NETWORK-BASED INTEGRATIVE ANALYSIS OF MULTI-OMIC DATA

by

Samuel Joel Hinshaw

B.S., The University of Pittsburgh, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis/dissertation entitled:

Network-Based Integrative Analysis of Multi-Omic Data

submitted by Samuel Joel Hinshaw in partial fulfillment of the requirements for

the degree of Master of Science

in Bioinformatics

Examining Committee:

Dr. Robert E.W. Hancock
Supervisor

Dr. Jennifer Gardy
Supervisory Committee Member

Dr. Steven Hallam
Supervisory Committee Member

Dr. Paul Pavlidis
Supervisory Committee Member

Abstract

The rise of high-throughput biology has brought an increase in generation of large datasets such as genomics, transcriptomics, proteomics, and metabolomics: "omics" data. While many biological studies now assay multiple omics types to assess biological function, the analysis of these datasets is typically undertaken separately, contrary to our understanding of how biological systems function. While efforts have been undertaken to integrate these data types, intuitive methodologies that take advantage of modern curated biological databases are lacking.

Here I present a methodology for network-based integrative analysis of multi-omic data. This method leverages the power of curated interactome databases and biological network analysis to produce multi-omic biological interaction networks for integrative analysis. The integration of metabolomics data with transcriptomics and proteomics data was enabled by identifying metabolite-protein interactions using MetaBridge, a novel tool that I developed, described here. Identification of these metabolite-protein interactions was shown to facilitate the leveraging of powerful curated protein-protein interaction (PPI) databases such as InnateDB to generate metabolome-centric PPI networks. Such PPI networks accurately encapsulate biological function and enable downstream analysis and dimensionality reduction using proven network analysis techniques. These metabolomics-derived PPI networks could then be integrated with proteomics and transcriptomics data to create multi-omic networks, which provided insights into biological function and could be mined for novel biological insights that would not otherwise be captured by any single omics type.

I demonstrated two applications of this methodology to multi-omic datasets. First, I showed how separate gene expression and metabolite signatures for predicting sepsis could be integrated to reveal novel targets for study, demonstrating the utility of this method for hypothesis generation. Second, I demonstrated tri-omic integration of metabolomics, proteomics, and transcriptomics data from neonates in the first week of life. This revealed that network-based multi-omic integration provided consensus on commonly dysregulated biological functions and facilitated novel insights into biological changes.

Lay Summary

In recent years, biological experiments have been producing large amounts of data that assess the multifaceted aspects of biology. The technologies that produce this data are highly specialized and are built to examine one facet of biology in detail. To gain a broader picture of the mechanisms behind these datasets it is important to analyze the results of these specialized technologies in concert, joining the individual pieces of the puzzle back together to create a full biological picture. Here, a new method and tool is described for integrating the results of multiple modern biological assaying technologies, using preexisting biological databases to stitch together the puzzle pieces with specific examples provided as to how this can be used to gain insights into sepsis and early life. In contrast to current approaches for this type of analysis, my method was intuitive, transparent, and flexible. This makes it easily accessible to non-expert researchers, bringing them closer to understanding biology.

Preface

Dr. REW Hancock conceived of the project idea (methods for integrative analysis of multi-omic data) and the principal mechanism (network-based). Dr. Jianguo Xia suggested that MetaBridge be made available as an online tool and consulted on the methodology for PPI network generation from InnateDB. The idea for comparison of the methodology to an appropriate null was conceived by Dr. Paul Pavlidis. Dr. Daniel Pletzer and Arjun Baghela beta-tested the MetaBridge online tool. All development and analysis was conducted by myself with guidance and mentorship from Dr. Amy Lee and Dr. Erin Gill.

A version of Chapter 2 describing the design and development of MetaBridge was published in *Bioinformatics*:

Hinshaw SJ, Lee AHY, Gill EE, & Hancock REW. MetaBridge: Enabling Network-Based Integrative Analysis via Direct Protein Interactors of Metabolites. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty331

The HIPC study discussed in Chapter 3.2 has been revised and resubmitted:

Gill EE, Shannon C, Amenyogbe N, Ben-Othman R, Bennike TB, Diray-Arce J, Idoko O, Lee AH, Haren S van, Pomat WS, Cao KAL, Cox M, Darboe A, Falsafi R, Ferrari D, Harbeson D, He D, **Hinshaw SJ**, Ndure J, Njie-Jobe J, Pettengill MA, Richmond PC, Ford R, Saleu G, Masiria, G, Matlam JP, Kirarock W, Roberts E, Malek M, Sanchez-Schmitz G, Singh A, Angelidou A, Smolen KK, the EPIC Consortium, Brinkman RR, Ozonoff A, Steen H, Hancock REW, Tebbutt SJ, Biggelaar AHJ van den, Kampmann B, Levy O, Kollmann TR. Dynamic molecular changes during the first week of human life follow a robust developmental trajectory. Resubmitted.

Table of Contents

Abstract iii
Lay Summary iv
Preface v
Table of Contents vi
List of Tables ix
List of Figures x
List of Abbreviations xii
Glossary xiii
Acknowledgements xv
Dedication xvi
Chapter 1: Introduction 1
1.1 Integrative Analysis of Multi-Omic Data 1
1.2 Approaches for Integrative Analysis 2
1.3 Network-Based Integrative Analysis 4
1.3.1 Curated Data 4
1.3.2 Biological Interaction Networks 4
1.3.3 Network Types 5
1.3.4 Building Networks 6
Chapter 2: Network-Based Integrative Analysis of Multi-Omic Data with MetaBridge 7
2.1 Metabolomics Data Analysis 7
2.1.1 MetaCyc 9
2.1.2 KEGG 10
2.2 Output 11
2.3 Network Integration 12
2.4 Downstream Analysis 14
Chapter 3: Applications of MetaBridge 16
3.1 Sepsis Signature Integration 16
3.1.1 Integration 16

3.1.2 Biological Conclusions 19
3.1.3 Comparison to Integration of Random Networks 21
3.2 HIPC Project 22
3.2.1 Integration 23
3.2.2 Biological Conclusions 29
Chapter 4: Conclusion 31
4.1 Applications 31
4.2 Limitations 32
4.2.1 Reliance on Annotations 32
4.2.2 Promiscuous Proteins 32
4.3 Recent Developments 33
4.4 Future Directions 34
4.4.1 Promiscuous Protein Removal 34
4.4.2 Pathway Enrichment Comparisons 34
4.4.3 Random Discovery Rate 35
4.4.4 Further Integrations 35
4.5 Concluding Remarks 35
Bibliography 36
Appendices 45
Appendix A - 15-Metabolite Signature Mapping to Enzymes and Genes 45
Appendix B - Literature Review of Nodes of Interest 49
B.1 Nodes Common to All Networks 49
B.2 Nodes Unique to Integrated Network 49
Appendix C - Integrated Networks Generated with KEGG Mapping 52
Appendix D - Comparison to Null 55
D.1 Hypothesis 55
D.2 Connectivity 55
D.3 Network Overlap 58
D.4 Random Discovery Rate 60
D.5 Future Considerations 64

D.5.1 Gene Selection for Random Networks 64
D.5.2 Metabolite Selection for Random Networks 65
D.5.3 Integration of Further Gene Signatures 65

List of Tables

Table 1. 15-metabolite sepsis signature maps to human enzymes and genes via the MetaCyc database 45
Table 2. Randomly integrated networks were significantly more connected than the sepsis integrated network 56
Table 3. Nodes common to all networks were more connected than the network as a whole in randomly generated networks 56
Table 4. Nodes unique to the integrated network were less connected than the network as a whole 57
Table 5. Eighteen non-random nodes were unique to the sepsis integrated network 61
Table 6. Only one non-random node was common to all networks 61

List of Figures

Figure 1. MetaBridge provides a central step interconnecting metabolomics data and network generation in an integrative analysis pipeline 8
Figure 2. MetaBridge maps metabolites to direct protein interactors 9
Figure 3. Example of how a CSV file containing a list of metabolites is uploaded 10
Figure 4. Mapping metabolite IDs using e.g. the MetaCyc database provides results in a browser 11
Figure 5. Mapping metabolites using e.g. the KEGG database enables results to be seen in a browser 12
Figure 6. Mapping via KEGG allows for pathway visualization 13
Figure 7. Minimum-connected 99-gene endotoxin tolerance signature PPI network 17
Figure 8. Minimum-connected 15-metabolite sepsis signature PPI network 18
Figure 9. Minimum-connected integrated signature PPI network 19
Figure 10. Changes increased across all omics datasets during the first week of life 22
Figure 11. Multi-Omic Changes on Day 1 of Life vs Day 0 24
Figure 12. By day 3 of life, transcriptional changes overwhelmed a multi-omic network 26
Figure 13. On day 7 of life, transcriptional changes overwhelmed a multi-omic network 27
Figure 14. A bi-omic integrated network can be useful when there is a disproportionately large transcriptome signal 28
Figure 15. Bi-omic integration improved the proportions of omic representation but incorporated large numbers of first order-interactors 29
Figure 16. 99-gene Endotoxin Tolerance Signature PPI Network (KEGG) 52
Figure 17. 15-metabolite Sepsis Signature PPI Network (KEGG) 53
Figure 18. Integrated Signature PPI Network (KEGG) 54
Figure 19. The sepsis integrated network was not more connected than randomly integrated networks 57
Figure 20. Integrated networks tended to be composed of more gene seeds than metabolite seeds 58

Figure 21. The sepsis seed networks did not overlap more than the randomly generated networks 59
Figure 22. Nodes unique to the integrated network were less likely to occur at random than nodes common to all networks 62
Figure 23. The proportion of non-random nodes comprising a network varied least for the gene seeded networks 63

List of Abbreviations

CAS – chemical abstracts service

DIABLO – data integration analysis for biomarker discovery using latent components

DNA – deoxyribonucleic acid

ettx – endotoxin tolerance

GWAS – genome-wide association study

HMDB – human metabolite database

ICU – intensive care unit

KEGG – kyoto encyclopedia of genes and genomes

miRNA – micro RNA

MMRN – multi-scale, multifactorial response network

PBMC – peripheral blood mononuclear cell

PCST - prize-collecting steiner tree

PPI – protein-protein interaction

RNA – ribonucleic acid

TF – transcription factor

Glossary

Edges – Connections between nodes of a network. In a PPI network, edges represent interactions between proteins (the nodes).

First order network – A network whose nodes consist of seed nodes as well as curated interactors.

High-throughput biology – The practice of using modern technologies such as RNA-Seq, which can assay an entire transcriptome—thousands of genes—in a highly parallel manner.

Jaccard Index – A measure of overlap between two sets. Typically represented as a ratio, I have represented it as a percentage here for clarity.

NP-hard – A problem for which the solution algorithm is computationally costly to achieve when solving for instances of the problem with large numbers of observations.

Minimum-connected network – A first-order network which has been trimmed using an algorithm to include only specific nodes of interest.

Network – A structure representing connections (edges) between items (nodes).

Network Cohesion, Network Integrity – The ability for a network to remain intact without splintering into multiple subnetworks.

Nodes – Items in a network which are connected by edges. In a PPI network, nodes represent proteins.

Non-random node – A node which occurs in less than 5% of randomly generated networks.

Ome, Omics – A set of biological molecules which share common properties. The genome, proteome, transcriptome, and metabolome are several notable examples.

Peripheral interactor – Terminal nodes of a PPI network which are not seed nodes.

Prize-Collecting Steiner Tree Problem – A mathematical problem arising from graph theory. The solution, simply put, attempts to generate a subnetwork which maximizes the number of nodes while minimizing the number of edges.

Protein-Protein Interaction – An interaction between two proteins; direct interaction, association, or colocalization.

Protein-Protein Interaction Network – A network built from proteins (nodes) and protein-

protein interactions (edges). These networks represent a set of possible interactions between proteins in a cell and have been shown to accurately encapsulate biological function.

Random occurrence rate – The rate at which a node appears in a network generated from randomly selected proteins.

RNA-Seq – RNA Sequencing. See: high-throughput sequencing.

Seed network – A network generated from a single omics type.

Seed nodes – A list of proteins or protein-coding genes used to create a protein-protein interaction network.

Terminal node – A node of a network which is has only one edge connecting it to the graph. Typical network diagrams will show these nodes at the periphery of the network.

Zero-order network – A network consisting only of seed nodes.

Acknowledgements

I would like to thank Dr. REW Hancock for his support and advocacy for me throughout my studies. In particular, for steering me back in the right direction when I veered off the path or fell down a rabbit hole.

I owe a great deal to Dr. Jenny Bryan, Dr. Jennifer Gardy, Dr. Steven Hallam, and Dr. Paul Pavlidis for serving on my committee and providing countless insights into my research methodologies.

I acknowledge Susan Farmer for tirelessly supporting the lab and helping me through mountains of paperwork.

I greatly appreciate Dr. Amy Lee's incredible mentorship and our illuminating conversations.

I am obliged to Dr. Sarah Mansour and Arjun Baghela, as well as all of the members of my cohort for their endless moral support and friendship.

Finally, I would like to thank my parents for their unwavering support and many delicious Sunday night dinners.

Support from the Canadian Institutes for Health Research (FDN-154287, MOP-74493) is gratefully acknowledged. I was a recipient of The University of British Columbia's "Four Year Doctoral Fellowship".

Dedication

To Amelia—

It felt like this day would never come, but here we are.

Thank you for your unconditional love and support.

I don't know how I ever could have done this without you.

Chapter 1: Introduction

1.1 Integrative Analysis of Multi-Omic Data

As high-throughput technologies become more popular, research groups increasingly assess, globally different categories of biological molecules in an attempt to better define the biological system under interrogation.[1] These categories, such as the transcriptome, proteome, and metabolome, are often referred to as different "omics" types. Individually, each of these assay technologies has limitations and drawbacks. By performing multi-omic studies, biologists can supplement the individual weaknesses of a given platform.[2] This reflects the reality of biology, in that living organisms do not operate in isolation. Biology progresses through highly interconnected processes and on multiple information levels, and by simultaneously surveying multiple omic outputs, biologists intend to better capture that interconnectivity.[3]

Unfortunately, the current standard in the bioinformatics field is to analyze the results from each platform separately.[4] Because each data type is captured in different ways, it is challenging to bring the results together for a unified analysis. As a result, the integration and analysis of data from a variety of different omics platforms is a major goal in the bioinformatics community.[4,5]

Each of the major omics platforms has its own limitations. For example, transcriptomic strategies using RNA-Seq provide an extremely comprehensive picture of gene expression. However, it is known that the production of many proteins, which mediate most functions in cells, is also regulated post-transcriptionally and not all genes that are actively transcribed are also actively translated. Conversely, due to a lack of peptide amplification technologies, proteomics typically identifies only the most abundant proteins in the cell and is not as effective at identifying sequestered proteins (e.g. nuclear and transmembrane proteins). Metabolomics elucidates the chemical products of pathways but does not directly report on the events that gave rise to those products. Additionally, similarly to proteomics, it is not currently possible to amplify metabolites. Furthermore, metabolite identification is still a challenging problem in the field of metabolomics, and currently only 15-30% metabolites can be identified in an assay.[6]

The unique challenges that metabolomics data analysis presents makes it one of the most difficult omics types to integrate. In particular, unlike transcripts and proteins (the targets of transcriptomic and proteomic assays, respectively), which can be mapped back to a source gene, metabolites are often the end result of multiple complex discrete biochemical pathways.[7] Additionally, due to the current methods by which metabolites are identified, metabolomics data is often sparse and filled with missing values, leading to a secondary challenge for integration.[8–12] Nevertheless, integration of metabolomics data may prove to be the most rewarding due to its relative proximity to phenotype, particularly when compared to genomics or transcriptomics.[6,7]

There are many different tools available which attempt to solve this problem and integrate these multiple omics types in an effort to better capture how biological systems work. Integrative analysis of multi-omic data should provide a better picture of a biological system, allowing for greater enrichment of biological function.[4]

1.2 Approaches for Integrative Analysis

The challenge of integrative analysis has been tackled from many different angles.[13] Some approaches rely on biological annotation databases, while other approaches use purely statistical techniques. However, each approach suffers from unique drawbacks.

Some approaches such as mixOmics and integrOmics use multivariate statistics methods to identify correlations between omics types.[14–19] The strength of these methods lie in their ability to detect changes irrespective of knowledge of function, filling in the gaps where prior biological knowledge is lacking. However, many biologists do not have strong statistical knowledge, and lack the prerequisites to understand how these methodologies operate.[20] This leads to either lack of use, or worse, misuse, by many scientists.

Due to the explosion of multi-omic studies over the past decade, literature curation from omics experiments has been fruitful, leading to a large number of high-quality curated interaction databases.[21–23] The ability to draw upon a wealth of biological knowledge from a programmatically-accessible database means that biology-based methods are now a viable option for integrative analysis approaches. Existing biological methods fall into two main categories: pathway-based and network/interaction-based. Pathway-based integration methods such as IMPaLA are often overly broad, lacking the granularity to report on effects

within predefined pathways.[24,25] Other tools such as PaintOmics3 provide detailed insights into the enrichment results within a pathway, but both this approach and IMPaLA are confined to predefined pathways in biological databases, thus introducing bias.[13,26,27] In contrast, network/interaction-based approaches leverage preexisting knowledge-bases without restricting results to predefined pathways.

Network-based integrative analysis provides a useful framework for integration of multiple data types while simultaneously providing a framework for visualization of the results. By leveraging curated biological data to construct networks, the methods benefit from biological awareness, and have greater transparency than purely statistical methods. This is of key importance when developing tools for biological insight, as it improves the likelihood of correct interpretation of results and makes the methods accessible to a broader audience.

Currently available methods for network-based integrative analysis vary in their accessibility. While tools like SAMNetWeb are provided via a web interface,[28] other tools are implemented as downloadable software, or, like the tool MetScape, plug-ins to the popular network creation and visualization software Cytoscape.[29,30] Unfortunately, many tools are implemented as programming libraries. This is true for both biologically-based and statistically-based methods such as pwOmics and mixOmics. Implementation of these methods requires knowledge of a programming language (typically R or Python).[31] Finally, far too many tools suffer from poor documentation, implementation, or availability involving complex installation procedures, defunct web servers, or burdensome data preparation steps.[32–34]

Compounding the problem, many network-based integrative analysis tools are inflexible, and can only operate in a set manner.[17,29,35] For example, 3Omics allows the user to specify any 2-3 omics types: transcriptomics, proteomics, or metabolomics, but infers any missing types from literature searches; the user cannot opt out of this strategy.[34] Other tools, like integrOmics, can only operate on two omics types at a time.[18]

With all the caveats listed for the preexisting methodologies, it was apparent that there was a need for a transparent, intuitive, and flexible methodology for network-based integrative analysis of multi-omic data. Such a tool would be able to accept multiple types of omics data as inputs, and yield biologically-relevant, easily-interpretable results. Here I

present a method for network-based integrative analysis of multi-omic data that is transparent, intuitive, and flexible, and is executable with easy-to-use web-based tools.

1.3 Network-Based Integrative Analysis

1.3.1 Curated Data

The approach adopted here for network-based integrative analysis revolved around curated data from InnateDB, MetaCyc, and KEGG.[36–38] InnateDB curates protein-protein interactions (PPIs) to internationally acceptable standards as part of the International Molecular Exchange Consortium (IMEx).[36] The high-quality, curated PPIs report on potential functional interactions between proteins.

Curated information from MetaCyc and KEGG provide a method for linking metabolites to source proteins. By identifying direct protein interactors of metabolites, namely synthetic and degradative enzymes, metabolomics data can be linked to transcriptomic and proteomic data by way of PPI networks.

1.3.2 Biological Interaction Networks

PPI networks have been demonstrated in multiple studies to represent specific biological conditions and uncover novel, relevant information.[39] Any network consists of a set of nodes connected by lines termed edges, which are the connections between the nodes. In a PPI network, proteins represent the nodes of the network and PPIs represent the edges which are the interactions between proteins. Certain nodes of a network containing many connections are referred to as "hubs" and can represent proteins playing key biological roles in PPI networks (discussed further in Chapter 2.4). NetworkAnalyst, a browser-based tool, provides an efficient platform for generation, visualization, and analysis of PPI networks.[40]

For multi-omic integration specifically, PPI networks are useful in filling in gaps in proteomics and metabolomics data that occur due to the inherent limitations of these platforms. Unlike genomics and transcriptomics, which allow for coverage of the entire genome and transcriptome, proteomics and metabolomics are not able to report on every protein or metabolite level in the cell. Even when using scattershot approaches for metabolite profiling, various technologies pick up different sets of metabolites due to their unique separation characteristics, leaving gaps in reporting.[6] Curated metabolite-protein and

protein-protein interaction data can supplement these gaps in knowledge in a biologically relevant way, filling in the gaps in our reporting with biologically meaningful data. Here I have extended the capability of these PPI networks to model biological systems by incorporating further curated data in the form of metabolite-protein interactions from the MetaCyc and KEGG databases.

1.3.3 Network Types

When constructing PPI networks from a specific set of proteins, it is possible use the curated interaction data in two ways. The first way is to incorporate only information about the connections between supplied proteins, using the database only to draw the edges between the provided seed nodes. The result of this method is called a "zero-order" network. The second way is to use the PPIs to incorporate first-order interactors of the supplied proteins, resulting in a network consisting of not only the supplied proteins as nodes, but also proteins from the database as nodes. This type of network is termed a "first-order" network. The incorporation of these interactors is a double-edged sword. They can allow for edges to be drawn between two seed nodes through an intermediate interactor, and thus expand the possible network that can be drawn, but can introduce "noise" into the network, in the form of first-order interactors that are not situated between seed nodes ("peripheral interactors"). The minimum-connected network is one solution to reduce the downside of first-order interactor incorporation. A minimum-connected network is generated by trimming a first-order network according to a specific algorithm. Usually, this results in removal of peripheral interactors added to the network but retains interactors where they connect seed proteins. While this reduces the noise introduced to the network, it reduces the potential for discovering novel biological insight in first-order interactors.

NetworkAnalyst generates minimum-connected networks from first-order interaction networks by calculating the shortest path between seed nodes, approximating the NP-hard solution to the Prize-Collecting Steiner Tree problem.[41] The minimum-connected network is particularly well-suited to multi-omic integration, allowing interactors to fill in gaps in missing data from metabolomics or proteomics experiments. This is crucial in expanding the PPI network to include as many of the seed nodes as possible, ensuring that each omics method contributes meaningfully to the overall integration.

1.3.4 Building Networks

Building such biological networks enables the user to simultaneously gather evidence from multiple components of biochemical pathways. This method described here incorporates proteins derived from transcriptomic data, proteomic data, and metabolic data (catabolic and anabolic pathways and their reactions), including protein breakdown (metabolomics). With support from multiple sources, these networks are likely to more accurately reflect the biological state under investigation.

Curated interactome databases record biological connectedness and serve as a map by which to connect different omics data types. Specifically, the methods described here for network-based integrative analysis uses the properties of biological networks to facilitate integration between metabolites and other omics datasets by connecting metabolites to genes or proteins in a biological network via their biochemical reactions. MetaBridge, the tool developed in this thesis, identifies the reactions, enzymes, and proteins that interact with a given metabolite (i.e. ones involved with both production and degradation of the metabolite), providing a link back to a source gene (namely these proteins/enzymes).[42] Then, the identified proteins can be integrated with proteomics, transcriptomics, or genomics data through networks constructed from curated interaction data.[24,43]

Unlike correlation-based methods that are blind to the underlying biology, or pathway-based integration methods that are limited in scope, network-based integrative analysis is an intuitive and powerful tool for capturing the interconnected nature of processes occurring simultaneously within a cell. Altogether, this method serves as a powerful hypothesis generation tool, integrating multi-omic data to identify key targets of interest for future studies.

My working hypothesis was that this network-based approach would demonstrate that data derived from metabolomics profiling and gene expression profiling report on the same biological processes. Additionally, I proposed to demonstrate how the information captured from this integration also provides novel biological information about distinct processes.

Chapter 2: Network-Based Integrative Analysis of Multi-Omic Data with MetaBridge

The key to my methodology is its ability to incorporate metabolomics data along with transcriptomics or proteomics data in an integrative analysis. However, unlike proteomics and transcriptomics, metabolites cannot be directly mapped to a source gene. Therefore, I developed MetaBridge (https://metabridge.org) to provide a link between the metabolite and a protein, transcript or gene, by leveraging metabolite-protein interaction data from the KEGG and MetaCyc databases.[38,42,44,45]

2.1 Metabolomics Data Analysis

MetaBridge was developed as a web application to remove the need to know a programming language to use it. MetaBridge was developed in R, using the "Shiny" framework developed by RStudio.[46]

MetaBridge operates with lists of metabolites IDs. Therefore, metabolites must be preprocessed externally to identify metabolites of interest and their corresponding metabolite IDs. Use of MetaboAnalyst is recommended for this step, as depicted in Figure 1.[24] Metabolites must be provided as a tabular dataset, specifically, a text delimited file. This file is parsed and displayed by MetaBridge. The column containing the metabolite IDs is selected by the user; MetaBridge can accept KEGG, HMDB, PubChem, or CAS IDs (Figure 2, Figure 3). Metabolites are then mapped against the KEGG database or the MetaCyc database (Figure 4, Figure 5). Here the method was described by following the mapping of an example compound, Pyruvate (KEGG: C00022, HMDB: HMDB00243, PubChem: 107735, CAS: 127-17-3).

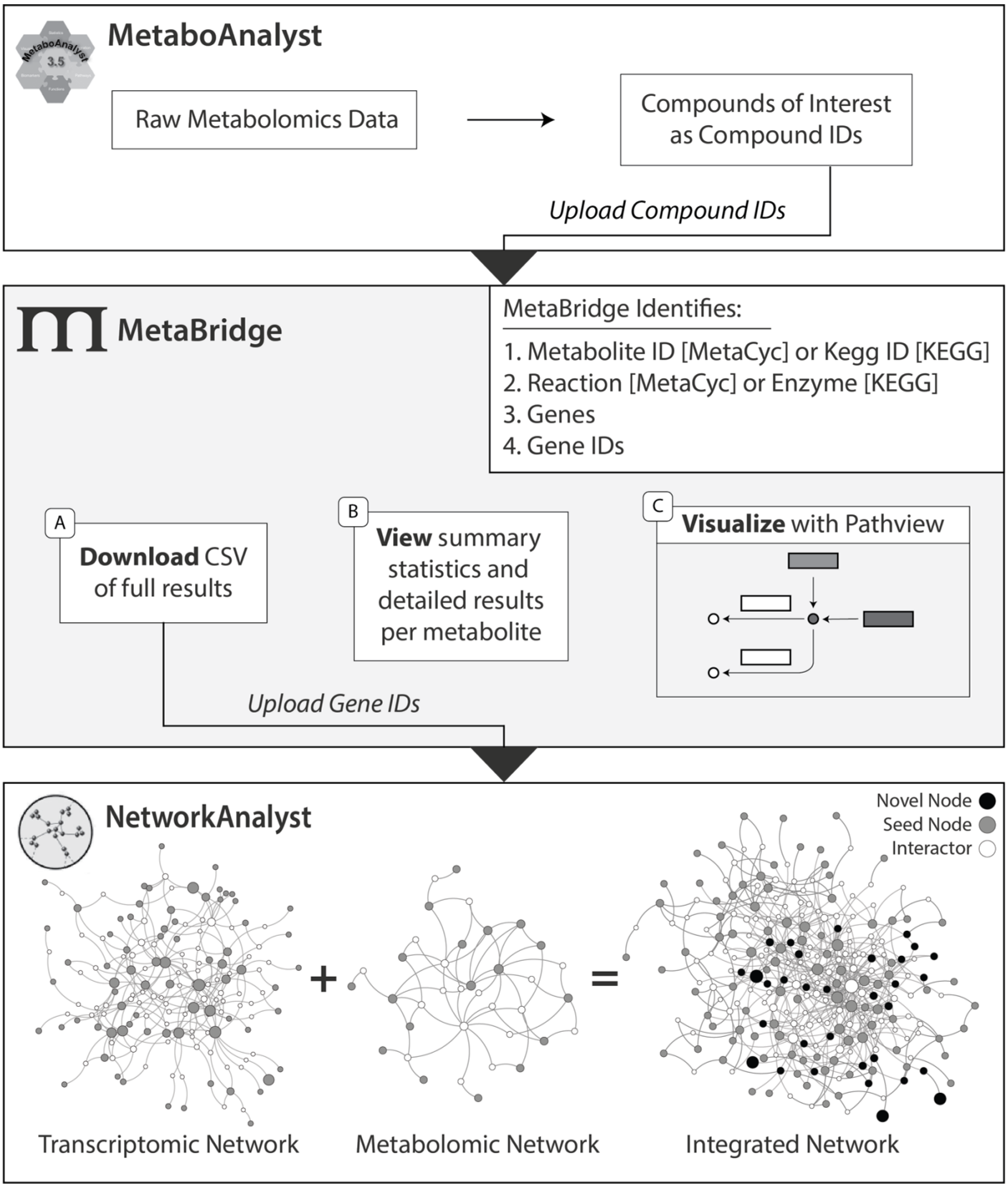

Figure 1. MetaBridge provides a central step interconnecting metabolomics data and network generation in an integrative analysis pipeline. Metabolomics data is preprocessed using MetaboAnalyst, identifying metabolites of interest. Metabolite IDs are uploaded to MetaBridge, outputting gene IDs of directly interacting enzymes. This gene list is uploaded to NetworkAnalyst for network-based integration and analysis. PPI networks are generated from each data type. Novel nodes are nodes present in the integrated network not found in either network generated from each individual data type.

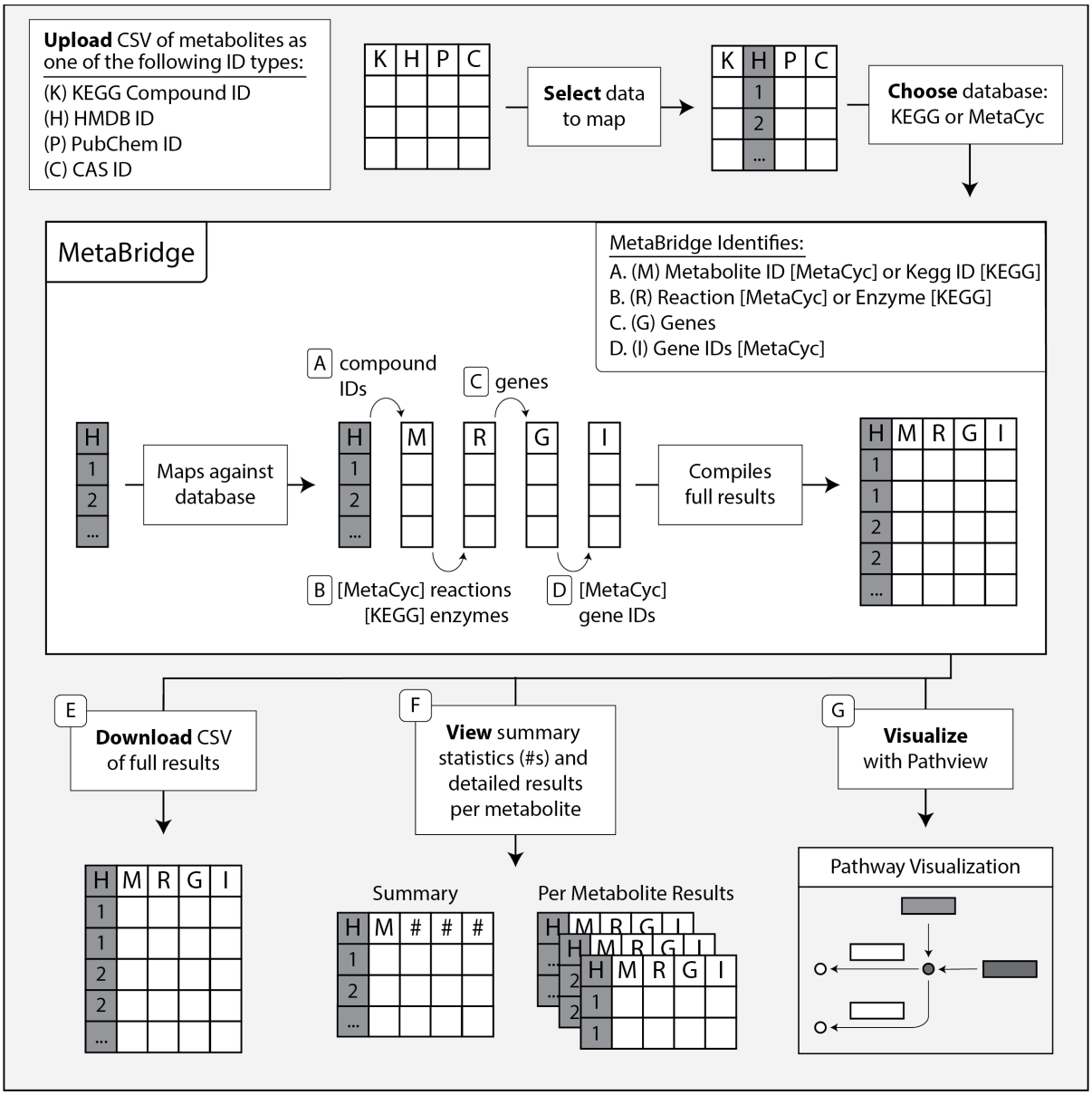

Figure 2. MetaBridge maps metabolites to direct protein interactors. Bolded keywords represent user actions. MetaBridge processes uploaded metabolites and offers the full results for download (A) and a summary for viewing (B). If metabolites were mapped via KEGG, the results can be visualized in the context of KEGG pathways (C). Here each grid represents tabular data either as a CSV or in-memory.

2.1.1 MetaCyc

MetaBridge maps provided metabolite IDs to their MetaCyc Object IDs (Figure 2A). In the above example, the ID is simply 'Pyruvate'. Then, using pathway-tools,[47] MetaBridge identifies all of the reactions in which the metabolite participates (Figure 2B). In the case of pyruvate, there are 26 reactions identified. Next, MetaBridge identifies all of the genes which

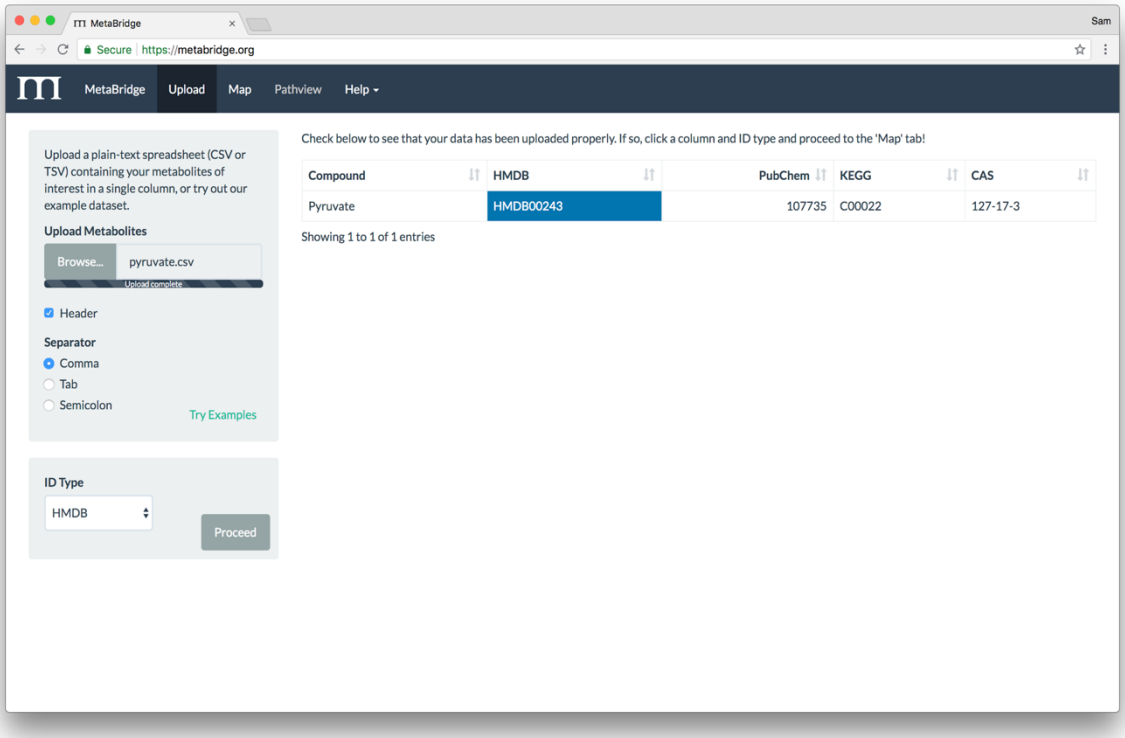

Figure 3. Example of how a CSV file containing a list of metabolites is uploaded. After a file is uploaded, the data is displayed to the user. The column containing the metabolite IDs and the ID type is selected.

encode for enzymes in these reactions (Figure 2C). Finally, MetaBridge identifies the Official Gene Symbols and Ensembl Gene IDs for each of the MetaCyc Gene IDs (Figure 2D). Taking the PEP-dephosphorylation reaction as an example, it maps to 2 genes—MetaCyc Gene IDs HS00906 and HS07088, two forms of pyruvate kinase. Finally, MetaBridge identifies the Official Gene Symbols (PKLR, PKM2) and Ensembl Gene IDs (ENSG00000067225, ENSG00000143627). The results are displayed to the user (Figure 2E, F).

2.1.2 KEGG

If the user did not upload KEGG IDs, MetaBridge converts the metabolite IDs to KEGG IDs (Figure 2A). Pyruvate's ID is C00022. Then MetaBridge identifies the KEGG-annotated interacting enzymes by EC number (Figure 2B). In the case of pyruvate, there are 23. Next, MetaBridge identifies the set of human genes that encode these enzymes (Figure

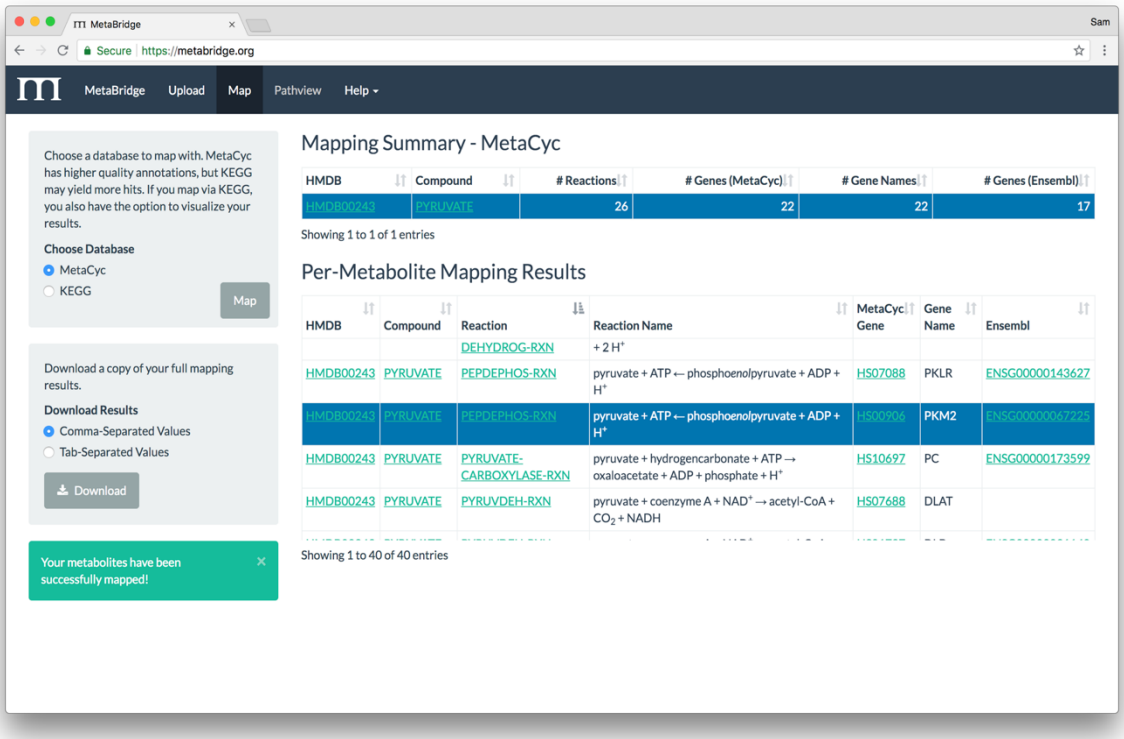

Figure 4. Mapping metabolite IDs using e.g. the MetaCyc database provides results in a browser. After database selection and mapping, a table with each metabolite's summary statistics is displayed. Each metabolite can then be clicked on to view further information.

2C). Taking EC 1.2.4.1, pyruvate dehydrogenase, as an example, MetaBridge identifies PDHA1, PDHA2, and PDHB as the genes which encode the subunits of the enzyme pyruvate dehydrogenase. For each metabolite uploaded, MetaBridge also identifies the KEGG pathways in which each metabolite participates. The results are then displayed to the user (Figure 2E, F).

2.2 Output

For each uploaded metabolite, the user can see how many unique reactions in which each metabolite participates (MetaCyc, Figure 4), or many unique enzymes with which each metabolite interacts (KEGG, Figure 5), and how many unique genes encode the identified enzymes. The user can select any summarized metabolite to see its full mapping details. With the summary table and per-metabolite mapping table, users can explore their results for each metabolite separately. The table can be sorted by any column and contains hyperlinks to the

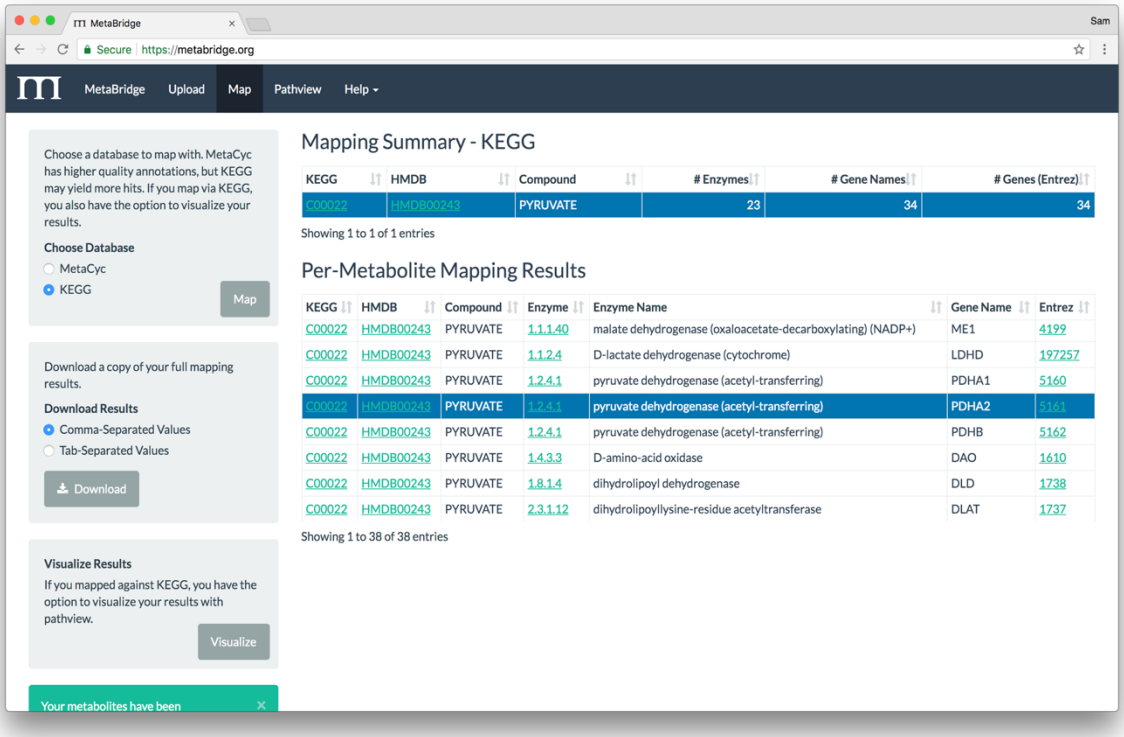

Figure 5. Mapping metabolites using e.g. the KEGG database enables results to be seen in a browser. After database selection and mapping, a table with each metabolite's summary statistics is displayed. Each metabolite can then be clicked on to view further information. With the KEGG database, a metabolite can be selected for visualization using Pathview (Figure 6).

relevant database for the entry. KEGG and MetaCyc are included, as is HMDB, Ensembl, and Entrez. The user can download the complete results of MetaBridge mapping as a CSV spreadsheet containing each metabolite, the reactions or enzymes for each metabolite, and the genes and gene IDs of each reactions or enzymes (Figure 2E). If the user mapped via KEGG, they can choose a metabolite and visualize the reactions and enzymes enriched with Pathview (Figure 2G).[48] Figure 6 shows direct interactors of pyruvate in KEGG's "pyruvate metabolism" pathway.

2.3 Network Integration

PPI networks for each omics data type are then generated with NetworkAnalyst, as per Chapter 1.3. As an example, I generated a list of 11 (MetaCyc) or 23 (KEGG) genes that coded for proteins that interact with pyruvate. Additionally, the gene and/or protein list

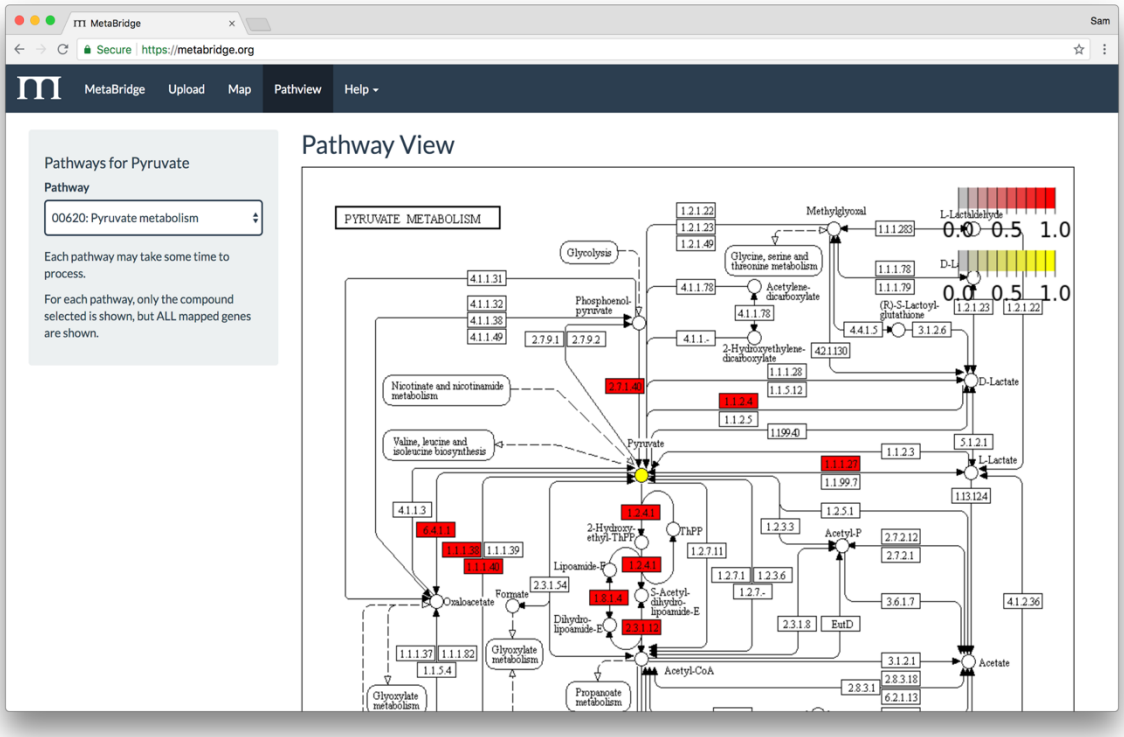

Figure 6. Mapping via KEGG allows for pathway visualization. Metabolites mapped against the KEGG database can be visualized in the context of KEGG pathways Pathview, a program which overlays pathway enrichment information onto KEGG pathways. Pathview color-codes uploaded metabolites and directly interacting enzymes. Here pyruvate and its direct interactors of are displayed.

based on transcriptomic and/or proteomic investigations should be uploaded. For integrative analysis these lists would be combined to generate integrated networks (Figure 1). In the case of a bi-omic integration, three networks can be generated:

1. Network seeded from metabolite interactors (metabolomics seed network)
2. Network seeded from differentially expressed genes (transcriptomics seed network)
3. Integrated network seeded from differentially expressed genes and metabolite interactors (metabolomics + transcriptomics)

In the case of a tri-omic integration, combinations of integrated networks of interest would be constructed, with up to 7 possible combinations: 3 individual “seed” networks, 3 bi-omic networks, 1 tri-omic network.

NetworkAnalyst builds PPI networks from these lists with a variety of parameters, as detailed by Xia et al., 2014, 2015.[43,49] Zero-order networks are always preferred as they only incorporate seed proteins as nodes. First-order networks are sometimes necessary to integrate smaller datasets, but minimum-connected networks should be used in place wherever possible, as detailed in Chapter 1.3.

2.4 Downstream Analysis

An analysis technique often used to enrich biological networks is pathway-enrichment analysis. Unfortunately, to assess the functional profile of the integrated network in comparison to the seed networks requires a complex comparison of ordered lists of pathway enrichment results. This problem has been the subject of intense study over the past decade.[50,51]

I suggest that to best facilitate hypothesis generation, identification of individual nodes of interest is preferable. However, picking out particular nodes of interest from a network is challenging. To assist in reaching this goal it is possible to narrow down a search to two main groups of interest. The first group is those nodes or first level interactors that are present in all generated networks. These are proteins of interest that are represented across all omics types. These nodes represent biological consensus between the omics types integrated. This group may consist of proteins not previously considered of interest to the biological condition under investigation, based on any single omics analysis. The second group is nodes that are exclusive to the integrated network. These nodes represent novel biological insights that would not be revealed by study of any single omics type in isolation.

Depending on the size of the inputs, the number of nodes common to all networks or unique to the integrated network may still be quite large and need to be narrowed-down to reach a reasonable list of novel targets for investigation. While many hubs of PPI networks play significant roles,[52–54] relying on node degree as the sole proxy for importance in the condition under investigation has several notable pitfalls.[55] Specifically, there are several confounding factors that may obfuscate the connection between connectivity and function. First, a protein can be highly connected due to nonspecific binding.[56,57] These proteins are commonly referred to as “promiscuous”. Second, due to the nature of a manually-curated database, a high node degree can be representative of a well-studied protein. Finally, it should

be noted that a protein need not interact with many other proteins to perform a vital cellular function. There have been several efforts in recent years to develop metrics specific to biology that can estimate the importance of a node in a network.[58–60] Chapter 3.1 describes one particular method for identifying nodes of interest while reducing noise.

Chapter 3: Applications of MetaBridge

3.1 Sepsis Signature Integration

The Hancock lab has previously shown that repeated exposure to LPS can induce an endotoxin-tolerant state in mononuclear cells.[61] This endotoxin tolerant state has been shown to be a key factor that drives the high mortality rate of sepsis. A transcriptome signature derived in part from these endotoxin-tolerant cells can accurately differentiate between patients with and without sepsis. This 99-gene signature indicative of cellular reprogramming/endotoxin tolerance was derived from reprogrammed peripheral blood mononuclear cells (PBMCs).[61]

Separately, a metabolite signature was shown by the Vogel lab to differentiate between patients in the ICU with confirmed cases of sepsis, and ICU controls. This 15-metabolite sepsis outcome differentiation signature, derived from NMR metabolite profiling of patients' blood, was shown to differentiate between patients with and without sepsis in hospital intensive care units (ICUs).[62] I integrated these two signatures to demonstrate network-based integration of multi-omic datasets with MetaBridge. 106 human genes were identified that encoded for proteins that directly interact with the 15 metabolites.

3.1.1 Integration

The two aforementioned sepsis signatures were integrated using the above strategy for network-based integrative analysis with MetaBridge and NetworkAnalyst, with one specific difference. To generate thousands of PPI networks at a time (for comparison to a null, as discussed in Chapter 3.1.3), it was necessary to generate PPI networks programmatically, rather than through NetworkAnalyst's web interface. Therefore, I developed a method for local generation of PPI networks directly from the InnateDB database that replicates the method employed by NetworkAnalyst.[40] Using a local copy of the InnateDB database (v5.5) downloaded via the EBI PSIQUIC registry,[21] I generated minimum-connected PPI networks using the same seed node list that would otherwise be uploaded to NetworkAnalyst. Minimum-connected networks were generated using a shortest-path approximation of the solution to the Prize-Collecting Steiner Tree problem in the "SteinerNet" R package.[63,64]

A handful of these networks generated with this method were selected at random and

compared to networks generated by NetworkAnalyst (data not shown). They were found to be over 90% similar. However, this comparison was conducted with interaction data from InnateDB v5.4, the version used by NetworkAnalyst. However, after the release of InnateDB v5.5, which included approximately 200,000 new curated interactions, the differences in locally-generated networks and networks generated on NetworkAnalyst increased. However, NetworkAnalyst is periodically updated with the latest copy of the InnateDB database, and

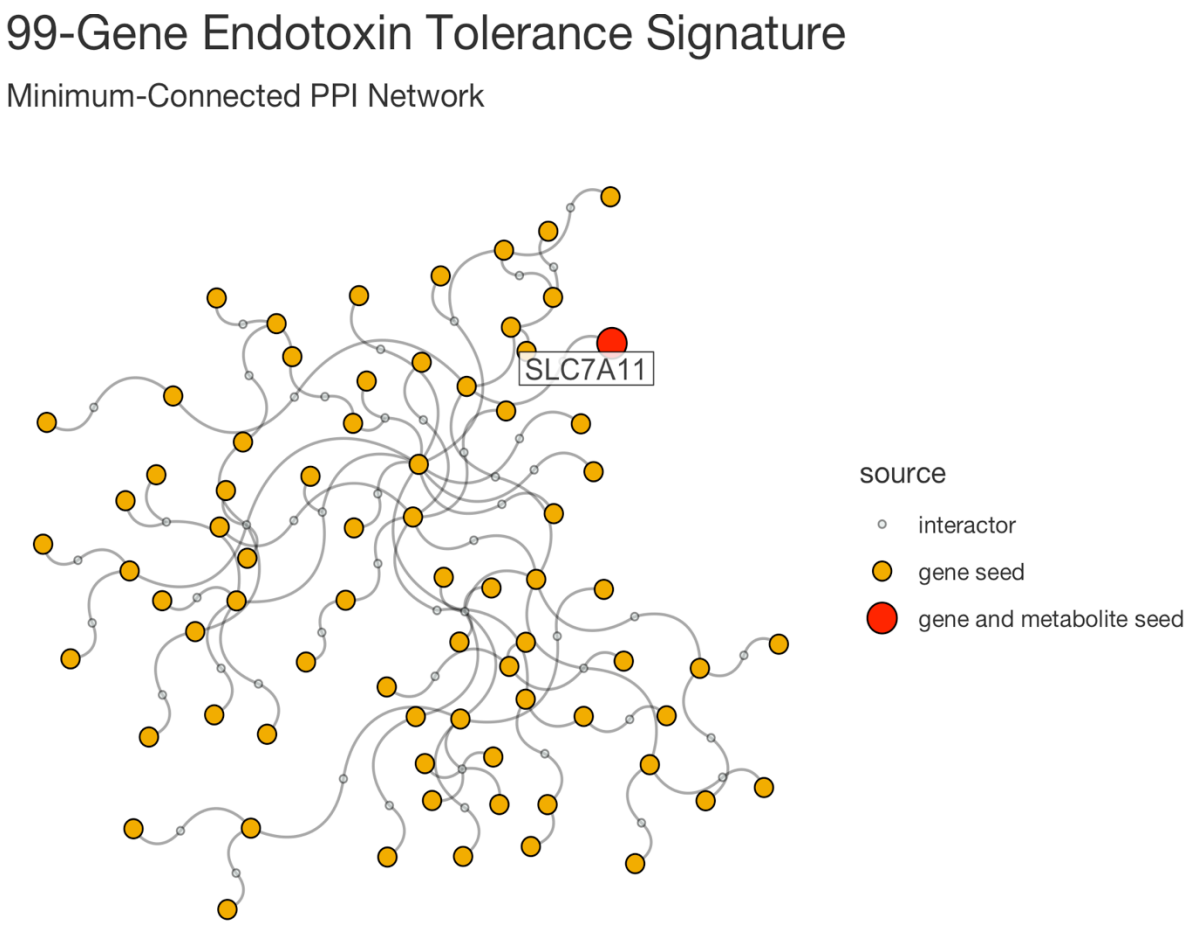

Figure 7. Minimum-connected 99-gene endotoxin tolerance signature PPI network. Minimum-connected PPI network of 99-genesepsis endotoxin tolerance signature. Nodes in the gene signature are highlighted in yellow. Nodes in the metabolites signature and gene signature are highlighted in red. Metabolite-protein mapping conducted with MetaCyc.

these networks should reflect networks generated with future versions of NetworkAnalyst.

Figure 7 shows a minimum-connected PPI network from the 99-gene endotoxin tolerance signature, the "transcriptomic network" (124 nodes). Using HMDB compound IDs from the 15-metabolite signature and mapping via MetaCyc with MetaBridge, I identified 42

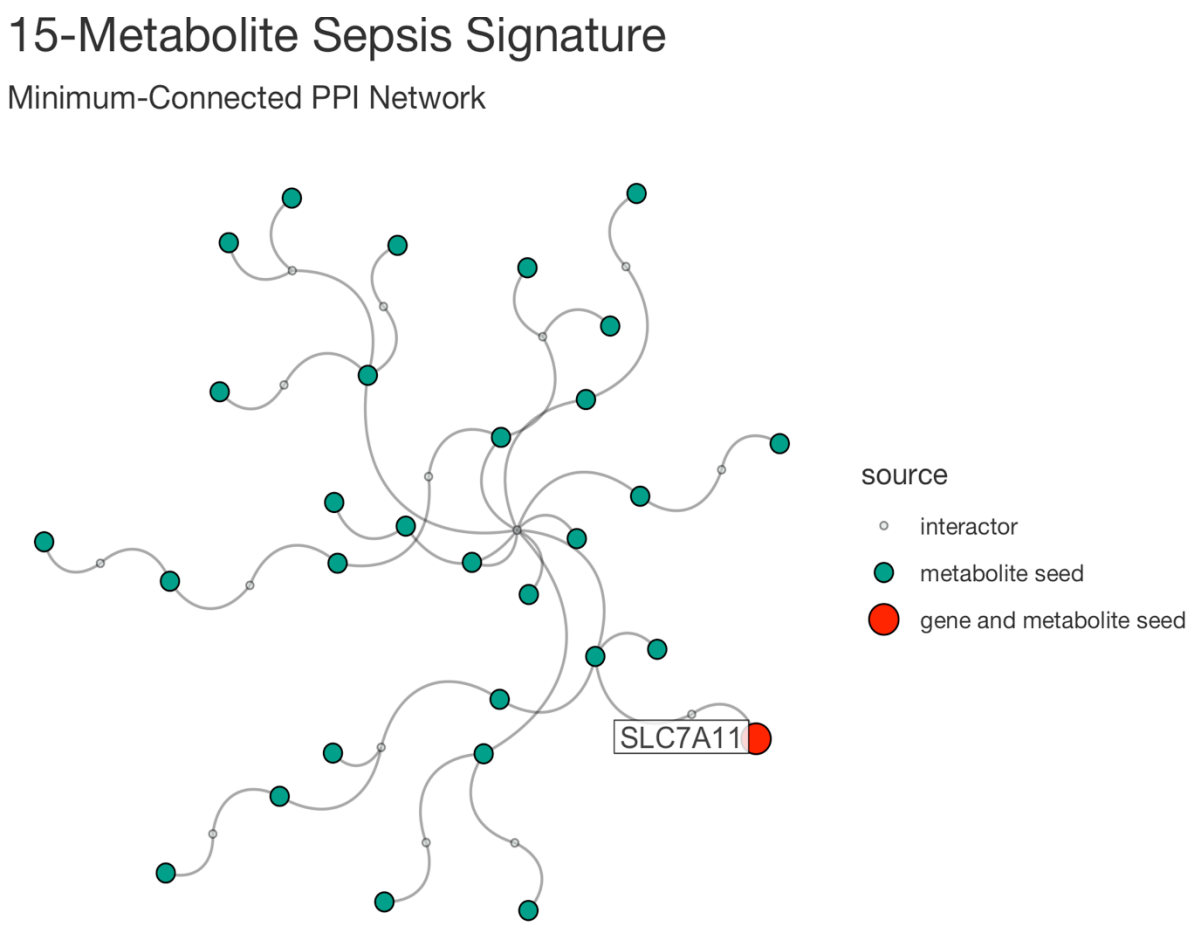

Figure 8. Minimum-connected 15-metabolite sepsis signature PPI network. Minimum-connected PPI network of 15-metabolite sepsis outcome differentiation signature. Nodes in the metabolites signature and gene signature are highlighted in red. Metabolite-protein mapping conducted with MetaCyc.

proteins that directly interacted with the 15 metabolites in the signature, Appendix A.[62] Figure 8 shows a minimum-connected PPI network from the directly interacting enzymes of the metabolite signature, the "metabolomic network" (45 nodes). Figure 9 presents a minimum-connected PPI network created by the union of the two gene lists, the "integrated network" (179 nodes). In each of Figures 7-9, the nodes were highlighted based on their source. Nodes that are found in the endotoxin tolerance gene expression signature are highlighted in yellow. Nodes that directly interact with compounds in the 15-metabolite signature are highlighted in green. Nodes that are found in both are highlighted in red.

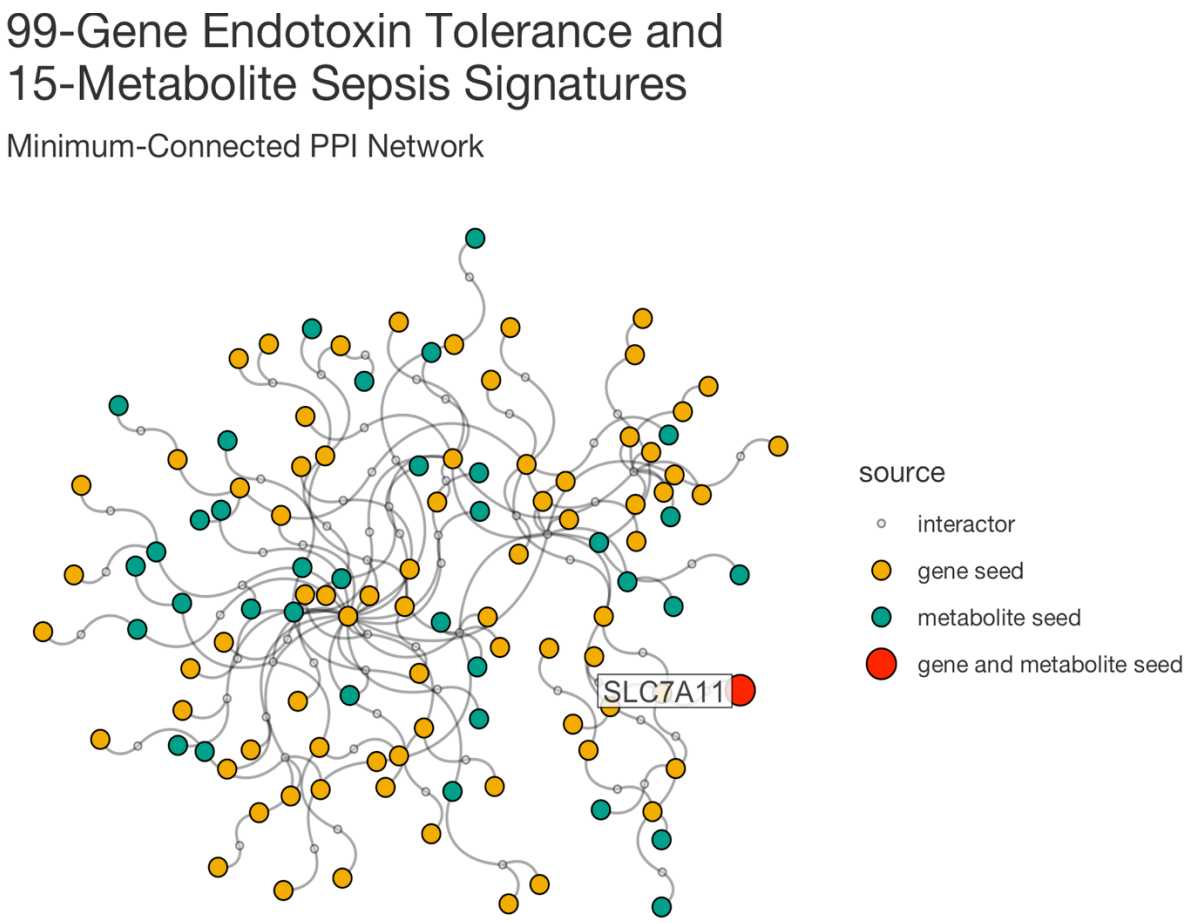

Figure 9. Minimum-connected integrated signature PPI network. Minimum-connected PPI network integrating 99-gene endotoxin tolerance signature and 15-metabolite sepsis outcome differentiation signature. Nodes in the gene signature are highlighted in yellow. Nodes in the metabolite signature are highlighted in green. Nodes in the metabolites signature and gene signature are highlighted in red. Metabolite-protein mapping conducted with MetaCyc.

3.1.2 Biological Conclusions

In the combined network generated from both the endotoxin tolerance signature and enzymes derived from the metabolic profile (Figure 9), I was able to make three notable observations about sepsis signature integration. First, there was contribution to the network from both data types, with 42.5% of nodes directly from the endotoxin tolerance signature and 27.6% of nodes directly arising from metabolite interactors. Second, a number of hubs not present in either initial network were present in the integrated network, indicating the potential for novel biological insights. Finally, I observed that SLC7A11, a gene in the endotoxin tolerance signature, was present in all generated networks.

To further examine both of these observations, I investigated the 40 nodes which were unique to the integrated network. 18 of these nodes occurred in randomly-integrated networks less than 5% of the time (methods detailed further in Chapter 3.1.3 and Appendix D). Of the 40 nodes, 8 were seed nodes; of the 18 nodes, 4 were seed nodes. A full literature review is included in Appendix B. Briefly, however, 14/18 (78%) nodes of potential interest had preexisting implications in sepsis in the literature. By identifying the random discovery rate of nodes in the PPI networks, it is possible to determine which nodes were less likely to occur by chance. This indicates unique biology represented within a given PPI network, and provides a method for identifying nodes of interest without relying on other network measurements, such as node degree. This helps filter out promiscuous proteins but does not exclude key proteins central to the biological condition under investigation, as seen by the inclusion of MYD88, a protein vital to innate immune signaling. As noted in Appendix B.2, Four of these fourteen hubs that have not yet been documented in connection with sepsis and may be novel targets for study.

These results underscore the reinforcement of biological understanding that can be captured by integration of multi-omic data, with many significant nodes of interest being linked to sepsis in the literature. Simultaneously, these results offer novel targets for study, demonstrating the ability of this methodology to generate hypotheses. While some of the nodes of interest highlighted here are well studied, others demonstrate links that are not as well-documented, and could provide useful targets for further study. Additionally, several nodes of interest have not been shown to be involved in sepsis in any capacity and could provide entirely new targets for study. Here I have shown that network-based integrative analysis of multi-omic data with MetaBridge provides consensus on biological function as well as novel information on a given biological condition. This method of integrative analysis is function based (since PPI interactions reflect functional relationships) and offers real-world benefits when identifying potential targets or biomarkers (key hubs) for further investigation. The intersection of these technologies reveals novel targets of interest that would not otherwise have been uncovered from either technology alone, namely those hubs exclusive to the integrated network.

3.1.3 Comparison to Integration of Random Networks

In an attempt to demonstrate that integration of multi-omic data is meaningful when compared to random noise, the integration of the two sepsis signatures was compared against multiple integrations of random genes and metabolites (termed here "random" integrations). This was a preliminary analysis intended to investigate the potential issues surrounding multi-omic integration. The results are described in Appendix D. Briefly, I randomly selected 99 genes and 15 metabolites, and integrated these pairs of lists to determine the occurrence of nodes in an integrated network generated from randomly selected genes and metabolites. I conducted this process 1000 times, and termed rate of occurrence of a node in an integrated network (out of the 1000 generated networks) as the "random occurrence rate" for that node. I then determined the random occurrence rate for each node of the integrated sepsis network and termed the nodes which occurred in less than 5% of the randomly integrated networks as "non-random nodes".

I found that most of the random integrated networks showed greater connectivity than the specific sepsis network (Appendix D). On the other hand, the integrated sepsis network had among the lowest mean node degree and one of the smallest numbers of nodes, indicating that it was likely more compact than the random networks (Figure 19). However, the integrated sepsis network had a high proportion of non-random nodes (Figure 23). This analysis remains incomplete but is included here for context.

One issue that should be addressed in future studies is the ability of the metrics utilized here to accurately characterize the "success" of a network-based integration. As observed above, the integration of gene expression and metabolomics data is somewhat dominated by the gene expression data, since transcriptomics reports on entire pathways rather than discrete end points. On the other hand, a single metabolite can bridge several discrete enzymes from different pathways. This might explain the high degree of variance amongst randomly integrated networks containing as many as 800 nodes, in contrast to the relatively compact integrated sepsis network. Since I observed integration of specific metabolite interactors with the gene expression network, it is likely that the integration process would still provide biological insights.

3.2 HIPC Project

With the integration of the sepsis signatures, I demonstrated integration of multi-omic data with respect to specific signatures developed to distill significant biology from concise sets of molecules. However, experimental datasets are often far richer, containing a significant level of noise along with the data points of interest.

A collaboration with the international Expanded Program on Immunization Consortium (EPIC) group of the Human Immunology Project Consortium (HIPC) provided

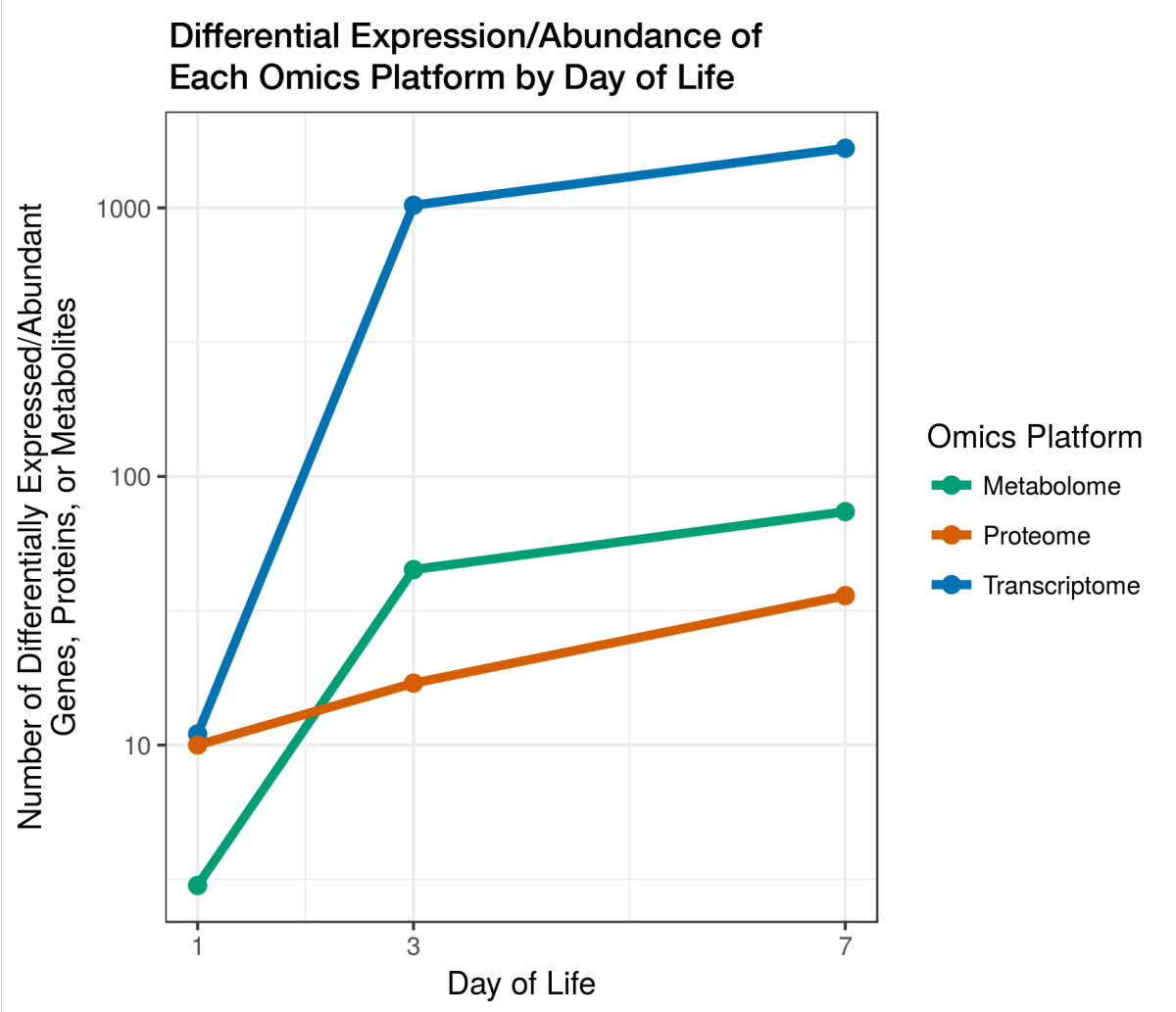

Figure 10. Changes increased across all omics datasets during the first week of life. Differential expression or abundance of the blood cell transcriptome, plasma proteome, and plasma metabolome as a function of day of life, when indexed to day 0 (as soon as possible after birth). Changes increased across all omics platforms during the first week of life, but at days 3 and 7 of life there were an order of magnitude more differentially expressed genes than differentially abundant proteins or metabolites—note the log-scaled Y-axis.

a further opportunity to test network-based integrative analysis of multi-omic data with MetaBridge. This project provided datasets on early life transcriptional, metabolomic and proteomic responses in Gambian neonates (day 0-7 of life). With this project I was able to attempt a three-way multi-omic integration with metabolomics, proteomics, and transcriptomics data. Importantly, this provided an opportunity to apply network-based analysis of multi-omic data with MetaBridge to experimental data, rather than the signatures previously used.

The HIPC datasets provided the challenge of an overwhelming transcriptomic signature when compared to the metabolome and proteome. Figure 10 highlights these differences. Changes to the transcriptome, proteome, and metabolome increased during the first week of life. On days 3 and 7 of life there was more than an order of magnitude more differentially expressed genes than differentially abundant proteins or metabolites. The difference in magnitude of results from these omics types is largely due to the sensitivity of the methods employed for metabolomics and proteomics and the body compartment (blood plasma) assayed. However, the differences are exacerbated further by limitations in experimental design when investigating neonates—only a small amount of blood can be drawn. While this small volume is not as impactful for transcriptomics, where small amounts of RNA can be easily detected and analyzed using modern sequencing technology, it presents a significant challenge for proteomics and metabolomics, where identification is performed based on the concentrations that exist in samples. This underscores the difference in integration of complete experimental data versus integration of signatures. Whereas with signature integration, 99 genes and 15 metabolites represent a more balanced signal from each method, with experimental data this is often not the case as discussed in Chapter 1.1 and summarized above.

3.2.1 Integration

When comparing day 1 of life to day 0 of life, comparable numbers of genes, proteins, and metabolites warranted a straightforward approach. For this comparison I integrated all three data types within a first order PPI network, as shown in Figure 11. Here, we can see relatively equal support in the network from the transcriptome, proteome, and metabolome.

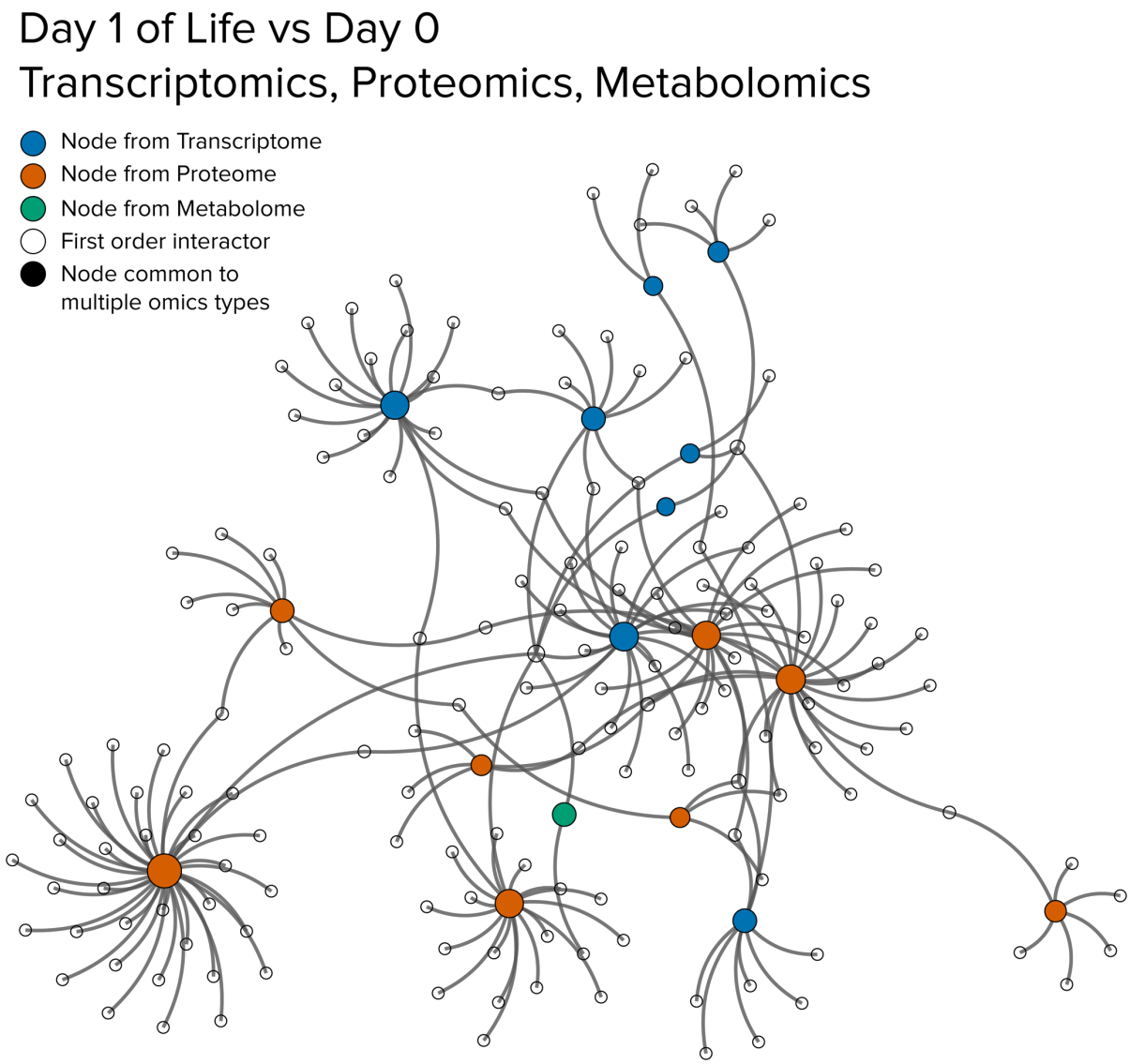

Figure 11. Multi-Omic Changes on Day 1 of Life vs Day 0. Shown is the integrated first-order PPI network of multi-omic changes on day 1 of life vs day 0 of life. With only modest changes in expression/abundance across all datasets, a relatively small and poorly connected PPI network was generated, even when first-order interactors were included.

However, when comparing days 3 and 7 of life to day 0, the disproportionately large signal from the transcriptome presented a challenge. This is likely related to the fact that transcriptomic reports on all the genes in the cell while proteomics and metabolomics report on more abundant proteins and endpoints of metabolism. I tackled this challenge with two approaches. First, I integrated all three omics types directly, utilizing zero-order networks to limit the number of nodes in the network (Figures 11-13). This technique provided limited success, however, as the disproportionately large transcriptomic seed list dominated the

signal of the network. Therefore, I also performed a three-way integration by integrating the proteome and metabolome, then looking for transcriptome hits within these networks (Figure 14).

Despite the lack of transcriptome signal seeding the network, I observed a proportion of the network represented by the transcriptome signature. Pictured in blue, these nodes were present both as first-order interactors as well as connections between proteome and metabolome nodes, showing significant overlap with the transcriptome as well. This indicates that the data from these omics platforms report on the same biological phenomena. Figure 15 provides a quantitative breakdown of the composition of each of these networks. By performing a three-way integration creating a zero-order network, I reduced the amount of noise in the network by eliminating incorporation of first-order interactors. However, this came at the expense of a transcriptome-dominated network, with less proportional support from each omics type. Additionally, this approach reduced the overlap between omics types, as demonstrated by the reduced number of nodes represented by multiple omics types (from 15 down to 8). This illustrates one of the trade-offs of the different network types discussed in Chapter 1.3.3.

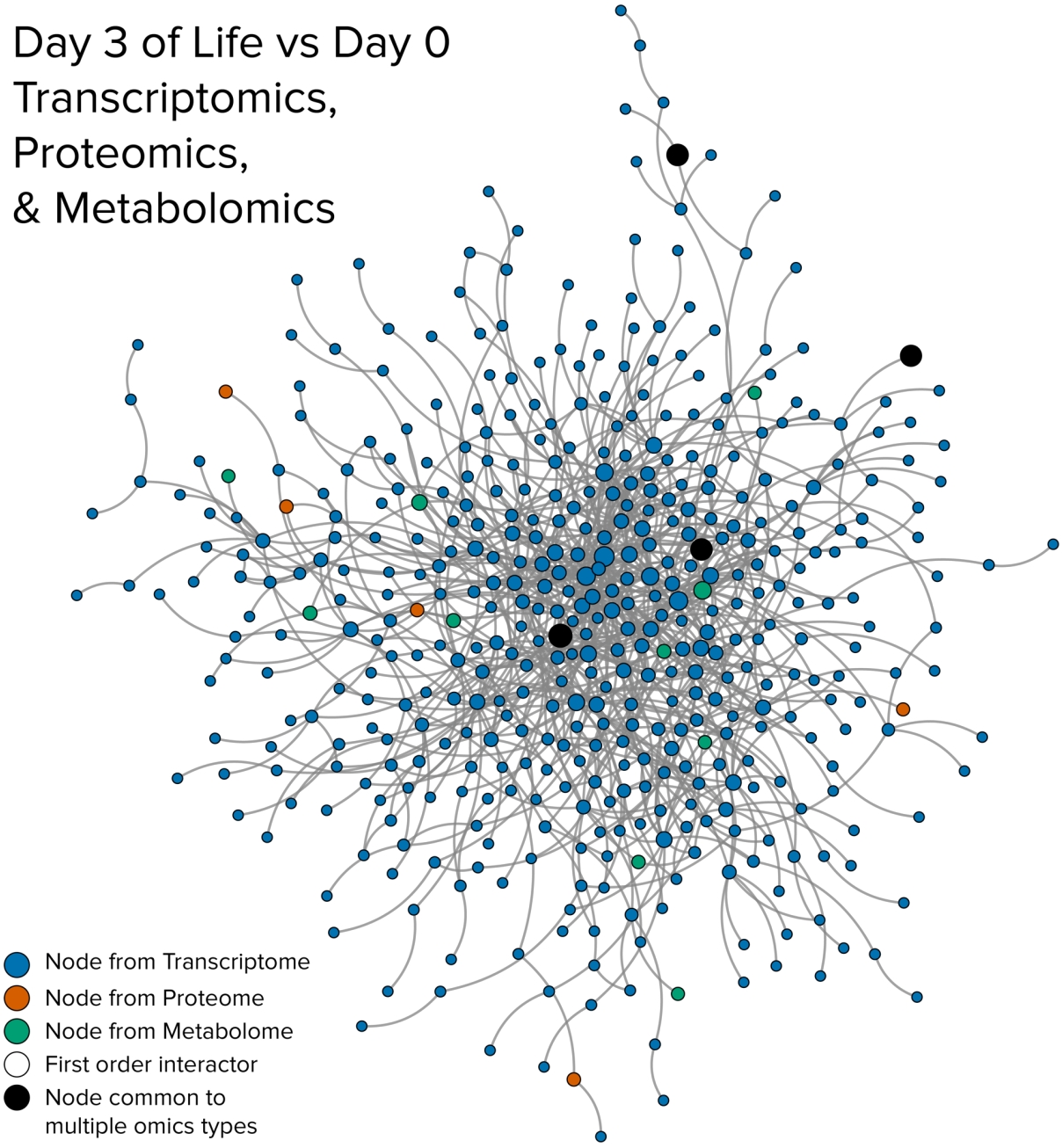

Figure 12. By day 3 of life, transcriptional changes overwhelmed a multi-omic network. Shown is an integrated zero-order PPI network of multi-omic changes on day 3 of life vs. day 0 of life. Even without first-order interactors, a large network was generated, dominated by the signal from the transcriptome.

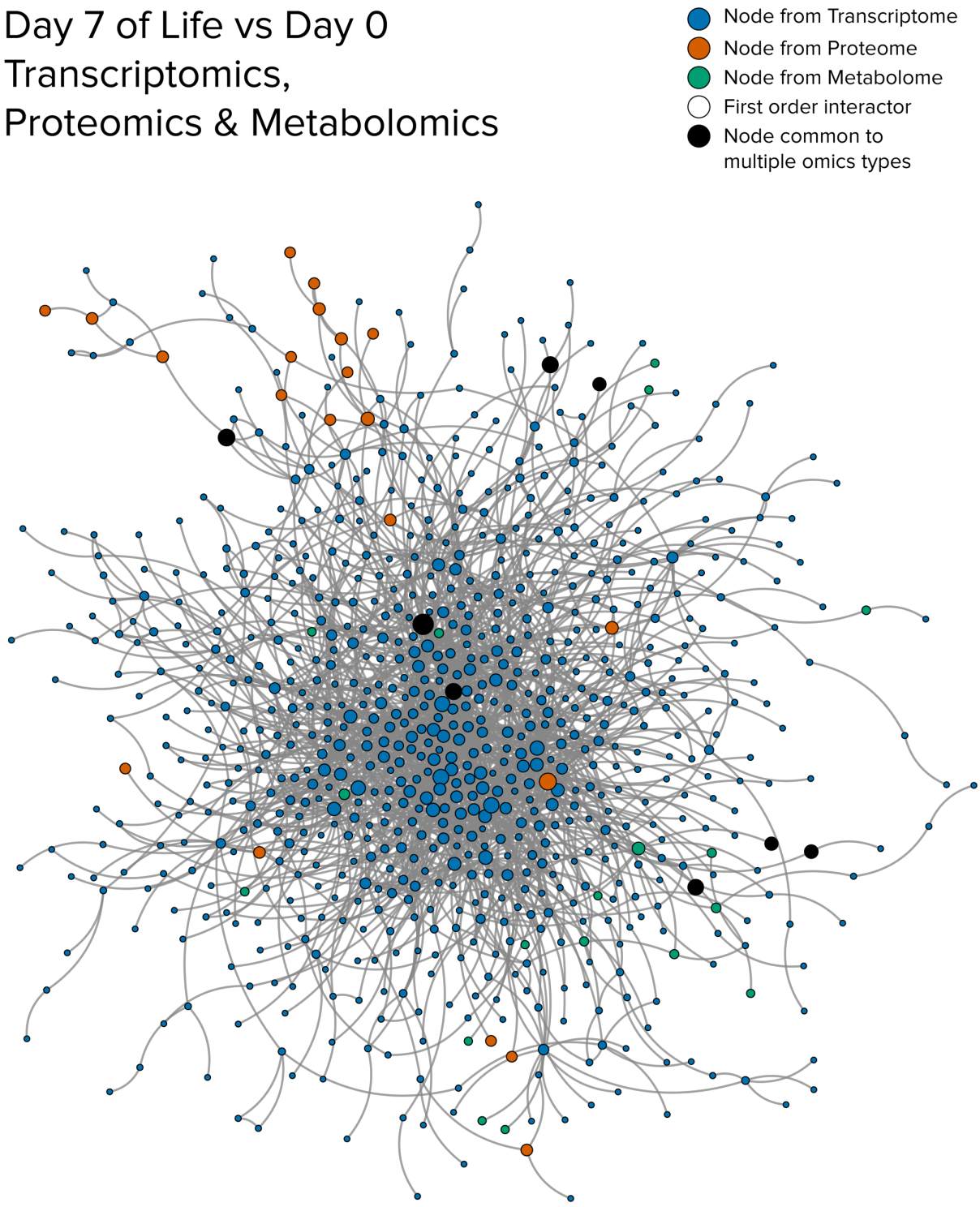

Figure 13. On day 7 of life, transcriptional changes overwhelmed a multi-omic network. Shown is an integrated zero-order PPI network of multi-omic changes on day 7 of life vs day 0 of life. Even without first-order interactors, a large network was generated, dominated by the signal from the transcriptome.

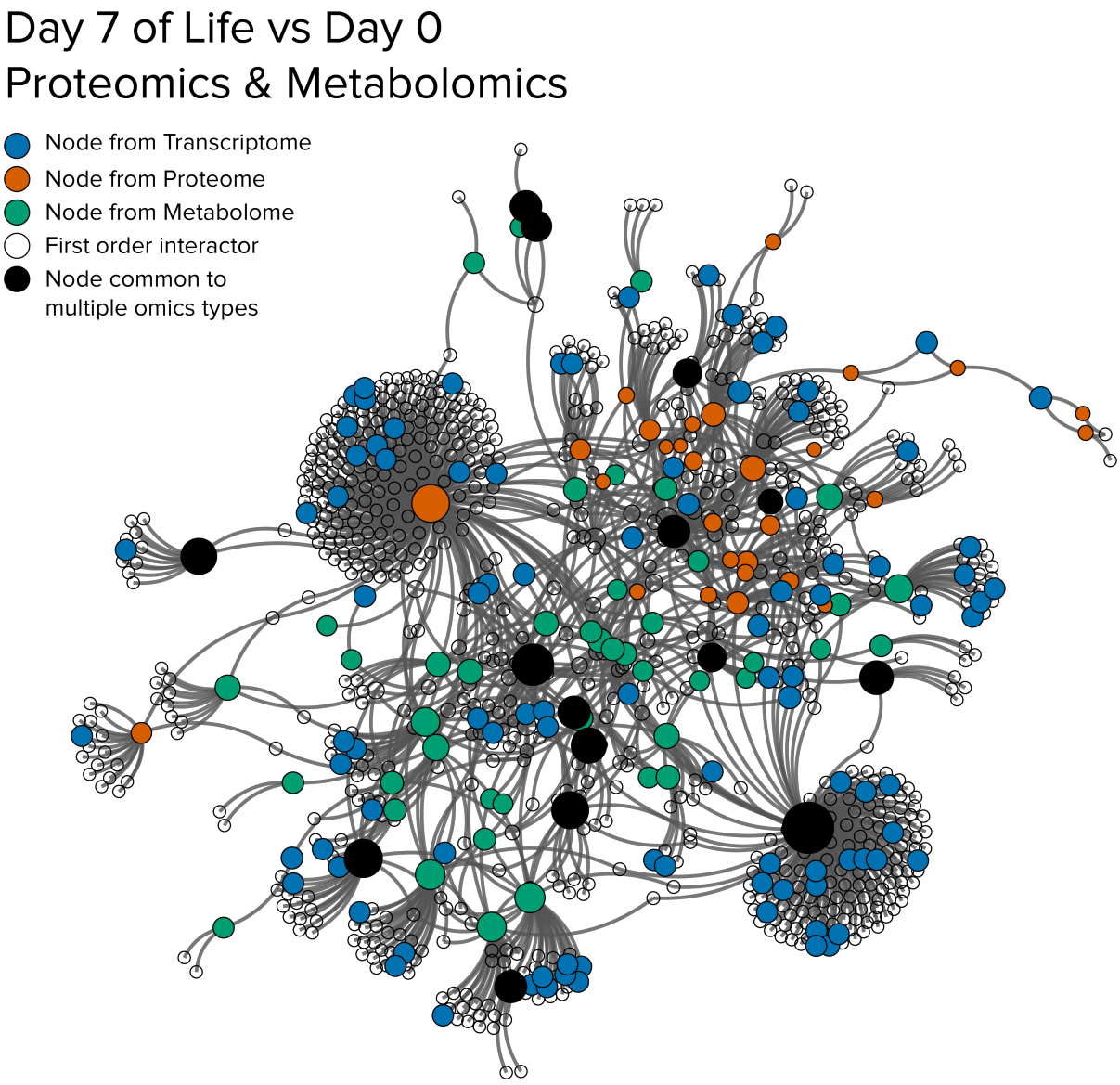

Figure 14. A bi-omic integrated network can be useful when there is a disproportionately large transcriptome signal. Shown is a first-order PPI network generated from integrated metabolome and proteome data comparing day 7 of life to day 0 of life.

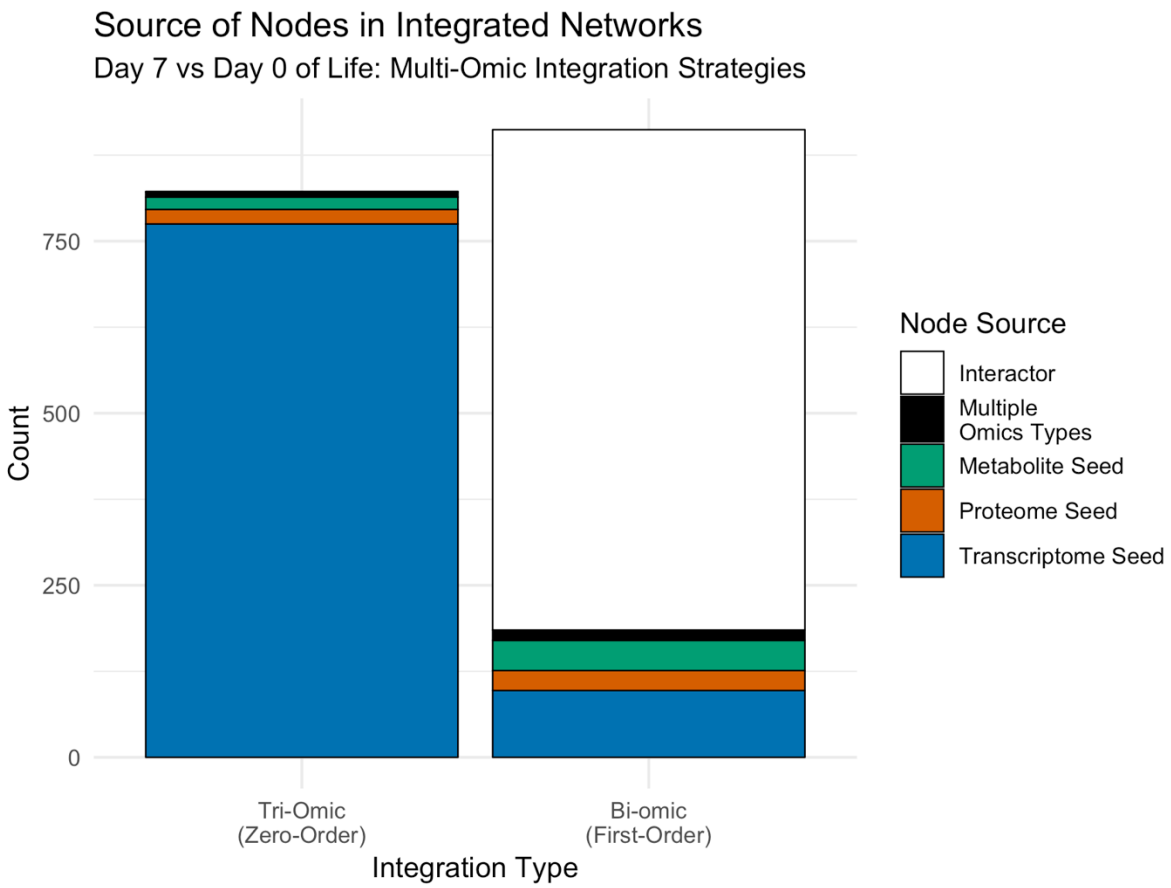

Figure 15. Bi-omic integration improved the proportions of omic representation but incorporated large numbers of first order-interactors. Shown is a comparison of the node composition from the two types of networks generated to integrate multiple omics types when comparing day 7 to day 0 of life. Notable, the number of seeds from multiple omics types was increased in the bi-omic integration, compared to the tri-omic integration (from 8 to 15).

However, the bi-omic integration strategy is not without issues, as the number of interactors incorporated was large, increasing the amount of noise (nodes with only a single interaction) in the network.

3.2.2 Biological Conclusions

The integrated networks shown here provided corroboration of other observations within the EPIC study, and also provided novel biological insights. From each day of life, the integrated networks were compared to each omics type in isolation. Additionally, collaborators applied data-driven statistical methods for integrative analysis such as those discussed in Chapter 1.2 involving multivariate ("data integration analysis for biomarker

discovery using latent components") and multifactorial ("multi-scale, multifactorial response network") approaches (data not shown).[14,19,65]

The integrated networks were enriched for interferon and complement pathways across the first week of life, confirming data obtained from individual omics types. Additionally, the integrated networks were enriched for cellular replication and creatine metabolism on day three of life, as well as fibrin clotting, adaptive immunity, and phagosome activity on day seven of life. These pathways were not significantly enriched in any omics type alone (data not shown).

Importantly, when the results from the network-based integration were compared to other aforementioned integrative analysis methods, there was significant overlap in their results. Specifically, 249 pathways were enriched across all methods, with 34 pathways enriched in all three methods ($p < 0.001$). Among these consensus-enriched pathways were the previously mentioned interferon and complement pathways. This indicates significant biological consensus from all integrative analysis approaches.

This demonstrates that network-based integrative analysis is a valuable tool for informing biological consensus and uncovering novel biological insights. Crucially, I have shown that the methodology presented here is effective when applied to complex experimental datasets, and not just when applied to signatures.

Chapter 4: Conclusion

In this thesis I have presented a novel integrative analysis method for multi-omic data, with a focus on integration of metabolomics data. Network-based integrative analysis with MetaBridge offers meaningful biological integration of multiple data types due to its emphasis on interaction-based mapping via PPIs, an accurate proxy for biological function.

Furthermore, I have shown that integrative analysis of multi-omic data with MetaBridge is applicable to integration of multi-omic datasets from multiple sources. Network-based integrative analysis is powerful when leveraging curated PPIs, but the flexibility of this methodology allows researchers to use other interactome and metabolome sources to expand the scope of their investigation. One can even do so directly in NetworkAnalyst by choosing a different interactome source.

4.1 Applications

Here I have shown how network-based integrative analysis of multi-omic data with MetaBridge and NetworkAnalyst can be used as a powerful, intuitive, and flexible web-based tool for hypothesis generation. I have shown how the methodology can integrate multi-omic datasets—in particular, metabolomics data—with curated metabolite-protein and PPI data. Additionally, I have demonstrated how the method can be adapted for use with two and three, and potentially even more omics types.

Despite the preliminary findings that networks integrated from biologically-similar multi-omic signatures did not differ qualitatively from randomly integrated networks in key network statistics, I showed that the biology represented in such networks remained unique. Furthermore, investigation of these nodes of interest showed a striking enrichment in genes coding for proteins pertinent to sepsis, as well as several which could serve as novel targets for investigation.

Finally, I illustrated how data of varying orders of magnitude can be meaningfully integrated to provide relevant biological insight, albeit involving certain tradeoffs. I demonstrated how results from the multi-omic integration can be filtered for relevance and promiscuity, ensuring results to not simply consist of the most high-connected proteins in a given interactome.

4.2 Limitations

Network-based integrative analysis of multi-omic data with MetaBridge and NetworkAnalyst is not without its limitations.

4.2.1 Reliance on Annotations

The first limitation of the methodology described here is its reliance on curated metabolite-protein interactions and PPIs. MetaBridge can use MetaCyc or KEGG to identify proteins which interact with a given set of metabolites and my results show that even using these two databases for metabolite interactor identification causes differences in output. Additionally, the methodology relies on InnateDB for curated PPIs to construct its PPI networks. Therefore, this integrative analysis technique de-emphasizes poorly-studied proteins, and is unable to identify uncharacterized targets, a common limitation in all bioinformatics analyses of omics datasets. However, I believe the strength of the method lies in its ability to identify novel functions of known targets (pathways and perhaps ontologies) based on these annotations. Furthermore, the use of random discovery rate as a filtration mechanism should eventually allow for less-studied proteins surface as potential targets of interest.

4.2.2 Promiscuous Proteins

Another limitation of network-based integrative analysis is the propensity for "promiscuous" proteins to be overrepresented. Promiscuous proteins are those that by their nature interact with numerous other proteins in cells and thus serve as a type of glue that can interconnect proteins with disparate functions. The method presented here integrates PPI networks, which can feature highly-interconnected proteins as network hubs.

These proteins may be key players in the biological condition under investigation, or they may interact with many things in the cell in a manner irrelevant to the biological question under investigation. One example of this is the protein ubiquitin, which occurs, as its name suggests, ubiquitously. Ubiquitin tags proteins for degradation, and thus curation has shown it, and the enzymes mediating its addition and removal, to interact with many proteins in the cell.

Unfortunately, aside from well-known promiscuous proteins (e.g. UBC, HNF4A) it can be difficult to determine what edges of a PPI network are driven by nonspecific binding, and thus determine which hubs represent promiscuous proteins.[66] Therefore, aside from these well-characterized promiscuous proteins, highly-connected proteins cannot be unilaterally removed from a given PPI network, as the edges could represent significant, specific binding. However, strategies can be employed to downplay their significance of suspected promiscuous proteins in the results generated from such PPI networks.

Methodologies such as the calculation of random discovery rate described here can be employed to prevent presentation of these promiscuous proteins as potential targets of interest. In this regard, it might be very useful to generate a reference list of these proteins. Additionally, other network analysis techniques could be implemented to reduce the effects of these promiscuous proteins on the users' hypotheses. For example, the use of pathway enrichment analysis, as described in Chapter 4.4.2.

4.3 Recent Developments

Since the inception of this project, more multi-omic integration methods targeted to individuals with little technical experience have come onto the market. In particular, OmicsNet was recently released, employing many of the methods discussed in this thesis.[67] Currently, OmicsNet is principally a tool for visualization of multi-omic datasets, rather than analysis. However, in contrast to the methods detailed in this thesis, OmicsNet provides a single web interface for integration of multiple data types in a network-based manner, also accepting transcription factor binding data and microRNA data.

OmicsNet is further limited in is to metabolite-protein mapping, using only the the KEGG and Recon2 databases, not incorporating data from the richly-curated BioCyc platform, one of the strengths of MetaBridge. Furthermore, OmicsNet does not describe methods for meaningful extraction of targets of interest from the resultant networks, as described here. Currently, OmicsNet is principally a tool for visualization of multi-omic datasets, rather than analysis.

4.4 Future Directions

4.4.1 Promiscuous Protein Removal

One undertaking of particular interest would be to examine the role of promiscuous proteins in integrated networks. Certain promiscuous proteins recur in PPI networks without pertinence to the biological state under investigation. Therefore, it would be interesting to examine how networks change when these proteins are removed, and what qualities of a network predict how it will change. Would the network structure remain mostly intact, indicating these proteins were not integral to the network? Or would the network splinter into multiple subnetworks, indicating the protein was integral to the network's structure, and thus, might impact the biological conclusions? Would networks generated randomly splinter more or less often than networks generated from real biological data? If the former, a network surviving promiscuous protein removal could be a good indication of network cohesion.

In pursuing this question, it would be very important to define a list of promiscuous proteins as those proteins which are the most highly connected in the interactome database InnateDB. If such a list was curated, it would also be possible to control for these promiscuous proteins in future analyses, rather than remove them from networks.

4.4.2 Pathway Enrichment Comparisons

Another undertaking that could prove fruitful would be expanding the analysis methods of the resulting integrated networks. One such method of analysis would be pathway enrichment. As mentioned in Chapter 2.4, meaningful comparison of pathway enrichment analyses is currently a major goal in the field. It would be interesting to apply some of the more recent concepts and pathway analysis tools developed in order to facilitate a deeper understanding of the biology represented by these integrated networks, particularly when comparing two similar networks. Specifically, how does the pathway enrichment of a PPI network alter when metabolomics data is incorporated? Which pathways become more significantly enriched, and what new pathways appear?

4.4.3 Random Discovery Rate

One key to the success of this strategy will be filtering out nodes with a high likelihood to occur at random to identify nodes of interest. Preliminary exploration of this issue was performed in Appendix D. Unfortunately, this rate likely needs to be calculated separately for every different set of inputs using computationally-demanding network simulations. Therefore, it would be a useful addition to such an integrative analysis tool to provide pre-calculated metrics for random discovery rates with various parameters. This could aid researchers in deciding which nodes of their networks are unique, and not likely to have shown up at random.

4.4.4 Further Integrations

Another direction would be further applications of network-based integrative analysis with MetaBridge to provide further biological comparisons to appropriate nulls. Here I applied the methodology to two datasets—one signature integration and one experimental data integration. However, it would be useful to apply the methodology to more, varied, datasets to further examine its advantages and drawbacks.

4.5 Concluding Remarks

The research described in this thesis provides novel approaches for the integrative analysis of multi-omic data. I detailed the landscape of tools currently available for integrative analysis of multi-omic data and described how this methodology can be utilized to obtain biological insights. Finally, I established the effectiveness of this methodology in multiple use cases and specified how the methodology could be improved in the future.

Bibliography

1. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).
2. Wang, Y.-C., Peterson, S. E. & Loring, J. F. Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res.* **24,** 143–160 (2014).
3. Karahalil, B. Overview of Systems Biology and Omics Technologies. *Curr. Med. Chem.* **23,** 4221–4230 (2016).
4. Kaever, A. *et al.* Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets. *PLoS ONE* **9,** e89297 (2014).
5. Palsson, B. & Zengler, K. The challenges of integrating multi-omic data sets. *Nat. Chem. Biol.* **6,** 787–9 (2010).
6. Misra, B. B., Langefeld, C. D., Olivier, M. & Cox, L. A. Integrated Omics: Tools, Advances, and Future Approaches. *J. Mol. Endocrinol.* JME-18-0055 (2018). doi:10.1530/JME-18-0055
7. Trivedi, D. K., Hollywood, K. A. & Goodacre, R. Metabolomics for the masses: The future of metabolomics in a personalized world. *New Horiz. Transl. Med.* **3,** 294–305 (2017).
8. Kohl, M. *et al.* A practical data processing workflow for multi-OMICS projects. *Biochim. Biophys. Acta* **1844,** 52–62 (2014).
9. Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R. & Griffin, J. L. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* **40,** 387–426 (2011).
10. Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2,** 345–350 (2005).
11. Psychogios, N. *et al.* The human serum metabolome. *PloS One* **6,** e16957 (2011).
12. Nie, L., Wu, G., Culley, D. E., Scholten, J. C. M. & Zhang, W. Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications. *Crit. Rev. Biotechnol.* **27,** 63–75 (2007).
13. Wanichthanarak, K., Fahrmann, J. F. & Grapov, D. Genomic, Proteomic, and Metabolomic Data Integration Strategies. *Biomark. Insights* **10s4,** BMI.S29511 (2015).

14. Rohart, F., Gautier, B., Singh, A. & Cao, K.-A. L. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13,** e1005752 (2017).
15. Tuncbag, N., McCallum, S., Huang, S.-S. C. & Fraenkel, E. SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways. *Nucleic Acids Res.* **40,** W505-509 (2012).
16. Fukushima, A. DiffCorr: An R package to analyze and visualize differential correlations in biological networks. *Gene* **518,** 209–214 (2013).
17. Tuncbag, N. *et al.* Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLOS Comput. Biol.* **12,** e1004879 (2016).
18. Lê Cao, K.-A., González, I. & Déjean, S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* **25,** 2855–2856 (2009).
19. Singh, A. *et al.* DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. *bioRxiv* 067611 (2018). doi:10.1101/067611
20. Chong, J., Xia, J., Chong, J. & Xia, J. Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. *Metabolites* **7,** 62 (2017).
21. Breuer, K. *et al.* InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* **41,** D1228–D1233 (2013).
22. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
23. Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42,** D358–D363 (2014).
24. Xia, J., Sinelnikov, I. V., Han, B. & Wishart, D. S. MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Res.* **43,** W251–W257 (2015).
25. Kamburov, A., Cavill, R., Ebbels, T. M. D., Herwig, R. & Keun, H. C. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* **27,** 2917–2918 (2011).
26. Hernández-de-Diego, R. *et al.* PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.* **46,** W503–W509 (2018).

27. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45,** D353–D361 (2017).
28. Gosline, S. J. C., Oh, C. & Fraenkel, E. SAMNetWeb: identifying condition-specific networks linking signaling and transcription. *Bioinformatics* **31,** 1124–1126 (2015).
29. Karnovsky, A. *et al.* Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* **28,** 373–380 (2012).
30. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13,** 2498–2504 (2003).
31. Wachter, A. & Beißbarth, T. pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics* **31,** 3072–3074 (2015).
32. Grapov, D., Wanichthanarak, K. & Fiehn, O. MetaMapR: pathway independent metabolomic network analysis incorporating unknowns. *Bioinformatics* **31,** 2757–2760 (2015).
33. Sun, H. *et al.* iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics* **30,** 737–739 (2014).
34. Kuo, T.-C., Tian, T.-F. & Tseng, Y. J. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst. Biol.* **7,** 64 (2013).
35. Fan, J. *et al.* Galaxy Integrated Omics: Web-based Standards-Compliant Workflows for Proteomics Informed by Transcriptomics. *Mol. Cell. Proteomics* **14,** 3087–3093 (2015).
36. Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9,** 345–350 (2012).
37. Lynn, D. J. *et al.* Curating the innate immunity interactome. *BMC Syst. Biol.* **4,** 117 (2010).
38. Altman, T., Travers, M., Kothari, A., Caspi, R. & Karp, P. D. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* **14,** 112 (2013).
39. Yeung, A. T. Y. *et al.* Conditional-ready mouse embryonic stem cell derived macrophages enable the study of essential genes in macrophage function. *Sci. Rep.* **5,** (2015).

40. Xia, J., Benner, M. J. & Hancock, R. E. W. NetworkAnalyst - integrative approaches for protein–protein interaction network analysis and visual exploration. *Nucleic Acids Res.* **42,** W167–W174 (2014).
41. Takahashi, H. An approximate solution for the Steiner problem in graphs. *Math Jpn.* **6,** 573–577 (1990).
42. Hinshaw, S. J., Lee, A. H. Y., Gill, E. E. & Hancock, R. E. W. MetaBridge: Enabling Network-Based Integrative Analysis via Direct Protein Interactors of Metabolites. *Bioinforma. Oxf. Engl.* (2018). doi:10.1093/bioinformatics/bty331
43. Xia, J., Benner, M. J. & Hancock, R. E. W. NetworkAnalyst - integrative approaches for protein–protein interaction network analysis and visual exploration. *Nucleic Acids Res.* **42,** W167–W174 (2014).
44. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–D462 (2016).
45. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44,** D471–D480 (2016).
46. Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. *shiny: Web Application Framework for R*. (2017).
47. Karp, P. D. *et al.* Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief. Bioinform.* **17,** 877–890 (2016).
48. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29,** 1830–1831 (2013).
49. Xia, J., Gill, E. E. & Hancock, R. E. W. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.* **10,** 823–844 (2015).
50. Donald, M. R. & Wilson, S. R. Comparison and visualisation of agreement for paired lists of rankings. *Stat. Appl. Genet. Mol. Biol.* **16,** 31–45 (2017).
51. Lottaz, C., Yang, X., Scheid, S. & Spang, R. OrderedList—a bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics* **22,** 2315–2316 (2006).

52. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).
53. Raman, K., Damaraju, N. & Joshi, G. K. The organisational structure of protein networks: revisiting the centrality–lethality hypothesis. *Syst. Synth. Biol.* **8,** 73–81 (2014).
54. Batada, N. N., Hurst, L. D. & Tyers, M. Evolutionary and Physiological Importance of Hub Proteins. *PLoS Comput. Biol.* **2,** (2006).
55. He, X. & Zhang, J. Why Do Hubs Tend to Be Essential in Protein Networks? *PLOS Genet.* **2,** e88 (2006).
56. Patil, A., Kinoshita, K. & Nakamura, H. Hub Promiscuity in Protein-Protein Interaction Networks. *Int. J. Mol. Sci.* **11,** 1930–1943 (2010).
57. Bader, J. S., Chaudhuri, A., Rothberg, J. M. & Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* **22,** 78–85 (2004).
58. Mistry, D., Wise, R. P. & Dickerson, J. A. DiffSLC: A graph centrality method to detect essential proteins of a protein-protein interaction network. *PLOS ONE* **12,** e0187091 (2017).
59. Qin, C., Sun, Y. & Dong, Y. A New Method for Identifying Essential Proteins Based on Network Topology Properties and Protein Complexes. *PLOS ONE* **11,** e0161042 (2016).
60. Zhang, X., Xiao, W., Acencio, M. L., Lemke, N. & Wang, X. An ensemble framework for identifying essential proteins. *BMC Bioinformatics* **17,** 322 (2016).
61. Pena, O. M. *et al.* An Endotoxin Tolerance Signature Predicts Sepsis and Organ Dysfunction at Initial Clinical Presentation. *EBioMedicine* **1,** 64–71 (2014).
62. Mickiewicz, B. *et al.* Integration of metabolic and inflammatory mediator profiles as a potential prognostic approach for septic shock in the intensive care unit. *Crit. Care* **19,** 11 (2015).
63. Sadeghi, A. *SteinerNet: Steiner Tree Approach for Graph Analysis*. (2018).
64. Sadeghi, A. & Fröhlich, H. Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics* **14,** 144 (2013).
65. Li, S. *et al.* Metabolic Phenotypes Of Response to Vaccination in Humans. *Cell* **169,** 862-877.e17 (2017).

66. Schreiber, G. & Keating, A. E. Protein Binding Specificity versus Promiscuity. *Curr. Opin. Struct. Biol.* **21,** 50–61 (2011).
67. Zhou, G. & Xia, J. OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res.* **46,** W514–W522 (2018).
68. Russell, J. A. *et al.* Vasopressin versus Norepinephrine Infusion in Patients with Septic Shock. *http://dx.doi.org/10.1056/NEJMoa067373* (2009). doi:10.1056/NEJMoa067373
69. Vary, T. C., Siegel, J. H., Nakatani, T., Sato, T. & Aoyama, H. A Biochemical Basis for Depressed Ketogenesis in Sepsis. *J. Trauma Acute Care Surg.* **26,** 419 (1986).
70. Takeyama, N., Takagi, D., Matsuo, N., Kitazawa, Y. & Tanaka, T. Altered hepatic fatty acid metabolism in endotoxicosis: effect of L-carnitine on survival. *Am. J. Physiol.-Endocrinol. Metab.* **256,** E31–E38 (1989).
71. Yamamoto, T. Rat liver peroxisomal and mitochondrial fatty acid oxidation in sepsis. *Surg. Today* **23,** 137–143 (1993).
72. Wijnands, K. A. P., Castermans, T. M. R., Hommen, M. P. J., Meesters, D. M. & Poeze, M. Arginine and Citrulline and the Immune Response in Sepsis. *Nutrients* **7,** 1426–1463 (2015).
73. Winkler, M. S. *et al.* Markers of nitric oxide are associated with sepsis severity: an observational study. *Crit. Care* **21,** (2017).
74. Choe, C. *et al.* Homoarginine levels are regulated by L-arginine:glycine amidinotransferase and affect stroke outcome: results from human and murine studies. *Circulation* **128,** 1451–1461 (2013).
75. Piagnerelli, M., Boudjeltia, K. Z., Gulbis, B., Vanhaeverbeek, M. & Vincent, J.-L. Anemia in sepsis: the importance of red blood cell membrane changes. *Transfus. Altern. Transfus. Med.* **9,** 143–149 (2007).
76. Yoo, H., Ku, S.-K., Kim, S.-W. & Bae, J.-S. Early Diagnosis of Sepsis Using Serum Hemoglobin Subunit Beta. *Inflammation* **38,** 394–399 (2015).
77. Yang, G., Li, T., Xu, J., Peng, X. & Liu, L. Mitogen-activated protein kinases regulate vascular reactivity after hemorrhagic shock through myosin light chain phosphorylation pathway. *J. Trauma Acute Care Surg.* **74,** 1033 (2013).

78. Davidson, D. *et al.* Gene Expression Profile of Endotoxin-stimulated Leukocytes of the Term New Born: Control of Cytokine Gene Expression by Interleukin-10. *PLOS ONE* **8,** e53641 (2013).
79. Müller, M. M. *et al.* Global analysis of glycoproteins identifies markers of endotoxin tolerant monocytes and GPR84 as a modulator of TNFα expression. *Sci. Rep.* **7,** 838 (2017).
80. Unuma, K., Aki, T., Funakoshi, T., Yoshida, K. & Uemura, K. Cobalt Protoporphyrin Accelerates TFEB Activation and Lysosome Reformation during LPS-Induced Septic Insults in the Rat Heart. *PLoS ONE* **8,** (2013).
81. Ma, J. *et al.* Lysosome and Cytoskeleton Pathways Are Robustly Enriched in the Blood of Septic Patients: A Meta-Analysis of Transcriptomic Data. *Mediators Inflamm.* **2015,** (2015).
82. Giegerich, A. K. *et al.* Autophagy-dependent PELI3 degradation inhibits proinflammatory IL1B expression. *Autophagy* **10,** 1937–1952 (2014).
83. Fortunato, F. *et al.* Impaired Autolysosome Formation Correlates With Lamp-2 Depletion: Role of Apoptosis, Autophagy, and Necrosis in Pancreatitis. *Gastroenterology* **137,** 350-360.e5 (2009).
84. Mühl, D. *et al.* Dynamic changes of matrix metalloproteinases and their tissue inhibitors in severe sepsis. *J. Crit. Care* **26,** 550–555 (2011).
85. Yazdan-Ashoori, P. *et al.* Elevated plasma matrix metalloproteinases and their tissue inhibitors in patients with severe sepsis. *J. Crit. Care* **26,** 556–565 (2011).
86. Al-Haj, L. & Khabar, K. S. A. The intracellular pyrimidine 5′-nucleotidase NT5C3A is a negative epigenetic factor in interferon and cytokine signaling. *Sci Signal* **11,** eaal2434 (2018).
87. Zhang, Z. *et al.* Plumbagin Protects Mice from Lethal Sepsis by Modulating Immunometabolism Upstream of PKM2. *Mol. Med. Camb. Mass* (2016). doi:10.2119/molmed.2015.00250
88. Palsson-McDermott, E. M. *et al.* Pyruvate Kinase M2 Is Required for the Expression of the Immune Checkpoint PD-L1 in Immune Cells and Tumors. *Front. Immunol.* **8,** (2017).

89. Deng, W. *et al.* The Circadian Clock Controls Immune Checkpoint Pathway in Sepsis. *Cell Rep.* **24,** 366–378 (2018).
90. Huang, J. *et al.* AMPK regulates immunometabolism in sepsis. *Brain. Behav. Immun.* **72,** 89–100 (2018).
91. Palsson-McDermott, E. M. *et al.* Pyruvate Kinase M2 regulates Hif-1α activity and IL-1β induction, and is a critical determinant of the Warburg Effect in LPS-activated macrophages. *Cell Metab.* **21,** 65–80 (2015).
92. Panetta, R., Guo, Y., Magder, S. & Greenwood, M. T. Regulators of G-Protein Signaling (RGS) 1 and 16 Are Induced in Response to Bacterial Lipopolysaccharide and Stimulate c-fos Promoter Expression. *Biochem. Biophys. Res. Commun.* **259,** 550–556 (1999).
93. Hussein, S. *et al.* Characterization of human septic sera induced gene expression modulation in human myocytes. *Int. J. Clin. Exp. Med.* **2,** 131–148 (2009).
94. Igietseme, J. U. *et al.* Role of Epithelial-Mesenchyme Transition in Chlamydia Pathogenesis. *PLoS ONE* **10,** (2015).
95. Servin, A. L. Pathogenesis of Human Diffusely Adhering Escherichia coli Expressing Afa/Dr Adhesins (Afa/Dr DAEC): Current Insights and Future Challenges. *Clin. Microbiol. Rev.* **27,** 823–869 (2014).
96. Pena, O. M. Exploring the development of endotoxin tolerance during sepsis and a possible immunomodulatory therapy. (University of British Columbia, 2013). doi:10.14288/1.0074260
97. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44,** D471–D480 (2016).
98. Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* **5,** (2018).
99. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* (2013). doi:10.1038/ng.2653
100. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46,** D608–D617 (2018).

101. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44,** D463-470 (2016).

102. Haug, K. *et al.* MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41,** D781–D786 (2013).

Appendices

Appendix A - 15-Metabolite Signature Mapping to Enzymes and Genes

Table 1. 15-metabolite sepsis signature maps to human enzymes and genes via the MetaCyc database. Shown here is the 15-metabolite sepsis signature represented in HMDB IDs. It has been mapped to its directly interacting enzymes, and those enzymes human genes via the MetaCyc database.

HMDB	MetaCyc Compound ID	MetaCyc Reaction ID	MetaCyc Gene ID	Gene Symbol	Ensembl ID
HMDB00159	PHE	RXN66-445	HS09679	SLC3A2	ENSG00000168003
HMDB00159	PHE	RXN66-445	HS02481	SLC7A5	ENSG00000103257
HMDB00357	CPD-335	3-HYDROXYBUTYRATE-DEHYDROGENASE-RXN	HS08579	BDH1	ENSG00000161267
HMDB00357	CPD-335	3-HYDROXYBUTYRATE-DEHYDROGENASE-RXN	HS08987	BDH2	ENSG00000164039
HMDB00294	UREA	AGMATIN-RXN	HS04051	AGMAT	ENSG00000116771
HMDB00294	UREA	ARGINASE-RXN	HS04231	ARG1	ENSG00000118520
HMDB00294	UREA	ARGINASE-RXN	HS01388	ARG2	ENSG00000081181
HMDB00201	O-ACETYLCARNITINE	CARNITINE-O-ACETYLTRANSFERASE-RXN	HS01816	CRAT	ENSG00000095321
HMDB00516	GLC	3.2.1.106-RXN	HS03863	MOGS	ENSG00000115275
HMDB01875	METOH	RXN-13425	HS11616	CES1	ENSG00000159398
HMDB01875	METOH	RXN-13424	HS11616	CES1	ENSG00000159398
HMDB00148	GLT	RXN-2901	HS02477	ABAT	ENSG00000183044
HMDB00148	GLT	PSERTRANSAMPYR-RXN	HS05946	PSAT1	ENSG00000135069
HMDB00148	GLT	RXN-11430	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11430	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-11433	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11433	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-11432	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11432	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-11435	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11435	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-11434	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11434	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-11431	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11431	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-11429	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11429	HS01949	GGT5	ENSG00000099998

HMDB	MetaCyc Compound ID	MetaCyc Reaction ID	MetaCyc Gene ID	Gene Symbol	Ensembl ID
HMDB00148	GLT	RXN-18759	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-18759	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-11428	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11428	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-11664	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-11664	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-13675	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-13675	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-19572	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-19572	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	FORMYLTHFGLUSYNTH-RXN	HS06237	FPGS	ENSG00000136877
HMDB00148	GLT	2.6.1.22-RXN	HS02477	ABAT	ENSG00000183044
HMDB00148	GLT	4-HYDROXYGLUTAMATE-AMINOTRANSFERASE-RXN	HS04858	GOT2	ENSG00000125166
HMDB00148	GLT	ALANINE-AMINOTRANSFERASE-RXN	HS09610	GPT	ENSG00000167701
HMDB00148	GLT	ALANINE-AMINOTRANSFERASE-RXN	HS09332	GPT2	ENSG00000166123
HMDB00148	GLT	RXN-19604	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-19604	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-19627	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-19627	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-10721	HS03239	AADAT	ENSG00000109576
HMDB00148	GLT	RXN-10721	HS06422	CCBL2	ENSG00000137944
HMDB00148	GLT	PSERTRANSAM-RXN	HS05946	PSAT1	ENSG00000135069
HMDB00148	GLT	GABATRANSAM-RXN	HS02477	ABAT	ENSG00000183044
HMDB00148	GLT	5-OXOPROLINASE-ATP-HYDROLYSING-RXN	HS11319	OPLAH	ENSG00000178814
HMDB00148	GLT	PRPPAMIDOTRANS-RXN	HS05157	PPAT	ENSG00000128059
HMDB00148	GLT	RXN-19607	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-19607	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-19608	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-19608	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-6641	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-6641	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-18176	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-18176	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	FGAMSYN-RXN	HS11329	PFAS	ENSG00000178921

HMDB	MetaCyc Compound ID	MetaCyc Reaction ID	MetaCyc Gene ID	Gene Symbol	Ensembl ID
HMDB00148	GLT	NAD-SYNTH-GLN-RXN	HS10587	NADSYN1	ENSG00000172890
HMDB00148	GLT	RXN-12618	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-12618	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-19578	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-19578	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	2-AMINOADIPATE-AMINOTRANSFERASE-RXN	HS03239	AADAT	ENSG00000109576
HMDB00148	GLT	RXN-13697	HS04858	GOT2	ENSG00000125166
HMDB00148	GLT	DIHYDROFOLATESYNTH-RXN	HS06237	FPGS	ENSG00000136877
HMDB00148	GLT	GLUTKIN-RXN	HS00730	ALDH18A1	ENSG00000059573
HMDB00148	GLT	GLUTDECARBOX-RXN	HS05215	GAD1	ENSG00000128683
HMDB00148	GLT	GLUTDECARBOX-RXN	HS06208	GAD2	ENSG00000136750
HMDB00148	GLT	GLUTCYSLIG-RXN	HS00071	GCLC	ENSG00000001084
HMDB00148	GLT	GLUTCYSLIG-RXN	HS00434	GCLM	ENSG00000023909
HMDB00148	GLT	GLUTAMATE-DEHYDROGENASE-NADP+-RXN	HS07548	GLUD1	ENSG00000148672
HMDB00148	GLT	GLUTAMATE-DEHYDROGENASE-RXN	HS07548	GLUD1	ENSG00000148672
HMDB00148	GLT	GLUTDEHYD-RXN	HS07548	GLUD1	ENSG00000148672
HMDB00148	GLT	RXN-14116	HS14757	ALDH4A1	ENSG00000159423
HMDB00148	GLT	TRANS-RXN-211	HS09679	SLC3A2	ENSG00000168003
HMDB00148	GLT	TRANS-RXN-211	HS07701	SLC7A11	ENSG00000151012
HMDB00148	GLT	3.4.13.7-RXN	HS00367	DPEP1	ENSG00000015413
HMDB00148	GLT	3.4.13.7-RXN	HS09532	DPEP2	ENSG00000167261
HMDB00148	GLT	2.6.1.7-RXN	HS03239	AADAT	ENSG00000109576
HMDB00148	GLT	2.6.1.7-RXN	HS06422	CCBL2	ENSG00000137944
HMDB00148	GLT	2.6.1.7-RXN	HS04858	GOT2	ENSG00000125166
HMDB00148	GLT	ORNITHINE-GLU-AMINOTRANSFERASE-RXN	HS00832	OAT	ENSG00000065154
HMDB00148	GLT	1.5.1.9-RXN	HS00244	AASS	ENSG00000008311
HMDB00148	GLT	RXN-12825	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-12825	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	RXN-19602	HS01957	GGT1	ENSG00000100031
HMDB00148	GLT	RXN-19602	HS01949	GGT5	ENSG00000099998
HMDB00148	GLT	FOLYLPOLYGLUTAMATESYNTH-RXN	HS06237	FPGS	ENSG00000136877
HMDB00517	ARG	GLYCINE-AMIDINOTRANSFERASE-RXN	HS10376	GATM	ENSG00000171766

HMDB	MetaCyc Compound ID	MetaCyc Reaction ID	MetaCyc Gene ID	Gene Symbol	Ensembl ID
HMDB00517	ARG	ARGINASE-RXN	HS04231	ARG1	ENSG00000118520
HMDB00517	ARG	ARGINASE-RXN	HS01388	ARG2	ENSG00000081181
HMDB00517	ARG	ARGDECARBOX-RXN	HS06971	ADC	ENSG00000142920
HMDB00517	ARG	RXN66-448	HS09679	SLC3A2	ENSG00000168003
HMDB00517	ARG	RXN66-448	HS02481	SLC7A5	ENSG00000103257
HMDB00517	ARG	RXN66-448	HS02450	SLC7A6	ENSG00000103064
HMDB00517	ARG	ARGSUCCINLYA-RXN	HS10034	ASL	ENSG00000169910
HMDB00517	ARG	TRANS-RXN66-1231	HS09679	SLC3A2	ENSG00000168003
HMDB00517	ARG	TRANS-RXN66-1231	HS02450	SLC7A6	ENSG00000103064
HMDB00167	THR	RXN-15122	HS05952	SDS	ENSG00000135094
HMDB00167	THR	RXN-15122	HS06616	SDSL	ENSG00000139410
HMDB00167	THR	THREDEHYD-RXN	HS05952	SDS	ENSG00000135094
HMDB00167	THR	THREDEHYD-RXN	HS06616	SDSL	ENSG00000139410

Appendix B - Literature Review of Nodes of Interest

* Potential involvement in sepsis.
** Confirmed involvement in sepsis.

B.1 Nodes Common to All Networks

SLC7A11* – Solute Carrier Family 7 Member 11

- Present in the endotoxin tolerance signature.[61]

B.2 Nodes Unique to Integrated Network

AVP** – (Arginine) Vasopressin

- Blood pressure is a key factor in sepsis survival. Therefore, it should not be a surprise that vasopressin plays a key role in sepsis and is a key treatment option for septic patients.
- However, its exact role, and use as a treatment option has been highly controversial.[68]

CRAT* – Carnitine O-Acetyltransferase

- Related to fatty-acid oxidation and ketogenesis in the liver, which may play a role in sepsis.[69–71]

DMWD – DM1 Locus, WD Repeat Containing

- No documented link to sepsis.

ERG28 – Ergosterol Biosynthesis 28 Homolog

- No documented link to sepsis.

GATM* – Glycine Amidinotransferase

- May be related to arginine, homoarginine availability, which is key to nitric oxide production during septic immune response.[72,73]
    - Has been shown to play a role in stroke outcomes.[74]

HBB** – Hemoglobin Subunit Beta

- Sepsis affects red blood cells dramatically—low hemoglobin concentration is an indicator of sepsis.[75]
- HBB has been used for early diagnosis of sepsis.[76]

ILK* – Integrin Linked Kinase

- Mitogen-activated protein kinases regulate vascular reactivity after hemorrhagic shock through myosin light chain phosphorylation pathway[77]
    - "*It has been reported that integrin-linked kinase (ILK) … can regulate the calcium sensitivity of smooth muscle…*"
    - "*The protein expression and activity of ILK of [superior mesenteric arteries] were significantly reduced after the 2-hour shock…*"

ITGB8** – Integrin Subunit Beta 8

- Present in the endotoxin tolerance signature.[61]
- Upregulated in monocytes after LPS treatment.[78,79]

LAMP2* – Lysosomal Associated Membrane Protein 2

- Associated with heart disease and key in autophagy.
- "*Lysosomal membrane-associated protein-2 (LAMP2), which is essential to the maintenance of lysosomal functions in the heart, is depleted transiently but restored rapidly during LPS administration in the rat.*"[80]
- Present in elevated amounts in the blood of septic patients.[81–83]

MMP2* – Matrix Metallopeptidase 2

- Matrix metalloproteinases and their inhibitors have been implicated in sepsis.[84,85]

MYD88** – Myeloid Differentiation Primary Response 88

- Integral in the activation of NF-κB, MYD88 is used my almost all TLRs.

NT5C3A* – 5'-Nucleotidase, Cytosolic IIIA

- While not directly implicated in sepsis, NT5C3A was "…*induced by type I interferons (IFNs) in multiple cell types and that NT5C3A suppressed cytokine production through inhibition of the nuclear factor κB (NF-κB)*"[86]

PAPLN* – Papilin

- Present in the endotoxin tolerance signature.[61]
- An ADAMTS-like protein. Shares homology with metalloproteinases (see MMP2).

PKM* – Pyruvate Kinase M1/2

- "*…responsible for the final and rate-limiting reaction step of the glycolytic pathway*."[87]
- Required for expression of PD-L1[88]
- The role of PKM in sepsis has been both directly and indirectly implicated in multiple studies.[87–91]

RGS3* – Regulator of G Protein Signaling 3

- "*RGS1 and RGS16 are induced in response to bacterial lipopolysaccharide and stimulate c-fos promoter expression.*" However, RGS3 was not examined in this study.[92]
- RGS3 has been characterized in human septic sera.[93]

SRSF11 – Serine and Arginine Rich Splicing Factor 11

- No documented link to sepsis.

SVIL* – Supervillin

- May play a role in *Chlamydia* and *E. coli* pathogenesis.[94,95]
- Was found to be induced by exposure to LPS.[96]

YES1 – YES Proto-Oncogene 1, Src Family Tyrosine Kinase

- No documented link to sepsis.

Appendix C - Integrated Networks Generated with KEGG Mapping

MetaBridge allows for mapping metabolites to protein interactors via KEGG and MetaCyc. Although the use of MetaCyc was demonstrated in this thesis, the use of KEGG has been included here for thoroughness.

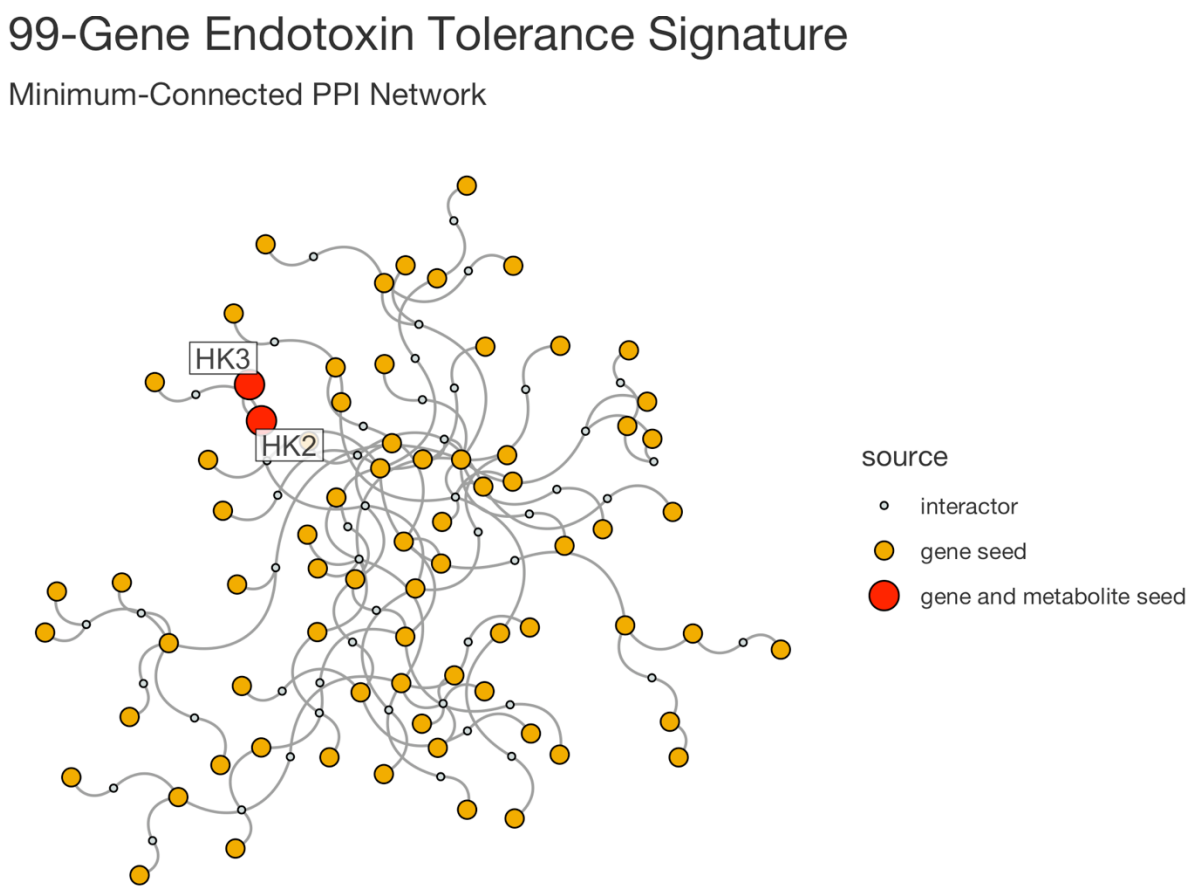

Figure 16. 99-gene Endotoxin Tolerance Signature PPI Network (KEGG). Shown is a minimum-connected PPI network of the 99-gene sepsis/endotoxin tolerance signature. Nodes in the gene signature are highlighted in yellow. Nodes common to both the metabolite signature and gene signature are highlighted in red. Metabolite-protein mapping was conducted with KEGG.

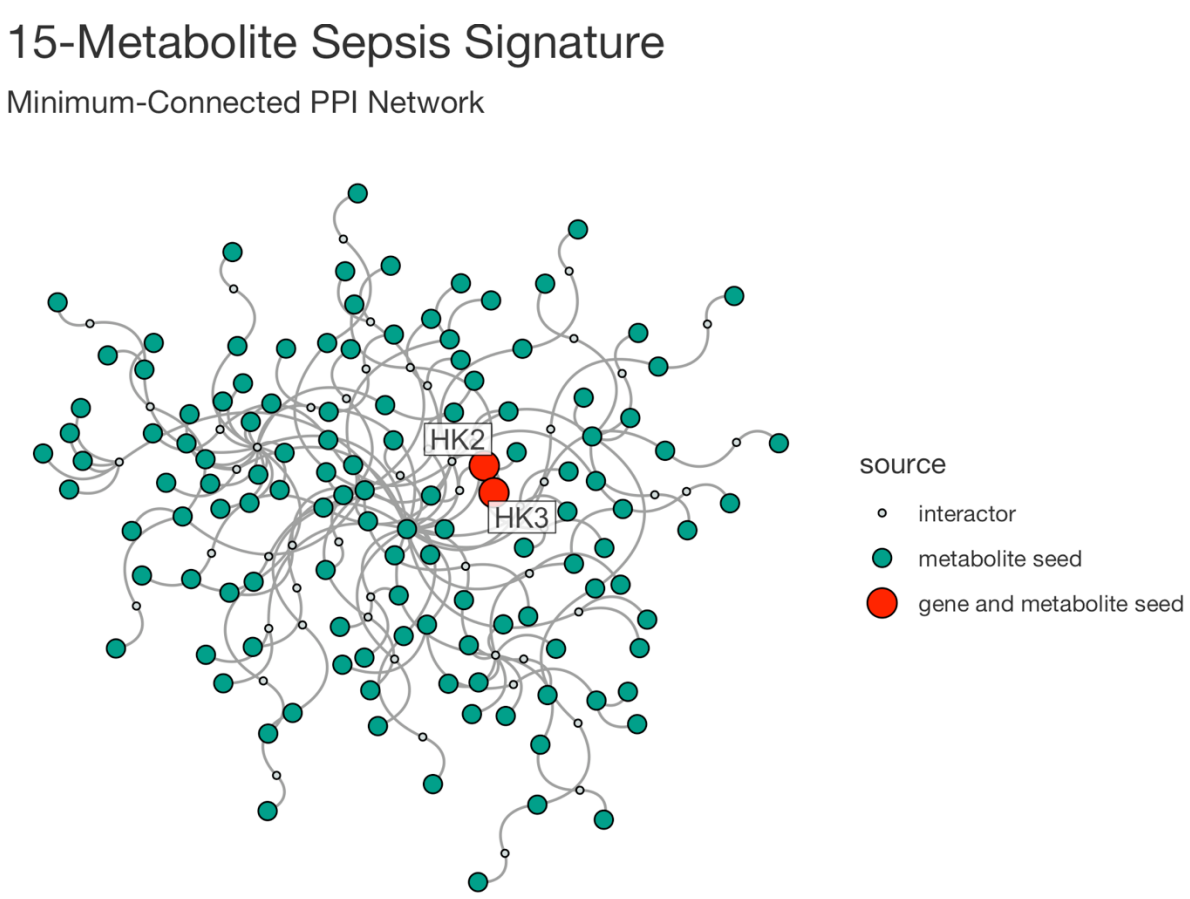

Figure 17. 15-metabolite Sepsis Signature PPI Network (KEGG). Shown is a minimum - connected PPI network of the 15-metabolite sepsis outcome differentiation signature. Nodes in the metabolite signature are highlighted in green. Nodes common to both the metabolite signature and gene signature are highlighted in red. Metabolite-protein mapping was conducted with KEGG.

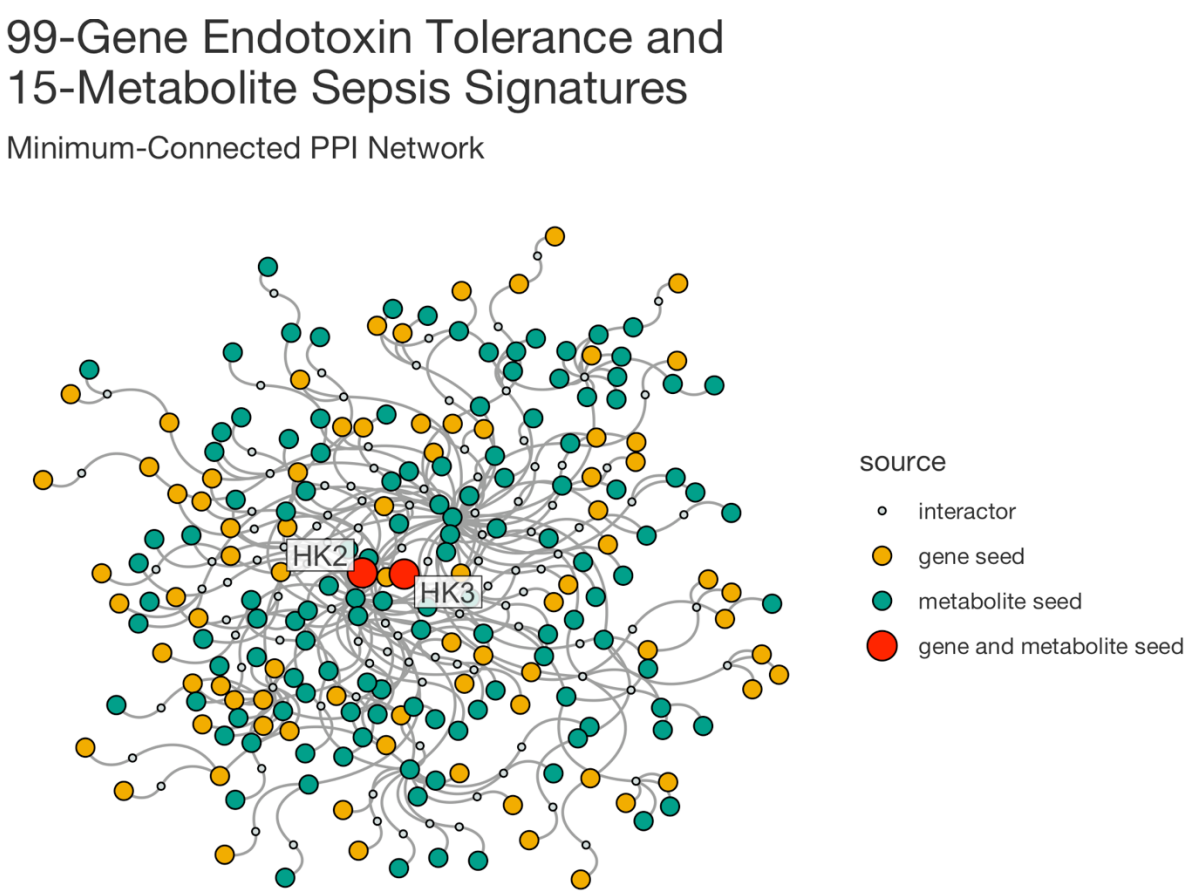

Figure 18. Integrated Signature PPI Network (KEGG). Shown is a minimum-connected PPI network integrating the 99-gene endotoxin tolerance signature and the 15-metabolite sepsis outcome differentiation signature. Nodes in the gene signature are highlighted in yellow. Nodes in the metabolite signature are highlighted in green. Nodes common to the metabolite signature and gene signature are highlighted in red. Metabolite-protein mapping was conducted with KEGG.

Appendix D - Comparison to Null

To demonstrate that the integration of multi-omic data generated meaningful biological insights, the integration of the two sepsis signatures was compared against randomly generated combined lists of 99 genes and proteins interacting with 15 metabolites. Ninety-nine random genes were selected from the InnateDB database (v5.5) and fifteen random metabolites were selected from the MetaCyc database, then mapped via MetaCyc to their directly interacting proteins.[97] This random selection was repeated 1000 times, and PPI networks were generated from the resulting sets. Seed-initialized pseudorandom number generation was used to ensure reproducible randomness. These random networks were integrated, and various parameters of the resultant networks were compared. Certain biases that this selection process and caveats to this preliminary analysis are discussed in Appendix D.5.

D.1 Hypothesis

I hypothesized that the integration of the sepsis signatures would result in a more highly connected network, with the two seed networks overlapping more substantially than would random networks. The two signatures would have reported on the same biological phenomena, while randomly selected genes and metabolites would not. Furthermore, I hypothesized that these integrated sepsis networks would represent distinct biological phenomena, rather than random noise.

To test my first hypothesis, I examined the connectivity and overlap of the integrated networks. To test my second hypothesis, I calculated the random discovery rate of each node of the sepsis-integrated networks to determine which nodes would appear at random, and which represented significant biological phenomena.

D.2 Connectivity

To begin, I compared the connectivity of the integrated sepsis network with all of the randomly integrated networks. I found significantly less connectivity in the sepsis integrated networks than the random networks, as measured by both degree and betweenness scores, two widely-used network centrality scores (Table 2). It was not expected that randomly integrated genes would show high connectivity.

As mentioned in Chapter 2.3, there are two primary groups of interest in a bi-omic integration. The first is the nodes that are exclusive to the integrated network and are not present in either of the two seed networks. These nodes represent potential novel biology revealed by the integration. The second group is the nodes that are present in all networks, including the seed and integrated networks. These nodes represent potential consensus biology.

Table 2. Randomly integrated networks were significantly more connected than the sepsis integrated network. Welch Two Sample t-test Comparing Network Connectivity of Null and Sepsis Integrated Networks. Comparison in network connectivity between 1000 randomly integrated PPI networks and the integrated sepsis PPI network.

Centrality Score	Diff	μ_1 null	μ_2 sepsis	t	SE	df	$CI_{95\%}$	p	Cohen's d	Power
Degree	18.3	42.3	24.1	7.2	2.5	180.9	(13.3 - 23.3)	<0.001	0.17	0.63
Betweenness	30121	71751	41630	7.0	4327	180.9	(21584 - 38658)	<0.001	0.16	0.59

Table 3. Nodes common to all networks were more connected than the network as a whole in randomly generated networks. Welch two sample t-test comparing Network Connectivity of common nodes to whole network. Comparison between mean node degree of nodes common to all generated (seed and integrated) networks to mean node degree of all nodes in the integrated network.

Integration	Diff	μ_1 common	μ_2 whole	t	SE	df	$CI_{95\%}$	p	Cohen's d	Power
Sepsis	-3.7	20.3	24.1	-0.19	19.6	180	(-42.3 - 34.9)	0.85	-0.11	0.05
Null	32.4	74.7	42.3	7.2	4.5	3729	(23.6 - 41.3)	<0.001	0.29	1

Interestingly, in the case of the sepsis integrated network, this did not hold true. The mean node degree of nodes common to all networks was lower than the mean node degree of the entire integrated network. This may reflect the fact that sepsis concerns the immune response which reflects some of the most common elements represented within the gene expression networks.[61] According to some estimates, more than 5000 genes are involved in immunity.[21,98]

However, nodes unique to the integrated network in each of both the random networks and sepsis networks had a lower mean connectivity (Table 4), as expected. These findings are summarized in Figure 19 which reveals that the sepsis network had one of the lowest mean node degrees of all studied networks. One possible explanation for this would be that the spesis network was more compact.

Table 4. Nodes unique to the integrated network were less connected than the network as a whole. Welch two sample t-test comparing network connectivity of unique nodes to whole network. Comparison between mean node degree of nodes unique to the integrated network to mean node degree of all nodes in the integrated network

Integration	Diff	μ_1 common	μ_2 whole	t	SE	df	$CI_{95\%}$	p	Cohen's d	Power
Sepsis	-19.2	4.8	24.1	-7.4	2.6	195.9	(-24.4 - -14.1)	<0.001	-0.63	0.95
Null	-34.3	8.1	42.3	-143	0.24	256998	(-34.7 - -33.8)	<0.001	-0.35	1

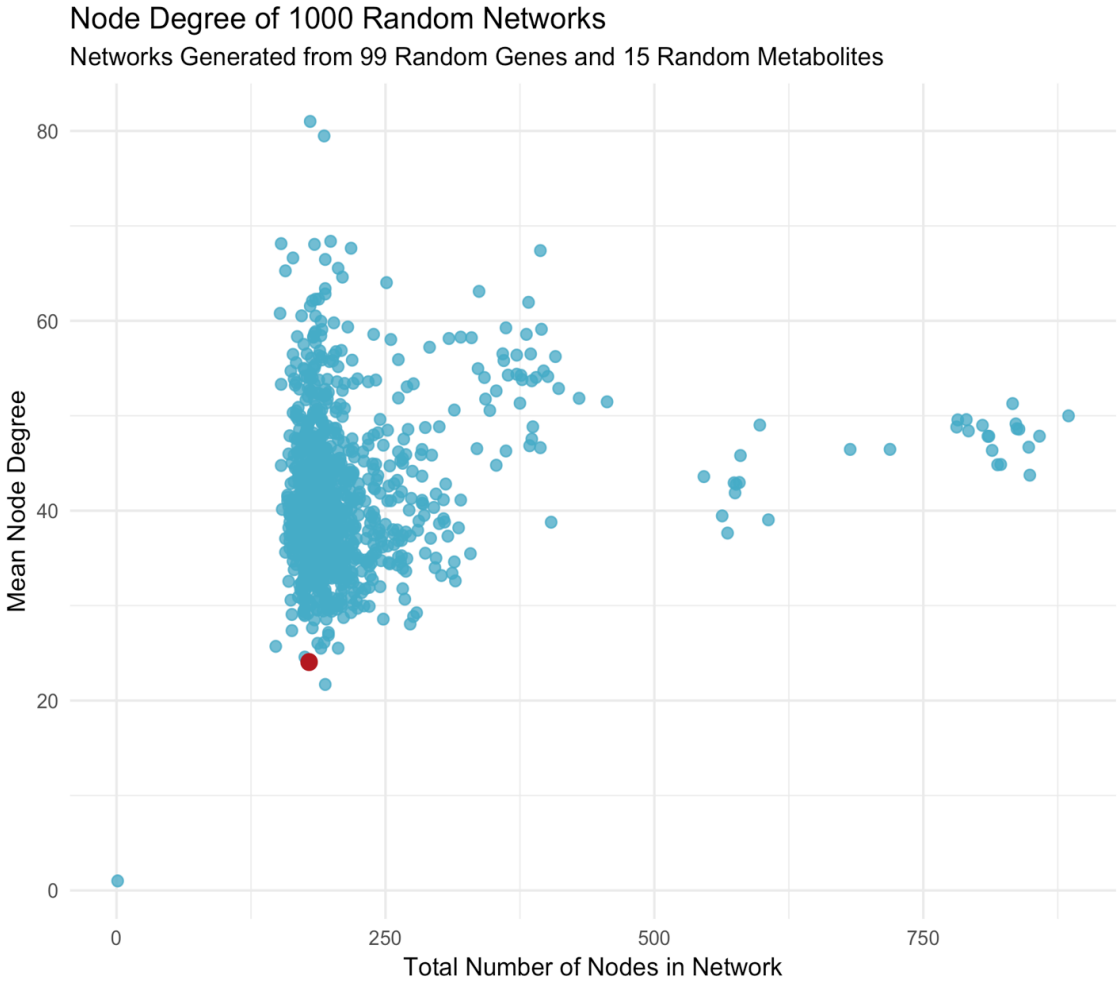

Figure 19. The sepsis integrated network was not more connected than randomly integrated networks. Each blue dot represents an integrated network generated from 99 random genes and 15 random metabolites. The red dot represents the integrated network generated from the 99-gene endotoxin tolerance signature and 15-metabolite sepsis signature.

These findings, subject to certain caveats, did not support the hypothesis that integrated PPI networks from similar biological sources will have greater connectivity than

randomly-integrated PPI networks with the same number of seeds.

D.3 Network Overlap

As shown in Figure 20, the average randomly generated networks shared very few nodes between the two seed networks, with a Jaccard index of 1.68%. The sepsis seed networks shared just three nodes, with a Jaccard index of 1.81%. However, the proportion of novel nodes in the integrated network (compared to both seed networks) was higher. On average, the integrated networks from randomly generated seed networks contained 18.4% novel nodes. The integrated network generated from the two sepsis signatures contained 22.3% novel nodes.

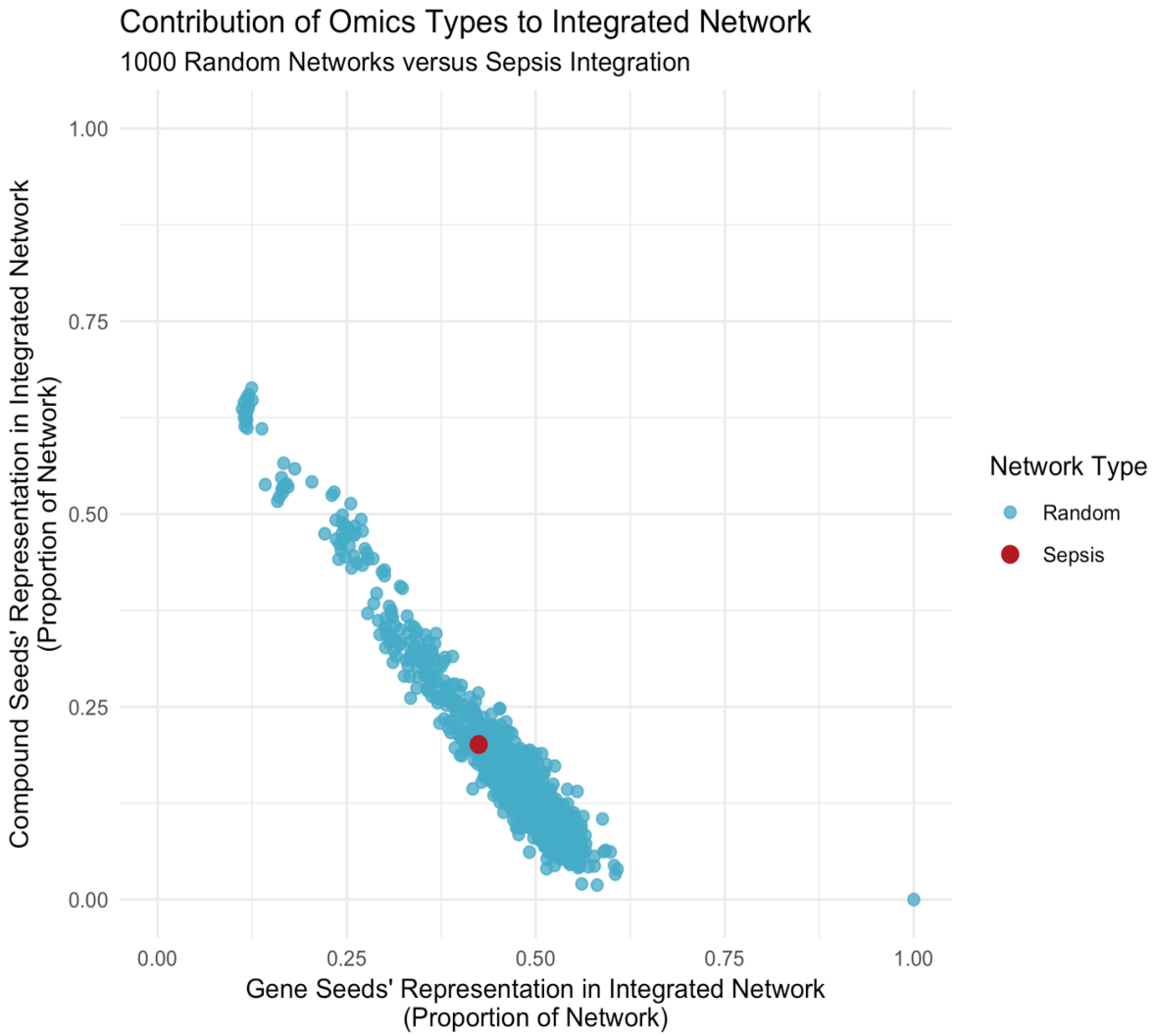

Figure 20. Integrated networks tended to be composed of more gene seeds than metabolite seeds. The representation of each omics type in the integrated, shown as the proportion of the integrated network that the seed nodes represent. Networks were skewed towards representation by gene seeds.

As shown in Figure 21, the integrated networks were skewed towards representation by the gene seeds. Despite the advantages of minimum-connected networks discussed in Chapter 1, the signal from transcriptomics still tended to be overrepresented, with an average of 45.2% of nodes in the network coming directly from the gene seeds. However, this can change dramatically based on the proportion of network seeds, and networks generated directly from experimental data were investigated further in Chapter 3. While the average contribution from metabolite seeds was only 18.4%, some integrations had more than 60%

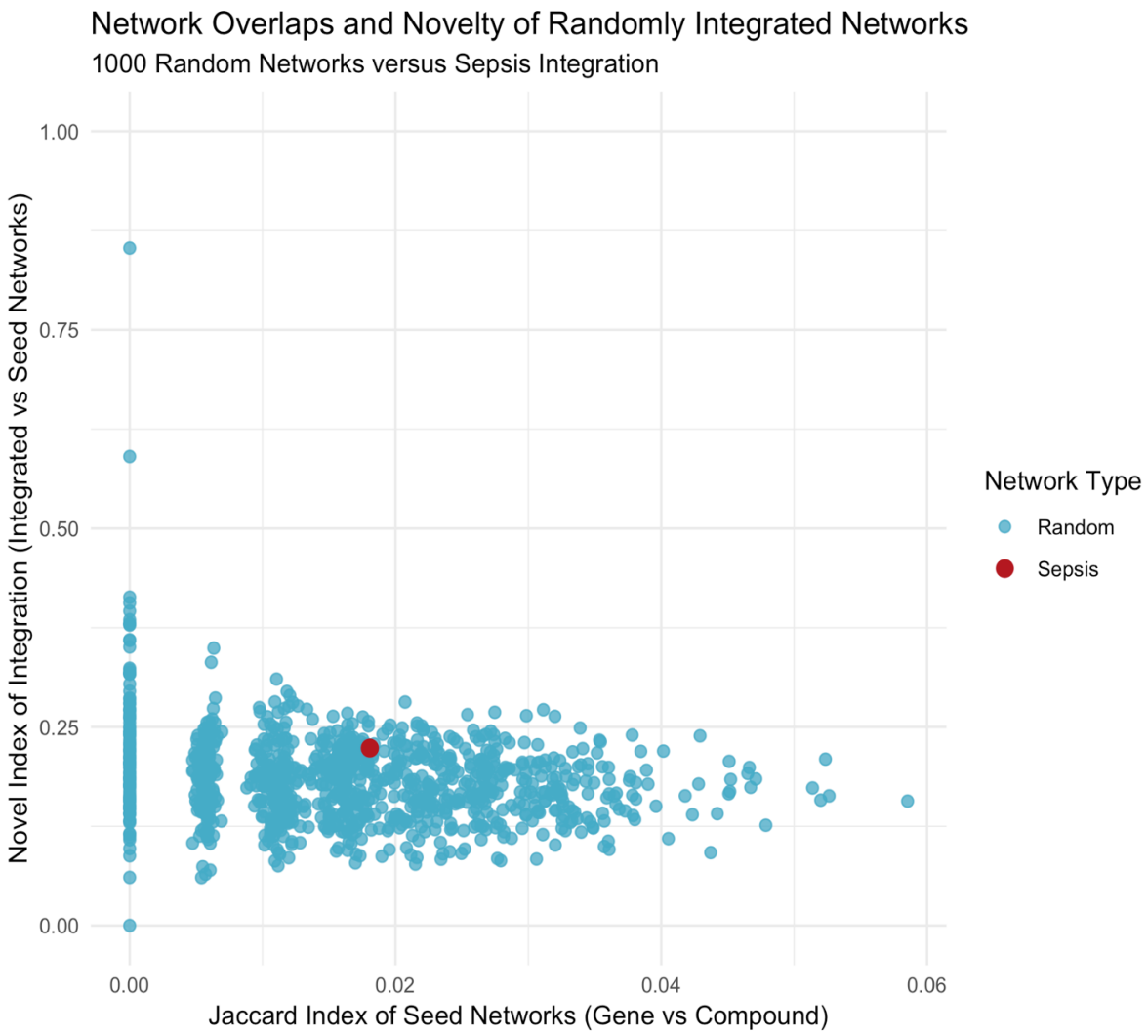

Figure 21. The sepsis seed networks did not overlap more than the randomly generated networks. Shown are 1000 randomly integrated networks compared to the integration of the 99-gene endotoxin tolerance signature with the 15-metabolite sepsis signature. The X-axis represents the overlap of the individual omics minimum-connected PPI networks. The Y-axis represents the proportion of novel nodes in the integrated network (versus both seed networks).

contribution from metabolite seeds consistent with the very large range of different interacting proteins from different metabolites. This large spread suggested that different experimental inputs would result in varying contributions. The sepsis signature had similar contributions to its integrated network as did the random networks with 42.5% contribution from the gene-seeds and just 27.6% from the metabolite seeds.

As demonstrated by Figure 20, these findings were not consistent with the hypothesis that multi-omic integration networks generated from similar biological sources will overlap more than networks generated from randomly-chosen data, given similar numbers of seeds. Moving forward, it will be essential to perform this same analysis with both random gene expression and random metabolite networks, and likely to limit these to proteins with known interactors (i.e. 87 as per the gene expression networks).

D.4 Random Discovery Rate

To determine whether the integrated sepsis networks represented unique biology versus randomly integrated networks, I calculated the rate of occurrence of the nodes in the sepsis networks in the 1000 randomly integrated networks. I defined this as the "Random Discovery Rate".

I filtered the two groups of interest previously mentioned, namely nodes common to all integrated networks and nodes unique to the integrated network for random discovery rate less than 5%. As shown in Table 5, 18/40 (45%) of the nodes that were unique to the integrated network had a random discovery rate of less than 5%. Table 6 shows that of the nodes common to all networks, only 1/3 had a random discovery rate of less than 5%. These findings support the hypothesis that nodes common to all networks in an integration are more likely to appear at random due to their high connectivity. The context of these values in reference to the null networks is displayed in Figure 22.

Table 5. Eighteen non-random nodes were unique to the sepsis integrated network. Shown are the 18 nodes present in the integrated network, but not present in the seed networks with a random discovery rate less than 5%. 40 nodes were unique to the integrated network, irrespective of random discovery rate. Node degree and betweenness are shown for the integrated network.

Gene Name	Node Degree	Betweenness	Seed Node?	Random Discovery Rate
CRAT	7	8170.1	TRUE	4.90%
ILK	4	7415.6	FALSE	4.80%
PKM	2	19080	FALSE	4.70%
GATM	2	2128	TRUE	4.10%
YES1	3	9296.7	FALSE	3.60%
LAMP2	2	4580.1	FALSE	3.40%
HBB	2	45928	FALSE	2.50%
MMP2	5	6391.9	FALSE	2.10%
AVP	2	19080	FALSE	1.70%
ITGB8	17	23489	TRUE	1.50%
RGS3	2	3993.8	FALSE	1.50%
DMWD	4	1628.1	FALSE	1.40%
SRSF11	2	8263.9	FALSE	1.40%
SVIL	2	28383	FALSE	1.40%
ERG28	2	5240.1	FALSE	1.00%
MYD88	2	3420.9	FALSE	0.50%
PAPLN	9	16996	TRUE	0.30%
NT5C3A	3	29510	FALSE	0.30%

Table 6. Only one non-random node was common to all networks. Nodes common to both seed networks and the integrated network in the sepsis signature integration. All but one node have ≥ 5% random discovery rate. Node degree and betweenness are shown for the integrated network.

Gene Name	Node Degree	Betweenness	Seed Node?	Random Discovery Rate
APP	12	122144	FALSE	98.50%
EGFR	34	230072	FALSE	82.60%
SLC7A11	15	11637	TRUE	2.80%

While these results do not show that the seed lists representing similar biology form networks significantly different in key statistics from randomly generated networks, they clearly indicate the presence of network nodes that represent unique biology. If we examine the proportion of network nodes that are non-random (Figure 23), we can make two important observations. First, we can see that the proportion of non-random nodes in the gene-seeded network is very tightly clustered around 75%. This is likely due to the network

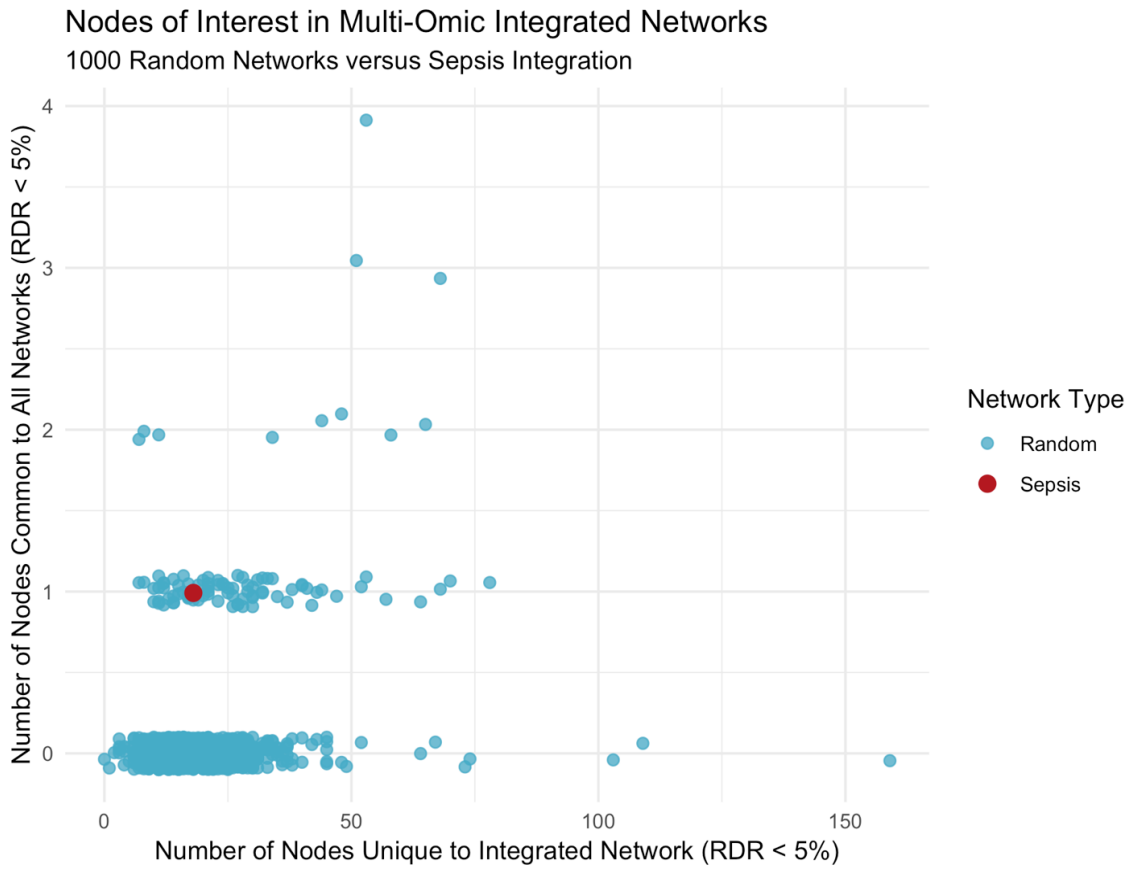

Figure 22. Nodes unique to the integrated network were less likely to occur at random than nodes common to all networks. The number of nodes present in each intersection of interest—nodes unique to the integrated network and nodes common to all generated (seed and integrated) networks—after filtering with a random discovery rate of 5%. Note that points are jittered along the y-axis to show the density of points at each integer value.

properties of a minimum-connected PPI network seeded with 99 elements, in that is there are enough elements to form a cohesive network without incorporating many first-order interactors. Second, we see that the metabolite-seeded networks are both more highly variable and involve a greater proportion of random nodes than the gene-seeded network. The former is likely due to the variable number of direct protein interactors identified for any given metabolite. The latter may indicate a need to use a greater number of first-order interactors in a metabolite-seeded network to obtain network cohesion. Finally, we see mixed results in the integrated networks, likely due to variability introduced by the metabolite seeds. Interestingly, we see that the sepsis signature-seeded networks had a higher than average number of non-random nodes. This might indicate a more cohesive, interconnected PPI

resulting from biologically cohesive metabolite signature as also indicated above. However, without statistical testing, it is not possible to draw a conclusion.

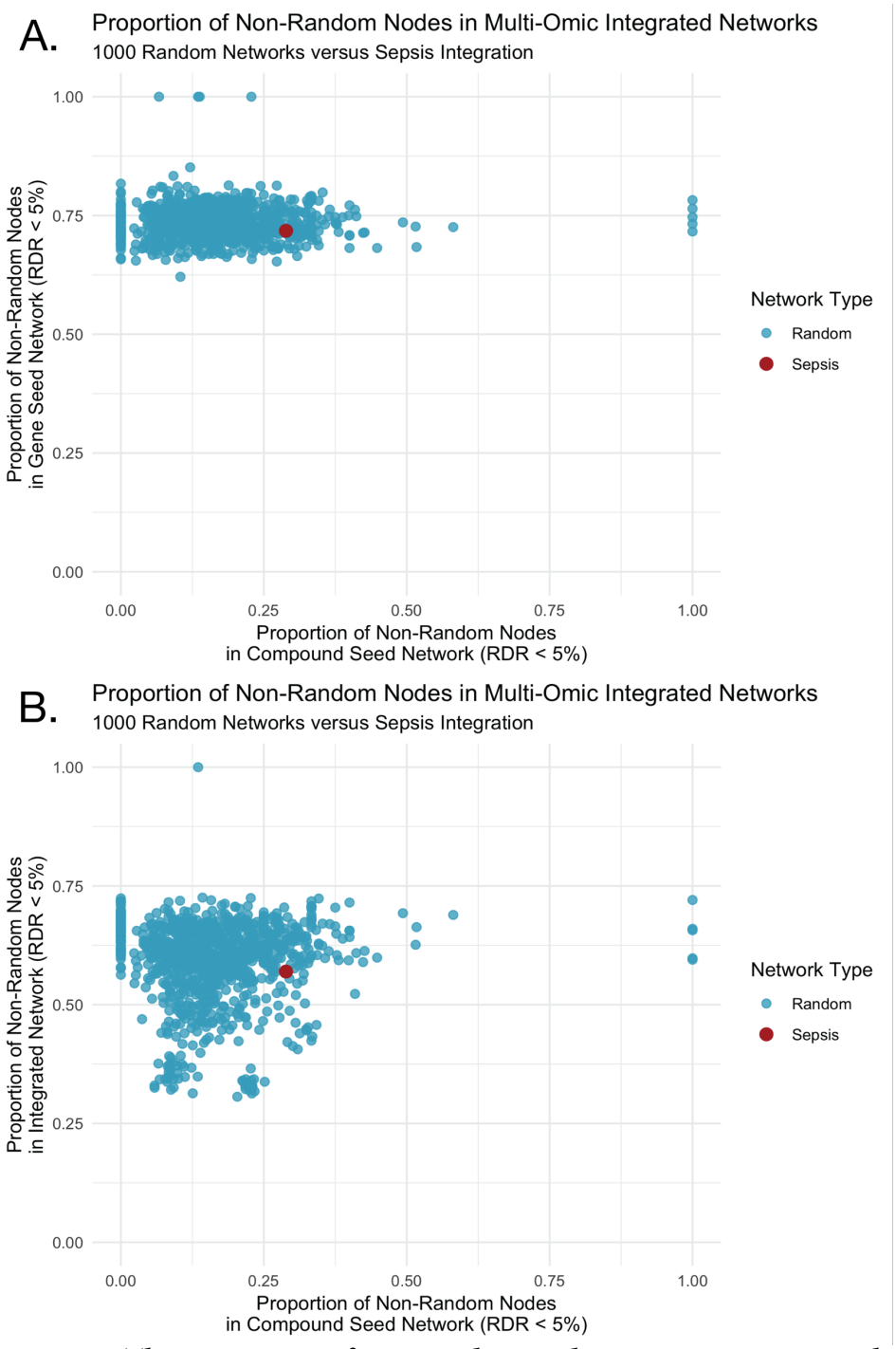

Figure 23. The proportion of non-random nodes comprising a network varied least for the gene seeded networks. A) The proportion of nodes in the compound seed network that are non-random (RDR < 5%) versus the proportion of nodes in the gene seed network that are non-random (RDR <5%). B) The proportion of nodes in the compound seed network that are non-random (RDR < 5%) versus the proportion of nodes in the integrated seed network that are non-random (RDR <5%).

These results support the hypothesis that integration of multi-omic data will reveal novel insights into a biological condition. While there are a substantial number of nodes novel to the integrated network that do not show up at random, we did not observe substantial seed network overlap, indicating a lack of biological consensus between the omics types, as mentioned in Appendix D.3. While the integrated sepsis networks do not differ significantly from randomly generated networks by the network statistics examined, this simply outlines the properties of integrated multi-omic networks of a specific size and nature (99 genes and 15 metabolites). More notably, for given networks of this size, we see consistent distinct biology represented in the resulting integrated network (Figure 22).

D.5 Future Considerations

D.5.1 Gene Selection for Random Networks

Ninety-nine random genes were selected from the InnateDB database (v5.5).[21] By selecting genes from the interactome database being used, I guaranteed that genes selected would have interactors present in the database. This could bias the resultant networks to be more highly-connected, as discussed above. This is of particular note, as only 87 of the 99 genes in the endotoxin tolerance signature were present in InnateDB as having interactors, whereas every random network generated would have 99/99 genes present in InnateDB. However, results were similar when genes were selected randomly from the HumanCyc database (data not shown).[45] Still, choosing genes from the HumanCyc database is biased as well; HumanCyc catalogues only functional genes, as its focus is on metabolic pathways.

Therefore, it would be of particular interest to reproduce this study with genes selected from a variety of sources. For instance, randomly selecting genes from an experimental source could more accurately represent a gene expression profile of a random cell. Using gene expression sequencing data from peripheral blood mononuclear cells would be a good candidate, as this subset of cells was used to identify the endotoxin tolerance signature. The genotype-tissue expression project could prove a useful source for future studies where gene expression profiles of specific tissue types could be used as a random gene source.[99]

Finally, it could be of interest to control for the level of connectivity in future comparisons. If genes of a similar level of connectivity profile within the InnateDB database

were selected at random, how would the resultant networks compare? Would similarly-connected random genes generate networks of similar properties as genes identified from a biological source?

D.5.2 Metabolite Selection for Random Networks

Fifteen random metabolites were selected from the MetaCyc database, then mapped via MetaCyc to their directly interacting proteins. This process of selection is perhaps less biased than the process of gene selection, as many metabolites in MetaCyc do not interact with human proteins. However, the process is still limited to the number of metabolites for which interactors are well known, and the large variability in the numbers of interactors for any given metabolite. While there are over 90,000 metabolites endogenous to humans now listed within the Human Metabolite Database, MetaCyc catalogues under 15,000, not all of which map to human protein interactors.[97,100]

Nevertheless, it would be of interest to search out other sources for selection of random metabolites. The KEGG database could serve as another source of random metabolites. Finally, repositories of experimental metabolomics data such as Metabolomics Workbench and MetaboLights could be used as a source of human metabolome profiles for random metabolite selection.[101,102] However, it should be noted that identifying an appropriate control might depend on the metabolomics assay technology used in addition to the tissue type assayed.

D.5.3 Integration of Further Gene Signatures

Integration of further gene signatures may also prove insightful. Integrating additional gene signatures and comparing such integration to randomly integrated networks could shed light on the differences between randomly integrated networks and biologically-connected integrated networks. In particular, it could provide additional insights into whether randomly-connected networks tend to be more highly connected than networks representing specific biological conditions.