roles in RNA metabolism, including splicing, stability and decay, transport and localization, translation, as well as editing and modification.

RBPs play a significant role in the regulation of metabolism. For instance, iron regulatory proteins - IRP1 and IRP2 - exemplify essential RBPs that govern cellular iron homeostasis. They bind to iron-responsive elements (IREs) within the
untranslated regions (UTRs) of target mRNAs, including those encoding ferritin, the iron storage protein, and transferrin receptor 1, responsible for iron uptake. Depending on cellular iron concentrations, IRPs either hinder ferritin mRNA's translation or stabilize the transferrin receptor 1 mRNA, ensuring a finely-tuned coordination of iron uptake and storage\textsuperscript{58–60}. Consider also human antigen R, another RBP, which is involved in glucose and lipid metabolism regulation. It influences these metabolic processes by modulating the stability of mRNAs encoding key proteins, such as the GLUT1 (glucose transporter 1) and LPL (lipoprotein lipase)\textsuperscript{61,62}. The Fragile X Mental Retardation Protein (FMRP), an RBP primarily known for its role in neurons, has also been linked to the regulation of glucose and lipid metabolism. Its disruption can impact metabolic regulation, leading to neurodevelopmental disorders like Fragile X syndrome\textsuperscript{63,64}. Cytoplasmic Polyadenylation Element Binding (CPEB) proteins, a family of RBPs, regulate mRNA translation impacting several metabolic pathways, including insulin signaling and lipid metabolism. CPEB proteins can either promote or inhibit the translation of target mRNAs, depending on the specific CPEB family member and the target mRNA involved\textsuperscript{65,66}. Taken together, these examples illustrate the many ways RBPs can exert influence over metabolic processes. They do this by regulating the stability, processing, localization, and translation of mRNA molecules that encode key metabolic enzymes and transporters, underscoring their integral role in metabolic homeostasis.
Transcriptional regulation of metabolism

Metabolism can also be regulated transcriptionally by changing the transcription (activating or repressing the expression) of genes encoding metabolic enzymes or transporters (Figures 1.2 and 1.3). Unlike allostery and post-translational modifications that provide rapid, direct influences on enzyme activity and stability, transcriptional regulation modifies metabolism over extended periods by modulating cellular enzyme concentrations. This form of regulation allows the organism to adapt to changing cellular or environmental conditions.

Transcriptional regulation of metabolism is influenced by several factors including transcription factors (TFs), co-activators and co-repressors, and epigenetic modifications. Numerous other proteins, including various signaling molecules, chromatin remodelers, and components of the basal transcriptional machinery, also contribute to the transcriptional regulation of metabolism (Figure 1.3). Their actions allow for the precise control of gene expression in response to a wide array of signals and conditions.

Recent analyses have revealed that large fractions of the transcriptome respond to changes in metabolism\textsuperscript{67,68}. For instance, many metabolic genes change in expression in a circadian manner to enable appropriate metabolism in our 24-hour cycle \textsuperscript{69,70}. While allostery is mechanistically relatively simple, the transcriptional control of metabolism is further convoluted by the requirement of elaborate multi-compartment cooperation because metabolism and transcription
mostly occur in different cellular compartments. Indeed, a majority of metabolic enzymes are encoded by nuclear genes, while most metabolic reactions occur outside the nucleus or in the cytosol or mitochondria or other organelles. Therefore, a change in metabolism has to be relayed to the relevant TFs, resulting in TF relocalization and/or activation.

Figure 1.3: Interaction between transcription and metabolism
Nutrients are taken in from the environment and converted into energy or biomass. Regulating metabolic flux is essential to maintain metabolite pools in balance. This can be accomplished allosterically or transcriptionally. Metabolites can bind to TFs, which go on to regulate target genes. In addition, metabolism also produces cofactors of chromatin-modifying enzymes that regulate chromatin state and, therefore, transcription.
**Transcriptional initiation, elongation, and termination**

Transcription, the process of converting DNA to RNA, is the first step in gene expression. It is primarily driven by RNA polymerase II (Pol II), a key eukaryotic enzyme, but Pol II cannot bind to DNA independently. Instead, other proteins known as TFs recognize specific DNA sequences and bring Pol II to them. This regulated binding is what differentiates expression across cells and life stages.

The transcription process initiates when Pol II, aided by general transcription factors (GTFs), attaches to target DNA at the core promoter, a small DNA section. Although no sequence motif is universal to all promoters, common elements like the TATA box motif exist. Sequence-specific TFs and GTFs contribute to spatiotemporal specificity by adjusting expression under appropriate conditions.

After binding, Pol II proceeds to elongation, the transcription phase where it transcribes the entire gene. Pol II can, however, pause or stall shortly after the promoter. Elongation factors like CDK-9/cyclin T and SPT-4/SPT-5 help Pol II continue along the gene. SPT-4 and SPT-5's actions depend on Pol II C-terminal domain phosphorylation. Other proteins such as the Integrator complex also help Pol II.

Finally, Pol II completes transcription by dissociating from the DNA strand, a process involving cleavage and polyadenylation factors that recognize sequence motifs within the mRNA's 3' UTR. Mis-transcribed RNAs are targeted for
degradation at this stage, underscoring the importance of termination in gene regulation. The transcription rate from DNA to RNA is controlled by various factors, including TFs, highlighting the complexity and precision of gene expression regulation.

Role of transcription factors (TFs) in regulating metabolism

TFs are proteins that play a critical role in the regulation of gene expression. They function as activators or repressors, thereby turning genes on or off, and their activity can be regulated by different signals. The transcriptional control of metabolic gene expression through the action of specific TFs is prevalent. They bind to specific DNA sequences, either promoting or inhibiting the transcription of target genes into mRNA, which is subsequently translated into proteins. In the context of metabolism, TFs help to maintain cellular homeostasis by modulating the expression of genes involved in various metabolic pathways. A well-known example of this process in mammals is the sterol regulatory element-binding protein-2 (SREBP-2), a TF that functions to maintain cholesterol levels and is vital to lipid and membrane homeostasis. SREBP-2 is integrated within the endoplasmic reticulum membrane. When cholesterol levels are low, SREBP-2 is released from the ER and translocates to the nucleus where it activates the transcription of cholesterol biosynthesis genes. The Hypoxia-Inducible Factor-1 (HIF-1) is another notable TF involved in metabolic regulation. HIF-1 is sensitive
to oxygen levels, and, in hypoxic conditions, it regulates the expression of genes involved in glucose metabolism, angiogenesis, and cell survival mechanisms. Its role is paramount in enabling cellular adaptation to hypoxic conditions, ensuring cell survival and functionality under stress\textsuperscript{72–75}. In summary, various TFs regulate a complex network of metabolic processes, ensuring equilibrium within the cell and highlighting their fundamental role in cellular dynamics. Despite extensive studies, our knowledge of TFs and their role in transcriptional regulation of various metabolic pathways remains an area of active investigation.

**Metabolic rewiring**

Transcriptional regulation of metabolism can cause metabolic rewiring, where the metabolic pathways are rerouted to alternative pathways to satisfy changing metabolic needs (Figure 1.4). This can be achieved by either downregulation of genes involved in canonical pathway or by upregulation of the alternative pathway, or both. A well-studied example of this transcriptional rerouting is seen in propionate metabolism in *C. elegans*\textsuperscript{76}. Such processes are usually regulated by TFs, such as by *nhr-10* and *nhr-68* in the case of propionate degradation. It is crucial to recognize that such regulatory mechanisms, while pivotal in maintaining homeostasis, are not devoid of associated costs.
Figure 1.4: Metabolic rewiring- Transcriptional rerouting that occurs in response to a changed condition.
For instance, a pathway that is catalyzed by Gene 1, goes from metabolite A to B in the presence of a certain cofactor X. However, in the absence of the cofactor X, an alternative pathway that goes from A to C, catalyzed by Gene 2, is favored to enable consumption of metabolite A.

Cost versus benefits

While transcriptional regulation is important in sustaining cellular and systemic homeostasis, it also entails costs to the organism as it is an energy-demanding process along with intricate temporal dynamics. The process necessitates substantial energy consumption, encompassing the synthesis and transportation of TFs, DNA unwinding and rewinding, mRNA synthesis, processing, nuclear export, and eventual protein translation, all of which require ATP. Additionally, the
non-instantaneous nature of transcriptional regulation may present drawbacks in instances where the organism requires rapid environmental response. However, the benefits conferred by transcriptional regulation of metabolism are many. This regulatory mechanism facilitates gene expression selectivity in response to varied conditions, thus enabling cells to adapt metabolic activity according to current needs and environmental shifts. Transcriptional regulation of metabolic genes aids in producing only the currently necessary enzymes, thereby mitigating resource wastage associated with unnecessary protein synthesis. Moreover, it allows metabolic pathway coordination, enabling simultaneous upregulation of DNA precursor synthesis and energy production during cell growth and division.

Empirical evidence of these benefits can be observed in unicellular organisms like *Escherichia coli* (*E. coli*) and yeast, where an individual's fitness can be gauged by its growth rate in culture\textsuperscript{77–79}. Numerous studies have explored the influence of metabolic pathway regulation on microbial growth rate, under steady state and local environmental changes. Typically, in such experimental setups, *E. coli* is cultured in a chemostat with one or two perturbed carbon sources over a given period.

Across evolutionary timescales, the fitness benefits derived from regulation must counterbalance the costs of maintaining the regulatory system which includes the synthesis of TFs and DNA replication. However, the cost of suboptimal gene
expression regulation, typified by excessive enzyme production compared to TFs, may significantly exceed the cost of maintaining the regulatory system. An additional entropic cost stems from random mutations that potentially disrupt TFs and binding sites. As postulated by Savageau, to maintain functioning regulation, a genotype population must counterbalance mutation-induced losses with a growth rate higher than mutants with defective regulation\textsuperscript{80,81}.

A case in point: the lac operon of \textit{E. coli}, a robust system for detecting and metabolizing lactose. The unnecessary expression of lac genes has been observed to cause a growth rate reduction of up to 5\%. However, lactose utilization boosts growth rate by around 10-15\%\textsuperscript{79,82}. Given this cost-benefit asymmetry, regulation significantly influences long-term growth rate if the resource is intermittently available. Fundamentally, gene regulation's utility is more obvious when environmental conditions fluctuate over time. Empirical evidence suggests a repressive mode, disabling DNA repressor binding in resource presence, is preferable when expression demand is infrequent, while an activating mode is preferred otherwise\textsuperscript{80,81}. For the lac system, entropic effects are significant, permitting only a repressive regulatory mode. Evidence demonstrates even the most rudimentary transcriptional network can substantially augment an organism's fitness; an effect sustained under narrow nutrient level variations\textsuperscript{83}.

In summary, despite energy and timing-related costs, transcriptional regulation of metabolism is overwhelmingly advantageous for an organism.
Studies illustrating transcriptional regulation of metabolic processes

There have been many studies displaying evidence of transcriptional regulation of specific metabolic processes in different organisms. This section goes through these different studies in unicellular and multicellular organisms.

Studies in unicellular organisms

Transcriptional regulation of metabolism in unicellular organisms is essential for their survival, growth, and adaptation to different environmental conditions. These organisms, such as bacteria and yeast, modulate their metabolic activities by altering the expression of specific genes in response to changes in nutrient availability, stress, or other environmental cues. For instance, intricate regulation is seen in the lac operon in *E. coli*. Comprising three contiguous genes - lacZ, lacY, and lacA involved in lactose metabolism, the operon’s expression is modulated by a single promoter and regulatory region. This regulation involves the lac repressor protein (LacI) and the inducer molecule allolactose. In the absence of lactose, the lac repressor binds to the operator sequence, inhibiting transcription. However, the presence of lactose facilitates allolactose binding to the lac repressor, causing a conformational change that prevents operator binding, thereby allowing transcription of the lac operon genes. Therefore, the lac operon
exemplifies an elegant feedback mechanism that allows the cell to conserve energy by producing metabolic enzymes only when the corresponding substrate in this case, lactose is available in the environment\textsuperscript{84,85}.

Another example of dynamic metabolic regulation is observed in many bacteria, including \textit{E. coli}, through a mechanism known as catabolite repression. This regulatory strategy allows bacteria to preferentially utilize glucose over other carbon sources. When glucose is plentiful, the intracellular concentration of cyclic AMP (cAMP) decreases, resulting in the dissociation of the CRP-cAMP complex and repression of genes involved in the utilization of alternative carbon sources. However, when glucose levels drop, the CRP-cAMP complex forms again, activating the transcription of these genes\textsuperscript{87–94}.

In the yeast \textit{Saccharomyces cerevisiae} (\textit{S. cerevisiae}), the GAL system is responsible for the transcriptional regulation of genes involved in galactose metabolism. The key regulators in this system are the transcriptional activator Gal4p and the repressor Gal80p. In the absence of galactose, Gal80p binds to Gal4p, inhibiting its transcriptional activation function. When galactose is present, it binds to another protein, Gal3p, which in turn interacts with Gal80p, releasing Gal4p to activate the transcription of GAL genes\textsuperscript{95,96}. Another prominent example in \textit{S. cerevisiae} is the nitrogen catabolite repression system. Yeast cells preferentially use ammonium and glutamine as nitrogen sources over other less-preferred nitrogen sources, such as proline or urea. This preference is regulated
by the nitrogen catabolite repression system, which involves the global transcription factors Gln3p and Gat1p. In the presence of preferred nitrogen sources, Gln3p and Gat1p are sequestered in the cytoplasm by the protein Ure2p. When preferred nitrogen sources are depleted, Gln3p and Gat1p are released from Ure2p, translocate to the nucleus, and activate the transcription of genes involved in the utilization of alternative nitrogen sources\textsuperscript{97–100}. 

Finally, it has also been shown that metabolic enzymes themselves can sometimes regulate transcription, for example, by binding to DNA. For instance, the yeast metabolic enzyme Arg5,6, which acts in the arginine biosynthesis pathway, has been shown to bind to a specific DNA sequence and affect transcription\textsuperscript{101}. While studying transcriptional regulation in unicellular organisms offers several advantages, such as simplicity, ease of genetic manipulation, and rapid growth, it also comes with certain limitations. These limitations include differences in metabolic pathways, gene regulation mechanisms, environmental responses, and the lack of complex organ systems.

**Studies in multicellular organisms**

Multicellular organisms, have highly sophisticated regulatory mechanisms that control metabolic processes, given their intricate physiological requirements and vast tissue heterogeneity. In addition to SREBPs responsible for maintenance of cholesterol in mammals\textsuperscript{71}, and HIF-1’s response to hypoxia and its role in glucose
metabolism\textsuperscript{72–75}, there have also been other past studies of transcriptional regulation of metabolism in multicellular organisms. For instance, peroxisome proliferator-activated receptors (PPARs) are a family of nuclear receptor TFs that play critical roles in the regulation of lipid and glucose metabolism, as well as energy homeostasis in mammals. There are three PPAR isoforms, α, γ, and δ (or β), each with distinct tissue distribution and target genes. PPARs function as heterodimers with the retinoid X receptor (RXR) and bind to specific DNA response elements called peroxisome proliferator response elements (PPREs) in the promoters of target genes. The transcriptional activity of PPARs is regulated by binding to specific ligands, such as fatty acids or synthetic agonists, which modulate their conformation and interaction with coactivators or corepressors\textsuperscript{102–108}.

The TF cAMP Response Element-Binding (CREB) protein, too, exerts considerable influence on gene expression in response to a range of extracellular signals, such as hormones, growth factors, and neurotransmitters, in mammals like humans and mice. CREB binds to cAMP response elements (CREs) in the promoters of target genes, governing processes like gluconeogenesis and fatty acid oxidation\textsuperscript{109–117}.

Plants, also employ transcriptional regulation mechanisms. For instance, in Arabidopsis thaliana (A. thaliana), TFs such as bZIP (basic leucine zipper) and MYB (myeloblastosis) play crucial roles in the regulation of sugar metabolism and
signaling by responding to changes in cellular glucose levels and regulating the expression of associated genes\textsuperscript{118–126}. bZIP TFs also extend their reach to the regulation of amino acid metabolism. For example, AtbZIP11, an Arabidopsis bZIP TF, plays a role in regulating the expression of genes involved in amino acid biosynthesis and transport, in response to changes in amino acid levels\textsuperscript{127,128}. The second major family of TFs affecting plant metabolism, MYB TFs are characterized by a highly conserved DNA-binding domain, known as the MYB domain. They are involved in various aspects of plant growth, development, and stress responses, including the regulation of secondary metabolism, such as biosynthesis of flavonoids, anthocyanins, and lignin, which are important for plant defense, pigmentation, and structural support\textsuperscript{129–134}. Therefore, these individual studies have exemplified transcriptional regulation of metabolism in multicellular organisms.

In our study, we focus on \textit{Caenorhabditis elegans} as the model organism to study transcriptional regulation of metabolism in animals as it offers various advantages.

\textbf{Caenorhabditis elegans as a model organism}

The nematode \textit{C. elegans}, has emerged as a highly tractable multicellular model organism with many advantages making it a suitable model to study transcriptional regulation of metabolism.
Advantages

The hermaphrodite *C. elegans* has an invariant body plan of 959 somatic cells in which the lineage of each cell has been determined\(^{135}\). It is a bacterivore that can easily be maintained in the laboratory on different bacterial diets. Its small size, short life cycle, and transparent body make it highly tractable for large-scale genetic studies and phenotypic analysis. The *C. elegans* genome encodes \(~20,000\) proteins, many of which are conserved from nematodes to humans\(^{136}\). *C. elegans* can be fed double-stranded RNA–containing bacteria to induce RNA interference (RNAi), making large-scale RNA interference screens both easy and cost-effective\(^{137}\). Transcriptional and translational fluorescent protein reporters can be used to assess gene regulation and protein localization *in vivo*\(^{138}\). When combined with RNAi, such reporters can be used to gain insight into transcriptional regulation of promoter activity\(^{139}\) or the metabolic rewiring of gene expression\(^{140}\). Although bacteria and yeast are well established as excellent models for studying regulation of metabolism, as a multicellular organism, *C. elegans* provides the opportunity to study cell nonautonomous regulation and tissue-specific gene expression. *C. elegans* also has a high-quality GEM called iCEL1314 that enables metabolic network modeling\(^{18, 19}\).
iCEL1314 - a C. elegans metabolic network model

iCEL1314 is a GEM for *C. elegans* \(^{18,19}\). This model contains 1314 genes, 2230 reactions and 907 unique metabolites. This network can mathematically convert bacterial diet into worm biomass. It also enables predictions of gene essentiality and other predictions. This model can be integrated with different gene expression datasets to further improve upon mechanistic flux predictions\(^ {18,19,141}\). For instance, metabolic rewiring in dauer animals versus growing animals has been demonstrated by integration of gene expression data with the model\(^ {18}\). Further, in a different study, functional metabolic network models of seven major tissues were derived for which transcriptomes were generated by aggregating scRNA-seq data from thousands of individual cells that originated from the same tissue in a population of animals. This study recapitulates known tissue functions, reveal metabolic properties that are shared with similar tissues in human, and predict numerous novel metabolic functions\(^ {19,142}\). The annotation database of iCEL1314 (wormflux.umassmed.edu), and its predictive power shown by multiple validation tests make it a suitable metabolic model for *C. elegans*.

Given the tractability of the organism, the homology with human genes\(^ {143}\), the availability of genome-scale or specific RNAi collections, and a high-quality GEM, *C. elegans* makes an ideal model system to elucidate the relationships between metabolic and gene regulatory networks.
Examples of transcriptional regulation of metabolism in *C. elegans*

In *C. elegans*, transcriptional regulation of metabolism plays an essential role in controlling various aspects of development, reproduction, and stress responses. Two well-studied examples of transcriptional regulation of metabolism in *C. elegans* are influenced by the presence of cofactor vitamin B12 \(^{76,144,145}\). They are known as B12 mechanism I and II.

*Transcriptional rewiring of *C. elegans* metabolism: a propionate persistence detector (B12 mechanism I)*

Propionate is a short-chain fatty acid that is produced upon the breakdown of the branched-chain amino acids valine and isoleucine, the amino acids threonine and methionine, as well as odd-chain fatty acids and cholesterol. In both *C. elegans* and humans, propionate is toxic when it accumulates\(^{76,146}\). Similar to humans, *C. elegans* uses vitamin B12 as a cofactor in a canonical propionate breakdown pathway, and under B12-replete dietary conditions, propionate is converted to succinate, which anaplerotically can enter the TCA cycle. Remarkably, however, *C. elegans* not only is able to grow on bacteria that synthesize vitamin B12 but also can thrive on bacteria that do not \(^{147,148}\). Indeed, the vast majority of research on *C. elegans* has been conducted with animals fed with the standard laboratory diet of *E. coli* OP50, a strain that cannot synthesize vitamin B12. On bacterial diets low in vitamin B12, *C. elegans* transcriptionally
activates the expression of five metabolic genes that function together in an alternate propionate breakdown or shunt pathway\textsuperscript{76}. The first gene in this pathway is virtually off when vitamin B12 levels are sufficiently high. This indicates that the animal favors the canonical, vitamin B12–dependent propionate breakdown pathway over using the shunt. The propionate shunt may be less favorable because the first intermediate in the pathway is acrylyl-CoA, which after removal of the CoA is converted into highly toxic acrylate. This idea is substantiated by the observation that knockdown of the enzyme that consumes acrylyl-CoA, \textit{ech-6}, renders the animals very sick. Double perturbation of \textit{ech-6} with the enzyme that produces acrylyl-CoA, \textit{acdh-1}, however, rescues this phenotype\textsuperscript{76}.

Recently, it was found that a specific type of regulatory circuit known as a type I coherent feedforward loop with an AND logic gate is required for the activation of propionate shunt gene expression\textsuperscript{144}. This circuit consists of two nuclear hormone receptor (NHR) TFs: NHR-10 and NHR-68, both of which are required for propionate shunt activation. In addition, NHR-10 activates the expression of NHR-68, and NHR-68 autoactivates, which likely provides tunability to the system (\textbf{Figure 1.5}). Propionate shunt expression is activated with a delay of several hours and only when propionate supplementation is sustained. This is precisely what modeling of this type of circuit predicted more than 15 years ago\textsuperscript{149}. Here, the delay is caused by the time it takes to express sufficient levels of NHR-68, which is under the control of NHR-10, and because sufficient levels of both TFs
are required. Therefore, this circuit functions as a true persistence detector to rewire propionate breakdown metabolism in an animal only when propionate accumulates persistently. This propionate-dependent mechanism is referred to as “B-12 mechanism I”\textsuperscript{145}. However, the molecular mechanism by which propionate accumulation is relayed to NHR-10 and, perhaps, NHR-68 remains to be elucidated. Gene expression of metabolic genes are sometimes influenced by metabolic enzymes \textsuperscript{101}. While the binding to DNA and regulation of gene expression by metabolic enzymes has not yet been directly shown in \textit{C. elegans}, ECH-6, has been found to bind the \textit{nhr-68} promoter in yeast one-hybrid assays\textsuperscript{150}. Future studies are required to validate this finding and to determine whether it provides an additional layer of regulation to the propionate persistence detector.
Figure 1.5: Activation of the propionate degradation shunt pathway is controlled by a persistence detector.

The nuclear hormone receptor transcription factors NHR-10 and NHR-68 act in a type I feedforward loop with an AND logic gate that causes a delayed response of the propionate shunt pathway only when there is a persistent, not a transient, level of propionate.
**Transcriptional regulation of methionine/ S-adenosylmethionine cycle (B-12 mechanism II)**

The second major example of transcriptional regulation influenced by cofactor vitamin B12 has been observed in one-carbon metabolism\(^{145}\). One-carbon metabolism produces various building blocks for cellular growth and repair, including nucleotides and methionine/ S-adenosylmethionine (Met/SAM), the major methyl donor of the cell. Methionine synthase, an enzyme that requires vitamin B12 as cofactor, converts homocysteine into methionine in the Met/SAM cycle. In addition to Met/SAM cycle, one-carbon metabolism also constitutes folate metabolism and parts of purine and thymine biosynthesis. SAM is critical for synthesis of phosphatidylcholine, an important component of cellular membranes, as well as the methylation of DNA, RNA and histones\(^{151}\).

It was found that vitamin B12 represses the expression of Met/SAM cycle genes by a propionate-independent mechanism which is referred to as ‘B12-mechanism-II’ \(^{145}\). B-12 mechanism II is activated upon perturbation of the Met/SAM cycle, either genetically, or nutritionally due to low dietary vitamin B12. B12-mechanism-II requires \(nhr-114\) to activate the expression of the vitamin B12 transporter \(pmp-5\), and to adjust influx and efflux of the cycle by activating the methionine sulfoxide reductase \(msra-1\) and repressing the expression of cystathionine beta synthase \(cbs-1\), respectively (Figure 1.6). These findings
indicate the low Met/SAM cycle activity is sensed and transcriptionally adjusted to be maintained in a tightly controlled regime\textsuperscript{145}.

**Figure 1.6: Schematic diagram of B12-mechanism II**
Vitamin B12 modulates the transcription of Met/SAM cycle genes and controls in/efflux through the sensing of SAM by \textit{nhr-114}. Dashed arrows indicate regulation of metabolic activity. Solid arrows indicate transcriptional regulation. Vitamin B12 increases Met/SAM cycle metabolic activity producing SAM, which represses \textit{nhr-114} transcription thereby reducing the expression of Met/SAM cycle related genes.

Therefore, in \textit{C. elegans}, the two vitamin-B12 dependent pathways are sensed by two different transcriptional mechanisms through different NHRs. Both these mechanisms display regulation of metabolism to enable animals to adapt to changing dietary conditions.
Other studies of transcriptional regulation of metabolism in C. elegans

There are few other studies that demonstrate the transcriptional regulation of metabolism in *C. elegans*. DAF-16, a member of the FOXO family of TFs, plays a central role in regulating metabolism, longevity, and stress resistance in *C. elegans*. It is known to regulate target genes involved in metabolic pathways, such as glycolysis, gluconeogenesis, and fatty acid metabolism. Studies have shown that insulin/insulin-like growth factor signaling (IIS) modulates the activity of DAF-16, affecting the expression of genes involved in metabolism and stress response\textsuperscript{152}.

The best-studied examples of TFs are NHRs, which can regulate gene expression upon physically binding to ligands such as fatty acids\textsuperscript{153}. Notably, the *C. elegans* encodes more than 270 NHRs, whereas the human genome encodes 48 NHRs\textsuperscript{154,155}. Interestingly, NHRs often physically interact with the promoters of metabolic genes, including those involved in compound detoxification\textsuperscript{150,156}. The transcriptional and physiological function has been characterized for only a small subset of NHRs. For instance, NHR-49 has been shown to modulate the shift between fat consumption for energy and fat composition for storage by activating *acs-2* and *ech-1*, genes involved in mitochondrial beta oxidation, *fat-7*, a gene that encodes a fatty acid desaturase\textsuperscript{157}. Other examples include NHR-8, which regulates cholesterol homeostasis, bile acid production, and fat metabolism \textsuperscript{158}, and NHR-25, an important factor in *C. elegans* development and molting, which
responds to long-chain fatty acids produced by acs-3 and regulates fat uptake and storage\textsuperscript{159}. However, only two NHRs, DAF-12 and NHR-33/HIZR-1, have known ligands.

Under growth conducive conditions, insulin/transforming growth factor beta (TGF-b) pathways result in the production of the bile acid like dafachronic acid, which directly binds as a ligand to DAF-12\textsuperscript{160}. Once dafachronic acid is bound, DAF-12 activates growth programs and represses dauer formation. NHR-33/HIZR-1 directly binds zinc and accumulates in the nucleus where it activates zinc-detoxifying and zinc-transporting enzymes\textsuperscript{161}.

Metabolic network is also transcriptionally influenced via chromatin modifications. Examples of these include histone methylation, acetylation, and crotonylation\textsuperscript{162}. A recent study found that the metabolite S-adenosyl methionine is required for H3K4me3-mediated transcriptional responses to \textit{C. elegans} infection by the pathogenic bacterium \textit{Pseudomonas aeruginosa}\textsuperscript{163}. Several examples of metabolites and enzymatic cofactors regulating \textit{C. elegans} metabolic gene expression are known, including vitamin B12\textsuperscript{147}, the short-chain fatty acid propionate\textsuperscript{144}, zinc\textsuperscript{161}, and nitric acid\textsuperscript{164}.

\textbf{Knowledge gaps and thesis objectives}

The current understanding of transcriptional regulation of metabolism primarily stems from studies examining individual genes or distinct pathways, as discussed
before. However, acknowledging the immense interconnectivity, crosstalk and dynamic nature of biological pathways, it is critical to widen the scope of inquiry and study transcriptional regulation at a genomic scale.

While there are handful existing genome-scale studies, they largely concentrate on single-cell organisms like *E. coli* and *S. cerevisiae*, with a limited number in plant species. The extent of transcriptional regulation of the entire metabolic network remains largely unknown.

**Genome-scale studies in unicellular organisms, and plants**

The few genome-scale studies carried out in unicellular organisms have varied from one another, each conducted with unique objectives in mind. In yeast, the metabolic network model has been utilized to understand how and to what extent the modulation of enzyme expression determines functional metabolic units and how this affects the global properties of the metabolic network. Several recurring regulatory patterns were observed that may represent general design principles of metabolic gene regulation\(^\text{165}\). Similarly, another investigation in *S. cerevisiae* explored the correlation between metabolic network distance and mRNA expression profiles, demonstrating that different topological motifs in the metabolic network reveal dynamic design principles\(^\text{166}\). In *E. coli*, it was studied on a genome-scale how the transcriptional regulatory system controls small-molecule
metabolism, and rules that are generally applicable to bacterial metabolic system were established\textsuperscript{167}.

While studies in plant metabolism have largely centered on individual pathways\textsuperscript{168–173}, one particular study constructed a TF-promoter interaction network of primary and specialized metabolism in \textit{A. thaliana}\textsuperscript{174} providing insights into transcriptional coordination across multiple metabolic pathways influenced by developmental and stress conditions.

**Importance of the study in animals over unicellular organisms**

In contrast to \textit{S. cerevisiae} and \textit{E. coli}, the genomes of multicellular organisms encode many more genes including TFs and enzymes. In multicellular organisms, different tissues and organs communicate with each other to coordinate metabolic processes at an organismal level. For example, the liver, adipose tissue, skeletal muscle, and brain all communicate to regulate energy homeostasis. This kind of interorgan communication cannot be studied in unicellular organisms. The advent of multicellularity also allowed for the allocation of functions across different cell types, likely leading to more intricate transcriptional and metabolic regulation compared to single-cell organisms. That is why studying transcriptional regulation of metabolism in a highly tractable and suitable animal model such as \textit{C. elegans} is important.
Importance of “target-centered” study at a systems-level

Evidence exists for specific cases of transcriptional regulation of individual genes or pathways in animals, including *C. elegans*\(^{76,144,145,152,157–161,163,175,176}\). They are usually one of the two kinds of study: a ‘regulator-centered’ study that investigate the mechanisms of a regulator of interest, such as DAF-2\(^\text{177}\) and DAF-16\(^\text{177,178}\); or a ‘biological-process-centered’ study that investigate the metabolic regulations of a particular process such as starvation or infection\(^\text{179–181}\). However, the extent to which overall metabolic activity is under transcriptional control in animals remains largely unclear (Figure 1.7). Our research fills this gap by being the first ‘target-centered’ study to systematically investigate all metabolic genes and pathways as regulatory targets in an animal model.

Missing annotations in *C. elegans* metabolic network

Even though there is significant advancement in the sequencing of the *C. elegans* genome, a sizeable challenge remains in the lagging pace of functional annotation, particularly for the *C. elegans* metabolic genes. A substantial proportion of these metabolic genes, called orphan genes, remain poorly characterized in terms of their enzymatic activities or their associations with metabolic pathways. Furthermore, numerous transporters annotated in the *C. elegans* genome remain unlinked to any metabolic pathway, and many transporter reactions do not have
any annotated transporter. Many TFs and RBPs are yet to be functionally deciphered in terms of their role in regulating metabolic pathways. Consequently, there is an extensive knowledge gap in understanding *C. elegans* metabolism. Our study aims to fill this gap, focusing primarily on the investigation of orphan genes, TFs, transporters, and RBPs.

*Figure 1.7: Diagram showing the major objective of study: ‘What is the overall extent of transcriptional regulation of metabolism in *C. elegans*?’*
Thesis objectives

In view of the above, my thesis seeks to address the following objectives:

- To quantify the prevalence of transcriptional regulation of metabolism in *C. elegans*.
- To ascertain which known metabolic pathways exhibit coordinated expression (potentially co-regulated) in *C. elegans*.
- To find tightly co-regulated sub-pathways in metabolic pathway in an unbiased manner.
- To uncover the activation or repression conditions of metabolic (sub-)pathways.
- To annotate “missing” gaps in *C. elegans* metabolism.
Preface

This chapter is adapted from “Systems-level transcriptional regulation of Caenorhabditis elegans metabolism”, first-authored by me (Nanda et al., 2023), published in Molecular Systems Biology 141.

Background

All organisms regulate their metabolism during development and to maintain homeostasis under fluctuating dietary and environmental conditions. In humans, failure to maintain homeostasis can lead to a variety of metabolic disorders such as inborn errors in human metabolism, obesity, hypertension, and diabetes 182,183. Metabolism can be regulated through different mechanisms. One well-known mechanism is allostery, a fast-acting mechanism where metabolites directly modulate enzyme activity. For instance, the enzyme phosphofructokinase, which regulates the conversion of fructose 6-phosphate to fructose 1,6-biphosphate, is allosterically regulated during glycolysis. This reaction is coupled to ATP hydrolysis where ATP binding to phosphofructokinase inhibits enzyme activity by decreasing
its affinity for fructose 6-phosphate, while conversion to AMP reverses the inhibitory effect and increases the activity of the enzyme\cite{41,42}. Metabolism can also be regulated transcriptionally by activating or repressing the expression of genes encoding metabolic enzymes or transporters. This mechanism is relatively slow and allows the organism to adapt to changing cellular or environmental conditions.

Well-known examples of the transcriptional regulation of metabolism include induction of the lac operon in \textit{E. coli} in response to a switch from glucose to lactose as a carbon source\cite{84,85,184}; the Leloir pathway in \textit{S. cerevisiae}, which is transcriptionally activated by galactose\cite{185,186}; and mammalian cholesterol biosynthesis genes, which are activated by the TF SREBP when cholesterol levels are low\cite{71,17}. Another example of transcriptional rewiring of metabolism involves propionate degradation in the nematode \textit{C. elegans}. Like humans, \textit{C. elegans} utilizes a vitamin B12-dependent pathway to breakdown this short-chain fatty acid. When dietary vitamin B12 is low, propionate metabolism is transcriptionally rewired to an alternative degradation pathway referred to as the propionate shunt, thereby preventing toxic propionate accumulation\cite{76,144,148}.

The contribution of transcriptional regulation of metabolism has mostly been studied at a systems, or network, level, in single cell organisms such as \textit{E. coli} and \textit{S. cerevisiae} and to a lesser extent in plants\cite{165–167,188,189}. However, the extent to which overall metabolic activity is under transcriptional control in animals remains unclear.
C. elegans is an excellent multicellular animal model to study the transcriptional regulation of metabolism at a systems level: its fixed lineage of 959 somatic cells was fully described\textsuperscript{135}, its metabolism shows extensive conservation with human metabolism\textsuperscript{143,190}, many gene expression datasets are available, and a genome-scale metabolic network model has been reconstructed\textsuperscript{18}. The most up-to-date metabolic network model, iCEL1314, contains 907 metabolites, 2,230 reactions and 1,314 genes\textsuperscript{19}. By using FBA, iCEL1314 can be used to gain insight into the metabolic state of C. elegans during different nutritional conditions or in different tissues. An additional set of metabolic genes has been predicted based on homology with known metabolic enzymes in other organisms or based on the presence of domains found in metabolic enzymes\textsuperscript{18,191}.

Guilt-by-association is a powerful concept in systems biology that can be used to identify genes with shared functions. One way this can be done is by coexpression analysis where a functional association can be predicted when genes are coexpressed in many transcriptomic datasets\textsuperscript{192–196}. In C. elegans, coexpression analysis has been used to study global, temporal, and spatial gene expression\textsuperscript{196–201}.

Here, we investigated the extent of transcriptional regulation of C. elegans metabolism. We developed a computational pipeline to identify genes of which the corresponding mRNA varies significantly during development, in different tissues, and across a gene expression compendium consisting of different conditions.
Using both a supervised and a semi-supervised method, we identified coexpressed metabolic pathways and sub-pathways. Overall, we found that three-quarters of metabolic genes exhibit variation in expression, which is comparable to the proportion in non-metabolic genes. Further, we found that most annotated metabolic pathways contain genes that are significantly coexpressed. With a custom-made semi-supervised method, we identified clusters of genes that define coexpressed sub-pathways or combinations of sub-pathways that likely form functional metabolic units. We extracted conditions where coexpressed clusters of genes are coordinately activated or repressed, revealing how these clusters may contribute to metabolic homeostasis. We developed a web application we named ‘WormClust’ that is available on WormFlux website. WormClust enables querying of *C. elegans* genes to identify metabolic (sub-) pathways with which these genes are coexpressed. Altogether, our findings show that transcriptional regulation of metabolic genes and pathways is ubiquitous in *C. elegans*, indicating that this principle is broadly conserved from single cell organisms to metazoans. Finally, our analyses and tools provide a platform for similar studies in other organisms, including humans.
Results

Three-quarter of metabolic genes are transcriptionally regulated

mRNA levels are determined by a combination of synthesis and degradation. Here, we used variation in mRNA levels as a first approximation for transcriptional regulation. We evaluated the expression of metabolic genes during development, in different tissues, and under different conditions to identify metabolic genes that are highly variant and therefore likely transcriptionally regulated. We used all annotated metabolic and grouped them into four classes based on current annotation (File 2.1): Class A, iCEL1314 genes \((N = 1,308; \text{after removal of six pseudogenes, see Materials and Methods})\); class B, genes annotated to reactions that cannot yet be connected to the iCEL1314 model \((N = 192)\); class C, genes encoding proteins with homology to metabolic enzymes in other organisms \((N = 860)\); and class D, genes encoding proteins with a domain found in known metabolic enzymes \((N = 132)\). Hereafter, we refer to the 1,308 genes in class A as “iCEL1314 genes” and the remaining 1,184 as “other metabolic genes”.

We first identified metabolic genes that vary in expression during larval development by using a high-quality postembryonic time-resolved RNA-seq dataset, hereafter referred to as the “development dataset”\(^{202}\); Figures 2.1A and 2.4A, and File 2.2). Briefly, this dataset contains expression profiles of stage-synchronized animals that were collected every 2 h after hatching for 48 h at 20°C. In the original paper, genes were grouped into 12 clusters based
on similarity in developmental expression profiles. One of these clusters contains 5,045 genes, including 995 metabolic genes, with relatively invariant temporal expressions. We will refer to this cluster as the “flat cluster.” However, although the expression levels of most of the flat cluster genes are relatively stable during development, we noticed that some did exhibit considerable variation. Additionally, many invariant genes from other clusters were not included in the flat cluster. Therefore, we used an unbiased statistical method, called variation score (VS) to stringently define variation in developmental gene expression. This included calculating deviation from the flat cluster genes’ expression and then empirically establishing a conservative VS threshold (0.169; Figure 2.1A and B, see details in Materials and Methods). We excluded 3,552 genes, including 213 metabolic genes, because they were expressed at levels too low for variability analysis. For the remaining metabolic genes, we found that 754 (31.4%, VS ≥ 0.169) are highly variant, and 98 were invariant (4%, VS = 0; Figure 2.4B). The remaining 1,332 metabolic genes (0 < VS < 0.169) were annotated as moderately variant (Figs 2.1C and 2.4B). About a quarter of iCEL1314 genes (329, or 26%) are highly variant, which is lower than the proportion of other metabolic genes (37%) and nonmetabolic genes (41%; Figure 2.1C and D, and File 2.2). The percentage of highly variant metabolic genes is lower than that of nonmetabolic genes across most VS thresholds (Figure 2.4C).
Figure 2.1: Transcriptional Regulation of Metabolism during Development

(A) Computational pipeline to identify highly variant metabolic genes during development. Genes that showed either moderate or high expression at least one time point were selected, reducing the number of genes from 18,113 (2,397 metabolic) to 14,561 (2,184 metabolic). For each gene, VS was calculated using the deviation from a flat reference profile at each time point (see Methods). The red line indicates the mean value and light red shaded area the standard deviation of the flat reference profile. The profile of acdh-1 is plotted in blue as an example of a developmentally regulated gene. d_s^5 denotes the deviation of acdh-1 expression from the flat reference profile at the 5th data point.
(B) Distribution of VS of genes belonging to the flat cluster versus those belonging to other clusters. The vertical lines at 1 and 3 represent the iCEL1314 genes *nduf*-7 and *ugt*-42, that have the lowest and highest VS of the flat set, respectively. The vertical line at 2 indicates the gene *C27A7.3* with selected threshold of VS at the 97% quantile of the flat cluster (VS = 0.169).

(C) Diagram showing criterion of categorizing metabolic and non-metabolic genes into four categories across development: lowly expressed, invariant, moderately variant and highly variant.

(D) Bar graph shows the distinction of low expressed, invariant, moderately variant and highly variant genes during development in iCEL1314 and other metabolic genes. Color legend as indicated in (C).

To identify metabolic genes that exhibit differential expression across tissues, we selected a high-quality single-cell RNA sequencing dataset that measured gene expression during L2 stage of *C. elegans* across seven major tissues: body wall muscle, glia, gonad, hypodermis, intestine, neurons, and pharynx\textsuperscript{142} (Figure 2.2A, Figure 2.4A, File 2.3). This dataset is hereafter referred to as the ‘tissue dataset’. Unlike the development dataset, the tissue dataset does not have a defined cluster of invariant genes. Therefore, we used the less sophisticated Coefficient of Variation (CV) measure to identify variation in gene expression across the seven tissues (Figure 2.2A). We previously found that the five genes comprising the propionate shunt are differentially expressed in different tissues, and each of these genes had a CV greater than 0.75 (Figure 2.2B). Visual inspection of genes with CV values from 0.15 to 1.2, indicates that CV=0.75 provides a clear, yet conservative threshold to annotate highly variant genes across tissues (Figure 2.2C). We further classified genes with a CV less than 0.75 but greater than or equal to 0.3 as moderately variant, and genes with a CV less
than 0.3 as invariant (Figure 2.2D, see Figures 2.2B and 2.2C for examples). A total of 6,370 genes, including 348 metabolic genes, were not included in this analysis because they are expressed at low levels (details in Methods). We identified ~60% and ~25% of metabolic genes as highly and moderately variant, respectively. These include 781 highly variant and 405 moderately variant iCEL1314 genes (Figures 2.2E and 2.4B). A very small number of metabolic genes (13, or 1%) were invariant across tissues. Even though the analysis of the different datasets used a different statistical approach, these results suggest that more metabolic genes are variant and, therefore, likely transcriptionally regulated in different tissues than during development (Figure 2.4B). In contrast to development, the percentage of metabolic genes that are highly variant across tissues at any CV cutoff is greater than non-metabolic genes (Figure 2.4D), indicating that metabolic processes exhibit a relatively high level of tissue specificity.

Overall, metabolic gene expression showed higher variation across tissues at a fixed time point (L2) than larval development. However, because we used two different statistical methods for the development and tissue datasets, we confirmed that it held true when we applied the same CV measure we used in the tissue dataset to the development dataset (Figure 2.3A).
Figure 2.2: Transcriptional Regulation of Metabolic Genes across Different Tissues

(A) Computational pipeline to determine highly variant metabolic genes across tissues. Genes that show either moderate or high expression in at least one tissue...
are selected for analysis reducing the number of *C. elegans* genes from 19,675 (2,491 metabolic) to 13,305 (2,143 metabolic). The coefficient of variation (CV) of each gene was calculated by dividing the standard deviation of expression across tissues by the mean expression. Different thresholds of CV were titrated to select a stringent CV to categorize genes as variant and therefore potentially transcriptionally regulated. Examples for each threshold are provided in (B).

(B) Tissue expression and CV values of propionate shunt genes.

(C) Bar graphs showing expression profiles of example genes across tissues with CV 0.15, 0.3, 0.45, 0.6, 0.75, 0.9, 1.05 and 1.2. Numbering of examples is according to the corresponding threshold lines in (A).

(D) Diagram showing criterion of categorizing metabolic and non-metabolic genes into four categories across tissues: lowly expressed, invariant, moderately variant and highly variant.

(E) Bar graph shows the distinction of low expressed, invariant, moderately variant and highly variant genes across tissues in iCEL1314 and other metabolic genes. Color legend as indicated in (D).

The two datasets also have different resolution: the development dataset has great temporal but no spatial resolution because it was measured by bulk RNA-seq while the tissue dataset, which was measured by single cell RNA-seq, has great spatial but no temporal resolution. Therefore, we examined genes that are highly tissue-specific, because they are highly expressed in a single tissue in the tissue dataset, and found that only 57% of these are also highly variant in the development dataset (Figure 2.3B). Therefore, we conclude that transcriptional regulation of metabolic genes more frequently establishes spatial than temporal gene expression patterns.
Figure 2.3: Transcriptional Regulation of Metabolism Across Development using alternative criterion

(A) Diagram showing alternative criterion using CV for categorizing metabolic and non-metabolic genes into four categories across development: lowly expressed, invariant, moderately variant and highly variant (left). This analysis was done to make sure the result that metabolic genes are more highly variant across tissues than development is not driven by the use of different statistical approaches (i.e., CV in tissue data set and VS in development). To eliminate this effect, we reevaluated the development dataset with the CV approach using the same threshold as with tissues, and found that, the percentage of highly variant genes during development were lowered from 31% to 15% with this method. Thus, using the same metric resulted in an even greater difference between metabolic genes exhibiting variation across tissues versus those exhibiting variation during development. Pie charts showing metabolic (center) and non-metabolic (right) gene expression variation in the development dataset using CV as criterion.

(B) Pie chart showing percentage of tissue-specific genes showing high variation in expression across development dataset.

To directly compare metabolic gene expression in tissues and development, we plotted VS values of metabolic genes across development versus CV values across tissues and found that these two parameters are moderately correlated (Figure 2.4E). We divided the scatter plot into four quadrants, based on the thresholds used in each dataset (File 2.4).
Figure 2.4: Analysis of Metabolic Gene Expression during Development, in Different Tissues and in a Gene Expression Compendium

(A) Computational pipeline to identify C. elegans metabolic genes that change in expression during development, across tissues, and compendium of multiple conditions. Statistically significant differences in gene expression were calculated in the developmental dataset using a variation score (VS), in the tissue dataset...
using coefficient of variation (CV) and by number of datasets with CV ≥ 0.75 in the compendium (collection of 177 datasets).

(B) Pie charts of metabolic and non-metabolic gene expression variation in the three different datasets: development, tissue, and compendium separately and combined. Bar graph shows metabolic genes in iCEL1314 and other (predicted) metabolic genes.

(C) Comparison of percentage of highly variant metabolic versus non-metabolic genes at different VS thresholds.

(D) Comparison of percentage of highly variant metabolic versus non-metabolic genes at different CV thresholds.

(E) Scatter plot of VS (development) versus CV (tissue) of metabolic genes. The plot is divided into four quadrants: Q1 with moderate/low VS and moderate/low CV; Q2 with high VS and moderate/low CV; Q3 with moderate/low VS and high CV; and Q4 with high VS and high CV. The Pearson and Spearman correlation coefficients and the corresponding p-values are indicated. Examples of Q4 genes that are highly variant both during development and in different tissues include \textit{ugt-13}, \textit{ugt-18} and \textit{ugt-34} (UGT enzymes); \textit{gcy-3} (guanylate cyclases); and \textit{ddo-3}, \textit{gln-2}, \textit{argk-1} and \textit{phy-3} (amino acid metabolism).

(F) Comparison of percentage of highly variant metabolic versus non-metabolic genes at different cutoffs of number of datasets with high CV (≥ 0.75).

(G) Comparison of percentage of invariant metabolic versus non-metabolic genes at different cutoffs of the fraction of datasets with low CV (<0.3).

(H) Comparison of percentage of highly variant metabolic versus non-metabolic genes at different CV cutoffs in at least three datasets in the compendium.

(I) Venn diagram of highly variant metabolic genes in the different datasets.

To determine if there are any functional enrichments, we performed pathway enrichment analysis (PEA) on the metabolic genes for each quadrant using the tool provided on the WormFlux website \textsuperscript{203}. The first quadrant (Q1) consists of genes with moderate/low developmental variation and moderate/low tissue variation. It has 595 metabolic genes, including 385 iCEL1314 genes that are enriched in several metabolic pathways, such as the electron transport chain (ETC), aminoacyl-tRNA biosynthesis, the tricarboxylic acid (TCA) cycle, the pentose phosphate pathway and glycolysis/ gluconeogenesis (\textbf{Figure 2.5A}). The
second quadrant (Q2), with high developmental variation and moderate/low tissue variation, consists of only 176 genes, including 68 iCEL1314 genes that are enriched in sulfur, cysteine, and methionine metabolism (Figure 2.5A). The 891 genes in the third quadrant (Q3) consist of genes with moderate/low developmental variation and high tissue variation. They include 504 iCEL1314 genes that are highly enriched in lipid metabolism. Notably, genes involved in peroxisomal fatty acid (FA) metabolism vary more in expression than mitochondrial FA degradation (Figure 2.5A). Finally, the 577 genes in the fourth quadrant (Q4) show high developmental variation and high tissue variation. They include 261 iCEL1314 genes which are enriched in UDP-glucuronosyltransferases (UGT) enzymes, guanylate cyclases, glyoxylate and dicarboxylate metabolism, and amino acid metabolism, such as arginine and proline metabolism and glutamate/glutamine metabolism (Figure 2.5A). Interestingly, there are differences among different types of metabolic genes. For instance, amino acid metabolism genes are variant in both development and in tissues, while lipid metabolism genes are mostly variant in tissues, and growth and energy metabolism are relatively invariant in both development and in tissues.
Figure 2.5: Pathway and Phenotypic Enrichment Analysis of Metabolic Genes Based on Expression Variation

(A) Bar graph showing enriched WormPaths pathways/categories for iCEL1314 genes in the four quadrants Q1, Q2, Q3, and Q4 in Figure 1H. The significance...
levels are indicated by asterisks or ‘ns’ (not significant). Not significant (ns) pathways are not shown.

(B) Bar graph showing phenotypes enriched for highly variant metabolic genes (FDR-corrected p-value ≤ 0.001).

(C) Bar graph showing WormPaths pathways/ categories enriched for moderately variant metabolic. The significance levels legend as indicated in (A).

(D) Bar graph showing phenotypes enriched for moderately variant metabolic genes (FDR-corrected p-value ≤ 0.001)

To evaluate metabolic gene expression more broadly, we combined 177 expression profiling datasets into an expression compendium, an earlier version of which we have used to study TF paralogs\(^{155}\) (Figs 2.4A and 2.6A, Files 2.5 and 2.6, see Methods). Using a CV threshold ≥ 0.75 in at least three datasets, we found that 775 of the 2,492 metabolic genes (~31%), including 284 iCEL1314 genes, are highly variant in the compendium, which is lower than non-metabolic genes (44%, Figs 2.4B and 2.6B). This difference holds true for different cutoffs of the number of datasets showing high variation (Figure 2.4F), and across different CV thresholds (Figure 2.4G). However, the percentage of invariant genes is similar between metabolic and non-metabolic genes using different CV cutoffs (Figure 2.4H).

When we compared highly variant genes in development, tissue, and compendium we found that a total of 1,867 metabolic genes (75%) are highly variant in at least one of the three datasets and that 283 metabolic genes are highly variant across all three datasets (Figures 2.4B, 2.4I). Using phenotypes provided in WormBase WS282\(^{204}\), we found that the 75% highly variant metabolic genes
are enriched in conditional response variants such as chemical and pathogen response, and depleted in essential phenotypes such as lethality, larval arrest, slow growth, and sterility (Figure 2.5B). Finally, the remaining 624 (25%) of metabolic genes that are not highly variant in any dataset are similar in pathway enrichment as the Q1 genes discussed above and are enriched in the essential phenotypes (Figure 2.5C-D).

**Figure 2.6: Transcriptional Regulation of Metabolic Genes across a Compendium**

(A) Computational pipeline to determine transcriptionally regulated metabolic genes across the gene expression compendium. Datasets were evaluated for batch effects using correlation distribution. Some skewed datasets were corrected for batch-effect by splitting into separate datasets, while some were removed if the source of skewness was not clear. This resulted in 177 datasets for analyses. CV of each gene was calculated. The criterion of classifying genes is based on the number of datasets with high CV and fraction of datasets with low CV.
Altogether, our analyses indicate that at least 75% (1,867 out of 2,492) of metabolic genes vary in mRNA levels and are therefore likely transcriptionally regulated, including 902 iCEL1314 genes (~69%) (Figure 2.4B). This is similar to the proportion of varying non-metabolic genes (~79%), indicating that metabolic genes are overall at least as much under transcriptional control as other genes (Figure 2.4B).

A supervised approach shows widespread coexpression of genes comprising metabolic pathways

We previously found that the five genes comprising the propionate shunt pathway are coordinately activated in response to propionate accumulation\textsuperscript{76,144}. In addition, we found strong coexpression of genes functioning in the methionine/S-adenosylmethionine (Met/SAM) cycle, for instance when flux through this pathway is perturbed\textsuperscript{145}. To systematically test which \textit{C. elegans} metabolic pathways exhibit coexpression, we developed a custom pathway enrichment analysis pipeline (Figure 2.7A) based on gene set enrichment analysis (GSEA) (see Methods\textsuperscript{205} and applied it to the compendium. We ran this pipeline using metabolic pathways, enzyme complexes, and enzyme families as defined in WormPaths\textsuperscript{203}. Henceforth,
we use ‘category’ to refer to a group of metabolic genes that best fit in an enzyme complex or related set of enzymes such as amino-acyl-tRNA synthetases. We calculated an enrichment score (ES) that defines the enrichment of relatively high coexpression within that set. A normalized ES (NES) indicates relative strength of this enrichment compared to randomized tests, the significance of which is measured as a false discovery rate (FDR) (Figure 2.7A). With an FDR cutoff of ≤ 0.05, 52 of 84 metabolic pathways or categories (~61%) exhibit coexpression, which is significantly more than expected by chance (Figure 2.7B, 2.7C, File 2.7). As expected, the 52 coexpressed metabolic pathways and categories include the propionate shunt and the Met/SAM cycle (Figure 2.7D-E). When we examined coexpression in pathways and categories separately, we found that 78% of categories showed significant coexpression compared to 58% of pathways (Figure 2.7B). Examples of metabolic pathways that exhibit high coexpression include peroxisomal FA degradation and starch and sucrose metabolism (Figures 2.7F and 2.7G). Examples of coexpressed categories include vacuolar ATPases, ETC complex I, and aminoacyl-tRNA synthetases (File 2.7). There are 32 categories and pathways that do not exhibit self-enrichment, including pantothenate and CoA biosynthesis and mevalonate metabolism (Figs 2.7H and 2.7I). Such pathways may either not be regulated at all, may be regulated by allosterly, or only one or a few genes in these pathways are transcriptionally regulated and may therefore function as key regulatory genes. Alternatively, these
pathways maybe coregulated in conditions that were not yet profiled and therefore are not included in the compendium.

The extent of within-pathway coexpression of metabolic genes can be potentially confounded because metabolic reactions are often associated with multiple genes in gene-protein-reaction (GPR) annotations. There are two reasons for this. First, some metabolic reactions are catalyzed by enzyme complexes comprising two or more proteins. In such cases, all genes need to be expressed for the reaction to take place and are therefore annotated here as ‘AND’ genes. Second, some genes are part of larger families (paralogs) that encode isozymes or highly similar proteins. Metabolic network reconstruction efforts use protein sequence homology to associate genes with reactions. As a result, multiple highly homologous paralogs may be associated with the same metabolic reaction. Such paralogs are annotated here as ‘OR’ genes. Some reactions are associated with a combination of AND and OR genes (Figure 2.8A). Finally, for some gene families it may be that one member catalyzes one reaction and another member catalyzes another. Paralogs that are associated with distinct reactions are referred to here as “other paralogs” (Figure 2.8B). Pathways can be associated with multiple types of AND and OR genes (Figure 2.7J).

AND genes encode proteins that function together in complexes, and such genes are often strongly co-expressed. For example, genes encoding ETC complex members are coexpressed and coregulated. Therefore, we wondered
if this holds true for AND genes in iCEL1314 and, if so, whether this would inflate pathway coexpression enrichment. To test this, we systematically assessed coexpression of different types of gene pairs. As expected, we found that AND genes are significantly more coexpressed than random gene pairs, OR genes and other paralogs (Figs 2.7K, 2.8C, 2.8D, and 2.8E). Both OR genes and other paralogs are also more coexpressed than random metabolic gene pairs in all three datasets (Figs 2.7K, 2.8C, 2.8D, and 2.8E). Surprisingly, OR genes are more coexpressed than other paralogs across tissues (Figs 2.7K, 2.8C, 2.8D, and 2.8E).
Figure 2.7: Supervised approach to investigate coexpression of metabolic pathways

(A) Custom computational pathway enrichment analysis pipeline that determines coexpressed genes functioning in the same metabolic pathway. Pairwise
coexpression was based on the gene expression compendium. For every annotated metabolic pathway, the coexpression of pathway genes (columns) to all metabolic genes (rows) was extracted. A ranked list of genes was obtained for each pathway by taking the mean of coexpression values in rows while ignoring self-correlations. Weighted gene-set enrichment analysis was then performed to find significantly enriched pathways. If a pathway is self-enriched with FDR ≤ 0.05, it is annotated as coexpressed.

(B) Bar graph indicating the percentage of metabolic pathways and categories that show significant coexpression compared to ones that are not self-enriched for coexpression.

(C) Histogram denoting the number of significantly coexpressed metabolic pathways obtained by 1000 randomizations while maintaining the structure of the data.

(D-I) Mountain plots showing self-enrichment of (D) propionate shunt pathway, (E) Met/SAM cycle, (F) peroxisomal fatty acid degradation pathway, (G) starch and sucrose metabolism, (H) pantothenate and CoA biosynthesis, and (I) mevalonate metabolism.

(J) Metabolic pathways often consist of reactions catalyzed by single genes, OR genes and AND genes. All genes involved in the same pathway are collectively annotated as all pathway genes. Genes that are associated with distinct reactions are annotated as PW genes. PW gene pairs exclude AND and OR gene pairs. Met/SAM cycle pathway, which consists of 13 metabolic genes, is shown as an example.

(K) Percentages of pairs of AND genes, OR genes, other paralogs, operon genes and random metabolic genes categorized as coexpressed using different coexpression values as cutoffs. Coexpression values are based on the gene expression compendium.

(L) Percentage of random, all pathway, PW, and PO gene pairs categorized as coexpressed using different coexpression values as cutoffs. Coexpression values are based on the gene expression compendium.

In *C. elegans*, ~18% of genes are transcribed from operons. In total, 26% of metabolic genes occur in operons. However, they most frequently occur as a pair with a non-metabolic gene. In total, 242 metabolic genes (~10% of all metabolic genes) occur in a pair with another metabolic gene in an operon (*File 2.1*). As expected, these operon gene pairs are more coexpressed than random
gene pairs, OR genes and other paralogs, thus, serving as a validation for our coexpression analysis. However, these pairs are less coexpressed than AND genes and there is no overlap between the two categories. This shows that enzyme complexes are strongly coregulated and their coregulation mechanism is largely independent of operonic organization (Figs 2.7K, 2.8C, 2.8D, and 2.8E).

Figure 2.8: Reaction-level analysis of metabolic pathways
(A) The conversion of succinate to fumarate, which is part of complex II of the ETC and of the TCA cycle, is carried out by succinate dehydrogenase. Diagram
showing that succinate dehydrogenase is composed of the OR genes \textit{sdha-1} and \textit{sdha-2} that each function together with the rest of the genes as AND genes. The GPR of this reaction is noted as \((sdha-1 \mid sdha-2) \& sdhb-1 \& mev-1 \& sdhd-1\). Color legend is indicated.

\textbf{(B)} Example of a gene family (\textit{aagr}) where members occur as paralogous OR gene pairs in separate reactions. Pairs of paralogs associated with different reactions are called other paralogs.

\textbf{(C)} Violin plot comparing coexpression for different populations of gene pairs including random, AND, OR, operon and other paralogs gene pairs in compendium of expression datasets.

\textbf{(D-E)} Percentages of AND, OR, other paralogs, operon genes and random metabolic gene pairs categorized as coexpressed using different coexpression values as cutoffs across developmental time (\textbf{D}) and tissues (\textbf{E}). Color legend as indicated in \textbf{(D)}.

\textbf{(F-H)} Violin plots showing coexpression of random gene pairs, all pathway genes, PW genes and PO genes across compendium (\textbf{F}), different developmental stages (\textbf{G}) and tissues (\textbf{H}).

Based on the analysis of AND, OR and operon genes, it is difficult to determine the contribution of the coexpression of such gene pairs to pathway enrichment. Therefore, we examined coexpression of gene pairs that are annotated with distinct reactions in a pathway, which we refer to as pathway (PW) genes (\textbf{Figure 2.7J}). We found that PW gene pairs are significantly more coexpressed than random gene pairs (\textbf{Figs 2.7L, 2.8F, 2.8G, and 2.8H}). We also examined coexpression of gene pairs that are not part of an operon, which we refer to as pathway excluding operon (PO) genes. There are only three pathway gene pairs that are part of operon, hence there is no significant difference between pathway genes and PO genes coexpression (\textbf{Figs 2.7L, 2.8F, 2.8G, and 2.8H}). Therefore, pathway coexpression is not just driven by AND, OR and operon genes,
indicating that pathway genes’ coexpression is a true feature of many metabolic pathways.

A semi-supervised approach extracts coexpressed sub-pathways

Our finding that metabolic pathways and categories exhibit extensive coexpression was based on previously annotated pathways. However, these pathways connect into the larger metabolic network and the definition of the start and ending of each pathway is somewhat arbitrary. Since there is extensive coexpression of genes that function together in pre-defined pathways, we reasoned that we may be able to use coexpression to extract metabolic (sub)-pathways in an unbiased manner. To specifically focus on metabolic genes that function in connected reactions in the metabolic network, we developed a ‘coflux’ metric that calculates flux dependency between metabolic genes using the network model (see details in Methods) (File 2.8). Reactions in linear pathways have complete flux dependence, (i.e., coflux = 1), while in branched pathways flux dependency may be partial (coflux = between 0 and 1), and in uncoupled reactions there is no dependence (coflux=0). We then used a custom semi-supervised approach that multiplies coflux and coexpression values, and clustered the resulting product matrix with a relatively stringent set of parameters (Figs 2.9A, 2.9B, and 2.10A, File 2.9, see Methods for details).
Figure 2.9: A semi-supervised approach to extract coexpressed and flux-dependent metabolic genes

(A) Computational pipeline to extract tightly coregulated units in the metabolic network: Functional relationships are provided through theoretical flux associations (coflux) calculated using C. elegans metabolic network model iCEL1314, while expression correlations come from the compendium of 177 expression datasets. Hierarchical clustering on the product of coflux and coexpression matrix gives coexpressed metabolic pathways.

(B) Heatmaps showing coflux and coexpression of iCEL1314 genes and clustered heatmap showing added modularity to coexpression space by product of coexpression and coflux. Color legend is indicated.
Distinct clusters denoted by clustered heatmap of genes in canonical and shunt pathways of propionate degradation were extracted using dynamic cut tree algorithm with stringent parameters (deepSplit=2, minClusterSize=3). Color legend as indicated in (B).

As expected, the propionate shunt pathway genes formed a tight cluster (Figs 2.9C and 2.10B). Interestingly, while the first four genes, acdh-1, ech-6, hach-1, and hphd-1, occurred closely together, the fifth gene, alh-8, was not part of the same cluster. This could be explained in two ways. First, alh-8 encodes an enzyme that functions at a junction in the pathway where its substrate malonate semialdehyde is converted either to acetyl-coa or, potentially, to beta-alanine. Therefore, metabolic flux is divided in two directions, and is not linearly coupled with the shunt pathway flux like the first four reactions. Second, alh-8 is annotated to another reaction where 2-methyl-3-oxopropionate is converted to propionyl-CoA (Figure 2.9C). This approach also revealed another cluster comprising the canonical, vitamin B12-dependent propionate degradation pathway, indicating that, like the propionate shunt, this pathway may also be transcriptionally activated or repressed under specific conditions (Figure 2.9C, Dataset EV9).
Figure 2.10: Mean Silhouette Score (MSS) to Evaluate Cluster Quality of Stringent and Relaxed Clusters

(A) Plot showing the distribution of the MSS of all clusters obtained using dynamic cut tree algorithm with stringent parameters (deepSplit=2, minClusterSize=3). The green vertical line shows the MSS of shunt cluster.

(B) Scatter plot with individual silhouette scores of genes in the propionate shunt cluster, with stringent parameters. MSS of this pathway is shown.

(C) Plot showing the distribution of the MSS of all clusters obtained using dynamic cut tree algorithm with relaxed parameters (deepSplit=3, minClusterSize=6). The green vertical line shows the MSS of shunt cluster.

(D) Scatter plot with individual silhouette scores of the cluster containing propionate shunt genes, with relaxed parameters. MSS of this pathway is shown along with the overall MSS threshold.
The semi-supervised approach reveals pathway boundaries

The propionate shunt example above shows that the semi-supervised approach can extract sub-pathways (e.g., the propionate shunt) from previously defined pathways (e.g., propionate degradation) based on coexpression and coflux. We therefore used other clusters defined by the semi-supervised approach to better define starts and ends of different pathways.

An example of a pre-annotated WormPaths pathway that was fully captured with the semi-supervised approach is peroxisomal FA degradation (Figure 2.11A, File 2.9). The first reaction in peroxisomal beta-oxidation is catalyzed by acyl-CoA oxidases (encoded by acox genes). Only acox-1.1 and acox-3 in the acox-family genes are coexpressed with the other peroxisomal FA oxidation genes, indicating that they are more likely to function in this pathway than the other acox genes, which are coexpressed with each other, and with mitochondrial FA degradation genes (File 2.9).

Examples where only a subset of annotated pathway genes clustered together include tyrosine metabolism and histidine degradation. Tyrosine can be metabolized via different reactions, in different pathway branches (Figure 2.11B). In one pathway branch, tyrosine is degraded in five steps to produce fumarate and acetoacetate. The genes in this branch; gst-43, C31H2.4, Y53G8B.1, hpd-1, hgo-1, fah-1 & gst-42 form a tight cluster (Figure 2.11B, File 2.9). This cluster consists of OR genes hpd-1 OR C31H2.4; and Y53G8B.1 OR gst-42 OR gst-43, which
suggests that these genes are correctly annotated to this pathway branch. Histidine can also be degraded via two pathway branches: one converting histidine to glutamate through N-formyl-L-glutamate, and the other converting histidine to 3-methylimidazoleacetic acid. The four genes associated with the conversion of histidine to N-formyl-L-glutamate; haly-1, Y51H4A.7, amdh-1 and opin-1, form one of the top-ranked clusters (Figure 2.11C, File 2.9). However, the genes in the other branch are not coexpressed.

We also found clusters consisting of genes that traversed different pathways. For instance, alh-8 and gta-1, which are functionally associated but not strongly coexpressed with the propionate shunt (Figure 2.9D), cluster with T09B4.8 (alanine metabolism) and four other genes belonging to pyrimidine metabolism: dpyd-1, dhp-1, dhp-2 and upb-1 (Figure 2.11D, File 2.9). This observation functionally connects genes in what were heretofore separately annotated pathways, i.e., pyrimidine, alanine, and propionate metabolism. The coexpression of these genes suggests that thymine is degraded, leading to the formation of propionyl-CoA, L-3-amino-isobutanoate and acetyl-CoA. This observation also suggests that alh-8 levels have a stronger functional role in the conversion of 2-methylloxopropanoate to propionyl-CoA than in the propionate shunt. Remarkably, this further indicates that alh-8 may participate in both the generation and degradation of propionyl-CoA.
Figure 2.11: A Semi-supervised approach defines metabolic pathway boundaries

(A) Clustered heatmap denoting a distinct cluster consisting of at least one gene from every reaction in peroxisomal fatty acid degradation. Heatmap genes are shown in bold font. Color legend, indicated here, applies to all panels.
(B) Clustered heatmap showing a distinct cluster formed by the tyrosine degradation genes separate from the rest of the tyrosine metabolism. 
(C) Clustered heatmap showing a distinct cluster formed by a boundary within the histidine degradation pathway. 
(D) Clustered heatmap showing a distinct cluster formed by genes traversing pathway boundaries that are parts of propionate, alanine and pyrimidine metabolism. 

The stringent cluster derivation parameters we used favor small clusters, and as a result, interconnections between different pathways may be lost in the analysis. To unveil such connections, we relaxed the parameters to allow the derivation of larger clusters of coexpressed genes (File 2.10, Figure 2.10C). With these settings, the propionate shunt cluster expanded and included $bckd-1A$, $bckd-1B$, $dbt-1$, $Y43F4A.4$, $ard-1$, $acdh-3$, $acdh-9$ and $B0250.5$, which are annotated to branched chain amino acids (BCAA) isoleucine and valine degradation pathways, but not $alh-8$ or $gta-1$ (Figure 2.9C, 2.12A, File 2.10). Propionyl-CoA, the starting metabolite of the propionate shunt, is produced by the breakdown of valine and isoleucine. We recently proposed that the propionate shunt not only functions to detoxify excess propionate, but also to produce acetyl-CoA for ketone body and energy production $^{210}$. The coexpression of valine and isoleucine breakdown genes with the propionate shunt indicates a functional connection between these pathways to produce energy.

The Met/SAM cycle provides another example of different degrees of clustering that can be unveiled with different parameter settings (Figure 2.12B). The smaller clusters with stringent clustering captured different parts of one-carbon
metabolism with their connections to Met/SAM cycle, while relaxed clustering combined these genes into one single cluster. This cluster acts as a subsystem that connects the Met/SAM cycle on one side with glycerophospholipid metabolism, specifically phosphatidylcholine biosynthesis, which depends on methylation reactions using SAM \(^{145,203}\), as well as with purine metabolism. Second, Met/SAM cycle genes are highly coexpressed with the folate cycle gene which produces the methyl group that is used to convert homocysteine into methionine in the Met/SAM cycle \(^{145,211}\). Together, these results confirm that the Met/SAM cycle is overall coexpressed \(^{145}\) and show that additional co-functioning genes can be identified.

The semi-supervised approach also identified gene clusters that are not part of any coexpressed pathway identified by the supervised method above. An example is selenocompound metabolism, where a set of seven genes form a highly coexpressed cluster (FDR=0.025, NES=1.7) (Figure 2.12C). In comparison, the respective FDR and NES values for self-enrichment of the WormPaths pathway of selenocompound metabolism were 0.75 and 0.98 (File 2.7). Altogether, these results illustrate that not all genes in a pathway are coexpressed and further indicate that a subset of a pathway or a combination of subsets from multiple pathways may be under transcriptional control, illustrating the utility of semi-supervised approach as an addition to the predefined metabolic pathways in WormPaths.
Figure 2.12: Semi-supervised clustering of product matrix using relaxed parameters

(A) Distinct cluster of known coregulated metabolic pathway propionate shunt genes together with isoleucine and valine degradation genes obtained using...
relaxed clustering parameters (deepSplit=3, minClusterSize=6) with the dynamic cut tree algorithm. 

(B) Clusters of Met/SAM cycle genes and genes of related pathways obtained using relaxed (deepSplit=3, minClusterSize=6) and stringent (deepSplit=2, minClusterSize=3) parameters. Stringent clusters are shown near the drawn pathways of clustered genes.

(C) Pathway boundary (green shade) within selenocompound metabolism defined by genes in a high-quality cluster obtained by relaxed parameters and shown by the clustered heatmap (left). Mountain plots showing comparison of self-enrichment analysis statistics (NES, ES and FDR) of selenocompound metabolism from WormPaths with that of selenocompound cluster obtained by the semi-supervised approach (right).

**Metabolic pathway communities reveal coexpression among complexes and pathways**

To explore additional coexpression clusters than those that were captured by the relaxed settings described above, we visually inspected the product matrix and extracted three clusters we refer to as metabolic pathway ‘communities’ (Figure 2.13A). We analyzed these communities by WormPaths PEA ²⁰³. The first community is enriched in ETC complexes I, III and IV, indicating broad transcriptional control of energy production (Figure 2.13B, File 2.11). The second community is enriched in mitochondrial and peroxisomal FA degradation, FA biosynthesis, ascaroside biosynthesis, and BCAA degradation (Figure 2.13C, File 2.11). The connection between FA metabolism and BCAA degradation may reflect the fact that some FAs are synthesized from BCAA breakdown products. For instance, branched-chain fatty acids (BCFAs) are synthesized from the branched-chain alpha keto acids of valine, leucine and isoleucine such as isovaleryl-CoA
and isobutyryl-CoA after their decarboxylation and further chain elongation \(^{212-214}\).

The third community is enriched in aminoacyl-tRNA biosynthesis, N-glycan biosynthesis, collagen biosynthesis, iron metabolism, and mevalonate metabolism, all of which produce biomass precursors. While aminoacyl-tRNA synthetases play a major role in protein biosynthesis by linking amino acids to their cognate transfer RNAs (tRNAs), mevalonate metabolism provides precursors for glycan, collagen biosynthesis provides collagen for formation of the cuticle and other extracellular matrices, and iron metabolism is important for many aspects of metabolism, including the production of heme groups of heme proteins. This result points toward the possibility that growth is transcriptionally regulated by a central mechanism controlling pathways that produce biomass precursors and assemble biomass (Figure 2.13D, File 2.11). Taken together, we confirmed the coexpression of metabolic pathways, revealed coexpressed sub-pathways, as well as coexpression among pathways.
Figure 2.13: Extraction of metabolic communities

(A) Clustered heatmap indicating communities formed by multiplying coflux and coexpression values of iCEL1314 genes. B, C, and D define three major communities shown in the respective parts of this figure.

(B-D) PEA of communities B (B), C (C), and D (D).
Metabolic sub-pathways are activated or repressed under different conditions

The gene expression compendium is comprised of 177 expression profiling datasets that measure relative mRNA levels in a variety of experimental conditions and genotypes. Therefore, we next asked whether we could identify specific conditions in which different metabolic gene clusters are activated or repressed. Using a custom computational pipeline (Figure 2.14A), we first identified those datasets that best represent the coexpression of a particular cluster. We then manually investigated the top datasets for each cluster (see Methods). To validate this approach, we first examined the expression of propionate shunt cluster genes in top-scoring datasets. We previously showed that propionate shunt genes are repressed in animals fed Comamonas aquatica DA1877, a bacterium that (unlike the standard E. coli OP50 diet) produces vitamin B12 thus enabling flux through the canonical propionate degradation pathway. The dataset from that study, labeled as dataset 15 in the compendium, scored as most significant for propionate shunt gene coexpression, where the genes are expressed in animals fed E. coli OP50, but not in animals fed C. aquatica (Figure 2.14B, File 2.4). Interestingly, propionate shunt genes are also highly coexpressed in a dataset that measured expression in spr-5 mutants versus wild type animals across 1(f1), 13(f13) and 26(f26) generations (Figure 2.14B, dataset 139). Propionate shunt genes are more highly expressed in the N2 reference strain compared to spr-5
mutant animals (Figure 2.14B). Since spr-5 encodes a histone demethylase, this may indicate that the expression of shunt genes is regulated not only by TFs, but also by epigenetic mechanisms.

 Peroxisomal FA degradation genes were most significantly coexpressed in dataset 132, which measured gene expression in precisely staged embryos during the first quarter of embryonic development (Figure 2.14C). The time course included a stage of exclusively maternal transcripts (four-cell), the transition to zygotic transcription (28-cell), and the presumptive commitment to the major cell fates (55-, 95-, and 190-cell stages) \(^{215}\). Peroxisomal FA degradation genes were lowly expressed in four-cell embryos and their expression increased in later embryonic stages, which may reflect a change in carbon source for energy and biomass generation prior to hatching and feeding. Peroxisomal FA degradation genes are also upregulated in animals fed \(P. \ aeruginosa\) compared to the standard \(E. \ coli\) OP50 diet (Figure 2.14C, dataset 51). This may reflect the high energy demand during infection \(^{180}\).
Figure 2.14: Condition analysis of metabolic gene coexpression

(A) Computational pipeline to extract activating or repressing conditions of metabolic clusters. The mean coexpression of all gene pairs in each cluster in each dataset are calculated separately. To rank datasets that showed highest coexpression uniquely for each cluster, these mean coexpression values are normalized by z-scoring the mean coexpression of every cluster with respect to each dataset. For every cluster, find 30 best datasets showing highest z-score of mean coexpression. Rank the 30 datasets on mean coexpression of cluster genes. Investigate individual heatmaps of the best datasets for each cluster to find activating and/or repressing conditions.
normalized using z-scoring across each dataset (as shown by heatmap). 30 best
datasets that potentially represent activation/repression conditions of each cluster
are identified by the z-score values of mean coexpression. Then, for each cluster,
datasets are manually inspected in the order of decreasing mean coexpression
along with its associated published paper and the heatmap to understand
activation/repression conditions.

(B-F) Mean coexpression of the 30 best datasets for clusters of propionate shunt
(B), peroxisomal fatty acid degradation (C), histidine degradation (D), tyrosine
degradation (E), and canonical propionate degradation (F), followed by heatmap
examples from selected datasets as indicated by bold-blue dataset numbers. Color bar and heat map legend as indicated in (B).

Inspection of expression of the histidine degradation cluster discussed
above revealed that it is highly expressed in animals fed *C. aquatica* (Figure
2.14D, dataset 15). However, these genes were not affected in animals fed *E. coli*
supplemented with vitamin B12 \(^{144}\), suggesting that the effect of *C. aquatica* on this
cluster may be independent of this cofactor. Animals showed lower expression of
this cluster upon exposure to UV treatment (Figure 2.14D, dataset 33). In humans,
UV converts cis-uruconate (the product of first reaction of histidine degradation,
Figure 2.11C) to trans-uruconate, which has been proposed to play a protective
role in skin \(^{216}\). Our result indicates that, in *C. elegans*, UV exposure rewires
metabolic flux to avoid histidine degradation, for instance to preserve histidine that
could be converted to trans-uruconate.

The tyrosine degradation cluster was most significantly coexpressed in
dataset 11, wherein expression profiles of *gas-1* mutant animals, which are
deficient in mitochondrial respiration, were compared to wild type animals \(^{217}\)
(Figure 2.14E). This study revealed that free tyrosine levels are decreased in *gas-*
1 mutants. In addition, there is a failure of NAD+-dependent ketoacid oxidation in mitochondrial respiratory chain mutants. To compensate for this respiratory dysfunction, multiple pathways are upregulated, including the TCA cycle and ketone body metabolism. Since the end products of tyrosine degradation pathway cluster are the TCA cycle intermediate fumarate and the ketone body acetoacetate (Figure 2.11B), the function of the upregulation of tyrosine degradation during mitochondrial dysfunction may be to supply metabolites for compensatory pathways.

Altogether, these results show that the gene expression compendium can be used to gain insight into the conditions that most greatly affect the activation or repression of different metabolic gene clusters. However, one needs to be careful to manually inspect the conditions of interest because sometimes coexpression can be biased by an outlier experiment. An example of this is dataset 47, one of the top datasets in which canonical propionate degradation pathway genes are coexpressed. Even though the conditions in this dataset are not related to canonical propionate breakdown, this dataset falsely appears as one of the top datasets due to coexpression driven by one bad outlier sample. It is however also possible that this outlier sample was unknowingly contaminated to change the nutritional or environmental state, hence driving the variable expression of these genes (Figure 2.14F).
Overall, our systematic analysis revealed specific conditions of when metabolic gene clusters are activated or repressed, reinforcing our overall finding that transcriptional regulation plays an important role in the control of metabolism.

**WormClust web application enables gene-by-gene query to identify coexpression with metabolic (sub)-pathways**

A major premise of this study is the assumption that variance in mRNA levels results, at least in part, from transcriptional regulation, which in turn suggests that genes are coexpressed because they are coregulated. In reverse engineering of gene regulatory networks, coexpression of TFs with their target genes has been used to define causal relationships \(^{218,219}\). To make our data available to the community as well as to enable the easy identification of TFs and other *C. elegans* genes that are coexpressed with metabolic (sub)-pathways, we developed a web-application named WormClust, which is available on the WormFlux website \(^{18}\). This tool takes any *C. elegans* gene as input and evaluates its coexpression with metabolic (sub)-pathways. If the query gene is an iCEL1314 gene, the output is a clustered heatmap of the coexpressed genes in the model based on product matrix, and according to the selected level of stringency, i.e., relaxed, or stringent. If the query gene is not an iCEL1314 gene, then an association of the gene with annotated metabolic pathways is provided. The threshold for this association can be based on FDR and/or NES (Figure 2.15A).
We, and others, previously found that nuclear hormone receptor (NHRs) TFs frequently associate with metabolic genes in different types of assays and dataset \(^{156,157,191,220}\). To illustrate the utility of WormClust, we tested three NHRs with known metabolic pathway associations for coexpression with annotated metabolic pathways in the compendium of 177 \textit{C. elegans} expression datasets. All of these showed coexpression with their target metabolic pathways (\textbf{Figs 2.15B, 2.15C and 2.15D}): \textit{nhr-68} was highly coexpressed with the propionate shunt (FDR=0.02 and NES=2.1) \(^{144}\), \textit{nhr-31} associated with vacuolar ATPases (FDR=0.036, NES=2.04) \(^{221}\), and \textit{nhr-79} is coexpressed with peroxisomal FA degradation (FDR=0.019, NES=2.3) \(^{222}\). In addition to pathways, we also performed enrichment of clusters or sub-pathways from our semi-supervised analysis with TFs. We found that \textit{nhr-79} is enriched to cluster 16 (stringent), which contains of peroxisomal FA degradation genes; \textit{nhr-31} shows enrichment to cluster 5 (relaxed) that consists of vacuolar ATPases; and \textit{nhr-68} shows enrichment to cluster 12 (stringent) which contains propionate shunt genes, albeit with higher FDR. In addition, \textit{nhr-68} shows enrichment to cluster 40 consisting of \textit{mans-2, hex-2} and \textit{fut-8} (N-glycan biosynthesis), and cluster 51 consisting of \textit{bgal-1, gana-1} (galactose metabolism) and \textit{hex-1} (sphingolipid metabolism) (\textbf{Figure 2.16A-C, File 2.9}). This observation suggests that \textit{nhr-68} may play a broader role in the regulation of metabolic gene expression.
Figure 2.15: WormClust: A web application that enables querying of genes to identify coexpression with metabolic (sub)-pathways

(A) Diagram showing workflow of WormClust. A *C. elegans* gene is taken as input. If the gene is part of iCEL1314, a clustered heatmap of closely associated genes in product matrix of coflux and coexpression is displayed, based on stringency level of clustering. If the input gene is not an iCEL1314 gene, enrichment bar graphs of the gene to annotated metabolic pathways are displayed, based on selected FDR and NES thresholds.

(B) Bar graph of pathways that are significantly coexpressed with *nhr-68* with NES ≥ 2 and FDR ≤ 0.05.

(C) Plot showing significant coexpression of *nhr-31* with vacuolar ATPases (FDR ≤ 0.05 and NES ≥ 2).

(D) Plot showing significant coexpression of *nhr-79* with peroxisomal fatty acid degradation (FDR ≤ 0.05 and NES ≥ 2).
WormClust also provides an opportunity to annotate new metabolic genes. For example, \textit{vha-20}, a vacuolar ATPase that is not part of the original iCEL model because it was only recently annotated by WormBase and KEGG, shows significant enrichment to vacuolar ATPases (\textit{Figure 2.15E}). This result suggests that WormClust can be used to ‘deorphan’ unannotated metabolic genes. As another example, we found that \textit{acdh-11} is highly coexpressed with mitochondrial fatty acid degradation and isoleucine degradation genes (\textit{Figure 2.15F}). Therefore, we propose that \textit{acdh-11} can now be added to the iCEL model as another OR gene to reactions in these pathways. We envision future systematic studies of orphan metabolic genes and TFs to increase the annotation of metabolic genes and elucidate the transcriptional mechanisms that regulate their expression.
Figure 2.16: Enrichment of TFs to (Sub-)Pathways

(A) Plot showing enrichment to coexpression of nhr-79 with peroxisomal fatty acid degradation and cluster 16 (NES≥2, FDR≤0.05).

(B) Plot showing enrichment to coexpression of nhr-31 with vacuolar ATPases and cluster 5 (NES≥2, FDR≤0.05).

(C) Plot showing enrichment to coexpression of nhr-68 with propionate shunt (NES≥2, FDR≤0.05) and cluster 12 (NES≥2).
## Methods

**Table 2.1: Tools Table**

<table>
<thead>
<tr>
<th>Resource</th>
<th>Reference or Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software</td>
<td></td>
</tr>
<tr>
<td>Python 3.6.6</td>
<td><a href="https://www.python.org">https://www.python.org</a></td>
</tr>
<tr>
<td>MATLAB 2019a</td>
<td><a href="https://www.mathworks.com">https://www.mathworks.com</a></td>
</tr>
<tr>
<td>GSEApv 0.1.18</td>
<td></td>
</tr>
<tr>
<td>dynamicTreeCut 0.1.0</td>
<td><a href="https://CRAN.R-project.org/package=dynamicTreeCut">https://CRAN.R-project.org/package=dynamicTreeCut</a></td>
</tr>
<tr>
<td>Numpy 1.18.3</td>
<td>225</td>
</tr>
<tr>
<td>Pandas 1.1.0</td>
<td>226</td>
</tr>
<tr>
<td>Matplotlib 3.2.1</td>
<td>227</td>
</tr>
<tr>
<td>Scikit-learn 0.20.3</td>
<td><a href="https://scikit-learn.org">Scikit-learn: Machine Learning in Python</a></td>
</tr>
<tr>
<td>Scipy 1.4.1</td>
<td>229</td>
</tr>
<tr>
<td>Seaborn 0.9</td>
<td>230</td>
</tr>
<tr>
<td>Statannot 0.2.3</td>
<td><a href="https://pypi.org/project/statannot/">https://pypi.org/project/statannot/</a></td>
</tr>
<tr>
<td>Sleipnir</td>
<td>231</td>
</tr>
</tbody>
</table>
Preprocessing of genes

The master list of *C. elegans* genes considered for analysis were downloaded from WormBase public ftp site (release WS282)\textsuperscript{4}. The genes were filtered out to obtain only live and protein-coding genes, which amounted to 19,985 genes in total.

Development dataset

Post-embryonic expression profiles were based on published RNA-seq data\textsuperscript{2}. Briefly, the authors measured the transcriptome of wild-type (N2) animals from hatching to 48h post-hatching every two hours. This dataset includes 21,714 protein-coding genes including 2,405 metabolic genes. Genes were classified into 12 clusters based on their expression profiles\textsuperscript{2}. We refer to the cluster showing relatively invariant expression as the “flat cluster” and the genes within this cluster as “flat genes”. Selecting only live protein-coding genes (WS282) resulted in a total of 18,113 genes, including 2,397 metabolic genes. Of these, 4,689 are flat genes, including 995 metabolic genes. We generated a histogram of average gene expression across development using the logarithm of reads per kilobase per transcript (RPKM) values at base 2. This resulted in a bimodal expression distribution that was fitted by two superposed Gaussian curves, representing a high expression subpopulation and a low expression subpopulation (LES). Genes that showed expression values less than the mean plus the standard deviation of
LES at all the time points were filtered out to avoid false fluctuations in gene expression. After this step, a total of 14,561 genes were left, including 2,184 metabolic genes. The number of flat genes was reduced to 4,646, including 986 metabolic genes.

**Tissue dataset**

Tissue-level expression profiles were based on a single-cell RNA-seq dataset of animals at the second larval stage (L2)\textsuperscript{142}. This dataset provides gene expression as transcripts per million (TPM) for 20,271 protein-coding genes including 2,506 metabolic genes across seven major tissues: body wall muscle, glia, gonad, hypodermis, intestine, neurons, and pharynx. Selecting only live genes (WS282) reduced the number of genes to 19,675 including 2,491 metabolic genes. The dataset was previously processed to label gene expression in every tissue according to the level of expression into four categories: high, moderate, low, and rare\textsuperscript{19}. Genes that showed rare or low expression in all seven tissues were filtered out in this study, resulting in 13,305 genes including 2,143 metabolic genes.

**Gene expression compendium**

A compendium of gene expression datasets was generated using a combination of public datasets. First, 374 microarray, RNA-Seq and tiling array datasets related
to *C. elegans* were downloaded from WormBase \(^{204}\). Then, only those datasets that consisted of at least ten conditions were selected, resulting in 169 datasets. These datasets were individually examined for batch effects, since many were obtained from multiple microarray experiments where total RNA was not normalized. Initially, histograms of expression values were analyzed, and sixteen microarray datasets that displayed abnormal distributions where the correlation distributions were skewed towards +1, hence suggesting that the data may consist of samples that are highly distinctive from each other or are from separate experiments altogether. Such datasets were selected for further examination (**Figure 2.3A**). Twelve of these datasets were found to be composed of two subsets of data, where all genes in one subset were up- or down-regulated with respect to the other except for a few. The samples forming each subset were independent of the other and seemed to have different amounts of total RNA or a similar batch effect. Therefore, these datasets were divided into two separate datasets to correct for batch effects. The remaining four datasets were removed since the source of abnormalities in their distributions of expression was not clear. This processing resulted in a total of 177 datasets. For each dataset, the expression of every gene was normalized by converting the expression values to z-scores based on expression across all conditions using the Normalizer function of Sleipnir library \(^{231}\). Once all the datasets were z-normalized, they were combined to form a compendium with 4,796 conditions (sum of multiple conditions within 177
datasets) using the Combiner function of the Sleipnir library, which took a union set of all genes across the different datasets and converted missing values to NaN (not a number) for subsequent processes.

Calculation of Variation Score in the development dataset

We define Variation Score (VS) as a measure of the deviation of a gene’s expression profile from a flat reference over time in the development dataset. Prior to any analysis, expression values of every gene were normalized by total expression in all time points using equation (1),

$$x_{i\text{norm}} = \frac{x_i}{\sum x_i}$$  \hspace{1cm} (1)

where $x_i$ indicates the expression value of gene $x$ at time $i$. To define a reference profile of invariant expression, a line was constructed in time by joining the mean normalized expression value of the flat cluster at every time point. An envelope around this line was then defined by adding and subtracting the standard deviation of each point. A deviation from this envelope, referred as Variation Score (VS), was then computed by taking the average distance between an individual gene profile and the flat reference profile according to Eq.2,

$$VS_g = \frac{\sum d_i}{n}$$  \hspace{1cm} (2)

with $d_i = \begin{cases} 0 & \text{if } x_{i\text{norm}} \in [\mu_i - \sigma_i, \mu_i + \sigma_i] \\ \min (|x_{i\text{norm}} - (\mu_i + \sigma_i)|, |x_{i\text{norm}} - (\mu_i - \sigma_i)|) & \text{else} \end{cases}$
where \( n \) is the number of observations for the gene \( g \), \( d_i \) is the distance at time \( i \) between the normalized level of expression of the gene \( x_i^{\text{norm}} \) and the closest border of the reference flat profile, and \( \mu_i \) and \( \sigma_i \) are mean and standard deviations of normalized expression values of flat genes at time point \( i \) respectively. A graphical example of this calculation is provided in Figure 2.1A. With this definition, a \( \text{VS} = 0 \) means that the profile of a given gene stays within the envelope of the flat cluster and is therefore perfectly flat, or invariant. To define highly variable genes, we empirically established a conservative VS threshold value of 0.169 based on the distribution of VS between flat genes and all other genes, such that 97% of flat genes were not annotated as variant (Figure 2.1B).

**Calculation of Coefficient of Variation**

Coefficient of Variation (CV) is a statistical measure that is used to calculate the dispersion of data. For every gene, CV was calculated by dividing standard deviation of expression across different samples (\( \sigma \)) (e.g., different tissues in case of tissue dataset) to the mean of expression across samples (\( \mu \)). CV was empirically thresholded using the CV of known propionate shunt genes to keep the approach conservative.

\[
CV = \frac{\sigma}{\mu} \quad (3)
\]
Categorizing genes based on expression variation in compendium

To be consistent with the approach used with the tissue dataset, and to be conservative in our assessment, a CV threshold of 0.75 was selected and required that highly variant genes had a CV greater than or equal to this threshold in at least three datasets. Genes that showed CV<0.3 in at least 95% of the datasets were labeled as invariant, and genes that fit into neither category were annotated as moderately variant (Figure 2.3A). It was further assumed that genes that are not present in a dataset are lowly expressed, and were removed from the analysis.

Calculation of coexpression of gene pairs

The correlation in expression of metabolic gene pairs during development, across tissues, and across the compendium of gene expression studies were calculated based on Pearson Correlation Coefficient (PCC) using the Distancer function of Sleipnir library. These correlations defined pairwise coexpression. Differences in the distribution of coexpression values between random, AND genes, OR genes, other paralogs, all pathway genes and PW genes were evaluated using Mann-Whitney U test. 

232
Custom pathway enrichment analysis pipeline

Pathway-to-gene annotations from level 4 of WormPaths were used as input gene sets. Each metabolic pathway (or category such as an enzyme complex; hereafter referred to as pathway for simplicity) consists of two or more annotated metabolic genes. The coexpression of genes in each metabolic pathway with all other genes in the metabolic network was extracted from the compendium. Subsequently, the mean of the correlations of pathway genes with all other metabolic genes (excluding the self-correlations) was calculated. The mean values were used to define a ranked list of metabolic genes for every metabolic pathway. GSEA was then performed on the pre-ranked list of each pathway using the PreRank module. Enrichment score (ES) is the degree to which the genes in a gene set are overrepresented at the top or bottom of the entire ranked list of genes. Since the genes that are functionally related are mostly positively correlated, we only consider the genes at the top of the list, hence the ones positively contributing to ES. Leading edge subset enlists the gene hits before the peak while calculating ES, therefore consisting of genes that contribute the most to the enrichment score.

NES was derived for each pathway by normalizing the ES values to mean ES for all permutations of the gene sets. This accounts for differences in gene-set size. FDR is the estimated probability that a gene set with a given NES represents a false positive finding. The significance cutoff for the GSEA was set at an FDR value of ≤ or equal to 0.05. If a pathway was found to be significantly self-enriched,
it was categorized as coexpressed. We validated this result by running the custom pathway enrichment analysis pipeline on 1,000 randomized gene sets. For these randomizations, the structure of the data was maintained such that the correlation matrix, the number of genes in each pathway and the number of times an individual gene was repeated across pathways all remained the same.

**Semi-supervised approach that combines coflux and coexpression**

The first part of the approach involves using FBA to simulate reaction rates (fluxes) in the metabolic network and then using a flux dependency metric, referred to as coflux, to measure pairwise associations of genes in the *C. elegans* metabolic network model\(^{18,19}\). We examined all reaction pairs to see if constraining the flux of one of the reactions to zero reduces the flux of the other (see below for the algorithm). The coflux value is zero for independent reactions and one for reactions that are fully coupled. Reactions that are connected by a junction to another reaction are usually partially dependent. After generating coflux values for each pair of reactions, we converted the reaction matrix to a pairwise gene coflux matrix using GPR associations. A high coflux value for a gene pair indicates that the genes encode enzymes acting in the same metabolic process. For the second part of the approach, a coexpression matrix was derived from the *C. elegans* gene expression compendium described above. All negative correlations were converted to zero to be consistent with the coflux matrix. The coflux and
coexpression matrices were multiplied to obtain a product matrix. Since both coexpression and coflux values are between 0 and 1, a product takes a high value only if both coflux and coexpression values are high. Hierarchical clustering was then performed on the product matrix using dynamic cut tree algorithm using cutreehybrid package\textsuperscript{224}.

**Coflux algorithm**

The coflux value for each gene pair was calculated using FBA with iCEL1314. First, the standard bacterial diet was amended with a minimum set of nutrients (\textit{i.e.}, by allowing uptake through exchange reactions in the model as indicated in File 2.8) that warranted non-zero flux in all reactions of the model. Then the following steps were taken to calculate coflux values:

- For every irreversible reaction \(i\),
  - Calculate \(v_{\text{max},i}\), the maximum flux that can be achieved with the intact network.
  - For every reaction \(j\), calculate \(v_{\text{max},ji}\), which is the maximum flux observed in reaction \(i\) when reaction \(j\) is constrained to a flux of zero. If \(i\) is included in a predefined set of 15 redundant reaction pairs (\textit{i.e.}, reactions with similar reactants and products except for differences such as the use of NADP instead of NAD as electron carrier, File
the flux of the corresponding reaction in the pair was also constrained to zero.

- For every reaction \( j \), calculate the coflux with \( i \) \((c_{ij})\) using Eqn 4.
  \[
c_{ij} = \frac{v_{\text{max},i} - v_{\text{max},ij}}{v_{\text{max},i}}
  \]  (4)

- For every reversible reaction \( i \),
  - For every reaction \( j \), repeat the above steps to calculate the coflux with \( i \) in forward direction \((c_{ij, \text{forward}})\).
  - Calculate \( v_{\text{min},i} \), the minimum \((i.e., \) the most negative, as negative flux indicates flux in reverse direction\) flux that can be achieved with the intact network.
  - For every reaction \( j \), calculate \( v_{\text{min},ij} \), which is the minimum flux observed in reaction \( i \) when reaction \( j \) is constrained to a flux of zero. Once again, the reaction redundant with reaction \( i \) is also constrained to zero flux, if applicable.
  - For every reaction \( j \), calculate the coflux with the reverse direction of \( i \) \((c_{ij, \text{reverse}})\) using Eqn 5.
    \[
c_{ij, \text{reverse}} = \frac{v_{\text{min},i} - v_{\text{min},ij}}{v_{\text{min},i}}
    \]  (5)

- For every reaction \( j \), calculate final coflux with \( i \) as the maximum of \( c_{ij, \text{forward}} \) and \( c_{ij, \text{reverse}} \) (Eqn. 6).
  \[
c_{ij} = \max (c_{ij, \text{forward}}, \, \text{abs} (c_{ij, \text{reverse}}))
  \]  (6)
Since \( c_{ij} \) and \( c_{ji} \) are not necessarily equal, calculate final coflux value for every reaction pair using Eqn. 7.

\[
c_{ij,\text{final}} = \max (c_{ij}, c_{ji})
\]  

(7)

- Convert the reaction coflux matrix to a gene coflux matrix based on gene-reaction associations. If a gene pair is associated through multiple reaction pairs (i.e., when at least one of the genes is associated with multiple reactions), take the maximum of coflux values between reactions to calculate gene coflux.

**Hierarchical clustering**

Hierarchical clustering was performed using average method of linkage on the dissimilarity matrix generated by 1 minus the product matrix value (coflux*coexpression). Dynamic cut tree algorithm from cutreehybrid package was used to cut the dendrogram generated by this clustering with stringent parameters deepSplit=2 and minClusterSize=3 and relatively relaxed parameters deepSplit=3 and minClusterSize=6 \(^{224}\). The stringent setting was thresholded based on the occurrence of propionate shunt genes together in one cluster while keeping the size of the smallest cluster to be at least 3. The relaxed setting was chosen to capture larger clusters, such as the Met/SAM cycle genes in a single cluster.
Quantifying cluster quality through Silhouette Score

Silhouette score determines the quality of clustering by measuring the cohesiveness of genes within the same cluster and separateness from the genes in the neighboring clusters \(^{233}\). It was calculated using scikit-learn package\(^ {228}\). We first calculated silhouette score of each metabolic gene based on its placement in each cluster and then calculated mean silhouette score (MSS) for every cluster. Stringent clustering led to 197 clusters, ranging in size from three to 27 genes. We ranked these stringent clusters using MSS, where few of the top-ranked were inspected in more detail (Figure 2.10A). For relaxed clustering, we followed the same approach (Figure 2.10C).

Finding activation and repression conditions of metabolic clusters

To find activation and repression conditions of each cluster, we first calculated the mean coexpression of all gene pairs in that cluster in each dataset separately. To rank datasets that showed highest coexpression uniquely for each cluster, we normalized these mean coexpression values of all clusters using z-scoring across each dataset. We then identified the 30 best datasets that potentially represents activation/repression conditions of each cluster by the z-score values of mean coexpression. After this, we manually inspected each dataset in the order of
decreasing mean coexpression, its associated published paper, and the heatmap to understand activation/repression conditions.

**Gene-centric coexpression with metabolic (sub)-pathways by WormClust**

We developed a custom computational pipeline that identifies coexpression of *C. elegans* genes with metabolic genes used in this study. The pathway-gene sets were generated using WormPaths as a GMT (Gene Matrix Transposed) file, a tab-delimited file of gene sets. The ranked coexpression list of metabolic genes was extracted for each queried gene, from the global coexpression matrix generated using compendium of 177 datasets. The ranked list of each queried gene was used to run Gene Set Enrichment Analysis (GSEA) on the custom metabolic pathway-gene sets.

**Data availability**

Gene clusters from semi-supervised approach and pathway enrichment of all protein coding genes outside the iCEL model, including but not limited to TFs, orphan metabolic genes, and transporters are available using WormClust on the WormFlux website ([http://wormflux.umassmed.edu/WormClust/wormclust.php](http://wormflux.umassmed.edu/WormClust/wormclust.php)). Other data can be found in files 2.1-2.11. We also created a Github repository
(https://github.com/WalhoutLab/WormClust) for this project, which includes scripts that generated results presented here.
CHAPTER III : SYSTEMATICALLY DECIPHERING THE “MISSING” KNOWLEDGE IN C. ELEGANS METABOLISM

Background

In 1998 the C. elegans genome sequence was published and predicted to contain, approximately 20,000 protein-coding genes\textsuperscript{234}. However, despite the rapid progress in genome sequencing, the functional annotation of the C. elegans genome, like many other organisms, has not kept pace.

One particular challenge lies in unraveling the functions of metabolic genes, a vital aspect of metabolism research. Many of these genes encode enzymes with unknown functions, presenting a significant obstacle in understanding the metabolic pathways. Moreover, functionally annotated genes may also participate in multiple metabolic pathways, further complicating the delineation of their roles.

The C. elegans genome contains many TFs whose regulatory targets are yet to be linked. There are many transport reactions in the metabolic network model lacking annotated transporters, and many transporters in the genome are not associated with specific metabolites. Additionally, there are several RBPs whose regulatory roles in metabolism are unknown. Resolving these knowledge gaps requires a comprehensive approach that integrates data from diverse expression datasets, encompassing various spatial, temporal, dietary, and environmental conditions.
Large-scale gene expression datasets offer a wealth of information, providing an opportunity to extract insights through meta-analysis. While individual expression profiles are often obtained to address specific research questions, a broader exploration of these datasets can yield valuable findings. Developing a well-defined pipeline is crucial to effectively harness the vast amount of expression data and infer the intricate relationships between genes and metabolism. One powerful approach is the study of coexpression profiles, as genes exhibiting coordinated expression are likely involved in the same biological processes, shedding light on the underlying cellular mechanisms.

Figure 3.1: Possible “missing” knowledge to C. elegans metabolism:
1) Transporter genes that encode transporters responsible for transporting metabolites, 2) Orphan genes that can be annotated to metabolic reactions but are not connected to iCEL1314 network yet, 3) TFs that regulate different metabolic genes/ pathways, 4) RBPs associated to metabolism, and 5) rest of the non-metabolic genes.
Despite significant progress, there remains a plethora of "missing" knowledge when it comes to comprehending *C. elegans* metabolism (Figure 3.1). In our study, we have employed a custom supervised approach to functionally annotate the *C. elegans* genes, specifically focusing on transporters, orphan metabolic genes, TFs, and RBPs. By establishing the links between these gene categories and metabolism, we aim to contribute to a deeper understanding of *C. elegans* metabolism and its regulatory networks.

**Transporters**

Transporters play a crucial role in cellular processes by facilitating the movement of molecules across cellular membranes. There are 889 potential annotated transporters in *C. elegans*\(^{236}\). However, the specific metabolites transported by many of these transporters remain unknown. Simultaneously, the metabolic network model of *C. elegans* contains many transport reactions for which the corresponding transporter proteins have not yet been determined. Bridging this gap between transport reactions and transporter-coding genes is important for unraveling the complete landscape of metabolic processes in *C. elegans*. By establishing the associations between transport reactions and transporter genes, we aim to further enhance our knowledge of metabolic pathways and their regulation.
Orphan metabolic genes

Among the protein-coding genes in *C. elegans*, approximately 2491 out of the total 19983 genes (13%) are annotated as metabolic genes\textsuperscript{141}. Many of these metabolic genes, at least half (1184 genes), remain poorly characterized in terms of their enzymatic activities or their associations with metabolic pathways. These genes are referred to as "orphan" genes, as they lack clear functional annotations in the context of metabolism. The designation of orphan genes as metabolic is based on two criteria: the presence of encoding proteins that exhibit homology to metabolic enzymes found in other organisms (N=860) or the possession of domains commonly associated with known metabolic enzymes (N=132). Despite their classification as metabolic genes, the specific roles and functions of these orphan genes within the metabolic framework of *C. elegans* remain elusive.

To address this knowledge gap, we have employed the WormClust\textsuperscript{141} approach, which is promising in elucidating the functionality of orphan genes and establishing their connections within the metabolic network of *C. elegans*. Notably, our previous investigations have demonstrated the effectiveness of this approach in "deorphaning" genes such as *vha-20* and *acdh-11*, providing valuable insights into their integration into specific metabolic pathways (Figs 2.12E and 2.12F)\textsuperscript{141}.

Expanding our understanding of the *C. elegans* metabolic network by linking additional orphan genes holds great promise. The incorporation of newly annotated reactions derived from orphan genes can significantly improve the
existing genome-scale network model. There are many SACURE-annotated metabolic reactions, which are linked to metabolic enzymes, but they are not yet connected to the metabolic network\textsuperscript{18}. The newly annotated reactions associated with orphan genes can potentially create links between the model and these SACURE-annotated reactions. These refinements will further enhance the model's predictive capabilities\textsuperscript{18,19,141}.

**TFs**

In the *C. elegans* genome, there are currently 941 annotated TFs, representing approximately 5\% of all protein-coding genes\textsuperscript{237}. This is roughly in line with humans, where approximately 7\% of protein-coding genes are TFs\textsuperscript{237}. However, one striking difference is the expansion of the NHR family in *C. elegans*, which comprises 284 annotated members, constituting over 30\% of all TFs. In contrast, humans possess only 48 NHRs, and flies have 2\textsuperscript{1,238,239}. Many *C. elegans* NHRs are orthologous to a single human NHR, hepatocyte nuclear factor 4 (HNF4)\textsuperscript{238}. This evolutionary expansion of NHRs in *C. elegans* highlights the unique regulatory landscape of this organism.

NHRs, like other TFs, possess DNA-binding domains that enable them to interact with specific DNA sequences in gene promoters. In addition to DNA binding domains, NHRs typically possess a ligand-binding domain that recognizes signaling molecules, which can be diverse, including steroids, fatty acids, and
metal ions, among others. Remarkably, the majority of *C. elegans* NHRs have no known ligands, which suggests the presence of novel ligands or unique regulatory mechanisms that remain to be elucidated.

While the ligands for most *C. elegans* NHRs are yet to be identified, several NHRs have been characterized for their roles in regulating physiological responses to environmental perturbations or xenobiotics. The expanded NHR family in *C. elegans* may serve as environmental sensors, enabling the organism to adapt and respond to various external cues and stressors. Unraveling the regulatory functions of these NHRs holds promise for understanding the interplay between environmental signals and gene expression in *C. elegans* to regulate metabolism.

**RBPs**

TFs and miRNAs are often the main focus, when it comes to gene regulation, due to their key roles in controlling gene expression. However, the role played by RBPs, is also critical but has not been studied as extensively. In our study we investigate the potential role of RBPs in *C. elegans* in regulating metabolic processes.

In the *C. elegans* genome, a total of 887 RBPs have been annotated, representing a diverse array of proteins involved in RNA processing, localization, stability, and translation. While the precise functions of many RBPs in *C. elegans* are yet to be fully elucidated, emerging evidence suggests that
approximately 250 of these RBPs may function in a gene-specific manner, exerting post-transcriptional control over specific subsets of genes\textsuperscript{243}. RBPs play pivotal roles in various aspects of gene regulation, including mRNA splicing and stability. Their interactions with target mRNAs enable fine-tuning of gene expression, influencing critical cellular processes and developmental programs. Given their diverse functions, RBPs hold great potential for shaping the regulatory landscape of metabolic genes and pathways.

Investigating the interplay between RBPs and metabolic genes in \textit{C. elegans} may uncover novel regulatory networks, shedding light on the mechanisms that govern metabolic processes beyond transcriptional control. This will enhance our understanding of post-transcriptional gene regulation of metabolism.

Overall, in this study, we aim to investigate transporters, orphan genes, TFs and RBPs in \textit{C. elegans} and establish their functional associations with metabolic processes. By leveraging the power of WormClust\textsuperscript{141} and multiple expression datasets, we seek to expand our knowledge of \textit{C. elegans} metabolism and its regulation.
Results

60% of protein-coding *C. elegans* genes are poorly annotated.

Figure 3.2: Pie chart depicting percentage of *C. elegans* genes with different levels of UniProt Annotation Scores (left); Bar graph showing percentage of different categories of genes with different annotation scores (right).

To find the extent of functional characterization of *C. elegans* genes, we extracted annotation scores of all live, protein-coding genes from UniProt\(^\text{244}\). The annotation scores range from 1 to 5, with 1 indicating the most basic annotation and 5 representing the best annotated entries, which are mostly accompanied by experimental evidence. A significant portion of the *C. elegans* genome, approximately 60% of genes, had an annotation score of only 1, indicating poor annotation, and only 13% of the genes have a score of either 4 or 5, representing good annotation (Figure 3.2). Hence, most of the *C. elegans* genome remains a black box when it comes to functional characterization.
Among different categories of genes, orphan genes and non-metabolic genes exhibited the lowest levels of annotation, with 61% and 70% of genes having an annotation score of 1, respectively (Figure 3.2). Notably, orphan genes, which have the potential to be involved in metabolism, showed a particularly low level of annotation, indicating a significant gap in our knowledge of *C. elegans* metabolism. This highlights the need for further investigation and characterization of these orphan genes to uncover their metabolic roles.

The iCEL1314 genes, which constitute the metabolic network model, exhibited a higher level of annotation compared to other gene categories, with approximately 28% of genes having an annotation score of 4 or 5. However, it is important to note that even within the iCEL1314 model, 18% of genes had poor annotation. These genes were included in the model based on shared sequence homology with well-characterized metabolic genes, but further refinement of their annotations is necessary.

Here, we aim to associate orphan genes, TFs, transporters, and RBPs with known metabolic genes and pathways. By integrating experimental data and computational analyses, we seek to improve the annotations of these genes and contribute to filling the "missing" knowledge gap in *C. elegans* metabolism.
Majority of *C. elegans* genes are associated with metabolic pathways

To associate orphan genes, TFs, transporters, and RBPs with metabolic pathways, we used a comprehensive approach with four different datasets that cover various temporal, spatial, and condition-specific gene expression. We first, utilized a WormClust-based approach\textsuperscript{141} for each dataset to get gene-pathway associations, and then employed a scoring system to assign confidence to these associations (see details in methods).

For the temporal dimension, we utilized two distinct datasets (Figure 3.3). Firstly, we utilized the post-embryonic development dataset\textsuperscript{202}, which has been previously employed in Chapter-II\textsuperscript{141}, to examine gene expression profiles throughout different developmental stages. This dataset offers valuable information on the dynamic changes in gene expression from L1 to L4, providing insights into the temporal regulation of metabolic pathways.

Secondly, we incorporated an embryonic development dataset, comprised of RNA-seq to generate transcription profiles for seven specific embryonic cell populations\textsuperscript{245}. These profiles span the crucial stages of gastrulation to the onset of terminal differentiation, encompassing major cell lineages and tissues within the nematode. This dataset offers a comprehensive understanding of both spatial and temporal dynamics of gene expression during embryogenesis, shedding light on the intricate regulation of metabolic pathways during development (Figure 3.3).
For spatial dimension, in addition to using the single-cell RNA-Seq tissue dataset\textsuperscript{142} that was used before in Chapter II\textsuperscript{141}, to evaluate expression across seven tissues, we also used the expression data spanning 27 cell-types from the same dataset (Figure 3.3).

Furthermore, to explore gene coexpression patterns under different conditions, we utilized the same compendium of expression data that was employed in Chapter II\textsuperscript{141}. This dataset captures gene expression profiles across diverse experimental conditions, enabling us to investigate the condition-specific regulation of metabolic pathways (Figure 3.3).

Through our comprehensive analysis, we made a significant observation: the majority of \textit{C. elegans} genes (~91\%) exhibit associations with metabolic pathways across various temporal stages, spatial locations, and conditions (Figure 3.3).
Figure 3.3: Pie charts depicting percentage of iCEL1314 genes, orphan genes, RBPs, TFs, transporters, rest of non-metabolic genes and all C. elegans genes significantly associated (FDR<=0.05, NES>=2) with metabolic pathways across time, space and conditions. The green part depicts the significant association, whereas the red portion indicates non-significant association.

RBP are highly coexpressed with energy and growth genes.

To identify the strongest associations between genes and metabolic pathways, we applied a scoring system based on their presence across the four datasets (see Methods for details). Our analysis revealed notable coexpression patterns between genes and specific metabolic pathways related to energy (ETC
complexes) and growth processes (aminoacyl-tRNA biosynthesis) (Figure 3.4A). In order to determine the genes primarily enriched in coexpression with energy and growth pathways, we focused on the subset of genes that showed significant enrichment in these metabolic pathways (Figure 3.4B). Notably, RBPs exhibited a remarkable enrichment in coexpression with aminoacyl-tRNA biosynthesis and electron transport chain (ETC) complexes, both of which are closely linked to energy production and cellular growth (Figs 3.4C-D). The high coexpression of RBPs with aminoacyl-tRNA biosynthesis and ETC complexes indicates that energy and growth are potentially also highly regulated on a post-transcriptional level, if we assume that RBPs are mainly involved in post-transcriptional regulation.

It is important to consider that ribosomal proteins are also categorized as RBPs. Considering the crucial role of ribosomal proteins in ribosome assembly and function, their coexpression with energy and growth genes is expected. Ribosome biogenesis directly impacts translation capacity, thereby influencing protein synthesis and cellular growth. Additionally, ribosome biogenesis and translation processes are energetically demanding and require substantial cellular resources, making them intimately connected to cellular energy metabolism. To address the potential confounding effect of ribosomal proteins on the enrichment of coexpression between RBPs and energy/growth genes, we examined the percentage of significantly enriched RBPs that are ribosomal proteins. Intriguingly,
we found that only 21% of the significantly enriched RBPs were ribosomal proteins (Figure 3.4E). This indicates that the coexpression enrichment of RBPs with energy and growth genes cannot be solely attributed to the presence of ribosomal proteins. Moreover, when we removed ribosomal proteins from the analysis, RBPs still exhibited substantial coexpression with growth and energy-related genes (Fig3.4F-G). This finding further supports the notion that RBPs are intricately involved in the regulation of energy and growth processes.

**Transporters are highly enriched in coexpression with energy-related genes**

In addition to RBPs, we also observed a significant enrichment of coexpression between transporters and energy-related genes, suggesting their involvement in energy metabolism (Figure 3.4G). This highlights the role of transporters in facilitating the transport of molecules crucial for energy production and utilization within the cell.

Overall, these findings provide valuable insights into the intricate interplay between RBPs, energy metabolism, and cellular growth in *C. elegans*. Additionally, the significant coexpression between transporters and energy-related genes underscores the importance of transport processes in cellular energy homeostasis.
Figure 3.4: Significant enrichment of C. elegans orphan genes, RBPs, TFs and transporters with metabolic pathways

(A) Clustered heatmap showing score (from 0 to 4) of genes with metabolic pathways. Aminoacyl-tRNA synthesis, and ETC complex I, III, IV and V genes are labeled with brown and pink dots respectively.

(B) Heatmap indicating enrichment score of categorized genes with only energy and growth metabolic pathways
(C) Bar graph indicating the enrichment of coexpression of categorized genes (including ribosomal proteins) with AA-tRNA biosynthesis

(D) Bar graph indicating the enrichment of coexpression of categorized genes (including ribosomal proteins) with ETC complexes I, III, IV and V.

(E) Pie chart depicting percentage of ribosomal proteins and remaining RBPs, that are significantly enriched to AA-tRNA biosynthesis and ETC complexes.

(F) Bar graph indicating the enrichment of coexpression of categorized genes (excluding ribosomal proteins) with AA-tRNA biosynthesis

(G) Bar graph indicating the enrichment of coexpression of categorized genes (excluding ribosomal proteins) with ETC complexes I, III, IV and V.

This result is further corroborated by examining the enrichment of coexpression of orphan genes, transporters, TFs, and RBPs genes with metabolic communities\cite{141} (refer to Figure 2.13A). Interestingly, among these categories, RBPs exhibited the highest enrichment in communities 1 (energy) and 3 (growth), providing additional evidence for their close association with energy and growth-related processes (Figure 3.5A-B).

![Heatmap showing scores of genes (categorized into orphan genes, RBPs, TFs and transporters) with clustered communities. The color legends are indicated on the rightmost corner of the figure.

![Biclustered heatmap showing score of clustered genes with metabolic communities.](image-url)

**Figure 3.5: Significant enrichment of orphan genes, RBPs, TFs and transporters to metabolic communities 1, 2 and 3 (Figure 2.13A)**

(A) Heatmap showing scores of genes (categorized into orphan genes, RBPs, TFs and transporters) with clustered communities. The color legends are indicated on the rightmost corner of the figure.

(B) Biclustered heatmap showing score of clustered genes with metabolic communities.
Transporters associated with metabolism

To find potential transporters associated with metabolism, we only focused on transporter-metabolic pathway relationships with a score of more than 1 i.e., when the same associations exist in more than one dataset (Figure 3.6). We further individually inspected all the pathways to determine its strong associations with transporters. We observed that many transporters are associated with Vacuolar ATPases, ETC complex-I, peroxisomal FA degradation and aminoacyl-tRNA synthesis (Figs 3.6 and 3.7A, File 3.1).
Figure 3.6: Heatmaps showing clustering of transporter genes significantly enriched (FDR ≤ 0.05 and NES ≥ 2) to the metabolic pathways, based on score, before and after filtering for score>1 criteria.
**Vacuolar ATPases facilitate the transport of multiple metabolites**

Vacuolar ATPases play a critical role in cellular physiology by acidifying various intracellular compartments, a process crucial for a wide range of metabolic transport activities\textsuperscript{247–249}. Our findings highlight the significant association of numerous transporter genes with these ATPases, further reinforcing the breadth of their functional implications. Of particular interest is the \textit{rhr-1} gene (Rhesus antigen related), which is associated with the highest score (4), indicating its prevalence across all datasets in our study (Figure 3.7B, Table 3.1). Our extensive investigation through WormBase and a comprehensive literature review reveals the involvement of \textit{rhr-1} in ammonium transmembrane activity. Intriguingly, its human orthologs – RHAG (Rh associated glycoprotein), RHBG (Rh family B glycoprotein), and RHCG (Rh family C glycoprotein) – have been implicated in various physiological and pathological conditions including Rh deficiency syndrome, Rh isoimmunization, and overhydrated hereditary stomatocytosis. In humans, the interplay between the NH\textsubscript{3} channel RhCG and the vacuolar-type H\textsuperscript{+}-ATPase pump is crucial in renal ammoniagenesis\textsuperscript{250}. Specifically, ammonium serves as a principal urinary proton buffer and is eliminated via the collecting duct. This intricate process is contingent on the concurrent secretion of ammonia (NH\textsubscript{3}) by RhCG and protons (H\textsuperscript{+}) by vacuolar ATPase, demonstrating a clear linkage between urinary ammonium content and urinary acidification\textsuperscript{251}. Interestingly, our
findings suggest a similar functional paradigm in the *C. elegans* model, where vacuolar ATPases likely acidify the environment to facilitate ammonium transport.

Furthermore, our study identified transporter genes associated with a diverse spectrum of metabolites, significantly coexpressed with vacuolar ATPases. These genes, exhibiting reciprocal and best-reciprocal blasts with human homologs, accentuate the broad physiological relevance of vacuolar ATPases (Figure 3.7B). Our results, therefore, provide evidence to postulate that in addition to their role in ammonium transport, vacuolar ATPases likely orchestrate the transport of various other metabolites by probably acidifying the environment.

Table 3.1: High-scoring transporters significantly enriched to metabolic pathways with score of 3 or 4

<table>
<thead>
<tr>
<th>WormBase ID</th>
<th>Gene</th>
<th>Pathway</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBGene00004061</td>
<td>pmp-4</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00004358</td>
<td>rhr-1</td>
<td>VACUOLAR ATP-ASE</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00006572</td>
<td>tin-9.1</td>
<td>ETC-COMPLEX I</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00008857</td>
<td>timm-23</td>
<td>ETC-COMPLEX I</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00012714</td>
<td>abce-1</td>
<td>AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00018339</td>
<td>abcf-3</td>
<td>AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00019187</td>
<td>H11E01.2</td>
<td>PROPIONATE DEGRADATION CANONICAL</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000178</td>
<td>aqp-10</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00002045</td>
<td>icd-1</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00002045</td>
<td>icd-1</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00006494</td>
<td>zipt-7.2</td>
<td>COLLAGEN BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00006512</td>
<td>abcf-1</td>
<td>AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00006573</td>
<td>tin-10</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Description</td>
<td>Count</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------------------------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>WBGene00006574</td>
<td>tin-13 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00008505</td>
<td>PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00008799</td>
<td>F14E5.1 PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00009092</td>
<td>tomm-20 PORPHYRIN METABOLISM</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00010644</td>
<td>K07G5.5 PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00011119</td>
<td>mpc-1 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00015186</td>
<td>misc-1 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00017798</td>
<td>slc-17.8 TAURINE AND HYPOTAURINE METABOLISM</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018765</td>
<td>mpc-2 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00019875</td>
<td>mca-2 VACUOLAR ATP-ASE</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00019900</td>
<td>vdac-1 ETC-COMPLEX V</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00021133</td>
<td>tomm-22 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00022042</td>
<td>icd-2 ETC-COMPLEX V</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00022252</td>
<td>Y73B6BL.31 PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00022722</td>
<td>tomm-70 AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene0000000003</td>
<td>aat-2 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene0000000003</td>
<td>aat-2 AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene0000000006</td>
<td>aat-5 GLYCOLYSIS/GLUCONEOGENESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene000000175</td>
<td>aqp-7 PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene000000175</td>
<td>aqp-7 GLYCOLYSIS/GLUCONEOGENESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene000000179</td>
<td>aqp-11 PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene000000533</td>
<td>clh-6 AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene000002975</td>
<td>lev-8 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene000002975</td>
<td>lev-8 ETC-COMPLEX V</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00003997</td>
<td>pgp-3 PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00006647</td>
<td>tsr-1 AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00006683</td>
<td>twk-31 GLYCOLYSIS/GLUCONEOGENESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00006692</td>
<td>twk-42 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Gene Name</td>
<td>Gene Description</td>
<td>Accession</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>WBGene00007686</td>
<td><strong>tomm-40</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00007686</td>
<td>PORPHYRIN METABOLISM</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00007686</td>
<td><strong>tomm-40</strong> AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00008396</td>
<td>D1086.9 PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00008774</td>
<td><em>F13H10.3</em> AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00008857</td>
<td><strong>timm-23</strong> ETC-COMPLEX IV</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00008998</td>
<td><strong>ncs-4</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00009092</td>
<td><strong>tomm-20</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00009092</td>
<td><strong>tomm-20</strong> AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00010834</td>
<td><strong>mct-3</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00011358</td>
<td><strong>lgc-16</strong> CHITIN DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00012804</td>
<td><strong>slc-36.1</strong> AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00012857</td>
<td><strong>pbo-5</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00013088</td>
<td><strong>Y51B9A.6</strong> VACUOLAR ATP-ASE</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00013486</td>
<td><strong>egas-1</strong> CHITIN DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00013740</td>
<td><strong>hut-1</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00015940</td>
<td><strong>zipt-3</strong> AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00016705</td>
<td><strong>slc-46.1</strong> PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00017681</td>
<td><strong>slc-17.5</strong> PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018270</td>
<td><strong>eas-1</strong> AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018934</td>
<td><strong>F56B3.11</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00019018</td>
<td><strong>abts-3</strong> GLYCOLYSIS/GLUCONEOGENESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00019493</td>
<td><strong>catp-5</strong> PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00019900</td>
<td><strong>vdac-1</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00020836</td>
<td><strong>lgc-34</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00021948</td>
<td><strong>kvs-5</strong> ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00022270</td>
<td><strong>Y73E7A.3</strong> PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00022281</td>
<td><strong>abtm-1</strong> PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Abbreviation</td>
<td>Description</td>
<td>Score</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------</td>
<td>-------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>WBGene00022281</td>
<td>abtm-1</td>
<td>ASCAROSIDE BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00045385</td>
<td>lgc-24</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
</tbody>
</table>
Many transporters are coexpressed with peroxisomal FA degradation

Intriguingly, a substantial number of transporter genes demonstrate a coexpression pattern with peroxisomal FA degradation. Among these, *pmp-4* is the top-ranking transporter gene associated with this metabolic pathway (Figure 3.7C, Table 3.1). Phylogenetic analyses have revealed that *pmp-4* in *C. elegans* is orthologous to mammalian peroxisomal transporters ABCD1 and ABCD2, effectively suggesting that *pmp-4* is an ancestral homolog. PMP-4 shares a considerable sequence similarity of 75% with ABCD1\textsuperscript{252,253}, further corroborating the credibility and efficacy of our approach for identifying transporters interconnected with metabolic pathways.

Another gene of considerable interest is *slc-17.5*, which, given its enrichment score of 3, could be a robust indicator of peroxisomal transporter
activity. The human ortholog of slc-17.5, SLC17A5, encodes the protein Sialin. In humans, sialin is a lysosomal membrane transporter, primarily responsible for extruding free sialic acid from lysosomes. A deficiency in sialin culminates in an abnormal accumulation of free sialic acid within lysosomes, leading to a condition known as Sialuria^{254}. However, considering that sialic acid is absent in C. elegans, it warrants further investigation to determine if slc-17.5 might be peroxisomal in C. elegans and the specific organic acid or a fatty acid that it might be transporting in this model organism. Additionally, a number of other genes are potential candidates in relation to peroxisomal or fatty-acid transport activity. These include abtm-1, Y73E7A.3, pgp-3, Y73B6BL.31, aqp-11, K07G5.5, F14E5.1, slcr-46.1, and catp-5 (Figure 3.7C, Table 3.1). These genes, each with a score of 3, present promising avenues for future research, warranting in-depth investigation to fully comprehend their roles and implications.

_Transporters enriched to aminoacyl-tRNA biosynthesis are mainly involved in transporting amino acids or tRNA import_

Our study has unveiled an intricate network of transporter genes linked with aminoacyl-tRNA synthetases, revealing the essential integration of transporter proteins in aminoacyl-tRNA biosynthesis (Figure 3.7B). A remarkable observation is the association of multiple high-scoring genes, specifically abcf-3, abce-3, F13H10.3, aat-2, abcf-1, laat-1, and abcf-2, with amino acid transport (Figure
This implicates these genes as potentially key players in the translocation of amino acids, which forms the backbone of protein synthesis.

Further enriching our findings, we have identified *tomm-20* and *tsr-1* genes, whose function is associated with tRNA import (Figure 3.7D, Table 3.1, File 3.1). tRNA import is an integral step in protein biosynthesis, facilitating the precise aminoacylation of tRNAs, thus enabling accurate translation of genetic information.

This analysis has not only validated our methodological approach, but it also brings into light the crucial role of transporter genes in aminoacyl-tRNA biosynthesis, specifically in the contexts of amino acid transport and tRNA import.

**Orphan genes associated with metabolism**

94% of orphan genes are associated with metabolism across at least one dataset (Figure 3.3, File 3.2). To view the highest-confidence associations, we constructed a network graph with genes and pathways as nodes, and score as edges, as well as clustered heatmap of gene-pathway enrichment scores (Figs 3.8 and 3.9). We observed that many unique genes are significantly coexpressed with peroxisomal fatty acid degradation, and not with any other pathway (Figs 3.8 and 3.9). To further filter the highest-confidence genes, we looked at orphan gene-pathway associations with a score of 3 or 4 only (Figure 3.10A, Table 3.2).
Figure 3.8: Network graph showing significant orphan gene-pathway associations with an score of more than 1.
The red and blue dots indicate pathways and genes respectively. The edges are colored according to enrichment score. Color legend is indicated.
Figure 3.9: Clustered heatmap showing the score of orphan genes based on their significant enrichment to coexpression, to metabolic pathways. Scores less than 2 have been masked.

Table 3.2: High-scoring orphan metabolic genes significantly associated with metabolic pathways (Score of 3 or 4 only)

<table>
<thead>
<tr>
<th>WormBase ID</th>
<th>Gene</th>
<th>Pathway</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBGene00000139</td>
<td>amx-3</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000970</td>
<td>dhs-6</td>
<td>MITOCHONDRIAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000981</td>
<td>dhs-18</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00003065</td>
<td>lpd-9</td>
<td>ETC-COMPLEX I</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00007228</td>
<td>acs-7</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00008787</td>
<td>F14B6.6</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Name</td>
<td>Description</td>
<td>Value</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------</td>
<td>--------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>WBGene00010993</td>
<td>vha-20</td>
<td>VACUOLAR ATP-ASE</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00013594</td>
<td>Y87G2A.2</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00013807</td>
<td>Y116A8C.30</td>
<td>ETC-COMPLEX I</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00015248</td>
<td>mai-2</td>
<td>ETC-COMPLEX I</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00015248</td>
<td>mai-2</td>
<td>ETC-COMPLEX V</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00016258</td>
<td>vha-16</td>
<td>VACUOLAR ATP-ASE</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00021073</td>
<td>nsun-1</td>
<td>AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00022260</td>
<td>Y73C8B.3</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000970</td>
<td>dhs-6</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003235</td>
<td>mif-2</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00006435</td>
<td>stdh-4</td>
<td>CHITIN DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00007099</td>
<td>trx-2</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00007099</td>
<td>trx-2</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00007099</td>
<td>trx-2</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00007143</td>
<td>hacl-1</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00007593</td>
<td>C14H10.3</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00008860</td>
<td>romo-1</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00009818</td>
<td>acbp-3</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00010983</td>
<td>mocs-1</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00012544</td>
<td>Y37D8A.2</td>
<td>PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00012860</td>
<td>acdh-11</td>
<td>MITOCHONDRIAL FATTY ACID DEGRADICATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00013029</td>
<td>glrx-5</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00013807</td>
<td>Y116A8C.30</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00015284</td>
<td>cest-5.2</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00015336</td>
<td>C02D5.2</td>
<td>COLLAGEN BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00015709</td>
<td>cyp-33A1</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00015755</td>
<td>C14B9.10</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00015755</td>
<td>C14B9.10</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016325</td>
<td>C32E8.9</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00017885</td>
<td>dip-2</td>
<td>INOSITOL PHOSPHATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Symbol</td>
<td>Description</td>
<td>Category</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>--------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>WBGene00018240</td>
<td>cox-17</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00020756</td>
<td>T24A6.20</td>
<td>VACUOLAR ATP-ASE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00021647</td>
<td>Y47G6A.22</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00021647</td>
<td>Y47G6A.22</td>
<td>PROPIONATE DEGRADATION SHUNT</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00022380</td>
<td>Y94H6A.8</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00022380</td>
<td>Y94H6A.8</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00000970</td>
<td>dhs-6</td>
<td>ASCAROSIDE BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003124</td>
<td>mai-1</td>
<td>GLYCOLYSIS/GLUCONEOGENESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003124</td>
<td>mai-1</td>
<td>PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00007551</td>
<td>C13C4.7</td>
<td>PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00009436</td>
<td>marb-1</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00010015</td>
<td>atad-3</td>
<td>PORPHYRIN METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00010721</td>
<td>K09E4.3</td>
<td>AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00011092</td>
<td>R07B7.8</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00011123</td>
<td>nuf-3</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00011676</td>
<td>cyp-13A2</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00011681</td>
<td>T10B10.4</td>
<td>PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00013804</td>
<td>Y116A8C.27</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00014244</td>
<td>ZK1307.1</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00015481</td>
<td>C05D11.1</td>
<td>AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016116</td>
<td>C25H3.7</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016442</td>
<td>C35D10.5</td>
<td>PORPHYRIN METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016630</td>
<td>acer-1</td>
<td>PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016630</td>
<td>acer-1</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016630</td>
<td>acer-1</td>
<td>GLYCOLYSIS/GLUCONEOGENESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016630</td>
<td>acer-1</td>
<td>GLYOXYLATE AND DICARBOXYLATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016655</td>
<td>acbp-1</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00016655</td>
<td>acbp-1</td>
<td>ASCAROSIDE BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Description</td>
<td>Score</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------------------------------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>WBGene00017065</td>
<td>D2092.4 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018046</td>
<td>coa-3 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018046</td>
<td>coa-3 AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018340</td>
<td>anmt-2 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018963</td>
<td>ucr-1 TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018963</td>
<td>ucr-1 ETC-COMPLEX IV</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018963</td>
<td>ucr-1 ETC-COMPLEX V</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018963</td>
<td>ucr-1 PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00018963</td>
<td>ucr-1 GLYCOLYSIS/GLUCONEOGENESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00019301</td>
<td>K02D10.1 PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00019843</td>
<td>R03E9.2 VALINE DEGRADATION</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00020113</td>
<td>R151.10 AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00020113</td>
<td>R151.10 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00021932</td>
<td>cox-18 AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00021933</td>
<td>Y55F3AR.2 N-GLYCAN BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00022195</td>
<td>Y71H2B.5 AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00022260</td>
<td>Y73C8B.3 ASCAROSIDE BIOSYNTHESIS</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00022516</td>
<td>mtx-2 ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

**Orphan genes associated with vacuolar ATPases**

As a validation, *vha-20* and *vha-16*, vacuolar ATPases that are not part of the original iCEL model because they were only recently annotated by WormBase and KEGG, shows significant enrichment to vacuolar ATPases with score 4 (Figure 3.10A, Table 3.2).
Figure 3.10: Orphan genes significantly associated with metabolic pathways

(A) Heatmap showing FDR values of orphan gene-pathway associations with only enrichment scores 3 and 4.
(B) Suggested addition of acs-7 gene to peroxisomal fatty acid degradation pathway

(C) Suggested addition of Y116A8C.30 and Y94H6A.8 genes as ETC Complex I subunits

(D) Suggested addition of acer-1 to propionate degradation, glycolysis/glucogenesis and glyoxylate/dicarboxylate metabolism pathways.

Orphan genes associated with peroxisomal FA degradation

We found that amx-3 exhibits strong coexpression with peroxisomal FA degradation, accompanied by a score of 4 (Figure 3.10A, Table 3.2). The amx-3 gene encodes amino oxidase/oxidative deaminase\textsuperscript{30,204}. In other nematodes like Caenorhabditis briggsae, as well as in humans and plants, the orthologous genes encode spermine oxidase, which catalyzes the conversion of spermine to spermidine, producing H2O2 as a by-product. Coincidentally, H2O2 is also generated during peroxisomal FA degradation (Figure 3.10B). Additionally, the enzyme catalase eliminates this H2O2 by converting it to O2 and H2O within peroxisomes. Consequently, we propose that the strong connection between amx-3 and peroxisomal fatty acid oxidation arises from the involvement of H2O2 in maintaining reactive oxygen species (ROS) homeostasis.

Furthermore, we identified acs-7 as highly coexpressed with peroxisomal FA degradation, displaying a score of 4 (Figure 3.10A, Table 3.2). The acs-7 gene in C. elegans is known to encode an acyl-CoA synthetase, which belongs to a family of enzymes involved in fatty acid metabolism. Acyl-CoA synthetases
activate fatty acids by converting them into fatty acyl-CoA esters, which can then participate in various metabolic processes, including beta-oxidation. Beta-oxidation is a multi-step process occurring in peroxisomes and mitochondria, responsible for breaking down fatty acids to generate energy. This association suggests the inclusion of acs-7 in the peroxisomal FA degradation pathway (Figure 3.10B).

Y73C8B.3 is another prominent example of an orphan gene associated with peroxisomal FA degradation and ascaroside biosynthesis with scores 4 and 3 respectively (Figure 3.10A, Table 3.2). It is a protein of completely unknown function and requires experimental intervention to understand its association to peroxisomal FA degradation and ascaroside biosynthesis, which share common metabolic genes. Similarly, acbp-1, an acyl CoA binding protein is strongly associated to peroxisomal FA degradation and ascaroside biosynthesis (Figure 3.10A, Table 3.2). ACBP-1 binds to long-chain fatty acyl-CoA molecules, facilitating their transport within the cell. By binding to and solubilizing acyl-CoA, ACBP-1 aids in shuttling fatty acids between different cellular compartments, including peroxisomes where fatty acid degradation occurs. This suggests that ACBP-1 may contribute to the efficient transport of fatty acids for peroxisomal degradation.

Y87G2A.2 is another strong candidate orphan gene for association with peroxisomal FA degradation with score of 4 (Figure 3.10A, Table 3.2). According
to WormBase, it is predicted to enable acyl-CoA hydrolase activity and involved in fatty acid catabolic process\textsuperscript{204}. It is also predicted to be located in peroxisomal matrix. Its human ortholog is ACOT8 (acyl-CoA thioesterase 8). Acyl-CoA thioesterase II, depending on its specific localization and substrate preference, could theoretically help control the levels of different acyl-CoA esters entering the  \(\beta\)-oxidation pathway, thus regulating the overall process of peroxisomal FA degradation. In the context of \textit{C. elegans}, the exact role and regulation of acyl-CoA thioesterase II and its relation to peroxisomal fatty acid degradation would need to be elucidated by experimental research.

Two \textit{C. elegans} orphan genes \textit{cyp-13A2} and \textit{cyp-33A1}, putative members of the cytochrome P450 family, are highly associated with peroxisomal FA degradation (\textbf{Figure 3.10A}, \textbf{Table 3.2}). As per WormBase, these genes are predicted to function in heme binding, oxidoreductase activity, and steroid hydroxylation, and are thought to participate in organic acid and xenobiotic metabolic processes\textsuperscript{204}. In mice, similar cytochrome P450 enzymes metabolize long chain fatty acids into toxic dicarboxylic acids, which undergo peroxisomal beta-oxidation\textsuperscript{257}. We hypothesize that \textit{cyp-13A2} and \textit{cyp-33A1} might fulfill comparable roles in \textit{C. elegans}.

In addition, \textit{dhs-18}, a dehydrogenase, shows a strong link to peroxisomal FA degradation (\textbf{Figure 3.10A}, \textbf{Table 3.2}). This process involves several enzymatic steps, including dehydrogenation, where DHS-18 could potentially
function. However, without experimental data, its exact substrate specificity remains uncertain. Another gene hacl-1, strongly linked to peroxisomal FA degradation, is predicted to bind thiamine pyrophosphate and participate in fatty acid alpha-oxidation (Figure 3.10A, Table 3.2). Located in the peroxisome, it's an ortholog of human HACL1. We hypothesize that hacl-1 may play a role in peroxisomal FA alpha-oxidation in *C. elegans*.

**Orphan genes associated to electron transport chain**

As seen before, there are many orphan genes associated with ETC complexes (Figures 3.8 and 3.9). Prominently, Y116A8C.30 is associated to ETC complex I with a score of 4 and to ETC Complex V with a score of 3 (Figure 3.10A, Table 3.2). Interestingly, according to WormBase, it is predicted to be involved in mitochondrial respiratory chain complex I assembly and located in membrane. Human ortholog(s) of this gene are implicated in nuclear type mitochondrial complex I deficiency. This is also corroborated by its best-scoring KO in WormFlux according to SACURE annotation. We propose that it should be added to ETC Complex I as a subunit in *C. elegans* (Figure 3.10C).

Similarly, the orphan gene Y94H6A.8 in *C. elegans* shows a strong association with ETC complexes I, III, and V (Figure 3.10A, Table 3.2), underscoring its significant role in the electron transport chain. It's predicted to be involved in the response to oxidative stress and is located within the
mitochondrion. As an ortholog of human NDUFA12, a subunit of the NADH:ubiquinone oxidoreductase complex (Complex I), it’s implicated in human mitochondrial Complex I deficiencies\(^{258-262}\). Given these associations, Y94H6A.8 warrants consideration as an integral subunit of ETC Complex I in \textit{C. elegans} (Figure 3.10C).

The orphan gene \textit{mai-2} is significantly linked to both ETC Complex I and V, each with a score of 4 (Figure 3.10A, Table 3.2). It exhibits ATPase inhibitor activity, implying a role in the negative regulation of ATP-dependent activity, and is located in the mitochondrion\(^ {204}\). As an ortholog of human ATP5IF1, the ATP synthase inhibitory factor subunit 1, it may have similar functions. Previous research has found that in various cell lines, when the mitochondrial electrochemical proton gradient is disrupted, IF1 (Inhibitor Factor-1) restricts the reverse hydrolytic activity of the F1Fo-ATP synthase. This allows cells to conserve ATP while sacrificing the mitochondrial membrane potential (\(\Delta \psi \text{m}\))\(^ {263-266}\). Given \textit{mai-2}’s substantial association with ETC Complexes I and V, we suggest that if ATP levels decrease due to increased \textit{mai-2} activity, these complexes might upregulate to counterbalance the energy deficiency. This suggests a potential cellular compensation mechanism in response to energy shortfalls.

\textit{lpd-9} (lipid depleted), surprisingly shows strong association to ETC Complex I with a score of 4 (Figure 3.10A, Table 3.2). In a previous single study, it was found that the \textit{C. elegans} gene \textit{lpd-5}, an ortholog of the human NDUFS4
gene, is crucial for Complex I function in the electron transport chain (ETC). Depletion of \textit{lpd-5} replicates biochemical, cellular, and neurodevelopmental aspects of human mitochondriopathies, emphasizing its role in these conditions\textsuperscript{267}. \textit{lpd-5} is also annotated as an ETC complex gene in WormFlux and highly coexpressed to other ETC genes in WormClust\textsuperscript{18,141}. We hypothesize that \textit{lpd-9} functions like \textit{lpd-5}. This warrants more experimental validation.

The orphan gene \textit{trx-2} in \textit{C. elegans}, a thioredoxin ortholog of human TXN2, is significantly linked (score 3) to ETC complexes I, III, and V (Figure 3.10A, Table 3.2), highlighting its crucial role in the electron transport chain. This gene, expressed in several tissues including AIY, ASEL, body wall musculature, and vulval muscle, is predicted to contribute to protein-disulfide reductase activity and cell redox homeostasis, and is located in the mitochondrion\textsuperscript{204}. While not a direct component of the ETC, \textit{trx-2} is integral to cellular redox balance and may influence ETC functionality. It could also safeguard ETC complexes from oxidative damage as the ETC is a major source of cellular reactive oxygen species\textsuperscript{268,269}. Studies in mice have found that the orthologous TrxR2 enhances both TCA cycle and ETC function, providing metabolic disease protection\textsuperscript{270}. Extrapolating from these findings, we propose that \textit{trx-2} might serve a similar function in \textit{C. elegans}, modulating ETC activity.

The gene C14B9.10 in \textit{C. elegans} shows a strong association with ETC complexes (Figure 3.10A, Table 3.2). This gene is predicted to be involved in
mitochondrial electron transport from ubiquinol to cytochrome c and is expected to be located in the mitochondrial inner membrane and respirasome. It's also anticipated to be a part of the mitochondrial respiratory chain complex III\textsuperscript{204}. As an ortholog of human UQCR10, a subunit of complex III, C14B9.10 should be considered an integral component of \textit{C. elegans}' ETC complex III, given its substantial associations (\textbf{Figure 3.10D}). It was previously not annotated by KEGG at the time of the reconstruction of our metabolic network model\textsuperscript{271}, but it is now annotated as part of complex III, hence also serving as validation.

\textit{ucr-1} is another gene that was previously not annotated and shows significant associations to ETC complexes (\textbf{Figure 3.10A, Table 3.2}). This gene, is now identified by KEGG as Cytochrome b-c1 complex subunit 1, a critical component of mitochondrial function. This annotation positions \textit{ucr-1} as an integral participant in the ETC processes. Interestingly, its human ortholog, UQRCRC1, is implicated in the operation of Complex III, a key player in oxidative phosphorylation, and is associated with Alzheimer's disease. This newly assigned identity of \textit{ucr-1} highlights its essential role in the ETC functionality within \textit{C. elegans} and also validates our approach.

Similarly, numerous orphan genes, including C25H3.7, C32E8.9, D2092.4, R151.10, Y116A8C.27, Y47G6A.22, \textit{acer-1}, \textit{anmt-2}, \textit{cox-3}, \textit{cox-17}, \textit{glrx-5}, \textit{marb-1}, \textit{mif-2}, \textit{mtx-2}, and \textit{romo-1}, exhibit strong associations with one or more ETC
complexes (Figure 3.10A). However, their precise roles within the electron transport chain remain to be elucidated.

**Orphan genes associated to Glycolysis/ Gluconeogenesis, canonical propionate degradation pathway, and glyoxylate/decarboxylate metabolism**

Glycolysis/ Gluconeogenesis, propionate degradation, and glyoxylate/decarboxylate metabolism are interconnected via shared metabolites. Pyruvate, a glycolysis product, can be converted into acetyl-CoA, which is involved in many metabolic pathways, including TCA cycle. Propionate degradation converts propionyl-CoA, a byproduct of branched chain amino acid breakdown, into succinyl-CoA, an intermediate in the TCA cycle and substrate for gluconeogenesis. The glyoxylate shunt, an alternative to the TCA cycle, conserves carbon units, synthesizing glucose from acetyl-CoA, thus linking all three processes.

The orphan gene acer-1 has notable associations with the canonical propionate degradation pathway, glycolysis/gluconeogenesis, and glyoxylate/dicarboxylate metabolism in *C. elegans* (Figure 3.10A, Table 3.2). *acer-1* is predicted to perform dual enzymatic functions: CoA-transferase and acetyl-CoA hydrolase activities. In its role as a CoA-transferase, *acer-1* likely facilitates the transfer of CoA groups between metabolic intermediates, significantly influencing processes such as propionate degradation, where it may assist in converting propionyl-CoA to propionate, and possibly also
glyoxylate/dicarboxylate metabolism (Figure 3.10D). Additionally, as an acetyl-CoA hydrolase, *acer-1* is integral to the conversion of acetyl-CoA to acetate. Given acetyl-CoA’s central role as a metabolic intermediate linking carbohydrate, protein, and lipid metabolism, *acer-1*’s enzymatic functions are potentially key to maintaining metabolic homeostasis and energy balance within *C. elegans* cells. Therefore, *acer-1* emerges as a crucial player in the interconnected metabolic networks of glycolysis/ gluconeogenesis, propionate degradation, and glyoxylate/decarboxylate metabolism, thanks to its multifaceted enzymatic activities.

*mai-1* shows strong association to canonical propionate degradation and glycolysis/ gluconeogenesis pathways (Figure 3.10A, Table 3.2). It is an ATPase inhibitor in *C. elegans*, may indirectly influence glycolysis/gluconeogenesis and propionate degradation by modulating ATP synthase activity and cellular ATP levels. Despite this potential connection, direct evidence of MAI-1’s involvement in these metabolic pathways remains unverified.

In conclusion, the strong associations observed between the orphan genes and key metabolic pathways highlight the potential significance of these orphan genes in regulation of metabolic pathways.

**TFs associated with metabolism**

Most TFs are collectively associated with metabolism across time, space and conditions (Figure 3.3, File 3.3). To view the most high-confidence associations,
we constructed a network graph with genes and pathways as nodes, and enrichment score as edges, as well as clustered heatmap of gene-pathway enrichment scores (Figures 3.11A and 3.11B). We observed that many unique genes are significantly coexpressed to different pathways, forming distinct clusters of pathways such as inositol phosphate metabolism, chitin degradation, vacuolar ATPases, and peroxisomal FA degradation and ascaroside biosynthesis (Figures 3.11A and 3.11B). Since we have seen past examples of NHRs being linked to metabolism, we examined only nhr genes first for their potential associations to metabolic pathways144,145,158,159,221,222,272,273 (Figure 3.12).
Figure 3.11: Significant TF-Pathway associations

(A) Network graph showing significant TF-pathway associations with a score of more than 1.

(B) Clustered heatmap showing the score(>2) of orphan genes based on their significant enrichment to coexpression to metabolic pathway
**NHRs associated with metabolism**

We found that *nhr-114* is significantly coexpressed to Met/SAM cycle, methionine salvage pathway, and folate cycle (connected to Met/SAM cycle), which is expected and known from a previous study\(^{145}\) (**Figure 3.12**). However, it is interesting to see that it is also coexpressed significantly with mitochondrial and peroxisomal FA degradation (**Figure 3.12, Table 3.3**).

*nhr-31* shows enrichment to vacuolar ATPases with a high score of 3 i.e. the association has been observed across three out of four datasets (**Figure 3.12, Table 3.3**). This is also known from a previous study\(^ {221}\). Similarly, other known TFs such as *nhr-68* and *nhr-79* show the expected enrichment to propionate shunt\(^ {144}\) , and peroxisomal FA degradation\(^ {222}\) respectively (**Figure 3.12**).

The association of *nhr-177* with peroxisomal FA degradation has the highest score of 4 (**Figure 3.12, Table 3.3**), which is more than what we see for our known *nhr*-pathway associations. Therefore the relationship of *nhr-177* definitely needs experimental intervention and validation. Other interesting associations to investigate and validate with scores of 3 are *nhr-254* with chitin degradation, *nhr-43* with chitin degradation, and surprisingly high number of associations of peroxisomal FA degradation with multiple *nhrs*, especially *nhr-13, nhr-121, nhr-114, nhr-68* and *nhr-22* (**Figure 3.12, Table 3.3**). In addition to NHRs, it might be interesting to see the top scores of other TF genes.
Figure 3.12: Significant NHR-metabolic pathway associations with a score of 2 or more
Other TFs associated with metabolism

To find more high-confidence associations of non-NHR TFs with metabolic pathways, we first looked at other TFs that are not NHRs, only with an enrichment score of 3 or 4 (Figure 3.13A, Table 3.3). The highest score is obtained by prx-5’s association to peroxisomal FA degradation (Figure 3.13A, Table 3.3). This highly validates our approach since PRX-5 is thought to be involved in the import of proteins into peroxisomes. As peroxisomes house enzymes vital to the degradation of fatty acids, defects in peroxisomal protein import can negatively impact fatty acid metabolism. A study reported that mutations in the prx-5 gene in C. elegans could lead to defects in peroxisomal protein import and consequential effects on fatty acid metabolism\textsuperscript{274}.

\textit{F23B12.7} shows high enrichment of coexpression to aminoacyl-tRNA biosynthesis with a score of 3 (Figure 3.13A, Table 3.3). Its human homolog CEBPZ encodes a ribosome biogenesis protein. This suggests a coordination between aminoacyl-tRNA biosynthesis and ribosome biogenesis, potentially regulated by F23B12.7, to ensure that the protein synthesis machinery can respond dynamically to the cell’s needs. Aminoacyl-tRNA synthetases provide the charged tRNA molecules that are used as substrates in protein synthesis on the ribosomes. In the ribosome, the anticodon of the charged tRNA molecule base-pairs with the codon on the mRNA molecule, ensuring that the correct amino acid is added to the growing polypeptide chain. Therefore, the availability and accuracy
of charged tRNA molecules, provided by the process of aminoacyl-tRNA synthesis, is critical for the proper function of the ribosomes and for accurate and efficient protein synthesis.

There are many TF genes associated to Inositol phosphate metabolism (Figures 3.11A, 3.11B, 3.13A, 3.13B, File 3.3). Amongst them, the high scoring ones are $F_{26}H9.2$, $attf-5$, $ceh-32$, $hmg-1.2$ and $hmg-11$, with score of 3 (Figure 3.13A, Table 3.3). Another interesting candidate to experimentally validate is $zip-3$ that shows enrichment to propionate degradation shunt and peroxisomal FA degradation, and ascaroside biosynthesis with scores 3 and 2 respectively (Figures 3.13A and 3.13B, File 3.3, Table 3.3). $zyx-1$ shows enrichment to energy generating and propionate degradation genes (Figures 3.13A and 3.13B, File 3.3, Table 3.3).

### Table 3.3: High-scoring TFs significantly enriched to metabolic pathways with a score of 3 or 4

<table>
<thead>
<tr>
<th>WormBase ID</th>
<th>Gene</th>
<th>Pathway</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBGene00004194</td>
<td>$prx-5$</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00017503</td>
<td>$nhr-177$</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00003625</td>
<td>$nhr-31$</td>
<td>VACUOLAR ATP-ASE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003633</td>
<td>$nhr-43$</td>
<td>HISTIDINE DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003704</td>
<td>$nhr-114$</td>
<td>METHIONINE SALVAGE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003711</td>
<td>$nhr-121$</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00006999</td>
<td>$zyx-1$</td>
<td>PROPIONATE DEGRADATION CANONICAL</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00009084</td>
<td>$F_{23}B12.7$</td>
<td>AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGeneID</td>
<td>Symbol</td>
<td>Description</td>
<td>Score</td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
<td>-----------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>WBGene00012330</td>
<td>zip-3</td>
<td>PROPIONATE DEGRADATION SHUNT</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00020214</td>
<td>attf-5</td>
<td>INOSITOL PHOSPHATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00022640</td>
<td>nhr-254</td>
<td>CHITIN DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00000453</td>
<td>ceh-32</td>
<td>INOSITOL PHOSPHATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00001972</td>
<td>hmg-1.2</td>
<td>INOSITOL PHOSPHATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00001976</td>
<td>hmg-11</td>
<td>INOSITOL PHOSPHATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003612</td>
<td>nhr-13</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003621</td>
<td>nhr-22</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003658</td>
<td>nhr-68</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003704</td>
<td>nhr-114</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004046</td>
<td>plp-1</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004857</td>
<td>sma-3</td>
<td>UGT ENZYME</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00006999</td>
<td>zyx-1</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00006999</td>
<td>zyx-1</td>
<td>GLYCOLYSIS/GLUCONEOGENESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00009174</td>
<td>F26H9.2</td>
<td>INOSITOL PHOSPHATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00012330</td>
<td>zip-3</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00012832</td>
<td>dmd-3</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
</tbody>
</table>
Figure 3.13: Significant associations of non-NHR TFs with metabolic pathways
\textbf{(A)} Clustered heatmap of non-NHR TF associations with metabolic pathways based on FDR, for score greater than 2. 
\textbf{(B)} Clustered heatmap of non-NHR TF associations with metabolic pathways based on FDR, for score equal to 2.

\textit{C30G4.4, vab-15, rnt-1 and pat-9} show enrichment to ETC complex genes, glycolysis/gluconeogenesis, canonical propionate degradation and BCAA degradation genes (\textbf{Figures 3.13A and 3.13B, File 3.3}). We know that propionate is produced by breakdown of BCAA, and propionate degradation end product is succinyl-CoA that feeds into TCA cycle, that produces energy. Similarly, glycolysis provides pyruvate to TCA cycle. Therefore, \textit{C30G4.4} and \textit{pat-9} putatively play a role in regulating either directly or indirectly BCAA degradation, propionate degradation, ETC complexes, and glycolysis, to ultimately increase energy generation.

To summarize this, various TFs (both NHR and non-NHR) displayed associations with numerous metabolic pathways. These findings provide a rich array of potential targets for further validation and investigation, thus broadening our understanding of metabolic regulation at the transcriptional level. As we transition, our focus will shift to the potential role of RBPs in the regulation of metabolic pathways.
**RBPs enriched to metabolic pathways**

94% of RBPs have been found to show associations with metabolic pathways across at least one dataset (Figures 3.3 and 3.14, File 3.4). Previous observations highlighted that many ribosomal proteins demonstrate enrichment towards energy and growth genes. To enhance the utility of this analysis and focus on non-ribosomal interactions, the ribosomal protein genes were excluded from the RBP-pathway analysis. Following this exclusion, the remaining associations were investigated via a network graph with a threshold of enrichment scores greater than 2 (Figure 3.15).

Even with the exclusion of ribosomal proteins, a considerable number of RBP genes remained associated with the aminoacyl-tRNA biosynthesis pathway (Figure 3.15). To gain further insights into this relationship, a detailed investigation was initiated to identify these associated RBPs.
Figure 3.14: Clustered heatmap of RBP associations with metabolic pathways based on FDR, for all scores
Figure 3.15: Network graph showing significant RBP-metabolic pathway associations with a score of more than 1, while excluding ribosomal proteins.

**RBPs associated with growth**

To delve deeper into the associations of RBPs with growth-related pathways, excluding ribosomal proteins, we explored enrichments in the aminoacyl-tRNA biosynthesis, collagen biosynthesis, iron metabolism, and N-glycan biosynthesis pathways (Figures 3.16A and 3.16B, File 3.4).
**Figure 3.16: Significant associations of RBPs with growth**

(A) Clustered heatmap of RBP associations with aminoacyl-tRNA biosynthesis, collagen biosynthesis, iron metabolism and N-glycan biosynthesis based on FDR, for score greater than 2.

(B) Clustered heatmap of RBP associations with aminoacyl-tRNA biosynthesis, collagen biosynthesis, iron metabolism and N-glycan biosynthesis based on FDR, for score equal to 2.

The 5'-3' exoribonuclease 2, XRN-2, showed the most substantial enrichment to the aminoacyl-tRNA biosynthesis pathway, exhibiting a score of 4, indicating associations across all analyzed datasets (**Figure 3.16A, File 3.4, Table 3.4**). While XRN-2 is not directly involved in aminoacyl-tRNA biosynthesis, it is integral to processes like ribosome biogenesis in eukaryotes and RNA degradation, according to information available from the WormBase\(^\text{204}\) and KEGG\(^\text{30}\) databases. These processes are critical for the synthesis and degradation of RNA, respectively, and while they do not explicitly include aminoacyl-tRNA biosynthesis, they contribute to the overall dynamics of protein synthesis where this process plays a pivotal role.

Several other RBPs, *including* C11D2.7, *hel-1, ssb-1, snr-3, snr-2, and rad-54.B*, among others, have strong associations with growth-related pathways (**Figures 3.16A and 3.16B, File 3.4, Table 3.4**). While these genes are involved in vital biological processes related to RNA metabolism and protein synthesis, their specific roles do not pertain directly to the enzymatic process of aminoacyl-tRNA biosynthesis. Their robust associations thus warrant further scrutiny for a comprehensive interpretation of their roles.
Following the investigation into growth-related pathways, we directed our attention towards all associations of RBPs with pathways beyond the scope of growth.

**RBPs associated to energy**

Several RBPs exhibit associations with genes involved in various aspects of energy metabolism, including pathways linked directly or indirectly to energy generation such as electron transport, glycolysis, the TCA cycle, and canonical propionate degradation (Figures 3.17A and 3.17B, File 3.4).

Among these RBPs, *fkb*-8 stands out due to its strong association with ETC Complex I, reflected by a score of 4 (Figure 3.17A, Table 3.4). FKB-8, a member of the FK506-binding protein family, has been recognized for its peptidyl-prolyl cis-trans isomerase (PPlase) activity, which is crucial for protein folding\(^ {204} \). Although a clear link between FKB-8 and ETC Complex I is yet to be firmly established, the influence of properly folded proteins and signaling processes on the functionality of ETC complexes suggests a plausible indirect connection. Given that some FKBP proteins localize within mitochondria—the site of the ETC—this further supports the possibility of a relationship between FKB-8 and ETC Complex I.
Figure 3.17: Significant associations of RBPs with energy
(A) Clustered heatmap of RBP associations with ETC complexes, glycolysis, TCA cycle, propionate degradation-canonical, and propionate degradation-other, based on FDR, for score greater than 2.

(B) Clustered heatmap of RBP associations with ETC complexes, glycolysis, TCA cycle, propionate degradation-canonical, propionate degradation-other based on FDR, for score equal to 2.

mtss-1 is another RBP that exhibits a notable enrichment to ETC Complex I (Figure 3.17A, File 3.4, Table 3.4). While not directly implicated in energy production, MTSS-1 is believed to contribute to maintaining mitochondrial integrity by participating in mitochondrial DNA replication. Since mitochondria are the primary sites of cellular energy production, the proper functioning of MTSS-1 could indirectly bolster energy generation by ensuring the health and function of mitochondria in C. elegans.

The unc-98 gene also displays a strong association with energy-related pathways, as it is enriched in propionate degradation, glycolysis/gluconeogenesis, and ETC Complexes I and V (Figure 3.17B, File 3.4, Table 3.4). Even though unc-98 does not directly contribute to energy generation, it plays an indispensable role in muscle function and contraction by contributing to sarcomere organization\textsuperscript{275}. These processes inherently rely heavily on energy utilization.

Further investigation reveals other RBPs showing associations with energy pathways (Figure 3.18B, File 3.4, Table 3.4), among which are cps-6, eif-1.A, drr-2, ral-1, and znf-706. Each of these RBPs plays a role in crucial cellular processes,
such as apoptosis, translation initiation, stress response, RNA splicing, and gene expression, which may indirectly influence energy metabolism. However, a firm validation of these associations necessitates further experimental investigations.

In conclusion, a significant number of RBPs are linked to genes involved in energy metabolism. While some associations seem to be indirect, the involvement of these RBPs in processes essential for cellular health, protein synthesis, and gene expression underscores the potential impact they may have on energy production, making them intriguing subjects for further research.

**RBPs associated to other pathways**

With respect to RBPs related to metabolic pathways beyond energy and growth, significant observations have emerged. A member of the Lon protease family, *lonp-2*, shows the highest enrichment in peroxisomal fatty acid (FA) degradation, marked by a notable score of 4 (Figure 3.18A, File 3.4, Table 3.4). Serving as a crucial component of protein quality control within peroxisomes, *lonp-2*, specifically LonP2, takes part in the degradation of damaged or misfolded proteins, thereby maintaining peroxisomal protein homeostasis.\(^{276}\)

Notably, LonP2 shares structural and functional characteristics with its mitochondrial counterpart, LonP1, including ATP-dependent proteolytic activity and the capability to recognize and degrade aberrant proteins. Such proteases are fundamental to preserving the overall health and functionality of their respective
organelles. Ongoing studies underscore the importance of Lon proteases in cellular homeostasis and their associations with age-related diseases, neurodegenerative disorders, and metabolic conditions. Despite these advances, the specific role of LonP2 in peroxisomal fatty acid degradation requires further exploration due to the limited existing knowledge on this topic.

Further examination has led to the identification of three genes - *exos-4.1*, *gfm-1*, and *mtss-1* - displaying associations with porphyrin metabolism, each with an enrichment score of 4 (*Figure 3.18A, File 3.4, Table 3.4*). This finding implies potential post-transcriptional regulation of porphyrin metabolism. The *exos-4.1* gene, an ortholog of human EXOSC4, is a component of the exosome complex implicated in various RNA processing and degradation pathways. Despite the absence of a clear direct link to porphyrin metabolism, *exos-4.1* may influence porphyrin metabolic enzyme levels by regulating RNA stability for proteins involved in this process. Similarly, *gfm-1*, an ortholog of the human GFM1 gene, exhibits GTPase activity and a role in mitochondrial translational elongation. This gene's impact on mitochondrial protein synthesis could subsequently affect the production of proteins integral to porphyrin metabolism. Lastly, the *mtss-1* gene, mirroring the human SSBP1 gene, is involved in mitochondrial DNA replication and the regulation of helicase activity. This gene could potentially modulate porphyrin metabolism indirectly by controlling the replication of mitochondrial DNA, which may code for proteins involved in the porphyrin metabolic process. Despite these
associations, additional research is required to better elucidate the complexities of these regulatory networks.

The gene \textit{mtss-1} appears to play a significant role in the regulation of BCAA degradation pathways, including those of valine, leucine, and isoleucine (Figure 3.18B, File 3.4). Intriguingly, there are additional potential gene-function associations worth investigating, including the connection of \textit{odd-1} and \textit{odd-2} with peroxisomal FA degradation; \textit{somi-1} with the degradation of proline, leucine, and valine; and \textit{ddx-27} with both peroxisomal FA degradation and ascaroside biosynthesis (Figure 3.18B, File 3.4).

It should be noted, however, that these observations remain largely predictive until empirically tested and validated. Experimental manipulation is necessary to conclusively uncover the potential regulatory roles of these RBPs in the different metabolic pathways.
Figure 3.18: Significant associations of RBPs with pathways other than energy or growth-related

(A) Clustered heatmap of RBP associations with other pathways (excluding energy and growth), based on FDR, for score greater than 2.  
(B) Clustered heatmap of RBP associations with other pathways (excluding energy and growth) based on FDR, for score equal to 2.

Table 3.4: High-scoring RBPs significantly associated with metabolic pathways with a score of 3 or 4

<table>
<thead>
<tr>
<th>WormBase ID</th>
<th>Gene</th>
<th>Pathway</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBGene00000105</td>
<td>alg-1</td>
<td>VITAMIN B12 METABOLISM</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000411</td>
<td>cdl-1</td>
<td>PORPHYRIN METABOLISM</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000472</td>
<td>cey-1</td>
<td>VACUOLAR ATP-ASE</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000472</td>
<td>cey-1</td>
<td>ETC-COMPLEX IV</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000472</td>
<td>cey-1</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00000472</td>
<td>cey-1</td>
<td>ETC-COMPLEX V</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00001230</td>
<td>eif-3.G</td>
<td>ETC-COMPLEX I</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00001433</td>
<td>fkb-8</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>4</td>
</tr>
<tr>
<td>WBGene00001433</td>
<td>fkb-8</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00001439</td>
<td>fkh-7</td>
<td>VITAMIN B12 METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003015</td>
<td>lin-29</td>
<td>HISTIDINE DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003845</td>
<td>odd-1</td>
<td>GLYCINE CLEAVAGE SYSTEM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003846</td>
<td>odd-2</td>
<td>PEROXISOMAL FATTY ACID DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00003902</td>
<td>pab-1</td>
<td>PYRUVATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004207</td>
<td>ptb-1</td>
<td>TAURINE AND HYPOTAURINE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004408</td>
<td>rla-0</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004408</td>
<td>rla-0</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004408</td>
<td>rla-0</td>
<td>PYRUVATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004408</td>
<td>rla-0</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004409</td>
<td>rla-1</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>Gene and Protein</td>
<td>Complex Component</td>
<td>Gene and Protein</td>
<td>Complex Component</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>WBGene00004409</td>
<td>rla-1</td>
<td>ETC-COMPLEX V</td>
<td></td>
</tr>
<tr>
<td>WBGene00004409</td>
<td>rla-1</td>
<td>ETC-COMPLEX IV</td>
<td></td>
</tr>
<tr>
<td>WBGene00004409</td>
<td>rla-1</td>
<td>ETC-COMPLEX III</td>
<td></td>
</tr>
<tr>
<td>WBGene00004409</td>
<td>rla-1</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004410</td>
<td>rla-2</td>
<td>ETC-COMPLEX I</td>
<td></td>
</tr>
<tr>
<td>WBGene00004410</td>
<td>rla-2</td>
<td>ETC-COMPLEX V</td>
<td></td>
</tr>
<tr>
<td>WBGene00004410</td>
<td>rla-2</td>
<td>ETC-COMPLEX IV</td>
<td></td>
</tr>
<tr>
<td>WBGene00004410</td>
<td>rla-2</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-1</td>
<td>ETC-COMPLEX I</td>
<td></td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-1</td>
<td>ETC-COMPLEX V</td>
<td></td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-1</td>
<td>ETC-COMPLEX III</td>
<td></td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-1</td>
<td>ETC-COMPLEX IV</td>
<td></td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-1</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-1</td>
<td>PORPHYRIN METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-1</td>
<td>UBIQUINONE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-2</td>
<td>ETC-COMPLEX IV</td>
<td></td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-2</td>
<td>ETC-COMPLEX I</td>
<td></td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-2</td>
<td>ETC-COMPLEX III</td>
<td></td>
</tr>
<tr>
<td>WBGene00004412</td>
<td>rpl-2</td>
<td>ETC-COMPLEX V</td>
<td></td>
</tr>
<tr>
<td>WBGene00004413</td>
<td>rpl-2</td>
<td>PYRUVATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004413</td>
<td>rpl-2</td>
<td>ETC-COMPLEX II</td>
<td></td>
</tr>
<tr>
<td>WBGene00004413</td>
<td>rpl-2</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004413</td>
<td>rpl-2</td>
<td>VACUOLAR ATP-ASE</td>
<td></td>
</tr>
<tr>
<td>WBGene00004413</td>
<td>rpl-2</td>
<td>AMINOACYL-TRNA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004413</td>
<td>rpl-2</td>
<td>GLYCINE CLEAVAGE SYSTEM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004414</td>
<td>rpl-3</td>
<td>ETC-COMPLEX IV</td>
<td></td>
</tr>
<tr>
<td>WBGene00004414</td>
<td>rpl-3</td>
<td>ETC-COMPLEX V</td>
<td></td>
</tr>
<tr>
<td>WBGene00004414</td>
<td>rpl-3</td>
<td>PUFA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004414</td>
<td>rpl-3</td>
<td>ETC-COMPLEX I</td>
<td></td>
</tr>
<tr>
<td>Gene ID</td>
<td>Accession</td>
<td>Description</td>
<td>Score</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------</td>
<td>------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>WBGene00004414</td>
<td>rpl-3</td>
<td>PROPIONATE DEGRADATION SHUNT</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004415</td>
<td>rpl-4</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004416</td>
<td>rpl-5</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004416</td>
<td>rpl-5</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004416</td>
<td>rpl-5</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004416</td>
<td>rpl-5</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004416</td>
<td>rpl-5</td>
<td>PORPHYRIN METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>PUFA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004418</td>
<td>rpl-7</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>PYRUVATE METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004419</td>
<td>rpl-7A</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004420</td>
<td>rpl-9</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004421</td>
<td>rpl-10</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004423</td>
<td>rpl-11.2</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004423</td>
<td>rpl-11.2</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004423</td>
<td>rpl-11.2</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004423</td>
<td>rpl-11.2</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004423</td>
<td>rpl-11.2</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene ID</td>
<td>Description</td>
<td>Count</td>
</tr>
<tr>
<td>------------------</td>
<td>----------</td>
<td>------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>WBGene00004423</td>
<td>rpl-11.2</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004423</td>
<td>rpl-11.2</td>
<td>VACUOLAR ATP-ASE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004424</td>
<td>rpl-12</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004424</td>
<td>rpl-12</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004424</td>
<td>rpl-12</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004424</td>
<td>rpl-12</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004424</td>
<td>rpl-12</td>
<td>PUFA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004424</td>
<td>rpl-12</td>
<td>ETC-COMPLEX II</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004425</td>
<td>rpl-13</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004425</td>
<td>rpl-13</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004425</td>
<td>rpl-13</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004425</td>
<td>rpl-13</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004425</td>
<td>rpl-13</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004425</td>
<td>rpl-13</td>
<td>PHENYLALANINE DEGRADATION</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004426</td>
<td>rpl-14</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004426</td>
<td>rpl-14</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004426</td>
<td>rpl-14</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004426</td>
<td>rpl-14</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004426</td>
<td>rpl-14</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004427</td>
<td>rpl-15</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004427</td>
<td>rpl-15</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004427</td>
<td>rpl-15</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004427</td>
<td>rpl-15</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004427</td>
<td>rpl-15</td>
<td>PUFA BIOSYNTHESIS</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004427</td>
<td>rpl-15</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004428</td>
<td>rpl-16</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004428</td>
<td>rpl-16</td>
<td>VACUOLAR ATP-ASE</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004428</td>
<td>rpl-16</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004428</td>
<td>rpl-16</td>
<td>PORPHYRIN METABOLISM</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004428</td>
<td>rpl-16</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
</tr>
<tr>
<td>WBGene00004428 rpl-16</td>
<td>PROPIONATE DEGRADATION SHUNT</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004429 rpl-17</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004429 rpl-17</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004429 rpl-17</td>
<td>PORPHYRIN METABOLISM</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004430 rpl-18</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004430 rpl-18</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004430 rpl-18</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004430 rpl-18</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004430 rpl-18</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004430 rpl-18</td>
<td>VACUOLAR ATP-ASE</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004431 rpl-19</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004431 rpl-19</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004431 rpl-19</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004431 rpl-19</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004431 rpl-19</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004432 rpl-20</td>
<td>ETC-COMPLEX I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004432 rpl-20</td>
<td>ETC-COMPLEX IV</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004432 rpl-20</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004432 rpl-20</td>
<td>ETC-COMPLEX III</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004432 rpl-20</td>
<td>TRICARBOXYLIC ACID CYCLE</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004432 rpl-20</td>
<td>ETC-COMPLEX II</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004432 rpl-20</td>
<td>PYRUVATE METABOLISM</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004432 rpl-20</td>
<td>ELECTRON TRANSFER FLAVOPROTEIN</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>WBGene00004433 rpl-21</td>
<td>ETC-COMPLEX V</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
Methods

Most of the tools and methods employed in this Chapter-III have been the same as used in Chapter- II. For the sake of non-redundancy, they have not been explained again in this section.

Datasets

In addition to the post-embryonic development dataset\textsuperscript{202}, tissue dataset\textsuperscript{142} and compendium dataset\textsuperscript{141,204}, two additional expression datasets have been employed for analysis i.e., the embryonic development dataset\textsuperscript{277} and the cell-types expression data from the same tissues dataset\textsuperscript{142}.

Scoring and Clustering of Gene-Pathway Enrichment Associations

To identify the most robust gene-pathway associations, significant enrichments of genes in metabolic pathways were selected from the datasets using thresholds of FDR \(\leq 0.05\) and NES \(\geq 2\). Each significant gene-pathway association was assigned a score based on its presence in one or more of the four datasets: the compendium, embryonic development dataset, post-embryonic development dataset, and tissue or cell-type expression dataset. A score of 1 was assigned if the association was detected in any of the datasets. For the tissue or cell-type expression dataset, the score was incremented by 1 regardless of whether the
gene-pathway association was found in tissue expression, cell-type expression, or both. Subsequently, the gene-pathway associations were clustered to identify the most robust and consistent associations.

**Building the network graph of gene-pathway associations**

The network graphs of gene-to-pathway associations have been constructed using the NetworkX library of Python\textsuperscript{278}. The edges are colored by the scores of gene-pathway associations, while masking the scores less than 1.
CHAPTER IV : DISCUSSION

Part I: Understanding the transcriptional regulation of metabolism at a systems-level in *C. elegans*

Our current understanding of transcriptional regulation of metabolism in animals, including *C. elegans*, predominantly stems from studies focusing on individual genes or distinct pathways. These studies are typically classified as either “regulator-centered” or “biological process-centered”. However, there remains a conspicuous gap in the literature when it comes to genome-scale investigations of transcriptional regulation of metabolism in animals. Addressing this, our research introduces a first-of-its kind “target-centered” approach, wherein we systematically evaluated all the metabolic genes and pathways for transcriptional regulation.

**Transcriptional regulation occurs extensively at the level of genes, reactions and pathways**

In the light of our findings, the transcriptional regulation of metabolism appears to be extensively playing out across the genomic landscape, enzymatic interactions, and broader metabolic pathways. A majority of metabolic genes demonstrates transcriptional regulation through discernible changes in their mRNA levels - a feature they share with their non-metabolic counterparts. This observation is particularly salient in the context of tissue environments, where metabolic genes...
exhibit a higher degree of transcriptional regulation across spatially diverse tissues at a single time point as opposed to variable developmental stages. Despite deploying distinct statistical methodologies for the tissue and developmental datasets and acknowledging the differential resolution inherent to each, we observed a consistent pattern in our findings. Moreover, our data revealed that only approximately half of the genes demonstrating high tissue-specificity also presented considerable variation across developmental stages.

A noteworthy revelation from our study is the widespread coexpression in metabolic pathways, suggesting a level of coordinated regulation among the constituting genes. Our data reveals a stratified hierarchy in coexpression, with the highest degree observed in genes forming protein complexes, and subsequently in operon genes, paralogous genes involved in the same reaction, and those not involved in the same reaction. All these categories, however, exhibit greater coexpression compared to random gene pairs. It is important to note that the coexpression observed among enzymes involved in the same pathway is not confounded by these phenomena.

By employing a semi-supervised approach, we have been able to identify coexpressed gene clusters or sub-pathways, sidestepping the traditional requirement of manually annotated pathways. This not only resulted in stronger coexpression within sub-pathways but also facilitated a more precise definition of pathway boundaries.
Our study unraveled that the coexpressed (sub-)pathways are constituted not just by the subunits of protein complexes, but also by independent enzymes functioning in a sequential order. Our extensive gene expression compendium served as a valuable resource, enabling us to identify specific conditions under which certain (sub-)pathways were activated or repressed.

Our findings show coordinate regulation within metabolic sub-pathways, because of existence of potential compensatory mechanisms. For instance, these mechanisms may serve to supplement metabolite deficiencies, as exemplified by the upregulation of tyrosine degradation in the face of mitochondrial insufficiency. Similarly, they could act to prevent the loss of crucial metabolites, like histidine, as evidenced by the downregulation of histidine degradation when exposed to ultraviolet radiation.

To make our research more accessible and useful to other researchers, we have developed an interactive web application called WormClust, available on our dedicated Wormflux platform. This tool incorporates a gene-specific search feature, enabling researchers to delve into metabolic (sub-)pathways that share close associations with their specific gene of interest. We envisage this application as an integral contribution to C. elegans research community.

The findings from our study appear to contest the traditional understanding of metabolic regulation, which has typically favored the role of allosteric or metabolite-protein interactions over transcriptional regulation. However, our
analysis highlights the potential significance and the huge extent of transcriptional regulation in shaping metabolism. This suggests a more intricate model of metabolic regulation in which alterations at the level of gene expression may wield considerable influence over metabolic dynamics.

Figure 4.1: Summary of systems-level study of transcriptional regulation of metabolism
**Caveats and future directions**

Despite the intriguing nature of these findings, it's crucial to interpret them within the broader context of metabolic regulation. A key direction to think about is how much of enzyme levels influenced by transcription, contributes to flux of entire metabolic network. This is a subject of ongoing debate in the field. Some argue that many changes in metabolic gene expression don't directly affect flux, attributing these changes instead to inaccuracies in gene expression programs. They contend that these alterations in gene expression can occur without observable effects, especially when an increase is noted in an enzyme that is already abundant\textsuperscript{279,280}. In contrast, a recent study demonstrated that both cognate and non-cognate reaction fluxes can be influenced by metabolic enzyme levels, bolstering the relevance of our investigation\textsuperscript{281}. It would be worthwhile to determine whether reactions known to be regulated by allostery or post-translational modifications are also subject to transcriptional regulation. Doing so will allow us to elucidate their individual and combined impacts on metabolic phenotype, presenting a more holistic view of metabolic regulation.

**Coexpression of metabolic genes functioning together- an evolutionarily conserved principle**

Even with our conservative approach, we found that most metabolic genes are under transcriptional control. By a combination of supervised and semi-supervised
methods, we found that genes in many metabolic pathways are coexpressed and identified gene clusters that represent parts of pathways, or combinations of pathways with strong coexpression. These results build upon earlier work in single cell organisms such as *E. coli* and *S. cerevisiae*\(^{165–167,188,189}\) implying that coexpression of metabolic genes that function in tandem is a principle that is evolutionarily conserved.

**Caveats of this principle and future directions**

Despite the observation that coexpression of metabolic genes functioning together is a principle that is evolutionary conserved, it is important to note that specific details and mechanisms of coexpression can vary widely among different organisms and metabolic processes. There are several caveats and exceptions to keep in mind, when relating the coexpression of genes across organisms. For instance, evolutionary conservation does not imply immutability. As organisms evolve, their metabolic networks can change, leading to new patterns of gene coexpression. A metabolic pathway that is co-regulated in one species may no longer be in another due to differences in their ecological niches or evolutionary histories.

Moreover, the coexpression of genes can be context-dependent, as evidenced by the specific activation or repression conditions of certain coexpressed metabolic modules. Factors such as gene pleiotropy or epistasis can
complicate the relationship between coexpression and metabolic function further.

While our study captured the tightly coregulated metabolic units using a compendium of 177 distinct expression datasets, it's plausible that few new modules might emerge if more conditions are added to the compendium. There could be a scenario where, under specific conditions, the initial stages of a metabolic pathway are significantly upregulated while, intriguingly, the later stages of the same pathway are downregulated. This can culminate in the emergence of intricate and multifaceted patterns of gene expression. Although one could perpetually add more datasets, the current study utilizes a large number of datasets, and hence samples, to guarantee the findings hold broad applicability and universal relevance.

Taking into account the context-dependent nature of these relationships, we now shift our focus to the integral concepts in metabolic regulation, termed ‘metabolic wiring’ and ‘metabolic rewiring’.

**Metabolic wiring and rewiring: the role of transcriptional regulation**

To understand the nuances of transcriptional regulation of metabolic genes, we must delineate two key concepts: ‘metabolic wiring’ and 'metabolic rewiring.' In this context, metabolic wiring refers to the baseline distribution of flux across the metabolic network, determined largely by the tissue-specific requirements and stages of development. Metabolic rewiring, on the other hand, represents the
dynamic changes in this network flux distribution under varying dietary or environmental conditions.

Our data indicates that metabolic genes are more prone to transcriptional regulation across tissues than during development. What is the purpose of transcriptional activation or repression of metabolic genes? We propose that the transcriptional regulation of metabolic genes can serve distinct yet interconnected purposes.

Firstly, the significant transcriptional regulation of metabolic genes across tissues is a manifestation of metabolic wiring. Tissues have diverse metabolic needs dictated by their unique functions. For instance, the C. elegans intestine serves as the entry point of bacterial nutrients and requires the expression of enzymes that aide digestion and metabolite transport to other tissues. Similarly, the animal's muscle needs to produce energy to support movement and is therefore highly catabolic.

Secondly, transcriptional regulation is evident during developmental stages, reflecting the temporal aspect of metabolic wiring. This indicates the need for altering metabolic functionalities as tissues differentiate and grow.

The third aspect pertains to metabolic rewiring, where metabolic gene activation or repression occurs under changing dietary or environmental conditions, owing to different metabolic needs. This can be for the breakdown of different nutrients, e.g., carbohydrates versus fats, versus protein, or to rewire
metabolic pathways when others are perturbed. An example of this is the propionate shunt, which is transcriptionally activated when flux through the preferred, vitamin B12-dependent pathway is perturbed\textsuperscript{76,144}.

Lastly, metabolic genes can also be transcriptionally activated when the flux through their respective pathway is hampered, as exemplified by the Met/SAM cycle in \textit{C. elegans} under low vitamin B12 dietary conditions\textsuperscript{145}. This further supports the concept of metabolic rewiring as an adaptive response to maintain metabolic homeostasis.

\textbf{Transcriptional and post-transcriptional regulation}

In this study, we performed a systems-level analysis of mRNA level variation, using it as a proxy for the transcriptional regulation in \textit{C. elegans}. Recognizing the intricacies of gene expression programs, we acknowledge that variables such as mRNA stability can significantly sway mRNA abundance, in addition to the activity of transcription.

Premature RNA, also known as pre-mRNA, refers to the initial transcript that is synthesized during the transcription process. This pre-mRNA is then processed through a series of modifications - including splicing, 5' capping, and 3' polyadenylation - to produce the mature mRNA. This mature mRNA is the transcript that is exported from the nucleus to the cytoplasm where it serves as a template for protein synthesis during translation.
The stability of this mature mRNA - that is, its rate of degradation versus its rate of synthesis - can significantly influence mRNA abundance levels. For instance, a gene exhibiting low transcriptional activity may nonetheless demonstrate high mRNA abundance if the resultant mRNA demonstrates high stability, marked by an extended half-life, and vice versa. Therefore, the quantification of premature and mature RNA is instrumental in segregating the influences of transcriptional and post-transcriptional regulation.

**Future directions to segregate transcriptional and post-transcriptional regulation**

Several techniques have been developed to estimate transcriptional activity. One such method is EISA (Exon Intron Split Analysis), which calculates the ratio of intronic reads to exonic reads from RNA sequencing data\(^\text{282,283}\). Another technique, Rembrandts, models transcriptional dynamics using sequential RNA sequencing data collected at multiple time points\(^\text{284}\). Similarly, INSPEcT infers rates of RNA synthesis, processing, and degradation using comparable datasets\(^\text{285}\). However, these methodologies often either overlook the complexities of RNA processing or overly rely on the assumption of constant processing rates, leading to potential inaccuracies\(^\text{286}\). Nevertheless, future studies could use these techniques on total RNA sequencing data and, retrospectively, on numerous
publicly available sequencing datasets to contrast the contribution of post-transcriptional regulation to the overall transcriptional output.

An alternative avenue for exploration might involve direct genome-wide measurements of nascent transcription. Indeed, when profiled over relatively short-time intervals, nascent transcription provides a direct measurement of the rate of premature RNA synthesis \(^{286,287}\). This is where additional techniques like GRO-Seq\(^{288}\) (Global Run-On sequencing), which directly measures gene transcription by sequencing nascent RNA, and NET-Seq\(^{289}\) (Native Elongating Transcript sequencing), which profiles RNA molecules in the process of being transcribed by RNA polymerase, could be utilized. These powerful techniques could enable more precise quantification of nascent transcription rates, providing a more accurate snapshot of the gene transcription landscape.

The disentangling of transcriptional and post-transcriptional regulation could lead to a deeper understanding of their distinct contributions to complex gene expression programs, particularly in tuning the magnitude and temporal or condition-specific variation of such programs. Moreover, examining the 3'UTR binding sites of transcriptionally regulated genes could provide insights into the potential role of RBPs in modulating gene expression. Additionally, determining RNA kinetic rates would be valuable in characterizing the functional role of RNA modifications \(^{290,291}\) and RBPs \(^{292}\), which are fundamental determinants in the fate of a transcript.
Another technique, Translating Ribosome Affinity Purification (TRAP), can shed light on the post-transcriptional scenario within a cell by focusing on actively translated mRNAs. However, it's worth noting that TRAP primarily targets ribosome-associated mRNAs, thereby revealing information about the actively translated mRNAs, rather than elucidating the age or stability of mRNAs. Techniques that can quantify synthesis and degradation rates of mRNA, such as 4-thiouridine (4sU) labeling or bromouridine (BrU) labeling, are therefore necessary to differentiate between newly transcribed and older, stable mRNA.

Combining metabolic labeling techniques with TRAP offers a more comprehensive picture, enabling the isolation and comparison of newly transcribed and actively translated mRNAs versus older, yet actively translated mRNAs. Such an approach identifies mRNAs that, despite their age, remain stable and are actively translated, thus offering insights into both mRNA stability and translation efficiency. It is, however, imperative to consider that both mRNA stability and translation efficiency can be influenced by numerous factors, including mRNA sequence features and cellular conditions. Consequently, a meticulous integration of data from varied methodologies and careful analysis would be effective in distinguishing and understanding the roles of transcriptional and post-transcriptional regulation in gene expression.
Coflux fine-tunes a simple guilt-by-association approach

We have developed a novel flux-coupling based algorithm to fine-tune the guilt-by-association approach based on only coexpression\textsuperscript{141}. Although coexpression establishes a robust basis for these investigations, it remains susceptible to the inclusion of false positives. Such inaccuracies commonly occur when genes, despite lacking shared molecular functions, are coexpressed due to their organization within identical operons or their close spatial positioning on the genome.

Briefly, we used coflux in addition to coexpression to obtain sets of coexpressed genes that function in connected reactions in the model. This approach provides a way to analyze coregulated metabolic genes that likely function together without the bias of prior pathway annotations. We are therefore most interested in clusters obtained by multiplying these matrices and are hence both interconnected and coexpressed, as the product cannot be large if one of these does not contribute.

Even though many genes that are part of connected reactions or pathways are coexpressed, these clusters or sub-pathways could not have been obtained by using the coflux matrix alone. A prominent example of this is one of the top clusters that captures the entire peroxisomal FA degradation genes. Only \textit{acox-1.1} and \textit{acox-3} out of seven \textit{acox} family genes are coexpressed with the other peroxisomal FA oxidation genes, indicating that these two genes are more likely to function in
this pathway than the other acox genes\textsuperscript{141}. The other acox genes are coexpressed with each other, and with mitochondrial FA degradation genes. If this clustering were only influenced by coflux, all the other acox genes would have clustered with the peroxisomal FA degradation cluster. Therefore, it is an important parameter that adds an extra layer of confidence on top of coexpression to extract tight coregulated sub-pathways. Further, for genes encoding enzymes that operate within multiple reactions and metabolic pathways (e.g., \textit{alh-8}), we can discern with which pathways they demonstrate stronger association with the help of this method.

Leveraging both coexpression and coflux, our novel algorithm improves accuracy in identifying gene functions and their associated metabolic pathways. However, metabolic pathway regulation also relies on distinct control points, adding another dimension to our understanding of metabolic regulation.

\textbf{Regulation of metabolic pathways influenced by control point(s)}

It has been noticed previously in the propionate shunt pathway, that first gene, \textit{acdh-1}, acts as a control point: its expression is induced several hundred-fold when vitamin B12 is limiting\textsuperscript{76,140,147,295}. These control points are pivotal as they modulate the flux of the entire pathway. For a linear pathway, which is insulated from other pathways, it is generally anticipated that the gene catalyzing the first reaction functions as a control point. However, it becomes intriguing to
determine whether this assumption is universally applicable to all linear pathways within the metabolic network model.

Furthermore, for pathways characterized by multiple intersections, the potential exists for several control points, each exerting an influence over the pathway flux. It is plausible that certain metabolic pathways do not demonstrate complete coexpression due to the existence of one or more control points. This concept of control point analysis has been discussed more at the protein level. For instance, in glycolysis, the involvement of glycolytic intermediates in multiple pathways necessitates the regulation of glycolysis at numerous points, thereby facilitating coordinated regulation of several interconnected pathways. This is demonstrated by the regulation of three enzymes: PFK, hexokinase and pyruvate kinase, all involved in glycolysis\textsuperscript{296–300}.

**Future direction to determine control points of metabolic pathways**

Looking ahead, a systematic examination of control points across all metabolic pathways represents a promising direction for future research. This can be achieved by leveraging gene expression data across temporal, spatial, and conditional dimensions. Both the absolute expression and variability in gene expression can serve as indicators for potential control points within these pathways. Moreover, it would be insightful to explore how these genomic control points correlate with the known protein-level control points. This would provide a
more holistic understanding of regulatory mechanisms across different molecular levels in metabolic pathways, potentially leading to novel insights into complex biological processes.

**Part II: Systematically deciphering the “missing” knowledge in *C. elegans* metabolism**

Following an exploration of the extensive metabolic regulation at gene, reaction, and pathway levels within a network framework, and the identification of conditionally regulated pathway boundaries, the research transitions to addressing "missing" annotations within the metabolic network. Despite the advancements made through the sequencing of the *C. elegans* genome, there is a significant lag in functional annotation, particularly for metabolic genes encoding enzymes of unknown function. Complications arise from potential multiplicity of function in annotated genes, unlinked transporters, and not yet functionally deciphered TFs and RBPs. This research aims to bridge this knowledge gap in understanding *C. elegans* metabolism, with a primary focus on orphan genes, TFs, transporters, and RBPs.
Most *C. elegans* genes are associated with metabolism

We observe that over 90% of *C. elegans* genes are associated with metabolic pathways in at least one of the four datasets, encompassing embryonic development, post-embryonic development, tissues or cell-types and compendium. This is not surprising even though only around 70% of total *C. elegans* genes are found to be transcriptionally regulated in a former analysis because of multiple reasons\textsuperscript{141}.

Firstly, only highly variant genes were considered to be potentially transcriptionally regulated in that analysis, ignoring the moderately variant genes to be conservative. Whereas, for finding the associations between *C. elegans* genes and metabolic pathways, GSEA-based approach considers coexpression patterns that can be captured even in moderately variant genes.

Secondly, for the former analysis, only three datasets were used. Whereas, for finding gene-pathway associations in the latter analysis, two more datasets: embryonic development as well as cell-types expression were added to broaden the spectrum and strengthen the confidence level of the analysis, thereby covering more dimensions.
Deriving gene-pathway associations: A scoring system approach

I adopted a conservative approach in establishing thresholds for gene-metabolic pathway associations within each dataset. Still, I acknowledge the possibility that many gene-pathway associations might be a result of their presence in a single dataset. To mitigate this, I implemented a scoring system to confirm that each gene-pathway association appears in at least two datasets. This makes sure that false associations are avoided. However, this stringent method could lead us to overlook context-dependent gene pathway associations, marking a trade-off between accuracy and comprehensiveness.

Indeed, it is possible that a single gene could be involved in different metabolic pathways under varying conditions. This adaptability is key to how metabolic networks function. Future studies could look more closely at these unique, condition-specific connections, which could help us understand how elements like TFs and RBPs help regulate these shifts in metabolic pathways. However, it is important to consider the challenges associated with the detection and validation of these condition-specific gene-pathway associations: high variability between conditions, limited reproducibility of specific conditions, and a sheer number of potential interactions can be difficult to analyze. Future studies should, can however, seek to bridge this gap and extend the understanding of gene-metabolic pathway interactions in a condition-specific context.
Many *C. elegans* genes are associated with peroxisomal FA degradation

Intriguingly, peroxisomal FA degradation shows strong associations with an array of TFs, orphan genes, and transporters. Many of these genes are uniquely associated with only peroxisomal FA degradation (Figures 3.7A, 3.8, and 3.11A). We also found through a supervised approach that peroxisomal FA degradation is highly self-enriched for coexpression\textsuperscript{141}. Not only that, but it is also the only pathway that is fully rescued as a tightly regulated cluster with both stringent and relaxed parameters using semi-supervised approach\textsuperscript{141}. The striking self-coexpression of peroxisomal FA degradation, along with its exclusive associations with numerous TFs, orphan genes, and transporters, makes it critical to delve into its regulatory network for a better understanding.

The ubiquitous presence of peroxisomal beta-oxidation across all organisms further underscores the importance of this metabolic process. Its primary role includes the processing of very-long-chain fatty acids (VLCFAs) before these compounds undergo further oxidation within the mitochondria\textsuperscript{301,302}. This pathway also showcases adaptability as it comes to the rescue when mitochondrial functions are compromised, such as during starvation periods in *hilh*-30 animals\textsuperscript{179}. In addition, peroxisomes play a crucial role in maintaining ROS homeostasis\textsuperscript{303}. Involvement of peroxisomes in handling a transient heat shock response throws light on their role in facilitating organellar communication during
stress response. A fascinating facet of peroxisomal beta oxidation, particularly within the context of the model organism *C. elegans*, is its necessity for ascarosides biosynthesis. Ascarosides are signaling molecules that modulate development, lifespan, and behavior.

A study employing comparative mass spectrometric analysis on peroxisomal beta-oxidation mutants and wildtype *C. elegans* revealed that peroxisomal beta oxidation contributes to the biosynthesis of several hundred previously unidentified metabolites. The peroxisomal beta oxidation-dependent metabolome was found to be remarkably diverse, indicating intersections with nucleoside and neurotransmitter metabolism. The tissue-specific differences in peroxisomal beta oxidation-dependent submetabolomes further underline its functional diversity, likening it to microbial natural product biosynthesis.

In the context of peroxisomal fatty acid degradation regulation, *nhr-79* and *nhr-49* are identified as regulators by WormClust as well as the scoring system to identify strong gene-pathway associations. This has been corroborated by a recent study. Yet, intriguingly, other NHRs score higher in their associations with peroxisomal FA degradation, necessitating further experimental validation, especially *nhr-177* with a full score of 4. The surprisingly high associations of peroxisomal FA degradation with other NHRs, like *nhr-13, nhr-121, nhr-114, nhr-68, and nhr-22*, also demand investigation. Such exploration may shed light on the
possibility of different TFs or their combinations regulating peroxisomal FA degradation under varying conditions.

In summary, peroxisomal FA degradation in *C. elegans*, characterized by its unique multiple gene associations and self-coexpression, plays diverse metabolic roles and appears to be regulated by multiple NHRs, necessitating further study.

**Ambiguous associations between paralogous genes and metabolic reactions**

Metabolic reactions are often associated with multiple genes in gene-protein-reaction (GPR) annotations \(^{20,306}\). There are two reasons for this. First, some metabolic reactions are catalyzed by enzyme complexes comprised of two or more proteins. An example of this is the first reaction of propionate degradation by propionyl-CoA carboxylase, which, in *C. elegans*, is composed of PCCA-1 and PCCB-1\(^{307,308}\). In these cases, both genes are required and are therefore annotated as ‘AND’ genes. Second, some genes are part of larger families and share great degrees of protein sequence homology. Since the enzymes catalyzing most metabolic reactions have been experimentally delineated in bacteria, sequence homology to a proven protein is used to associate a gene with a reaction in metabolic network reconstruction. As a result, multiple similar paralogs can be associated with the same reaction, and these genes are annotated as ‘OR’ genes.
For instance, genes belonging to trehalase family *tre*-1, *tre*-2, *tre*-3, *tre*-4 and *tre*-5 have been all annotated to the same reaction that catalyzes the conversion of trehalose to D-glucose\(^{271}\). Other reactions can be associated with a combination of AND and OR genes. For instance, the conversion of succinate to fumarate by succinate dehydrogenase is part of both the ETC complex II and the TCA cycle and this reaction requires *sdha*-1 OR *sdha*-2 AND *sdhb*-1, *mev*-1 AND *sdhd*-1 (Figure 2.8A). Finally, some gene families consist of paralogs where some family members are associated with one, and others with another reaction. For instance, all the genes belonging to *aagr* gene family are paralogs, however there are distinct OR gene pairs: *aagr*-1 OR *aagr*-2 and *aagr*-3 OR *aagr*-4 in different carbohydrate metabolism reactions (Figure 2.8B).

OR genes may truly catalyze identical reactions under varying conditions, such as different developmental stages, tissues, or environmental circumstances. Alternatively, these genes might catalyze reactions that are distinct yet analogous, particularly when the precise GPR association has yet to be empirically determined. A significant complexity emerges due to sequence homology-based annotations, as they can result in multiple reactions being associated with the same, or a similar set of OR genes. In such cases, the uncertainty lies in identifying which specific gene encodes the enzyme responsible for each distinct reaction. This obscurity poses a challenge in accurate functional assignment and metabolic modeling.
Future directions to resolve OR genes

To untangle the complexities of multiple 'OR' gene associations with metabolic reactions, a comprehensive understanding of the role of paralogs is fundamental. The existence of these paralogs is justified by their potential to modulate metabolic dynamics across varying spatial, temporal, and environmental conditions. For this, there is a need to examine these paralogous genes for differential expression across different developmental stages, tissues, or conditions. If a pair of paralogs demonstrates a distinct anti-correlation in their expression across these variables, it suggests their existence caters to the diverse metabolic demands in varying tissues, cell types, conditions or timeframes. However, when paralogs display coexpression across all of these dimensions, they likely represent truly redundant genes. In this case, their role could be to act as a backup system against potential genetic mutations or dysfunction, or to meet the elevated demand for a specific metabolite within the cell.

To resolve the ambiguities at the level of reactions and pathways, where paralogs could be functioning in different reactions, I propose two strategies: supervised, and semi-supervised. Initially, I will utilize a supervised approach to associate C. elegans genes with metabolic pathways based on their confidence scores. These scores are derived from the degree of association across multiple
datasets. Such an approach can effectively dissociate 'OR' genes into different reactions.

In the second step, a semi-supervised approach will employ an algorithm based on coflux values\textsuperscript{141}. Here, the paralog gene exhibiting the highest product of coflux and coexpression values will be matched to the corresponding reaction.

**Proposed OR gene resolution algorithm**

The proposed OR gene resolution algorithm proceeds as follows:

**Input:**

- A set of reactions $R = \{r_1, r_2, ..., r_n\}$, each associated with a paralog gene set $G = \{g_1, g_2, ..., g_m\}$
- Their respective coflux and coexpression values

**Output:**

- Revised sets of reactions ($R'$) and associated genes ($G'$)

**Steps:**

2. For each reaction $r$ in $R$:
   a. Compute coflux scores: For each pair of reactions, calculate their coflux scores.
   b. Derive gene-specific coflux values: Determine the gene-specific coflux values from the calculated coflux scores of associated reactions.
c. Compute coflux-coexpression values for the network: Iterate over each gene g in G. Compute the product of the gene-specific coflux (from association of reaction r with other reactions) and the gene's coexpression values with other genes associated with corresponding reactions.

d. Sum coflux-coexpression values: Accumulate the computed coflux-coexpression values across all reactions to obtain a summed value for each gene in G.

3. Repeat: Carry out the steps (2a to 2d) for all reactions in the set R.

4. Determine max and min score genes: For each reaction in R, identify the genes with maximum (max_score_gene) and minimum (min_score_gene) summed values.

5. Match high scoring genes: Associate genes with high scores to their respective reactions. The threshold for a high score can be empirically decided.

6. Eliminate low scoring genes: Remove genes with low scores from their associated reactions. The threshold for a low score can be empirically decided.

7. Resolve all reaction and gene sets: Follow the same process for all associated reactions and paralog gene sets.

8. Output: The revised sets of reactions (R') and associated genes (G').

201
This comprehensive approach affords a structured methodology to resolve the complexities associated with 'OR' genes in metabolic reactions, with further refinement based on specific requirements and empirical validation. While coexpression serves as a crucial tool in bridging gaps within the metabolic network, it's worth noting that functionally related genes don't always exhibit coexpression.

**Not all TFs are coexpressed with their regulated genes**

Genes that exhibit coexpression are frequently found to operate in unison and are often in sync with their transcriptional regulators \(^{192-196}\). Leveraging this underlying principle, we developed WormClust\(^{141}\) and the consequent scoring system to enable the exploration of metabolic pathways that exhibit coexpression with any given *C. elegans* gene, inclusive of TFs.

However, it is of critical importance to highlight that this coexpression does not uniformly apply to all TFs and their regulated metabolic genes or pathways. Intriguingly, the critical pivotal regulators - those that are primed to respond to the initial influx of information - are often not found to be coexpressed with their target genes.

Indeed, *nhr-10*, which is essential for activation of the propionate shunt in response to high levels of propionate, does not change much in expression under relevant conditions\(^{144}\). This intriguing aspect emphasizes the complexity of transcriptional regulation where the conspicuous absence of coexpression does
not necessarily imply the absence of control or influence over metabolic processes. There are few strategies that can be generally employed to improve TF-metabolic pathway associations.

**Future directions to improve TF-metabolic pathway associations**

To identify ‘first responders’ such as *nhr-10*, it will be useful to employ promoter-reporter strains with large scale RNAi screens\(^{139,191}\). Additionally, in vivo techniques like Chromatin Immunoprecipitation Sequencing (ChIP-seq) can provide direct evidence of specific TFs' target sequences, augmenting our understanding of gene regulation.

Recent scientific advances have led to the determination of the genomic distribution of binding sites for 92 TFs and regulatory proteins across diverse stages of *C. elegans* development. This has been achieved by conducting 241 ChIP-seq experiments\(^{309}\). This data opens the opportunity to align TF-target binding information with the coexpression data of TFs and metabolic genes or pathways across different dimensions like space, time, and conditions. Moreover, identifying the ligands for *C. elegans* NHRs would further clarify their gene regulatory effects.

By integrating these methods, it will become possible to understand how specific TFs regulate metabolic genes and pathways across different cell types, development stages, or conditions, and thus allow for the development of unique
regulatory logics for different dimensions, giving us a more detailed view of gene regulation.

**Experimental validation of predicted gene-metabolic pathway associations**

The prediction of gene-metabolic pathway associations in *C. elegans* can be validated using various experimental approaches. These methods provide evidence to support the predicted relationships between specific genes and metabolic pathways. Here is a more detailed explanation of each approach:

- **Gene Perturbation Studies:** Disrupting the function of a gene using RNA interference (RNAi) or CRISPR/Cas9 can alter the transcriptional profile of the associated metabolic pathway, as identified by RNA-Seq. This approach validates the association if the gene disruption causes the expected transcriptional changes within the pathway.

- **Overexpression Studies:** Conversely, researchers can overexpress the gene of interest, increasing its activity, and observe the resulting effects on the metabolic pathway. If the overexpression leads to an increase in pathway activity as predicted, it strengthens the association between the gene and pathway.
• **Metabolomics Analysis:** Metabolomics involves measuring the levels of various metabolites within an organism. By manipulating the activity of the predicted gene, researchers can assess whether the expected changes in metabolite levels occur. If the levels of metabolites known to be influenced by the gene change accordingly, it provides evidence for the gene-pathway association.

• **Biochemical Assays:** These assays are performed to determine whether the product of a gene (usually a protein) exhibits the predicted activity within a metabolic pathway. For example, enzyme assays can be employed to measure the enzymatic activity of a protein against its expected substrates. If the protein demonstrates the expected biochemical activity, it supports the gene's involvement in the pathway.

• **Genetic Interaction Studies:** If a gene is hypothesized to function within a pathway alongside other genes, studying the simultaneous mutation or knockdown of multiple genes in the pathway can reveal unique effects compared to manipulating each gene individually. This synthetic genetic interaction study provides robust evidence for gene-pathway associations and strengthens the predicted relationships.

It is important to note that while each of these approaches contributes valuable evidence, sometimes multiple lines of evidence from different experimental
techniques are often required to convincingly confirm predicted associations. Having gained confidence in predicted gene-metabolic pathway associations in *C. elegans* through experimental approaches, we now turn our attention to theoretical modeling approaches that integrate gene-regulatory and metabolic network data to broaden the applicability and enhance predictive power in understanding an organism’s behavior.

**Future computational approaches to further explore the systems-level view of metabolism and gene regulation**

**Integration of regulatory information with metabolic network model**

**Regulatory FBA (rFBA)**

One way to theoretically integrate and derive insights from gene-regulatory and metabolic network model is regulatory FBA (rFBA). The constraints-based framework, with FBA, has been used successfully to predict time course of growth and by-product secretion, effects of mutation and knockouts, and gene expression profiles. However, its predictive power may falter in scenarios where regulatory impacts are the primary determinants of an organism’s behavior. Consequently, there is a requirement to incorporate regulatory incidents within the FBA to widen its application range and predictive proficiency.
In such contexts, rFBA can serve as a valuable tool for the amalgamation of gene regulatory networks with metabolic networks. The objective here is to establish a comprehensive model of cellular behavior that incorporates both metabolic processes and their transcriptional regulation. With sufficient annotations or correlations of TFs and RBPs to the metabolic network, the regulatory information can be leveraged to refine the existing metabolic network model.

Past research efforts have explored the analysis of both metabolic and regulatory networks in organisms such as *E. coli* and yeast$^{310,311}$. The application of this approach can notably enhance mechanistic flux predictions by integrating the contextual regulatory structure into our existing *C. elegans* iCEL1314 network and facilitating the simulation of a wide array of experimental conditions.

**Limitations of rFBA**

Despite the benefits of rFBA, including its ability to deliver more accurate and dynamic predictions of metabolic behaviors, it does have potential drawbacks. These include the binary nature of its regulatory rules and the computational complexity it entails. rFBA typically employs Boolean logic to symbolize regulatory interactions, thereby representing each interaction as either entirely "on" or "off". This oversimplification can neglect the often intricate and nuanced nature of gene regulation.
Moreover, integrating regulatory interactions into FBA considerably elevates the complexity of the model. This augmentation can make the computational task more difficult to resolve, especially for large-scale networks. However, as our knowledge of regulatory networks improves and computational methods continue to advance, the accuracy and utility of rFBA are likely to increase.

**Machine-learning approaches to detect novel metabolic pathways**

The generation and integration of large-scale multi-omics data, comprising transcriptomic, proteomic, and metabolomic information, under varied conditions such as dietary alterations or perturbations of specific transcription factors or metabolic enzymes, are indispensable to achieving an all-encompassing understanding of metabolic processes and gene regulation. The role of computational methods in integrating these complex layers of data is paramount.

Machine learning methodologies offer promising tools for predicting metabolic flux levels and metabolite concentrations. Recent studies conducted on budding yeast illustrate the utility of machine learning in elucidating relationships between metabolic flux and enzyme protein levels across an array of kinase mutants. These studies underscore the potential of integrative approaches in decoding the intricacies of metabolic and gene regulatory networks.
The predictive capabilities of machine learning extend to the identification of novel metabolic pathways\textsuperscript{313–317}. The tight coregulated clusters obtained by our semi-supervised approach\textsuperscript{141}, for instance, can serve as the gold standard set to determine novel groups of genes that share functional similarity by using naïve Bayesian approach\textsuperscript{318}. By detecting patterns within large datasets, machine learning algorithms can predict gene-metabolite interactions, facilitating the inference of potential new pathways. A recent supervised machine learning approach involving deep neural networks and random forests has been developed to generate feature representations of metabolic pathways, predicting both known and unknown metabolic pathways with exceptional performance metrics (>97% accuracy, >95% recall, and >99% precision)\textsuperscript{317}.

As we look to the future, the ability of machine learning algorithms to integrate diverse types of biological data, such as gene expression, protein-protein interactions, metabolic flux data, and genomic sequence data, is particularly exciting. The resulting integrated models will allow for more precise predictions about metabolic pathways.

It is important to acknowledge that although every improved computational approach provides powerful predictions, these predictions must be eventually experimentally validated in the laboratory.
Applicability to other organisms, including humans

Finally, in this study, we have employed methodologies that should be widely applicable to any organism, including humans, provided that extensive gene expression profile compendiums and high-quality metabolic network models are accessible. The application of these approaches should yield profound insights into the transcriptional regulation of metabolism as well as the circumstances that trigger the activation or repression of metabolic genes and pathways.

It's worth noting that many core biological processes are conserved across different species, implying similar operational patterns in organisms such as *C. elegans*, other organisms, and humans. Therefore, understanding these processes within a relatively simpler model organism like *C. elegans* can offer valuable insights applicable to more intricate organisms.

Many diseases, including diabetes, obesity, cardiovascular disease, and cancer, are often the result of disruptions to metabolic pathways and their regulatory mechanisms. By examining the transcriptional regulation of these pathways, we can obtain a more in-depth understanding of disease mechanisms and potentially identify novel targets for therapeutic intervention. Once we grasp the regulatory processes governing a metabolic pathway, it becomes feasible to engineer drugs capable of modulating these pathways. For example, potential therapeutics could be devised to inhibit or activate specific TFs, thereby altering the expression of downstream metabolic genes.
Metabolic regulation undergoes transformation during development and aging. Deciphering these changes in *C. elegans* may offer insights into the intricacies of growth, development, and aging, and could potentially assist in devising interventions for age-related diseases. Moreover, *C. elegans* serves as an excellent model for studying the effects of environmental factors and diet on metabolic regulation, contributing to insights pertinent to human health and disease.

Nevertheless, it’s important to acknowledge the differences between species. While the study of *C. elegans* can yield valuable insights, not all findings will directly translate to humans due to the greater complexity of human biology and disparities in cell types, organ systems, and metabolic needs. Despite these differences, the study of *C. elegans* offers an indispensable foundation for understanding the basic biological processes that hold relevance to human health and disease.


114. Xing, L. & Quinn, P. G. Involvement of 3',5'-Cyclic Adenosine Monophosphate Regulatory Element Binding Protein (CREB) in Both Basal and Hormone-Mediated Expression of the Phosphoenolpyruvate Carboxykinase (PEPCK) Gene. (1993).


257. Hashimoto, T. *et al.* Peroxisomal and Mitochondrial Fatty Acid β-Oxidation in Mice Nullizygous for Both Peroxisome Proliferator-activated Receptor α


