Integration of genomic and metabolomic data for the prioritization of rare disease variants

by

Emma Graham

BSc Molecular Biochemistry and Biophysics, Yale University, 2016

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)

November 2018

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

Integration of genomic and metabolomic data for the prioritization of rare disease variants

submitted by **Emma Graham** in partial fulfillment of the requirements for the degree of **Master of Science** in **Bioinformatics**.

Examining Committee:

Sara Mostafavi, Statistics and Medical Genetics
Supervisor

Wyeth Wasserman, Medical Genetics
Supervisory Committee Member

Clara van Karnebeek, Pediatrics
Supervisory Committee Member

Martin Hirst, Microbiology and Immunology
Committee Chair

Abstract

Many inborn errors of metabolism (IEMs) are amenable to treatment, therefore early diagnosis and treatment is imperative. Despite recent advances, the genetic basis of many metabolic phenotypes remains unknown. For discovery purposes, Whole Exome Sequencing (WES) variant prioritization coupled with clinical and bioinformatics expertise is currently the primary method used to identify novel disease-causing variants; however, causation is often difficult to establish due to the number of plausible variants. Integrated analysis of untargeted metabolomics (UM) and WES or Whole Genome Sequencing (WGS) data is a promising systematic approach for prioritizing causal variants from a list of candidates. In this thesis, we present an automated network-based bioinformatics approach to the integration of WES with UM data from 13 neurometabolic patients with known IEMs and 25 controls. We perform label propagation on the STRING network initialized using an integrated genomic and metabolomic score, and use the results to rank candidate genes in order of their likely relevance to the disease. Integrated genomic and metabolomic evidence was able to prioritize the causative gene in the top 20th percentile of candidate genes for 61.5% (8 of 13) of patients, 75% of which achieved a percentile prioritization score at least one standard deviation above a permuted percentile. Combining genomic and metabolomic evidence resulted in the prioritization of the causative gene in 30.7% more patients than was possible with genomic evidence alone. The results of this study indicate that for diagnostic and gene discovery purposes, metabolomics can lend support to WES gene discovery methods. This is the first method that uses UM and WES data to rank candidate variants in order of their biological relevance. To improve this method, expansion of gene-metabolite annotations and metabolomic feature-to-metabolite mapping methods

are needed.

Lay Summary

Recent technological advances have made it possible to profile an individual's genetic code and metabolic processes with increasing ease and precision. These improvements have allowed scientists to probe an individual's biology, enabling the identification of the cause of previously undiagnosable rare diseases. Individually, personalized genetic and metabolic profiles have been used to successfully identify the cause of inborn errors of metabolism (IEM)–a group of rare pediatric diseases that affect metabolism. If left untreated, IEMs can result in severe damage to organs, including the brain, therefore early treatment is imperative. Each IEM patient's disease is caused by a different genetic "mistake"; identifying the location of this "mistake", or mutation, is challenging. Metabolic data represents a snapshot of our body's biological reactions, and can therefore help distinguish between the "quirks" in our genome–DNA changes that make each person unique–and the mutations that can have damaging consequences to our metabolism. In this thesis, a method that combines genetic and metabolic evidence to help physicians identify the location of the mutation causing a patient's disease was developed. Its merits and shortcomings highlight how genomic and metabolic profiling technologies can be utilized and improved for use in the clinic.

Preface

The introduction of this thesis–with the exception of the section on label propagation–as well as some parts of the discussion, are largely reproduced from a review I wrote on the genomics and metabolomics field (Graham et al. [2018]). The WES variant filtering pipeline introduced was created and maintained by the Wasserman lab at BC Children's Hospital Research Institute, specifically Maja Tarailo-Graovac, Allison Matthews, Jessica Lee and Phillip Richmond. Their contributions included all WES data generation, variant filtering and candidate variant list creation. The raw LC-MS metabolomic data was generated by the lab of Ron Wevers at Radboud University in Nijmegen, Netherlands, specifically Udo Engelke and Leo Kluiijtmans. I completed all other analysis performed in this thesis.

Table of Contents

Abstract . . . . . **iii**

Lay Summary . . . . . **v**

Preface . . . . . **vi**

Table of Contents . . . . . **vii**

List of Tables . . . . . **x**

List of Figures . . . . . **xii**

Glossary . . . . . **xiv**

Acknowledgments . . . . . **xv**

1 Introduction **1**

1.1 Introduction to this chapter 1

1.2 Inborn Errors of Metabolism 2

1.3 Whole exome sequencing for the identification of rare disease variants 3

1.3.1 Canonical approach to identifying the causative gene . . . 3

1.4 Metabolomic profiling 4

1.4.1 Generating Liquid Chromatography coupled Mass Spectrometry data 6

1.4.2 Processing LC-MS data 7

1.4.3 Normalizing LC-MS data . . . 8
1.4.4 Testing for significant features in IEM studies . . . 10
1.4.5 Annotating features: adducts, isotopes, and metabolites . . 11
1.4.6 Identifying IEMs through untargeted metabolomic analysis 13
1.5 Integrating genomic and metabolomic data . . . 14
1.5.1 Identification of metabolically active loci . . . 14
1.5.2 Gathering evidence for metabolic perturbation on a gene-by-gene basis . . . 15
1.6 Treatable Intellectual Disability Endeavour (TIDE) Exome Sequencing project . . . 16
1.7 Strategies for disease gene prioritization . . . 17
1.7.1 Gene Recommendation . . . 18
1.7.2 Label propagation . . . 20
1.7.3 Theoretical framework of label propagation . . . 20
1.7.4 Using label propagation algorithms for candidate gene prioritization . . . 21
1.8 Functional Linkage Networks and the STRING network . . . 22
1.8.1 Selection of the STRING network . . . 22
1.9 Aims of this thesis . . . 23

2 Methods . . . 25
2.1 Datasets . . . 25
2.2 WES variant filtering pipeline . . . 27
2.3 Processing and normalization of untargeted LC-MS metabolomics 27
2.3.1 LC-MS data generation . . . 28
2.3.2 LC-MS data normalization and filtering . . . 28
2.4 Integrative analysis of WES and LC-MS data . . . 29
2.4.1 Creation of combined WES and metabolomic score . . . 30
2.5 Label propagation . . . 31
2.6 Categorizing polarity of gene-associated metabolites . . . 31
2.7 Summary of method . . . 32

3 Results 33
3.1 Characterization of WES and LC-MS data 33
3.1.1 WES variants 33
3.1.2 Characterization of LC-MS metabolomics features 34
3.2 Assessing enrichment for gene-associated metabolites 34
3.3 Integration of genomic and metabolic data to prioritize causative genes 36
3.3.1 Using label propagation to rank candidate genes 39
3.3.2 Assessing the utility of metabolomic evidence 41
3.3.3 Permutation test to generate null model of percentile rank 41
3.3.4 Characterization of factors that effect gene prioritization . 42
3.4 Summary 44

4 Disscussion 48
4.1 Future Work 51

Bibliography 52

A Supporting Materials 74

List of Tables

Table 3.1 **Summary of WES data.** Percentage of each patient's variants that fall into one of four modes of inheritance and the average number of candidate genes to which these variants map. 34

Table 3.2 **Characterization of LC-MS features.** 35

Table 3.3 **Summary of the causative gene(s) identified for each patient through the WES variant filtering pipeline.** The enrichment of causative gene-associated metabolites, function of each causative gene and polarity of gene-associated metabolites are provided. 37

Table 3.4 **Metabolic enrichment profile.** Number of genes enriched for in the set of patient-specific differentially abundant metabolites. 39

Table 3.5 **Label propagation results.** Prioritization results after LP with both the combined genomic and metabolomic initial label scores and the genomic-only initial label scores. The percentile rank of each causative gene in the list of candidate genes, the number of candidate genes as well as the change in percentile rank of the causative gene after addition of metabolomic evidence to the initial label score (DeltaM) is provided, in addition to each gene's final genomic-only and combined prioritization category (PC). 40

Table 3.6 **Label propagation results with permuted initial labels.** Percentile rank of the causative gene in each patient's list of candidate genes, as well as the mean and standard deviation of the permuted (n=500) percentile for each causative gene. 43

List of Figures

Figure 1.1 **WES rare variant analysis pipeline.** This pipeline was used in Tarailo-Graovac et al. [2016]. Raw reads are aligned to the human genome. Variants are annotated using SnpEff as well as custom Perl and Python scripts. Variants that do not map to protein-coding regions, have MAF > 0.01 in several variant databases or that do not pass QC steps are removed. Variants that do not agree with multiple inheritance models and that would not agree with the observed phenotypic effect are also removed. 5

Figure 1.2 **Sample LC-MS metabolomics analysis pipeline.** Briefly, raw metabolomics data can be processed using freely available processing software (e.g. XCMS), annotated (e.g. CAMERA), normalized (e.g. through use of internal standards) and filtered. Differentially abundant metabolites can be isolated using univariate or multivariate tests. Biological interpretation such as pathway analysis can be performed using published metabolomic databases (e.g. HMDB, BioCyc, METLIN). . . 6

Figure 1.3 **Automated and manual steps in untargeted metabolomics pre-processing pipeline.** The algorithms listed are only examples of tools that could be used in each step. 9

Figure 2.1 **Summary of overall method.** A) Briefly, raw metabolomics data was processed using XCMS and CAMERA and subsequently normalized through linear baseline normalization. Differentially abundant metabolites were isolated based on z-score. Raw genomic sequencing reads for each patient were processed and SnpEff was used to identify a list of candidate genomic variants with MAF ≤ 0.01. B) Enrichment for each of the 5371 HMDB genes in the patient-specific set of DAMs was assessed using Fisher's Exact Test, generating a metabolomic score. To generate the WES (i.e genomic) score, each gene in the STRING network was assigned "1" if also in the patient-specific candidate gene list, and "0" if not. C) These scores were plotted on an x-y grid, and the inverse of the distance between (1,1) and the coordinate of each gene, i, was considered the combined genomic and metabolomic score. Label propagation was then performed with the combined initial label score. 26

Figure 3.1 **DeltaM of causative vs non-causative genes.** 42

Figure 3.2 **Effect of polarity of gene-associated metabolites on percentile ranking.** "None" indicates that the gene has no associated metabolites in HMDB. 45

Figure 3.3 **Utility of metabolomic evidence for causative genes associated with metabolites of varying polarities.** DeltaM of causative genes by polarity of gene-associated metabolites. "None" indicates that the gene did not have any associated metabolites in HMDB. 46

Glossary

DAM Differentially abundant metabolite

GBA Guilt by Association

IEM Inborn Error of Metabolism

LC-MS Liquid-Chromatography Mass Spectrometry

LP Label Propagation

RWR Random Walk with Restart

STRING Search Tool for the Retrieval of Interacting Genes/Proteins

TIDEX Treatable Intellectual Disability Endeavour - Exome

WES Whole Exome Sequencing

Acknowledgments

I would like to thank my supervisor, Sara Mostafavi, without whom this document would be a collection of gibberish. I would also like to thank my committee members, Wyeth Wasserman and Clara van Karnebeek, for the insights and conversations.

Thanks to my family and friends for allowing me to ramble on about metabolites and the power of data integration, for sharing in my joy when things worked out, and for buying me beers when it didn't.

Special thank you to all past and present members of the Mostafavi and Wasserman labs, especially: Bernard Ng, Farnush Farhadi, Mike Vermeulen, Will Casazza, Sasha Maslova, Sina Jafarzadeh, Peter West, Louie Dinh, Halldor Thorhallsson, Hamid Omid, Joanna Lubieniecki, Phil Richmond, Magda Price, Jessica Lee and Robin van der Lee. And of course, my bioinformatics buddies: Eric Chu, Annie Cavalla, Allison Tai, Arjun Baghela and Nivi Thatra.

Most of all, I would like to thank my husband, Evan, for always being there for one more cup of coffee.

Chapter 1

Introduction

To define a person' s medical essence, we have to look at more than just their sequence — Eric Toppel (2017)

1.1 Introduction to this chapter

In this chapter, we will review the acquisition, processing and diagnostic utility of genomic and metabolomic data and explore how these data types can be integrated to prioritize candidate genes and improve the diagnosis of rare monogenic metabolic diseases. Specifically, we will provide an overview of whole exome sequencing (WES) variant filtering and liquid chromatography coupled mass spectrometry (LC-MS) metabolomic processing and feature selection pipelines. We will then discuss previous approaches to the integration of genomic and metabolomic data for the generation of relevant biological insights and review candidate gene prioritization strategies, with a focus on those using network-based methods. Finally, we will introduce several relevant components of our integration method, namely functional linkage networks and the network-based method label propagation. All of this chapter, with the exception of Section 1.6-1.8, is included in Graham et al. [2018], with a few minor modifications.

1.2 Inborn Errors of Metabolism

Inborn errors of metabolism (IEMs) are the largest group of genetic diseases amenable to causal therapy, and are caused by genetic variants that disrupt the function of enzymes or other proteins involved in cellular metabolism, leading to energy deficit and/or accumulation of toxins (Van Bokhoven [2011]). Early detection, enabled by newborn metabolic screening programs and genetics profiling, is pivotal so that treatment can be initiated before the onset of irreversible progressive damage to the central nervous system, which in some cases can result in intellectual disability disorder (IDD) and damage to additional organ systems.

There are currently more than 100 treatable IEMs, but for many phenotypes the genetic basis remains to be discovered (Van Karnebeek and Stockler [2012]). Cases for which the causal gene was identified have in turn provided insights and opportunities for interventions targeting their downstream molecular or cellular abnormalities (Collins et al. [2010], Horvath et al. [2016], Karnebeek et al. [2016]). These efforts have been catalogued in the online resource IEMbase, which provides further information on the etiologies and treatment of over 500 IEM disorders (B. et al. [2014]).

WES is the primary tool for discovery of the genetic basis of IEMs, and thus establishment of a genetic-based diagnosis that, in some cases, can lead to improved outcomes through targeted interventions. The promise of this approach was illustrated by a recent neurometabolic gene discovery study (Tarailo-Graovac et al. [2016]), in which deep phenotyping and WES achieved a diagnostic yield of 68% in patients with unexplained phenotypes, identified novel human disease genes and most importantly enabled targeted intervention for improved outcomes in 44% of patients. Overall, published studies applying WES coupled with variant prioritization in patients with unexplained phenotypes are successful in identifying the underlying cause in 16-68% of patients (Tarailo-Graovac et al. [2016]).

However, with our current limited understanding of variant pathogenicity and the biological impact of rare variants, variant prioritization algorithms that aim to completely automate the process of prioritization fail to identify the causal variant in a substantial number of patients. Further, variants that are identified as plausible often have a low level of supporting evidence, and are thus not adequate to

establish a genetic-based diagnosis. Using multiple types of personalized "-omic" data is a promising approach to address the evidence gap in support of an IEM diagnosis. The integration of metabolomics data with WES/WGS data to identify genes causing IEM is a prime example of this approach. For example, a diagnosis of maple syrup urine disease can be supported by 1) pathogenic variants in either DBT, BCKDHB or BCKDHA, 2) high levels of amino acids such as alloisoleucine, isoleucine, leucine and valine, and 3) branched-chain oxoacids (Strauss et al. [2013]). These biochemical biomarkers can be detected individually (targeted metabolomics), or as part of a broader characterization of the metabolome (untargeted metabolomics). Recently, the unbiased approach afforded by untargeted metabolomics has increased in popularity due to decreasing costs, lack of required parameter tuning, and opportunities for pathway analysis (Johnson et al. [2016]).

1.3 Whole exome sequencing for the identification of rare disease variants

1.3.1 Canonical approach to identifying the causative gene

Genomic sequencing-driven variant prioritization involves multiple filtering steps that incorporate prior knowledge about allele population frequency and predicted pathogenicity. Databases such as ExAC, dbSNP and gnomAD provide information about allele frequencies seen in the general population, and are then used to filter out common and likely non-pathogenic variants in the patient (Exome Aggregate Consortium [2016], Smigielski et al. [2000], Lek et al. [2016]). Once identified as pathogenic through use of *in silico* pathogenicity prediction tools (such as Polyphen-2 and SIFT), genomic data from an individual's parents is then used to filter variants according to Mendelian models of inheritance, treating the parents of the individual as "controls"(Ng and Henikoff [2003], Adzhubei et al. [2013]). This allows for both the isolation of pathogenic variants, and the assignment of mode of inheritance. However, it should be noted that some studies have questioned whether genomic databases may in fact contain individuals with disease-associated genotypes but no clinical presentation of the underlying disease at the time of the

inclusion, as more than 2.8% of the ExAC population was found to carry likely/-pathogenic genotypes reported in ClinVar (Tarailo-Graovac et al. [2017]). Continued expansion of variant databases and variant filtering methods will play an important role in identifying pathogenic variants.

A sample WES variant filtering pipeline used in Tarailo-Graovac et al. [2016] is detailed in Figure 1.1. In the case of WES, in which around 20,000 to 50,000 variants are observed in protein coding regions per individual, standard filtering steps typically enable researchers to reduce the number of variants to 10 to 200 candidate variants depending on the WES study design (e.g., access to trio data and pedigree structure) (Yang et al. [2013], Belkadi et al. [2015]). For the challenging task of identifying the needle in the haystack, i.e., the single causative variant, clinical input and extensive discussion among physicians, genetic counselors and bioinformaticians is typically needed; for genes previously unreported to cause human disease, identification of other families with similar phenotypes and other variants in the same gene as well as in vitro functional studies are required as evidence for validation of etiology (Tarailo-Graovac et al. [2016]).

The reliance on clinical expertise throughout the variant filtering process results in long processing times, especially if trio data is not available. Further, causality is difficult to establish–especially for variants previously unreported in human disease, of poor sequencing quality or unknown significance (Bertier et al. [2016]). Integrating multiple types of "-omic" data is a promising approach to help address this challenge.

1.4 Metabolomic profiling

Since IEMs result from a malfunction of protein-coding genes, many of which control the concentration of a variety of metabolites, biochemical tests of known IEM-related metabolites have long been performed for IEM diagnosis. The simultaneous assay of many IEM biomarkers through the use of untargeted metabolomics is an active research area. In this section, we provide an overview of metabolomic profiling methods and review existing approaches for the processing and analysis of untargeted LC-MS metabolomics data for IEM diagnosis and discovery. This includes four critical components: 1) generation of LC-MS data, 2) identification

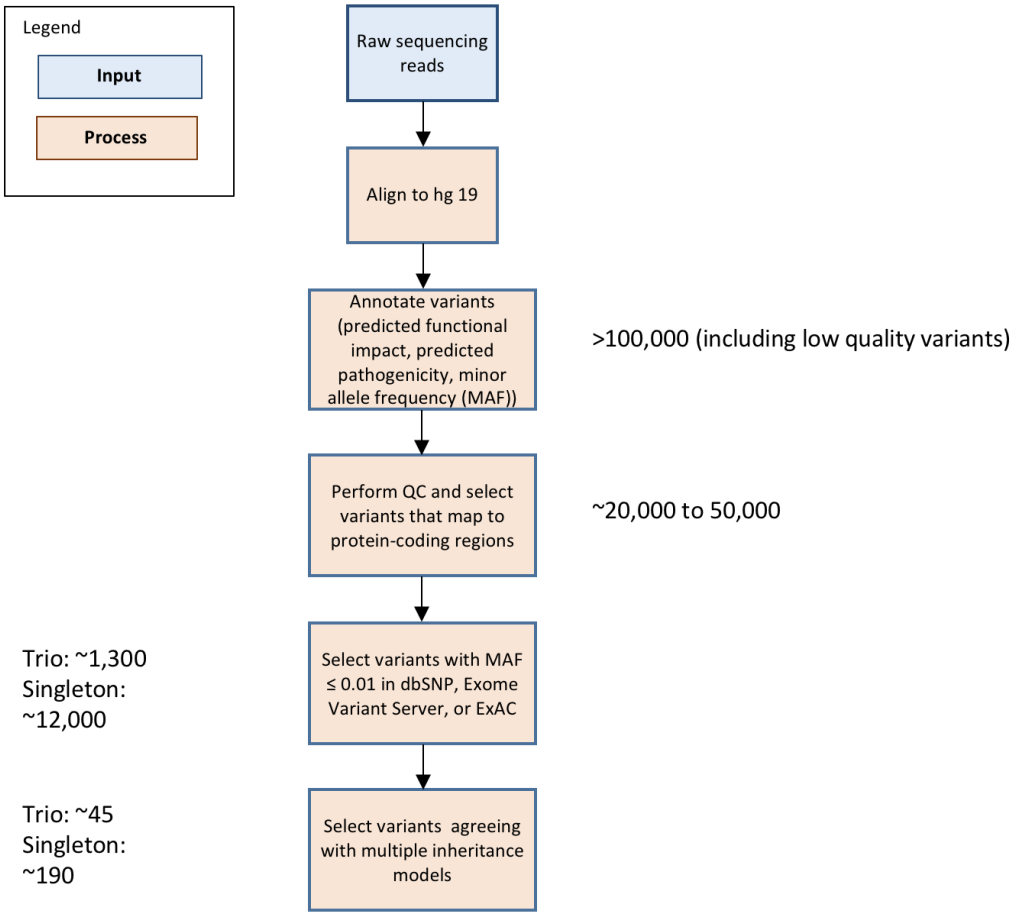

Figure 1.1: WES rare variant analysis pipeline. This pipeline was used in Tarailo-Graovac et al. [2016]. Raw reads are aligned to the human genome. Variants are annotated using SnpEff as well as custom Perl and Python scripts. Variants that do not map to protein-coding regions, have MAF > 0.01 in several variant databases or that do not pass QC steps are removed. Variants that do not agree with multiple inheritance models and that would not agree with the observed phenotypic effect are also removed.

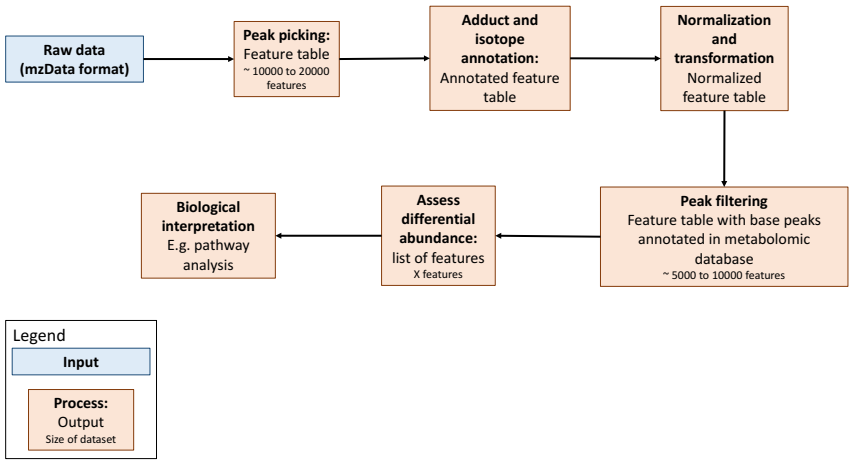

Figure 1.2: Sample LC-MS metabolomics analysis pipeline. Briefly, raw metabolomics data can be processed using freely available processing software (e.g. XCMS), annotated (e.g. CAMERA), normalized (e.g. through use of internal standards) and filtered. Differentially abundant metabolites can be isolated using univariate or multivariate tests. Biological interpretation such as pathway analysis can be performed using published metabolomic databases (e.g. HMDB, BioCyc, METLIN).

of units of analysis ("features") and normalization, 3) identification of significant features, and 4) association of significant features with known metabolites. An overview of a hypothetical untargeted LC-MS pipeline is provided in Figure 1.2.

1.4.1 Generating Liquid Chromatography coupled Mass Spectrometry data

In general, metabolomics quantifies a subset of small molecules (metabolites) in a tissue or body fluid using either nuclear magnetic resonance (NMR) spectroscopy or MS (Johnson et al. [2016]). NMR spectroscopy quantifies solution-state molecular structures based on atom-centered nuclear interactions. NMR spectroscopy is inexpensive, capable of high throughput analysis and highly reproducible; however, it lacks sensitivity and is generally only able to quantify metabolites of medium to high abundance. For this reason, MS based quantification has primarily

been used in the context of IEM diagnosis and discovery.

In mass-spectrometry (MS) based quantification, metabolites are first chromatographically separated and quantified in a semi-quantitative manner using high resolution mass spectrometers in detection modes that measure both positive and negatively charged ions produced through electrospray ionization (ESI). MS separation techniques include liquid chromatography, capillary electrophoresis, gas chromatography and ultra-performance liquid chromatography (Zhang et al. [2012]). No single chromatographic separation protocol can quantify all metabolites in a sample. Therefore, to completely capture all metabolites, multiple chromatographic methods must be used. For example, reverse-phase LC quantifies non-polar to slightly polar molecules, while hydrophilic interaction LC detects strongly polar to slightly polar molecules (Bajad et al. [2006], Roberts et al. [2012]). This review will focus on liquid chromatography coupled MS (LC-MS), as it quantifies the largest range of metabolite polarity, and is widely used. When coupled with LC, the most common means of separation are reverse phase liquid chromatography (RPLC) for separation of hydrophobic metabolites, and hydrophilic interaction chromatography (HILIC) for the separation of hydrophilic metabolites (Zhou and Yin [2016]). MS platforms commonly used for untargeted metabolomics studies include low resolution techniques such as triple quadrupole (QQQ), quadrupole-ion trap (QIT) and high resolution techniques such as quadrupole-time of flight (Q-TOF), quadrupole Orbitrap (Q-Orbitrap) and Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS).

1.4.2 Processing LC-MS data

An overview of the manual and automatic components of the LC-MS pre-processing pipeline are detailed in Figure 1.3. The first step is to convert an LC-MS-produced dataset for a single individual into a list of "features" (defined as the combination of mass to charge (m/z) ratio and retention time) and their intensities. A variety of software packages designed to process metabolomic data have been developed for this purpose, of which XCMS, Mzmine2 and MAVEN are among the most popular (Katajamaa et al. [2006], Tautenhahn et al. [2008], Melamud et al. [2010]). Each pipeline involves three steps: 1) "peak selection", in which features

are identified and quantified, 2) retention time alignment, whereby intensity profiles of consecutive samples are aligned to allow maximal feature overlap and 3) adduct and isotope annotation. The most prominent difference between existing packages involves their approach towards assessing peak quality during the peak selection step. Both XCMS and Mzmine2 define low quality peaks according to a user-defined signal-to-noise ratio cutoff threshold; in contrast, MAVEN uses a machine-learning (neural network) approach. Because an independent comparison of MAVEN, Mzmine2 and XCMS has not yet been completed, one recommendation is to analyze metabolomics data using several packages and remove peaks that are not robustly identified by multiple algorithms (Tautenhahn et al. [2008]). This is one of several methods that aims to minimize false positives, as it has been shown that up to 90% of features in an LC-MS experiment are non-biological noise or degenerate in a typical LC-MS experiment (Mahieu and Patti [2017]). Other methods include curating databases of confirmed features identified using different separation techniques, and removing features not profiled in the corresponding database (Mahieu and Patti [2017]). An additional approach is to confirm the presence of the feature in a technical replicate. In practice, it is difficult to identify the same metabolites across replicates, as retention times may differ, and is therefore most often done in targeted metabolomics, in which only a small subset of features are quantified (Crews et al. [2009]). In addition to the above, another method for identifying robust features involves removing features that are not detected in a set of quality control (QC) samples consisting of either a set number of defined metabolites, or a combination of all tested samples (pooled sample) (Brodsky et al. [2010], Godzien et al. [2015]).

1.4.3 Normalizing LC-MS data

To be biologically informative, raw intensities need to be corrected for a) batch effects, b) missing values, and c) inter-sample variation. This section describes standard approaches used for such normalizations.

As a first step, raw intensities of each feature produced from data processing packages typically need to be corrected for systematic variation due to batch effects. In metabolomics data, a common type of batch effect is “chemical drift”.

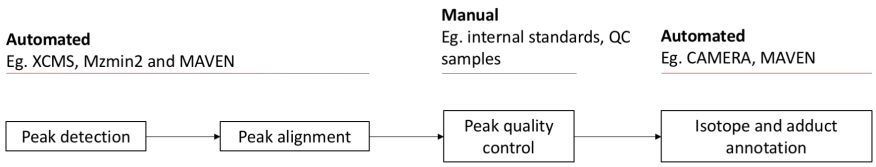

Figure 1.3: Automated and manual steps in untargeted metabolomics pre-processing pipeline. The algorithms listed are only examples of tools that could be used in each step.

This drift–caused by changes in signal that occur as metabolites interact with each other while waiting to be analyzed–can be corrected if QC samples are analyzed in between experimental samples (Vaikenborg et al. [2009], Shen et al. [2016]). While these corrections are not always performed, they have been shown to minimize inter-batch variation (Godzien et al. [2015]). Inter-batch variation and chemical drift can be visualized using dimensionality reduction approaches such as PCA and t-SNE (De Livera et al. [2015]).

Missing values can result from a variety of processes, and thus require a nuanced approach. Specifically, a missing value, which is an intensity of zero or infinity, can be created from a metabolite existing in one sample but 1) not existing in another, 2) existing at a concentration below an instrument's power of detection or 3) existing at a concentration above an instrument's power of detection. The problem of missing values can best be improved by increasing the sensitivity of detection of the MS platform. Numerous strategies have been developed to reduce missing values through a group of analytic techniques called missing value imputation (MVI). The utility of these techniques has empirically been found to depend on whether univariate or multivariate techniques are used to detect differentially abundant features (Karpievitch et al. [2012]).

Subsequently to above, both sample-wise and feature-wise normalization methods that concurrently consider multiple samples are typically applied to adjust for technical and biological variation. Sample-wise normalization methods

include quantile, linear baseline, total ion count (TIC) and LOESS normalization (Wu and Li [2015]). These methods adjust for technical factors that may have affected the entire sample. Feature-wise normalization methods involve constructing scaling factors for each feature, and include centering, scaling and transformations (Bolstad et al. [2003], van den Berg et al. [2006]). These approaches minimize the intensity differences between metabolites with low or high abundance, allowing relative perturbations of each metabolite to be compared. Usually, both sample-wise and feature-wise normalization methods are applied during pre-processing. However, because the type of normalization required is dependent on the separation technique and mass spectrometer used, no gold standard approach exists.

1.4.4 Testing for significant features in IEM studies

In a typical experimental design relevant to IEMs, one typically measures metabolomics data for a set of patients only or a set of patients and some controls (i.e., case/control design). Because each case is likely unique (i.e., may represent a unique disease caused by a rare genetic variant), data is usually analyzed for one patient at a time and compared against a) controls or b) other patients. Both parametric (e.g. t-test, ANOVA) and non-parametric (e.g. Mann-Whitney U-test, Wilcoxon-signed rank and Kruskal-Wallis) tests can be used to identify differentially abundant features in a given patient sample. When pursuing parametric tests, which typically have more statistical power compared to non-parametric tests, care must be taken to transform data so that it is distributed according to the expectation of the test (e.g., Gaussian for t-test). Correction for multiple testing is usually not performed due to lack of statistical power. When studying the genetic causes of rare diseases, in contrast to studies of common disease, one seeks to identify "outlier" features, as they represent abnormal metabolites that may be pathogenic. Availability of biological and technical replicates is important in confirming that a given metabolite value is a "biological" outlier, rather than an artifact of technical variation.

In metabolomic studies, selection of "control" samples (or comparators) that are as similar as possible to the patient being studied is paramount to reducing noise. This is difficult due to the numerous factors that influence the metabolome

(e.g age, sex, ethnicity, food consumption and time of day). Selection of controls often depends on patient availability, and the type of bio-fluid analyzed; finding suitable controls is much easier for analysis of urine samples, and much more difficult for plasma and CSF samples, due to the relative ease at which these samples can be provided. Estimations of the genetic component of metabolite variance vary between approximately 10 and 60%, with the largest determining factor being the the type of metabolite (Shah and Newgard [2015], Rhee et al. [2013]). Because of this moderate heritability, the trio structure has been suggested as a possible replacement for the classic case-control design, as it may enable the removal of metabolomic features attributable to non-disease related heritable phenotypes. For example, parents would likely show an abnormal profile for metabolites related to a heterozygous variant, which would be magnified in the bi-allelic patient (Long et al. [2017]). However, due to inherent uncertainty in quantification, the significant impact of age, gender and diet, and varying heritability of each metabolite, the human metabolome needs to be explored further before the trio structure can be robustly used in this manner. Overall, like any other -omics study of dynamic molecular traits, experimental designs that enable robust statistical adjustments for the effect of demographical and environmental factors are of key importance in identifying meaningful disease-associated metabolites. At the least, care should be taken to utilize metabolomic controls that share as many characteristics as possible with the population being studied.

1.4.5 Annotating features: adducts, isotopes, and metabolites

Once features have been identified, they can be annotated as an adduct or isotope of a particular metabolite. An adduct is an ionized metabolite that has become associated with another ion through electrospray ionization (ESI), most commonly H+, Na+, K+ and H20. An isotope is a metabolite that is composed of elements that are not in their most abundant form. A metabolite's most abundant isotopic form generally corresponds to its most abundant features. The most readily quantifiable adduct depends on the chromatographic separation performed (Keller et al. [2008]). Annotation of isotopes and adducts corresponding to a particular metabolite reduces the multiple testing burden by enabling the removal of features be-

longing to the same metabolite. Removal of redundant features is performed at the discretion of the researcher, as no standard filtering approach exists. In Mzmine2 and MAVEN packages, adduct and isotope annotation is performed automatically, whereas processing with XCMS requires use of an external package such as CAMERA to make these annotations (Kuhl et al. [2012]).

The putative metabolite mass annotated in the peak-annotation step described earlier is used to map a specific feature to known metabolite(s). Databases that include mass, adduct, spectra and structure data are then used to match metabolite masses to known metabolites that fall within the specific mass accuracy of the mass spectrometer used. Several such databases exist, such as the Human Metabolome Database (HMDB), Recon2, BioCyc and METLIN (Petri and Schmidt-Dannert [2004], Smith et al. [2005], Wishart et al. [2007], Thiele et al. [2013]). The HMDB in particular contains information on endogenous, food-based and drug-related metabolites found in human urine, CSF and plasma of humans. The human-specificity of this databases makes it particularly useful for mapping features identified through untargeted metabolomic methods, as the entire database can be utilized without a priori knowledge of each metabolite's origin. At the time of writing, it contains over 114,100 metabolites annotated with structure and chemical properties, a portion of which are also associated with specific genes (n = 5701). Of these metabolites, 19.5% have been detected in a bio-fluid, and 81.5% are predicted or expected. Its sister database, the Small Molecule Pathway Database (SMPDB), annotates a portion of genes and metabolites to specific small molecule pathways. Together, the HMDB and SMPDB facilitate biological interpretation at the gene and pathway level. Limitations of these databases include the relatively small number of detected and quantified metabolites, as well as the relevant paucity of genes annotated to both HMDB and SMPDB.

Identifying the "true" identity of a specific feature is challenging because each neutral mass can be annotated to multiple metabolites, so called isobaric compounds (i.e., their chemical properties result in them having the same mass and retention time upon ionization). Narrowing down the identity of a given feature is currently an active area of research (Li et al. [2013], Pirhaji et al. [2016]). Public databases that contain metabolite masses and MS/MS spectra can assist in confirming metabolite identities, in cases where mass spectra are available (Wishart et al.

[2007]). Additionally, “internal standards”, or radiolabeled compounds that can be easily identified through isotopic analysis, can be used for targeted metabolomics as well as untargeted lipidomics (detection of all lipids in the metabolome), as they allow researchers to benchmark when certain ions elute over time (i.e. their retention time), knowledge which can then be used to guide the interpretation of unknown features (Sysi-Aho et al. [2007], Ejigu et al. [2013], Weindl et al. [2015]). Validation of mapping between a feature and its assigned metabolite can be achieved by analyzing a purchased chemical standard through identical processing techniques, and comparing its m/z ratio, ion-source fragments and retention time to that of an experimentally-derived feature.

1.4.6 Identifying IEMs through untargeted metabolomic analysis

The creation of processing tools and metabolomic databases has greatly facilitated the use of untargeted metabolomics in diagnosing IEMs. As opposed to the narrow m/z range of targeted methods, untargeted methods aim to characterize a metabolome without any pre-conceived limitations on the m/z range under study. Both univariate and multivariate tests have been used to identify biomarkers of IEMs through untargeted metabolomics (Wikoff et al. [2007], Dercksen et al. [2013], Venter et al. [2014], Najdekr et al. [2015], L. et al. [2016], Kennedy et al. [2017], Pappan et al. [2017]). Recently, untargeted metabolomics was able to identify 20 of 21 IEMs, demonstrating its utility as a replacement for traditional newborn dry blood spot screenings (Miller et al. [2015]). Another more recent study found that untargeted metabolomics enabled the diagnosis of 42 of 46 IEMs (Coene et al. [2018]). Challenges with this method include separating noise (unrelated food and environmental influences) from disease signal, identifying isobaric compounds and attaining adequate quantification of polar and non-polar metabolites. Because of these challenges, untargeted metabolomics alone is unlikely to usurp traditional genomics-based methods that identify causative genes for novel IEMs (Tarailo-Graovac et al. [2016]). The benefits and drawbacks of integrating genetic and metabolomic data for the purpose of identifying both known and unknown IEMs are addressed in the subsequent section.

1.5 Integrating genomic and metabolomic data

Integration of genomic and metabolomic data has been performed for two primary purposes: to identify 1) metabolically active loci and 2) genes relevant to a metabolic disease phenotype. Most existing methods for combining genomic and metabolomic data conceptually follow from the former purpose.

1.5.1 Identification of metabolically active loci

Identification of metabolically active loci involves identifying variants that affect a metabolite's abundance. Population-based studies have combined genotyping microarray data (Gieger et al. [2008], Hicks et al. [2009], Illig et al. [2010], Suhre et al. [2011], Demirkan et al. [2012], Tukiainen et al. [2012], Kettunen et al. [2012], Shin et al. [2014], Draisma et al. [2015], Rhee et al. [2016], Long et al. [2017]) or WES/WGS data (Guo et al. [2015], Yazdani et al. [2016], Yu et al. [2016]) with metabolomic data through quantitative trait loci (QTL) analysis, in order to identify metabolically active genetic loci and to characterize the impact of common genetic variation on metabolite abundance (also known as heritability). In so-called metabolite QTL (mQTL) analysis, linear regression is used to associate genetic variants with individual metabolite intensities (or metabolite ratios), and significant variants are deemed metabolically active loci. To reduce spurious associations, most studies analyze between 2000 and 8000 subjects, and restrict their analysis to common SNVs (e.g. MAF $\geq$ 5%) or to pairs of variants and metabolic loci that are located nearby each other (i.e., "cis" association analysis) (Tukiainen et al. [2012]). mQTL analyses have identified disease biomarkers by associating variants in known disease-causing genes with metabolites (Yazdani et al. [2016], Gauba et al. [2012]. In these previous studies, loci strongly associated with metabolite intensities or ratios–termed "metabotypes"–have predominantly been found to map close to or in genes associated with enzymes, transporters and regulators of metabolism, facilitating biological interpretation (Shin et al. [2014]).

Consistent with transcriptomic-based QTL studies (e.g., eQTL studies), it has been reported that, on average, genetic variation is a stronger predictor of metabolite variance across individuals than demographic and symptom-based clinical covariates (Rhee et al. [2013], Shin et al. [2014]). Heritability estimates have varied

across classes of metabolites; Shin et al. found the heritability of amino acids (e.g. carnosine, h2 = 0.86, P = 6.8×10^{-4}) to be higher than lipids (e.g. lysophosphatidylcholine, h2 = 0.46, P = 2.0×10^{-7}), and that of essential amino acids (mean h2 = 0.29) to be lower than non-essential amino acids (mean h2 = 0.53), suggesting that some metabolites are more influenced by genomic variation than others, as one might expect (Shin et al. [2014]).

Rare SNVs (0.5% $\leq$ Minor Allele Frequency (MAF) $\leq$ 5%) have been found to have a larger effect size than common SNVs (MAF > 5% (Long et al. [2017]). However, because association analysis (e.g., mQTL analysis) is statistically challenging and underpowered in the rare variant setting, the effects of rare variants have primarily been studied on a case-by-case basis. Long et al. identified the effect of seventeen rare coding variants (SnpEff annotations such as "stop", "missense" or "frame") by first manually identifying an outlier metabolite that based on known biochemical reactions could be affected by the variant, and then confirming the presence of this putative rare variant and outlier metabolite combination in at least one other sample (Long et al. [2017]). Guo et al. examined the effect of rare coding variants by assessing the overlap of genes in perturbed metabolic pathways (i.e biochemical pathways with at least one outlier metabolite) with rare exonic variants (Guo et al. [2015]). These studies indeed show that rare variants can have a large effect on metabolic variation; however, the small number of rare variant-metabolite relationships yet identified suggest that clarifying their role in a systematic manner will likely require a more nuanced approach.

1.5.2 Gathering evidence for metabolic perturbation on a gene-by-gene basis

Studies aiming to explore rare disease using smaller sample sizes have used metabolomics in conjunction with curated biochemical knowledge to derive disease-specific biological insights. In a "pathway based approach", genes in enriched metabolic pathways were found to harbor variants that explained the patient's biochemical phenotype (Guo et al. [2015]). Several studies also reported how untargeted metabolomics could be used to quantify gene-associated metabolites to provide evidence a variant is disease-causing (Gauba et al. [2012], L. et al. [2016], Pappan et al. [2017]). An example of this approach is a study

that used untargeted metabolomics to demonstrate that a bi-allelic variant in N-acetylneuraminic acid synthase (NANS) in patients with infantile-onset severe developmental delay and skeletal dysplasia was reflected in high levels of N-acetylneuraminic acid (Karnebeek et al. [2016]). Confirmation of high levels of this enzymatic substrate of NANS suggested that the clinical phenotype was likely caused by an enzymatic deficiency in NANS. Normalization of skeletal dysplasia in a zebrafish model with knocked-out *nansa* and *nansb* (zebrafish orthologs for human NANS) occurred after supplementation with sialic acid, shedding light on a possible treatment. These findings support the idea than an integrated approach involving both genomics (i.e mircorarrays/WES/WGS) and metabolomics can be used to facilitate variant filtering and improve diagnosis and IEM discovery. As of yet, genome-wide integration has mainly been performed for exploratory purposes.

Integrating genomic and metabolomic data in a systematic manner is challenging because there is not a one-to-one mapping of genes to metabolites. Rather, some genes, such as enzymes involved in beta oxidation, will have many annotated metabolites, and other enzymatic genes will uniquely associate with only a handful of relevant metabolites. This dynamic makes it difficult to directly integrate genomic and metabolomic data at the variant and feature level, respectively.

1.6 Treatable Intellectual Disability Endeavour (TIDE) Exome Sequencing project

The TIDE study was conducted at BC Children's Hospital from 2015 until 2017 (UBC IRB approval H12-00067). The project aimed to identify the genetic cause of intellectual disability in patients with a neurometabolic phenotype. Patient inclusion criteria consisted of 1) a confirmed or potential neurodevelopmental disorder and 2) a metabolic phenotype, which could be a pattern of abnormal metabolites in urine, blood and CSF, abnormal results on biochemical functional studies or abnormalities in clinical history. Details of the WES bioinformatic pipeline and variant interpretation protocol have previously been published in Tarailo-Graovac et al. [2016] as well as in separate case reports (Collins et al. [2010], Horvath et al. [2016], Karnebeek et al. [2016]).

1.7 Strategies for disease gene prioritization

The ultimate goal of the TIDEX project was to identify the causative gene for each patient using a combination of WES and clinical expertise. Here, we do this through integrated use of personalized genomic and metabolomic data, with two assumptions:

1. Genomic data is static and metabolomic data is dynamic
    - A genetic mutation only perturbs the function of the gene in which it resides
    - In contrast, metabolites that differ between a single patient and all controls can reflect perturbations of the causative gene in addition to all genes with which it interacts
2. Genomic and metabolomic data are incomplete: each imperfectly captures one axis of an individual's biological processes

Given these assumptions, utilizing networks that depict functional and regulatory interactions between genes could help prioritize the causative variant by identifying subnetworks of disease-relevant regulation that could include the causative gene and its interaction partners. Networks could also mitigate the incompleteness of both genetic and metabolomic data by implicating genes closely connected to genes supported by patient-specific evidence.

The overall goal of this approach is to leverage interaction networks to identify genes that are likely related to an individual's disease given experimental evidence. Conceptually, this approach is similar to that taken by a class of algorithms known as "gene recommendation" algorithms, which identify new disease genes based on their relatedness to a set of "seed" genes already implicated in the disease. In our case, however, the overall goal is not to identify new disease genes, but to identify the relevance of each patient's "seed" gene to their disease based on the physical evidence supporting that gene and its proximity to other "seed" genes in an interaction network. Here, we will review the gene recommendation literature and discuss label propagation, a network-based computational algorithm that we use to perform gene recommendation in this thesis.

1.7.1 Gene Recommendation

Initial approaches performed gene recommendation based on similarity of genomic features. These features included those related to sequence (Marcotte et al. [1999], Adie et al. [2005], Aerts et al. [2006]), expression patterns (Marcotte et al. [1999], van Driel et al. [2003], Aerts et al. [2006], De Bie et al. [2007], Oti et al. [2008], Ala et al. [2008]), functional annotations (Freudenberg and Propping [2002], Perez-Iratxeta et al. [2002], Turner et al. [2003], Aerts et al. [2006], Li and Patra [2010a]), physical interactions (Aerts et al. [2006], Oti and Brunner [2007], Fraser and Plotkin [2007]) and literature descriptions (van Driel et al. [2003], Aerts et al. [2006], Li et al. [2006], Gaulton et al. [2007]). These initial approaches were moderately successful in prioritizing genes based on heterogeneous lines of evidence.

With the growth of ontologies defining gene function and streamlined access to large databases, integration of multiple types of evidence into functional linkage networks (FLN) became more feasible. Functional linkage networks (FLN) are the backbone of many network-based algorithms, as they model the interactions between genes through the incorporation of data from studies of coexpression, genetic interaction, protein interaction, coinheritance, colocalization, or shared domain composition. Many authors have noted the improvement in gene function prediction and prioritization when FLNs are used in this manner (Myers et al. [2005], Mostafavi et al. [2008], Tsuda et al. [2005], Deng et al. [2004], Mostafavi and Morris [2010], Lanckriet et al. [2004], Peña-Castillo et al. [2008], Pavlidis et al. [2002], Lancour et al. [2018]). Proximal genes in an FLN have a higher likelihood of performing similar biological functions and of belonging to similar disease pathways, making them ideal resources for gene recommendation/prioritization algorithms (Sharan et al. [2007], Oti and Brunner [2007]).

The first network-based algorithms to utilize FLNs were successful, encouraging more instances of their use (Weston et al. [2004], Xu and Li [2006]). Franke et al. integrated genomic features and coexpression data into a Bayesian network and used the distance between known disease genes and candidate genes for prioritization (Franke et al. [2006]). Information about disease-gene relationships was used to prioritize candidate genes (Lage et al. [2007], Wu et al. [2008]). Lage et al. scored each candidate in a PPI network based on its direct neighbors' associations

with diseases similar to the disease of interest. A more recent approach, termed CIPHER, assigned each candidate gene a score based on the correlation between a query phenotype's similarity with every other phenotype in a human phenotype network and a candidate gene's topological similarity to genes annotated to that phenotype in a protein-protein interaction network (Wu et al. [2008]). These efforts as well as others (e.g. bioPixie, STRING and HumanNet) demonstrated the utility of integrating heterogeneous sources of evidence in a network structure to prioritize relevant genes from a list of candidates (Myers et al. [2005], Peña-Castillo et al. [2008], von Mering et al. [2007], Lee et al. [2011]).

Several methods built highly specific networks to answer particular research questions (Itan et al. [2013], Greene et al. [2015]. Greene et al. enabled tissue-specific disease gene prioritization through use of a classifier on tissue-specific PPI networks. In another approach, Itan et al. built a condensed network of distances between a seed gene and candidate genes to enable disease gene prioritization.

Over time, computational disease gene prioritization algorithms became more efficient, allowing for deployment on web-based servers (Mostafavi et al. [2008], Linghu et al. [2009], Li and Patra [2010a,b], Itan et al. [2013]). GeneMania used ridge regression to integrate multiple heterogeneous networks and performed label propagation (discussed below) on the resulting network to identify genes of similar function to a query functional annotation. Due to its efficiency, label propagation (LP), as well as the related guilt by association (GBA), and random walk with restart (RWR), have been commonly used for function prediction and disease gene prioritization (Li and Patra [2010b,a], Lee and Lee [2018], Lancour et al. [2018], Qian et al. [2014]).

In general, gene function prediction and disease-gene prioritization involves three components:

- Functional linkage network (FLN) that depicts the relationship between genes using various types of evidence (Section 1.8)
- Genes related to process/disease of interest (i.e "seed" genes)
- Network-based algorithm

"Seed" genes generally have 1) similar function to genes we wish to identify

or 2) relevance to a disease of interest. They can be given a binary or continuous initial label score reflecting their relevance to the function/disease being studied.

1.7.2 Label propagation

Intuitively, LP can be thought of as an iterative procedure in which seed genes (nodes) propagate their initial label scores to their first degree neighbors, and then to their second and third, etc. Upon convergence of LP, nodes that are closer to seed genes have a higher score than nodes farther away. Nodes with a higher final score can be thought of as more relevant to a function/disease associated with seed genes. In this section, a general theoretical framework for Gaussian field label propagation will be provided. Previous implementations of label propagation and its related family of algorithms for the purpose of gene recommendation will then be reviewed.

1.7.3 Theoretical framework of label propagation

We assume that we are given a network, represented as a symmetric matrix, W. If there are n nodes (genes) in a network, then W is an $n \times n$ matrix. The weighted edges between node i and j is given the value w, with $w_{ij} = w_{ji} >= 0$. In a FLN, w can be binary (0 or 1) or continuous (weighted by the $-\log(pvalue)$ of the strength of evidence associating nodes (genes) i and j). The node labels can be represented as $y \in [0;1]^n$, with larger scores representing increased importance in the network.

The basic assumption of LP is that a score of node i, f, at iteration r can be written as a weighted combination of its neighbors so that

$$f_i^{(r)} = l \sum_{j=1}^{n} w_{ij} f_j^{r-1} + (1-l) y_i$$

where l is a constant such that $0 < l < 1$. In this formulation, nodes with high weighted degree with positive neighbors will have a relatively high impact on their neighbor's scores. To reduce the influence of these nodes, it is standard practice to normalize the matrix W. We choose to symmetrically normalize the matrix so that

$$S = D^{-1} W D^{-1}$$

or equivalently

$$f_i^r = l \sum_{j=1}^{n} \frac{w_{ij}}{\sqrt{d_i d_j}} f_j^{r-1} + (1-l)y_i$$

where D is a matrix with diagonal elements, d_i equal to the degree of node i.

Using the symmetrically normalized matrix, S, the final score vector can be written as

$$f = (1-l)(I - lW)^{-1}y$$

Given a symmetrically normalized network, the final score is dependent on only the *lambda* parameter, which controls the relative influence of neighboring nodes.

1.7.4 Using label propagation algorithms for candidate gene prioritization

Rather than directly using final scores of each node after convergence of a propagation algorithms to prioritize genes, several groups focused on combining the output of propagation algorithms with GWAS results to prioritize/recommend genes. Lee et al. integrated each node's score after GBA with its GWAS log odds to form a posterior score, which they then used to prioritize candidate genes (Lee et al. [2011]). Qian et al. applied LP to the same data sets to determine whether formally adjusting for node degree would uncover more disease-relevant genes, but found little overlap with Lee et al. (Qian et al. [2014]). Each group performed their prioritization on different FLNs, suggesting that both choice of FLN and selection of local vs global network propagation methods can have a large impact on gene prioritization. In a related approach, Lancour et al. combined LP final scores with gene level p-values from an Alzheimer's GWAS to identify relevant Alzheimer's genes (Lancour et al. [2018]). Ranking each gene by this score was found to increase the replication rate.

1.8 Functional Linkage Networks and the STRING network

The goal of integrating genomic and metabolomic data is to prioritize a list of candidate variants (and by association, genes) in order of their relevance to a patient's disease. As discussed above, using network-based methods on FLNs has been particularly successful at predicting disease genes as FLNs provide a framework through which prior knowledge about the relationship between genes and their association with disease can be introduced. In this section, we will introduce the FLN used in this thesis.

1.8.1 Selection of the STRING network

A recent performance comparison of 21 FLNs was conducted by testing their ability to recover half of a gene set when given the other half as seed genes (Huang et al. [2018]). Gene sets were constructed from Gene Ontology databases in addition to disease-specific GWAS and high-throughput gene expression assays. After performing random walk with restart, a form of label propagation, STRING was found to have the highest percentage recovery in both curated and gene-expression based networks. STRING's network also performed the best when evidence from MEDLINE abstract co-citation were removed, an indication that STRING exhibits the least amount of literature bias.

The STRING network is constructed from a combination of existing databases, including MINT, (Licata et al. [2012]), HPRD (Keshava Prasad et al. [2009], BIND (Alfarano et al. [2005]), DIP (Salwinski [2004]), BioGRID (Chatr-Aryamontri et al. [2017]), KEGG (Kanehisa et al. [2017]), Reactome (Fabregat et al. [2018]), IntAct (Kerrien et al. [2007]), EcoCyc (Keseler et al. [2013]), NCI-Nature Pathway Interaction Database and Gene Ontology (GO) protein complexes (Jensen et al. [2009]). Evidence of shared functions between genes is augmented with three additional sources of computational data:

- Evidence from evolution
  - Physical proximity of two genes (i.e intergenic distance) in prokaryotic genomes

  - Presence of gene fusions
  - Phylogenetic similarity between corresponding gene families. Genes that originated from a recent ancestor would be likely to share a function

- Evidence from high-throughput datasets
  - Similarity of transcriptional regulation profiles (i.e co-expression)

- Evidence from literature
  - Co-occurrence of both gene names in abstracts from SGD (Cherry et al. [1998]), OMIM (Amberger and Hamosh [2017]), The Interactive Fly (Yan et al. [2010]), and all abstracts from PubMed

In addition to these sources of evidence, knowledge of protein-protein interactions are also transferred between organisms based on their orthological hierarchy. Phylogenetically more similar organisms undergo stronger transfer. This transfer of information disproportionately benefits poorly characterized organisms.

1.9 Aims of this thesis

Candidate rare variants potentially causative for neurometabolic disease are commonly identified through the use of WES bioinformatics pipelines. Currently, the only method for identifying the causative gene from this list of candidate genes is through manual curation by clinical experts. The primary aim of this thesis is to develop a computational method for prioritizing the causative gene in a single IEM patient using evidence from personalized LC-MS and WES data. The sub-aims of this thesis are to:

- **Characterize LC-MS and WES data for IEM patients with neurometabolic disorders.**

- **Investigate whether LC-MS data can be used to prioritize the causative gene in a list of candidate genes.** Detecting enrichment for causative gene-associated metabolites in a patient's set of differentially abundant metabolites may provide evidence for the metabolic impact of certain mutations.

- **Determine whether performing label propagation on a functional linkage network initialized with a combined genetic and metabolomic score can prioritize a patient's causative gene.** As discussed in Section 1.7, assuming that each patient's causative gene impacts metabolism in some way, perturbation of metabolic function is more likely to occur in genes located proximally to the causal gene. Therefore, adding metabolomic evidence to an FLN could add signal to the causative gene's local neighborhood and help prioritize the causative gene.

Chapter 2

Methods

The goal of this integrative analysis was to use both WES and LC-MS data to prioritize the gene causative for each patient's IEM. Analyzing data from patients with known causative genes enabled us to assess whether and to what extent metabolomics could assist the prioritization of the causative gene from a list of 10-200 candidate variants. An outline of our method is provided in Figure 2.1.

2.1 Datasets

The 13 IEM patients in this study were genetically diagnosed through the TIDEX gene discovery project (UBC IRB approval H12-00067). Clinical and genomic variant information for each patient is available in the Appendix. The causative variants in four of the IEM patients were previously reported (Tarailo-Graovac et al. [2016]). Three of these patients were also profiled in separate case reports (Collins et al. [2010], Horvath et al. [2016], Karnebeek et al. [2016]). The IEM patients analyzed in this study met the patient inclusion criteria outlined in Section 1.6 and were found to have a validated or putatively causative variant through the WES variant prioritization pipeline used in this thesis and clinical expertise. An additional inclusion criteria was that the causative gene was included in the STRING database, as our method would not be able to prioritize the gene if absent from the STRING network. The WES and LC-MS methods used are detailed in the following sections.

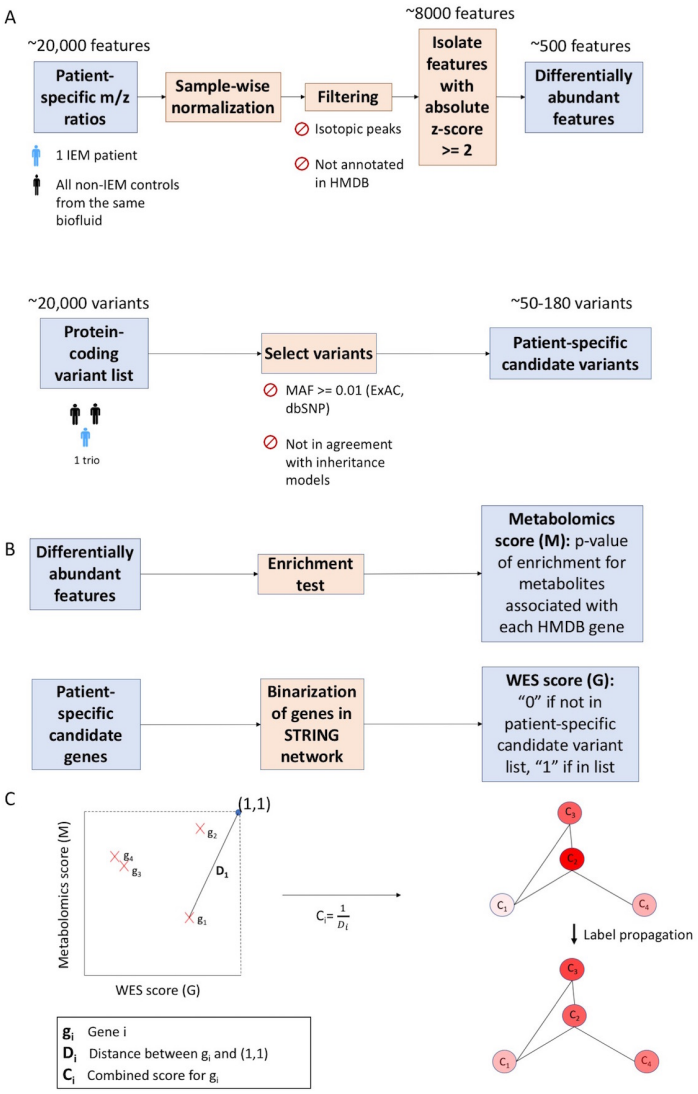

Figure 2.1: Summary of overall method. A) Briefly, raw metabolomics data was processed using XCMS and CAMERA and subsequently normalized through linear baseline normalization. Differentially abundant metabolites were isolated based on z-score. Raw genomic sequencing reads for each patient were processed and SnpEff was used to identify a list of candidate genomic variants with MAF ≤ 0.01. B) Enrichment for each of the 5371 HMDB genes in the patient-specific set of DAMs was assessed using Fisher's Exact Test, generating a metabolomic score. To generate the WES (i.e genomic) score, each gene in the STRING network was assigned "1" if also in the patient-specific candidate gene list, and "0" if not. C) These scores were plotted on an x-y grid, and the inverse of the distance between (1,1) and the coordinate of each gene, i, was considered the combined genomic and metabolomic score. Label propagation was then performed with the combined initial label score.

2.2 WES variant filtering pipeline

The WES data used in this study was from patients and, in some cases, their family members. WES data was generated using the Agilent SureSelect capture kit and the Illumina HiSeq 2000 sequencer. The obtained WES data was analyzed using a modified version of the semi-automated bioinformatics pipeline used in Tarailo-Graovac et al. [2016]. Figure 1.1 provides a visual overview of the pipeline. Briefly, the pipeline included 1) aligning the sequencing reads to the hg19 human reference genome and 2) annotating variants based on their predicted impact on protein function, predicted pathogenicity (CADD score), match to clinician-provided phenotype descriptions, and minor allele frequency (MAF) in dbSNP, NHLBI Exome Sequencing Project Exome Variant Server (EVS), and Exome Aggregation Consortium (ExAC) (Kircher et al. [2014], Smigielski et al. [2000], Exome Aggregate Consortium [2016], Lek et al. [2016]). Variants with an MAF $\leq$ 0.01 in dbSNP, EVS, or ExAC were removed. Using genetic information from each trio (mother, father, index), the variants were screened for agreement with multiple inheritance patterns in order to generate a stringent list of potentially pathogenic variants for each patient. Recessive mutations were defined as variants in which one allele was present in both parents and two alleles were present in the index. *De novo* variants were defined as variants only occurring in the index. These mutations were identified separately in the autosomal and sex chromosomes. If trio information was not available, all variants were classified as *de novo*. The genes to which each variant mapped were referred to as "candidate genes" (Figure 2.1, section A).

2.3 Processing and normalization of untargeted LC-MS metabolomics

In this section we describe the generation and processing of LC-MS data into a list of differentially abundant features.

2.3.1 LC-MS data generation

High-resolution untargeted metabolomics analysis of CSF and plasma was performed using UHPLC-QTOF mass spectrometry. Due to sample availability, plasma was analyzed for five of the IEM patients and 10 of the controls, and CSF was analyzed for eight of the IEM patients and 15 of the controls. Only samples profiled in the same bio-fluid were compared. CSF and plasma samples were deprotonated in methanol:ethanol solution (50:50; 100 microlitres of each sample plus 400 microlitres of methanol:ethanol solution). Samples were profiled in duplicate, however only one of each duplicate pair was analyzed in this study. A 2-microlitre sample was applied to an Acquity HSS T3 reverse-phase column (100 2.1 mm; 100 angstroms, 1.8 micrometers), and an Agilent 6540 UHD accurate mass UHPLC-QTOF mass spectrometer with acquisition in positive and negative modes was used. The buffers in positive mode consisted of buffer A (0.1 formic acid in water) and buffer B (0.1 formic acid in water:methanol solution (1:99)); in negative mode, the buffers consisted of buffer A (10 mM acetic acid) and buffer B (10 mM acetic acid in water:methanol solution (1:99)) (Coene et al. [2018]).

2.3.2 LC-MS data normalization and filtering

Once MS data was generated, the centwave and obiwarp methods from the XCMS package were used for peak detection and retention time correction, respectively, for both positive and negative electrospray ionization (ESI) detection modes (Tautenhahn et al. [2008]). Data-driven parameters were optimized separately for plasma and CSF samples using the IPO package (Libiseller et al. [2015]). CAMERA was used to annotate adducts and isotopes (Kuhl et al. [2012]). Linear baseline normalization was applied to each feature (Bolstad et al. [2003]). In linear baseline normalization, a baseline intensity profile is created from the median intensity of all features across all samples (hereby referred to as the ”baseline”), and all runs are assumed to be scalar multiples of the baseline intensity profile. For each metabolite i in sample j:

$$y'_{ij} = b_j y_{ij}$$

Where y'_{ij} is the log normalized abundance of a particular feature and y_{ij} is the log transformed unnormalized abundance. b is the per-sample scaling factor defined as the mean intensity of the baseline over the mean intensity of the sample (j):

$$b_j = \frac{y_{baseline}}{y_j}$$

Two filtering criteria were applied before analysis: removal of 1) features not annotated to any known metabolites in the HMDB and 2) features annotated as non-base isotopes (Wishart et al. [2007]). Z-scores based on the mean and standard deviation of a given metabolite across the IEM patient and controls were computed. Features for which the IEM patient had an absolute z-score greater than 2 (2 SD away from the mean) were isolated and called "differentially abundant metabolites" (DAMs). All DAMs found through both positive and negative mode analyses were annotated with compound identities within 15ppm of the compound mass using HMDB. Results from both positive and negative modes were combined for subsequent enrichment tests (Figure 2.1, section A).

2.4 Integrative analysis of WES and LC-MS data

Though the steps described above, an LC-MS metabolomics pipeline identified differentially abundant metabolites and a WES pipeline identified a set of candidate variants/genes (Figure 2.1, section A). The primary goal of subsequent analysis was to determine whether the output of these respective pipelines could be combined to prioritize the causative variant. To do this, as illustrated in section B and C of Figure 2.1, a combined metabolomic and genomic per-gene score was generated and propagated across the STRING network using label propagation. The final score of each candidate gene was used to rank candidate variants in order of their biological relevance to the network. Genes with higher relevance to the disease were hypothesized to have a higher final score. Our methods for devising a combined WES and metabolomic score and for performing label propagation are outlined below.

2.4.1 Creation of combined WES and metabolomic score

WES score

To generate the WES (i.e genomic) score, each gene in the STRING network was assigned a "1" if also in the patient-specific candidate gene list, and a "0" if not.

Metabolomic score

An enrichment test was performed to determine whether metabolites known to be associated with candidate genes were overrepresented in the patient-specific set of differentially abundant metabolites (Figure 2.1, section B). Curated sets of metabolites associated with each putative gene were parsed from files available from the HMDB web portal (hmdb.ca, Jan 1st 2017) (Wishart et al. [2007]). Enrichment was calculated using Fisher's Exact test. P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure, and reported as False Discovery Rate (FDR). Duplicate HMDB compound IDs were not removed, and it should be noted that significantly different results were obtained when they were.

The metabolomic score was computed as follows:

$$f_{met} = -log(p)$$

where p is the unadjusted enrichment p-value.

The metabolomic score for each gene was scaled to fall between 0 and 1.

Creation of combined score

The WES and metabolomics score were combined by defining each gene as a point on an (x,y) plane, where (x,y) = (WES score, metabolomics score), and calculating the inverse of the Euclidean distance between (1,1) and the coordinates of each gene.

The final combined score, f, for each gene, i, can be written as:

$$f_i = \cancel{p}\frac{1}{(1-G)^2 + (1-M)^2}$$

where G is the per-gene WES score (binary) and M is the scaled f_{met} score defined above. The inverse was taken because points closer to (1,1) should have higher scores, as their significance is supported by both genomic and metabolomic evidence.

2.5 Label propagation

Label propagation was performed as stipulated by Zhou et al. (Zhou et al. [2003]). The per-gene score, f_i, of each node at iteration, r, was determined by

$$f_i^{(r)} = l \sum_{j=1}^{n} w_{ij} f_j^{r-1} + (1 - l) y_i$$

where j is a connected node, l is a parameter between 0 and 1 that controls the degree of propagation between a node and its neighbors, w_{ij}, is the symetrically normalized edge weight between node i and node j and y_i is the label of node i.

Initial label values, y, were continuous between 0 and 1. Label propagation was run 30 times, although LP algorithms have been demonstrated to converge in less than 20 (Weston et al. [2004]). l was set at 0.99, as this is the parameter used in Zhou et al. [2003], and was not optimized for our data set due to limited sample size. However, the relative ranking of candidate genes were robust to small changes in this parameter. The final scores of each of the candidate genes were ranked to generate a prioritized candidate gene list.

2.6 Categorizing polarity of gene-associated metabolites

The utility of LC-MS metabolomic analysis in prioritizing the causative gene is dependent on whether metabolites associated with a particular gene are quantifiable by the LC-MS system. Reverse phase LC achieves the greatest chromatographic separation of semi-polar to non-polar metabolites, therefore these metabolites will be more readily detectable by the MS quantification system. Given these limitations, LC-MS analysis cannot accurately provide evidence for genes with highly polar metabolites.

Having a qualitative understanding of the polarity of metabolites associated with the causative gene is important for predicting whether or not LC-MS anal-

ysis can support its candidacy as the causative gene. The metabolites associated with each causative gene were evaluated for polarity. Briefly, if the majority of annotated metabolites ($\geq$ 50%) contained a phosphate group, or were ions, the metabolites were categorized as "very polar"; if the molecules contained saturated or unsaturated hydrocarbons ten or more carbons in length, the metabolites were categorized as "non-polar". All other genes that did not fit into either of these categories were categorized as "semi-polar".

2.7 Summary of method

Differentially abundant metabolites were identified through an untargeted LC-MS processing pipeline. A conservative list of pathogenic candidate variants were selected through a WES processing pipeline. A per-gene metabolic and genomic score was devised for each gene in the STRING network. Enrichment for HMDB gene-associated metabolites in the patient-specific set of differentially abundant metabolites was determined through a Fisher's Exact Test, and the resulting p-value was used as the basis for the per-gene metabolomic score. The per-gene genomic score was binary, with "1" reflecting the presence of that gene in the patient's candidate gene list. The combined genomic and metabolomic score was generated by first plotting the metabolomic and genomic scores as points on an (x, y) grid, with (x, y) = (Genomic score, Metabolomic score), and then taking the inverse of the Euclidean distance between (1,1) and the gene's coordinates. Label propagation was performed on the STRING network, with the per-gene combined scores as prior labels. The resulting final scores of each node/gene were then used to rank each candidate gene.

Chapter 3

Results

3.1 Characterization of WES and LC-MS data

In this chapter, we will first characterize the TIDE WES and LC-MS data used in this study, then detail the results of two approaches that use LC-MS data to support the prioritization of WES-identified causative genes. The first method uses metabolic enrichment directly, and the second method propagates a combined genomic and metabolomic score through the STRING network using the label propagation algorithm. Finally, we will characterize the factors that influence the success of these methods.

3.1.1 WES variants

Candidate WES variants in each patient mapped to an average of 108 genes (Table 3.1). Of these variants, 36.8% were autosomal recessive, 49.6% were autosomal *de novo*, 1.7% were recessive on sex chromosomes, and 2.4% were *de novo* on sex chromosomes. Definitions of these modes of inheritance can be found in Section 2.2. Only 29.9% of candidate genes were profiled in the HMDB, highlighting the challenges posed by our limited understanding of the metabolic involvement of all genes.

Table 3.1: Summary of WES data. Percentage of each patient's variants that fall into one of four modes of inheritance and the average number of candidate genes to which these variants map.

Variant category	Number
Number of candidate genes	109.6 ± 86.8
% recessive variants	36.8 ± 23.4
% recessive variants, Chr X or Y	1.7 ± 3.2
% de novo variants	49.6 ± 21.0
% de novo variants, Chr X or Y	2.4 ± 1.7
Number of genes annotated in HMDB (% of total implicated genes)	29 (29.9)

3.1.2 Characterization of LC-MS metabolomics features

LC-MS data from a single IEM patient was compared with those of controls in order to identify disease-relevant metabolites. CSF samples were available for eight individuals and 15 controls, and plasma samples were available for five individuals and 10 controls. A summary of the number of features identified in each bio-fluid is included in Table 3.2. On average, more features were detected in the ESI+ mode than in the ESI- mode. Out of all detected features, an average of 21.3% in the ESI+ mode and 24.7% in the ESI- mode mapped to known metabolite in the HMDB database. This low rate of mappability highlights a common shortcoming of existing analysis: a large number of potentially differentially abundant features are essentially discarded from further investigation. When each IEM patient was individually compared to controls, the number of features with absolute z-score greater than 2 in the IEM patient was on average larger when measured in CSF than in plasma (mean = 128 vs 662 in ESI+, 82 vs 390 in ESI-). This could be due to greater biological variability in CSF metabolites, or due to the reduced chemical stability of metabolites in CSF.

3.2 Assessing enrichment for gene-associated metabolites

Enrichment of differentially abundant metabolites (DAMs) for metabolites associated with Human Metabolome Database (HMDB) genes was assessed using a

Table 3.2: Characterization of LC-MS features.

Summary (Avg. number)	Plasma (n = 7)	CSF (n = 8)
Features in ESI+ mode	23405 ± 1343	23405 ± 2503
Features in ESI- mode	15720 ± 1377	16232 ± 2032
DAMs in ESI+ mode	128 ± 121	662 ± 268
DAMs in ESI- mode	82 ± 89	390 ± 125
Features that map to HMDB compounds in ESI+	5000 ± 471	11227 ± 1111
Compound assignments for each feature in ESI+	4 ± 6	6 ± 8
Features that map to HMDB compounds in ESI-	3884 ± 426	7012 ± 799
Compound assignments for each feature in ESI-	4 ± 7	4 ± 7
Compounds assigned to DAMs in ESI+	21795 ± 1979	31488 ± 839
Compounds assigned to DAMs in ESI-	10030 ± 654	12044 ± 199

Fisher's Exact Test. The total number of genes enriched for in each patient, and whether causative gene-associated metabolites were enriched is provided in Table 3.3 and Table 3.4. Each patient's DAMs were enriched for between 23 to 261 genes. The high degree of variability in the number of enriched genes reflects the heterogeneity of the underlying metabolic processes in this patient population. Two patients had multiple causative genes; for the purpose of analyses, these genes were considered independently, so that 15 genes were considered causative for 13 patients. For four of 15 causative genes, detection of enrichment in the causative gene was not possible because metabolite annotations were not available in the

HMDB. Enrichment was detected in two of the remaining 11 genes. The LC-MS detection method used is biased towards the detection of semi-polar to non-polar metabolites; to investigate whether polarity of causative gene-associated metabolites affected their enrichment, the polarity of metabolites associated with each causative gene was determined. Enrichment was detected in two of six causative genes associated with semi-polar and non-polar metabolites, but in no genes associated with very polar metabolites.

3.3 Integration of genomic and metabolic data to prioritize causative genes

To determine whether metabolomic information could help prioritize the causative gene, each patient's enriched gene profile was combined with their candidate variants to generate a combined genetic and metabolomic score. This score was then used as the initial label for label propagation on the STRING network, as outlined in Section 2.4.

Only a small percentage of nodes/genes were given an initial label. The initial label for each node/gene could be determined by genomic evidence, metabolomic evidence, or a combination thereof, depending on the availability of genomic/metabolomic evidence for that node/gene. Overall, $0.46\% \pm 0.45\%$ of nodes were assigned a score based on both genetic and metabolomic evidence, $0.47\% \pm 0.40\%$ of nodes were assigned a score based on only genetic evidence, $17.2\% \pm 10.6\%$ of nodes were assigned a score based on only metabolomic evidence, and $82.2\% \pm 10.5\%$ of nodes were not assigned an initial label.

Table 3.3: Summary of the causative gene(s) identified for each patient through the WES variant filtering pipeline. The enrichment of causative gene-associated metabolites, function of each causative gene and polarity of gene-associated metabolites are provided.

Patient number	Causative gene	Causative gene-associated metabolites enriched	Function of gene	Polarity of associated molecules
1	CPT1A	Yes	Transporter of fatty acids across the mitochondrial inner membrane	Non-polar
2	NANS	Yes	Generates phosphorylated forms of N-acetylneuraminic acid	Semi-polar
3	SCN2A	No	Functions in the generation and propagation of action potentials in neurons and muscle	Very polar
4	DYRK1A	No	Catalyzes its autophosphorylation on serine/threonine and tyrosine residues	Very polar
5	CACNA1D	No	Mediates the entry of calcium ions into cells and is involved in other calcium-dependent processes	None
6	CNKSR2	Not in HMDB	Encodes a multidomain protein that serves as a scaffold protein to mediate the mitogen-activated protein kinase pathways downstream from RAS	None

7	ECI1	No	Involved in beta-oxidation of fatty acids through transformation of enoyl-CoA esters	Semi-polar
8	IDS	No	Required for the lysosomal degradation of heparansulfate and dermatan sulfate	Semi-polar
8	HAL	No	Breaks down histidine to urocanic acid, which is further broken down in the liver to glutamic acid	Semi-polar
9	CHRNA1	Not in HMDB	Plays a role in acetylcholine binding as a membrane protein	None
9	DHFR	No	Catalyzes an essential reaction for de novo glycine and purine synthesis, and for DNA precursor synthesis	Semi-polar
10	ATP8A2	No	Transports aminophospholipids from the outer to the inner leaflet of various membranes and maintains asymmetric distribution of phospholipids, primarily in secretory vesicles	Very polar
11	MYO5B	Not in HMDB	May be involved in vesicular trafficking	None
12	KCNQ2	No	Part of ion channel complex	Very polar
13	VGLL4	Not in HMDB	Regulates alpha1-adrenergic activation of gene expression in cardiac myocytes	None

Table 3.4: Metabolic enrichment profile. Number of genes enriched for in the set of patient-specific differentially abundant metabolites.

Patient number	Number of enriched genes
1	86
2	88
3	28
4	45
5	84
6	197
7	91
8	107
9	23
10	55
11	143
12	48
13	261

3.3.1 Using label propagation to rank candidate genes

The final score of each node/gene was found using the label propagation algorithm. To simplify the results, the percentile rank of each category was sorted into two prioritization groups: high evidence (rank in 80th to 100th percentile) and low evidence (rank below the 80th percentile). The percentile rank of each causative gene(s) within each patient's candidate gene list is provided in Table 3.5. Overall, after LP with initial node labels defined by the combined genomic and metabolomic score, 8 of 15 (53.3%) causative genes were ranked in the "high" prioritization category. At the patient level, at least one causative gene was found in the "high" prioritization category for 8 of 13 (61.5%) patients, one of which was also prioritized using metabolomic enrichment alone (CPT1A).

Table 3.5: Label propagation results. Prioritization results after LP with both the combined genomic and metabolomic initial label scores and the genomic-only initial label scores. The percentile rank of each causative gene in the list of candidate genes, the number of candidate genes as well as the change in percentile rank of the causative gene after addition of metabolomic evidence to the initial label score (DeltaM) is provided, in addition to each gene's final genomic-only and combined prioritization category (PC).

Patient number	Causative gene	Rank with combined score	Rank with genomic score	Percentile with combined score	Percentile with genomic score	Percentile change (DeltaM)	PC with combined score	PC with genomic score
1	CPT1A	1/29	12/29	100	60	40	High	Low
2	NANS	56/151	91/151	63	40	23	Low	Low
3	SCN2A	5/45	12/45	91	74	17	High	Low
4	DYRK1A	6/68	6/68	94	94	0	High	High
5	CACNA1D	2/8	2/8	88	88	0	High	High
6	CNKSR2	31/53	30/53	42	44	-2	Low	Low
7	ECI1	30/271	63/271	89	77	12	High	Low
8	HAL	9/50	13/50	83	78	5	High	Low
8	IDS	19/50	20/50	63	61	2	Low	Low
9	CHRNA1	111/213	100/213	48	53	-5	Low	Low
9	DHFR	5/213	9/213	98	96	2	High	High
10	ATP8A2	92/167	107/167	45	36	9	Low	Low
11	MYO5B	11/81	11/81	87	87	0	High	High
12	MAST1	132/234	173/234	44	26	18	Low	Low
12	KCNQ2	61/234	89/234	74	62	12	Low	Low
13	VGLL4	50/55	51/55	10	8	2	Low	Low

3.3.2 Assessing the utility of metabolomic evidence

In order to determine whether the addition of metabolomic evidence was useful in the prioritization of the causative gene, the percentile ranking of the causative gene after LP with the combined initial score was compared to the percentile ranking after LP with the genomic-only score. The difference between the percentile ranking with the combined score and the percentile ranking with the genomics-only score will hereby be referred to as "DeltaM" where

$$DeltaM = P_C - P_G$$

if P_C is the percentile ranking with the combined score and P_G is the percentile ranking with the genomics-only score. A higher DeltaM signifies an increase in percentile ranking upon the addition of metabolomic evidence.

With the addition of metabolomic evidence to the initial score, the prioritization category changed from "low" to "high" for 3 of 13 patients and stayed the same for the remainder. In contrast, all genes without associated metabolites remained in the same prioritization category. Causative genes benefited more from the addition of metabolomic evidence more than non-causative genes (mean DeltaM for causative genes = 8.2%, mean DeltaM for non-causative genes = -0.11%, $p < 0.05$, Figure 3.1). In addition, DeltaM was positive for all causative genes, but not all non-causative genes, providing further evidence that metabolomics preferentially benefits the prioritization of causative genes.

3.3.3 Permutation test to generate null model of percentile rank

In order to put these rankings into context, we performed a permutation test to generate a null distribution of percentile rankings of each candidate gene. To generate this null distribution, the combined genomic and metabolomic labels were shuffled, and LP was performed. The average percentile ranking of the causative gene as well as their standard deviation across all permutations was calculated, and compared to the percentile ranking observed in real data (Table 3.6). Overall, 10 of 15 causative genes received a percentile ranking more than one standard deviation above the mean gene-specific permuted percentile. Six of these genes were

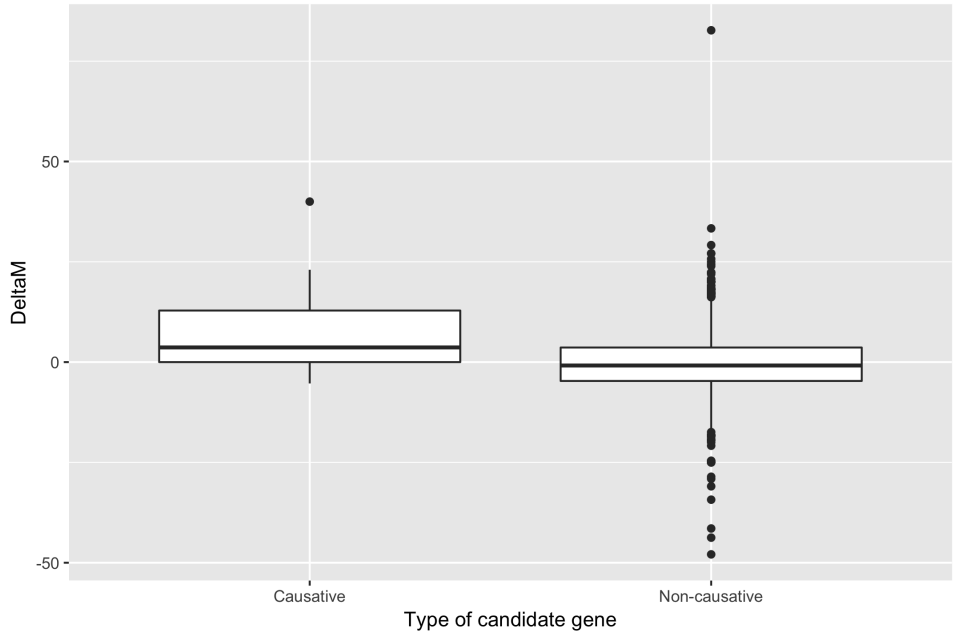

Figure 3.1: DeltaM of causative vs non-causative genes.

part of the 8 (71.4%) in the "high" prioritization category. Although more patients are needed to confirm this finding, these results suggest that genes with a high percentile rank may be less likely to be false positives.

3.3.4 Characterization of factors that effect gene prioritization

We next wanted to examine the influence of four characteristics of the causative gene on its percentile ranking: its centrality, the types of interactions occurring between it and its first degree neighbors, the number of its first degree neighbors profiled in the HMDB and the polarity of its associated metabolites. The significance of these factors in influencing the final prioritization percentile was evaluated through a linear model, with each of these factors (except polarity of associated metabolites due to uneven categorical representation) as independent variables and the percentile ranking as the dependent variable. The lack of association between centrality and prioritization in our data ($p > 0.05$) suggests that symmetric normalization effectively adjusts for network topology; however, prior literature has

Table 3.6: Label propagation results with permuted initial labels. Percentile rank of the causative gene in each patient's list of candidate genes, as well as the mean and standard deviation of the permuted (n=500) percentile for each causative gene.

Patient number	Causative gene	Betweeness centrality	Percentile ranking of causative gene with combined score	Permuted percentile ranking with combined score	Standard deviation of permuted percentile ranking with combined score
1	CPT1A	3823	100	61	7.6
2	NANS	3882	63	36.9	5.9
3	SCN2A	17960	91	78.3	6.4
4	DYRK1A	56248	94	94.2	0.7
5	CACNA1D	25905	88	75	8.2
6	CNKSR2	4086	42	45	7.3
7	ECI1	10523	89	78	3.8
8	HAL	2865	83	72.1	6.6
8	IDS	6455	63	57	7.9
9	CHRNA1	2759	48	53	5.1
9	DHFR	32071	98	95	1.3
10	ATP8A2	3370	45	35	4.8
11	MYO5B	38544	87	87.8	1.8
12	KCNQ2	7820	74	61.3	5.3
13	VGLL4	184	10	12.1	7.3

shown that the centrality of a node empirically influences its prioritization, therefore further research is needed to confirm this trend (Lee et al. [2011], Qian et al. [2014]). The STRING database sorts functional interactions (i.e. edges) between proteins into several categories: "activation", "binding", "catalysis", "expression", "inhibition", "post-translational modifications" and "reactions". The type and strength of the functional interactions between a causative gene and its first degree neighbors was not associated with its percentile ranking ($p > 0.05$). Similarly, the number of a causative gene's first degree neighbors profiled in HMDB had no effect on its percentile ranking, suggesting that the metabolic activity of a gene's neighborhood does not affect its prioritization ($p > 0.05$). In general, the increasing polarity of gene-associated metabolites negatively affected its percentile ranking, although no formal statistical test was performed (Figure 3.2).

To determine whether the above factors influenced the utility of metabolomic evidence, a linear model was used to test the association between these factors and DeltaM. Centrality and the number of HMDB-associated genes in the causative gene's first-degree neighborhood were not significantly associated with DeltaM. However, causative genes connected to other genes through the "activation" relationship were more likely to benefit from metabolomic evidence ($p < 0.05$). Again, the polarity of causative gene-associated metabolites was found to mildly influence the utility of metabolomic evidence; causative genes with non-polar metabolites exhibited a higher DeltaM than semi-polar or polar metabolites (Figure 3.3). Combined, these results suggest that genes associated with non polar metabolites or those involved in acting as/associating with enzyme activators may benefit preferentially from the addition of metabolomic evidence.

3.4 Summary

In this thesis, we assembled an untargeted LC-MS metabolomic analysis pipeline capable of taking in raw LC-MS data as input and returning a list of differentially abundant metabolites (DAM). Through characterization of processed LC-MS data, we found that metabolomics suffers from low feature to metabolite mappability. On average, only approximately one fifth of LC-MS features mapped to known metabolites in the HMDB database. We assessed enrichment in this list of DAMs

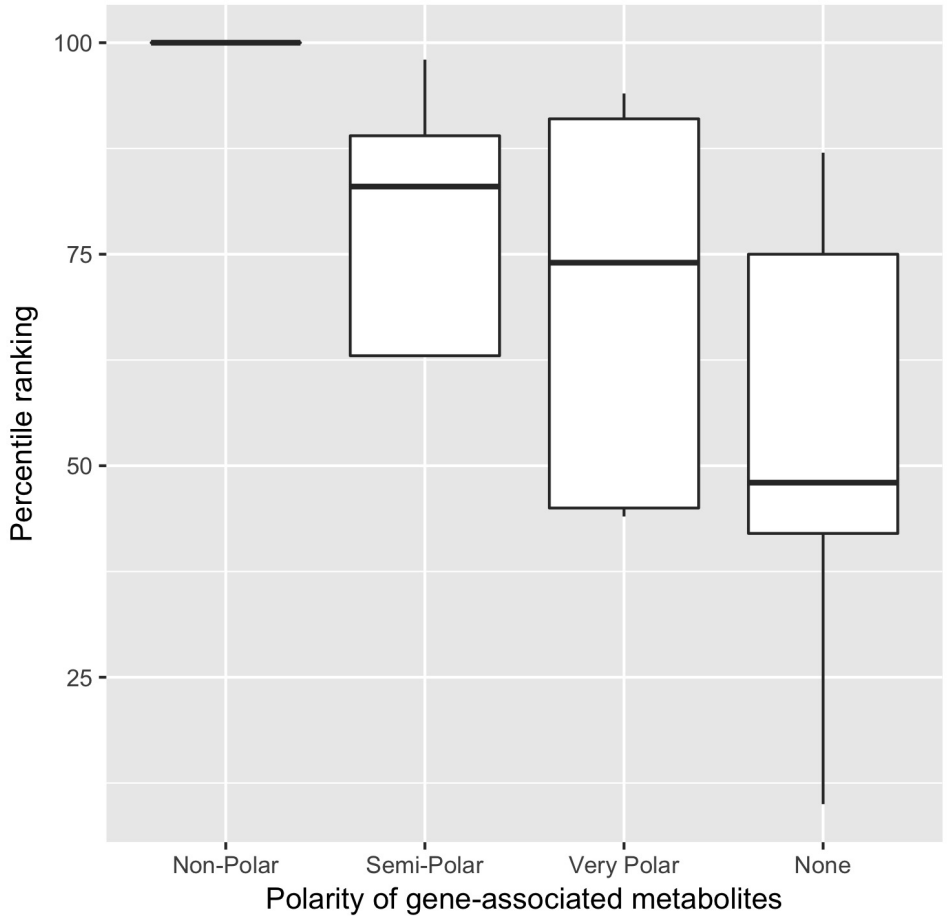

Figure 3.2: Effect of polarity of gene-associated metabolites on percentile ranking. "None" indicates that the gene has no associated metabolites in HMDB.

for gene-associated metabolites and combined this with candidate variant lists from a WES variant filtering pipeline to create a combined per-node score that was propagated through a FLN using a label propagation algorithm. The final propagated score of each node was used to rank each candidate gene in order of its relevance to each patient's disease. Integrated genomic and metabolomic evidence was able to prioritize the causative gene in the top 20th percentile of candidate genes for 61.5% (8 of 13) of patients, 75% of which achieved a percentile prioritization score

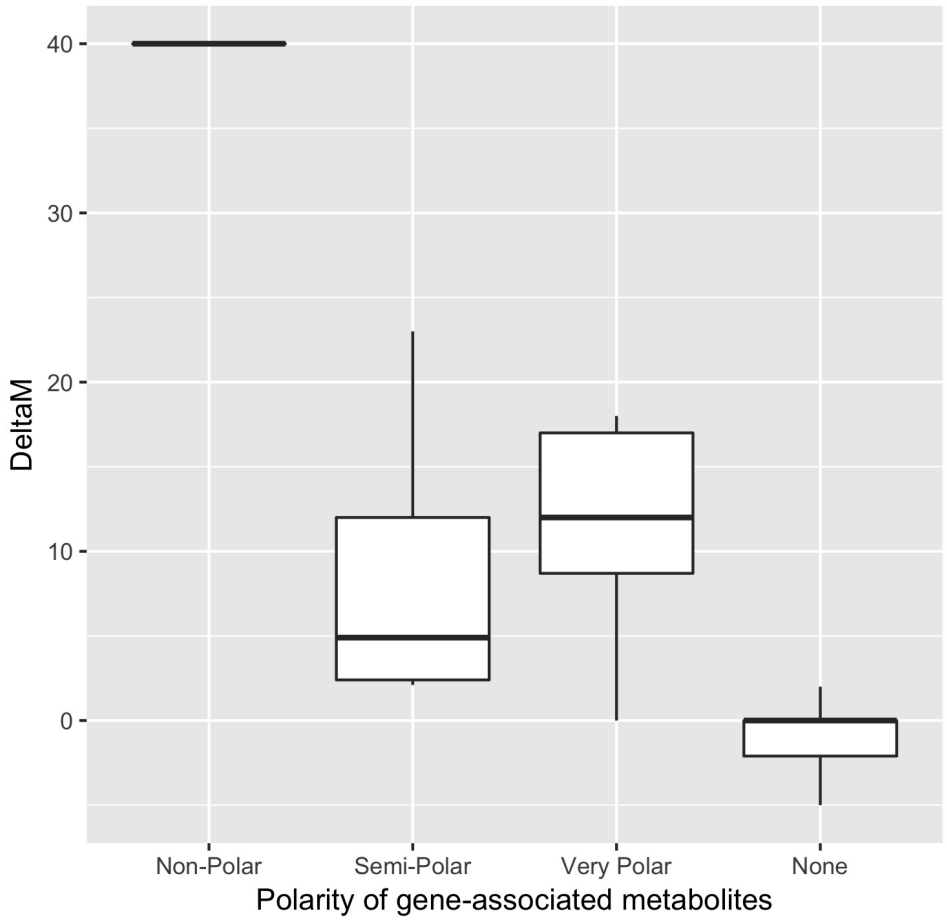

Figure 3.3: Utility of metabolomic evidence for causative genes associated with metabolites of varying polarities. DeltaM of causative genes by polarity of gene-associated metabolites. "None" indicates that the gene did not have any associated metabolites in HMDB.

at least one standard deviation above a permuted percentile. Combining genomic and metabolomic evidence resulted in the prioritization of the causative gene in 30.7% more patients than was possible with genomics evidence alone, and on average improved the percentile rank of the causative gene by 7.9%. Metabolomic evidence primarily benefited the prioritization of causative genes, although non-causative genes also saw an increase in DeltaM. Metabolomic evidence was particularly helpful for prioritizing genes with non-polar metabolites, eg. CPT1A, with an increase of 40 percentile points upon addition of metabolomic evidence. In addition, causative genes involved in activating roles with their first degree neighbors may preferentially benefit from the addition of metabolomic evidence.

Chapter 4

Disscussion

To the author's knowledge, the method presented in this thesis is the first to combine genomic and LC-MS data for the purpose of disease gene prioritization, although others have approached the broader problem of genomic/metabolomic data integration (Krumsiek et al. [2011], Li et al. [2013], Pirhaji et al. [2016]). Our network-based method enabled the prioritization of the causative gene in approximately 60% of the patients profiled. Polarity had an obvious impact on prioritization, as genes with non-polar metabolites experience a greater boost in prioritization when LC-MS data is used. The combination of multiple chromatography techniques would address this problem by enabling metabolites of diverse polarities to be profiled. Additionally, genes acting as or associating with enzyme activators benefited more from the addition of metabolomic evidence. Enzyme activators are often associated with metabolic pathways and processes (e.g. hexokinase-I and glucokinase), suggesting that a gene that is directly implicated in metabolism will be more likely t benefit from metabolomic evidence.

Several areas of our method could be further refined. The binary genomic score failed to reflect relevant characteristics of the variant, namely its frequency, pathogenicity and relevance to disease. Inclusion of such variables in the generation of label biases would enable a more data-driven approach to prioritization. In addition, integrating the nature of the genomic controls used (i.e singleton or trio) into the genomic score may provide additional context for the strength of genomic evidence.

The success of the method put forward in this thesis is limited by the precision of LC-MS metabolomic data, as we are unable to conclusively identify each feature. Challenges to integrating metabolomic data into the variant prioritization process can broadly be divided into those concerning the technical aspects of metabolite quantification and identification, and those concerning the biological interpretation of results. On the technical side, it is currently impossible to know the number of unique metabolites in the typical plasma/CSF/urine metabolome, as no LC-MS protocol is capable of identifying all metabolites. This means that for experiments aiming to capture an unbiased snapshot of the metabolome, a combination of chromatography techniques must be used. Comparisons across platforms are difficult to make, as little is known about how results from different analytic techniques can be compared, although some efforts have been made (Büscher et al. [2009], Yet et al. [2016], Leuthold et al. [2017]). Further, only approximately 65% of metabolites are quantifiable in all three body fluids (plasma, urine and CSF), indicating that care must also be taken to select the most disease-relevant bio-fluid (Kennedy et al. [2017]). Additionally, the choice of pre-processing algorithms may have a large effect on feature detection and adduct annotation. This renders analysis reproducibility difficult. On the biological interpretation side, there is a lack of established methods for mapping genomic perturbations to their downstream (directly and indirectly) impacted metabolites in the rare disease context. mQTL studies are underpowered, particularly for those caused by rare variants, making it difficult for them to identify novel gene-metabolite associations. Incomplete annotation of gene-metabolite associations in databases such as the HMDB limits our ability to use this data for patient diagnostics. The lack of methods available to assess the overall characteristics of metabolites, like polarity, challenges large scale omic studies, as it limits their ability to account for these variables in a quantitative and reliable manner.

The challenges facing the use of metabolomics in rare metabolic disease diagnostics are best illustrated through the exploration of four cases analyzed in this study and by Tarailo-Graovac et al., each with known IEM-causing variants in CPT1A, NANS, DYRK1A and SCN2A, respectively. CPT1A and NANS are enzymes that catalyze highly specific interactions, and do not share many metabolites with other genes. In contrast, SCN2A, a transmembrane sodium ion transporter,

interacts with the common metabolites ATP, sodium and water, and DYRK1A, a phosphotransferase, interacts with ATP and ADP. Metabolites associated with SCN2A and DYRK1A would be less likely to be identified as differentially abundant, as ATP and ADP are used in multiple metabolic pathways and are under strong homeostatic control. This is echoed by Nicholson et al, who notes that unlike in eQTL studies, there is not a one-to-one mapping between a metabolite and a gene (Nicholson et al. [2011]). Because more statistical tests are performed in mQTL studies, effect sizes must be larger to reach statistical significance. This suggests that even when a robust snapshot of the metabolome is procured using multiple chromatography methods, metabolomics may only be useful in confirming perturbations in genes that interact with metabolites under weak homeostatic control, as they are likely to have larger effect sizes. Metabolomics therefore might not be of use in the prioritization of SCN2A and DYRK1A. The finding that adding metabolomic evidence to the initial label bias does not assist the prioritization of these genes through our method supports this claim. Further work is needed to evaluate the impact of homeostatic control on gene prioritization using metabolomic data.

In this work, the identify of each differentially abundant feature could only be narrowed down to on average of 4 metabolites in ESI+ and ESI- modes, reflecting the high degree of uncertainty associated with the identity of a differentially abundant metabolite. A solution to this would be to restrict metabolite detection to a list of approximately 300 metabolites known to be detectable by this LC-MS system, as has been done previously (Coene et al. [2018]). However, this would limit the ability of this approach to characterize the metabolome in an unbiased manner. Using network structures to refine the true feature to metabolite mapping has been proposed as a viable approach to reduce uncertainty associated with metabolite identification, and should be explored further(Pirhaji et al. [2016]).

Given the increased technical reliability of WES as compared to untargeted LC-MS, methods that could dynamically weight either source of evidence based on its technical reliability deserve further exploration. Ideally, high confidence metabolite identifications would be weighed more heavily than low confidence metabolite identifications, thereby mitigating the effects of noise. Additionally, mapping all metabolomic features to the gene level in order to perform LP on a

homogeneous network may have the effect of oversimplifying the interactions between genes and metabolites. For example, some metabolites are known to be associated with a particular gene with a higher confidence than others. By utilizing an enrichment score, our method effectively considers all metabolites equally, when in reality some metabolites are more robustly associated. Applying LP to a heterogeneous network, which would include edges between genes and between genes and metabolites, may allow propagation to occur while allowing for weighting of specific gene-metabolite associations (Lotfi Shahreza et al. [2017]. Further quantification of the strength of gene-metabolite associations is needed before this weighting can occur in a robust manner.

4.1 Future Work

In order for the successful integration of genomics and LC-MS based metabolomics in the clinical diagnosis of IEMs, two major technical areas of improvement must be addressed. First, existing feature detection and adduct/isotope annotation methods must be refined and benchmarked for use in clinical metabolomics. Several publicly available databases with known chemical compositions have been generated for this purpose (Kenar et al. [2014]). Second, explorations of gene-metabolite associations through mQTL studies are needed to expand gene-metabolite annotations.

On the biological interpretation side, understanding the degree to which a particular metabolite is regulated (i.e by which genes) would help identify metabolites that are under strong homeostatic control, and by association, genes that may not benefit from metabolomic-guided prioritization. In addition, given that label propagation allows each node to be influenced by its neighbors, the local neighborhood surrounding a causative gene may be important in determining whether or not prioritization through label propagation will be effective. Preliminary characterization of the local neighborhood surrounding each causative gene in this thesis suggests that prioritization is not influenced by just one factor, but rather by a multitude of factors working in concert. Further investigation into network-based and metabolic factors that affect prioritization may be informative for future methods.

Bibliography

Emma Graham, Jessica Lee, Magda Price, Maja Tarailo-Graovac, Allison Matthews, Udo Engelke, Jeffrey Tang, Leo A.J. Kluijtmans, Ron A. Wevers, Wyeth W. Wasserman, Clara D.M. van Karnebeek, and Sara Mostafavi. Integration of genomics and metabolomics for prioritization of rare disease variants: a 2018 literature review, 2018. ISSN 15732665. → pages vi, 1

Maja Tarailo-Graovac, Casper Shyr, Colin J. Ross, Gabriella A. Horvath, Ramona Salvarinova, Xin C. Ye, Lin-Hua Zhang, Amit P. Bhavsar, Jessica J.Y. Lee, Britt I. Drögemöller, Mena Abdelsayed, Majid Alfadhel, Linlea Armstrong, Matthias R. Baumgartner, Patricie Burda, Mary B. Connolly, Jessie Cameron, Michelle Demos, Tammie Dewan, Janis Dionne, A. Mark Evans, Jan M. Friedman, Ian Garber, Suzanne Lewis, Jiqiang Ling, Rupasri Mandal, Andre Mattman, Margaret McKinnon, Aspasia Michoulas, Daniel Metzger, Oluseye A. Ogunbayo, Bojana Rakic, Jacob Rozmus, Peter Ruben, Bryan Sayson, Saikat Santra, Kirk R. Schultz, Kathryn Selby, Paul Shekel, Sandra Sirrs, Cristina Skrypnyk, Andrea Superti-Furga, Stuart E. Turvey, Margot I. Van Allen, David Wishart, Jiang Wu, John Wu, Dimitrios Zafeiriou, Leo Kluijtmans, Ron A. Wevers, Patrice Eydoux, Anna M. Lehman, Hilary Vallance, Sylvia Stockler-Ipsiroglu, Graham Sinclair, Wyeth W. Wasserman, and Clara D. van Karnebeek. Exome Sequencing and the Management of Neurometabolic Disorders. *New England Journal of Medicine*, page NEJMoa1515792, 2016. ISSN 0028-4793. doi:10.1056/NEJMoa1515792. URL http://www.nejm.org/doi/10.1056/NEJMoa1515792. → pages xii, 2, 4, 5, 13, 16, 25, 27, 49

Hans Van Bokhoven. Genetic and Epigenetic Networks in Intellectual Disabilities. *Annu. Rev. Genet*, 45:81–104, 2011. ISSN 1545-2948. doi:10.1146/annurev-genet-110410-132512. → page 2

Clara D M Van Karnebeek and Sylvia Stockler. Treatable inborn errors of

metabolism causing intellectual disability: A systematic literature review, 2012. ISSN 10967192. → page 2

Sorcha A. Collins, Graham Sinclair, Sarah McIntosh, Fiona Bamforth, Robert Thompson, Isaac Sobol, Geraldine Osborne, Andre Corriveau, Maria Santos, Brendan Hanley, Cheryl R. Greenberg, Hilary Vallance, and Laura Arbour. Carnitine palmitoyltransferase 1A (CPT1A) P479L prevalence in live newborns in Yukon, Northwest Territories, and Nunavut. *Molecular Genetics and Metabolism*, 101(2-3):200–204, 2010. ISSN 10967192. doi:10.1016/j.ymgme.2010.07.013. → pages 2, 16, 25

Gabriella A. Horvath, Michelle Demos, Casper Shyr, Allison Matthews, Linhua Zhang, Simone Race, Sylvia Stockler-Ipsiroglu, Margot I. Van Allen, Ogan Mancarci, Lilah Toker, Paul Pavlidis, Colin J. Ross, Wyeth W. Wasserman, Natalie Trump, Simon Heales, Simon Pope, J. Helen Cross, and Clara D.M. van Karnebeek. Secondary neurotransmitter deficiencies in epilepsy caused by voltage-gated sodium channelopathies: A potential treatment target? *Molecular Genetics and Metabolism*, 117(1):42–48, 2016. ISSN 10967206. doi:10.1016/j.ymgme.2015.11.008. → pages 2, 16, 25

Clara D M Van Karnebeek, Luisa Bonafé, Xiao-yan Wen, Maja Tarailo-graovac, Sara Balzano, Beryl Royer-bertrand, Angel Ashikov, Livia Garavelli, Isabella Mammi, Licia Turolla, Catherine Breen, Dian Donnai, Valerie Cormier, Delphine Heron, Gen Nishimura, Shinichi Uchikawa, Belinda Campos-xavier, Antonio Rossi, Thierry Hennet, Koroboshka Brand-arzamendi, Jacob Rozmus, Keith Harshman, Brian J Stevenson, Enrico Girardi, Giulio Superti-furga, Tammie Dewan, Alissa Collingridge, Jessie Halparin, Colin J Ross, Margot I Van Allen, Andrea Rossi, Udo F Engelke, and Leo A J Kluijtmans. NANS-mediated synthesis of sialic acid is required for brain and skeletal development. *Nature Publishing Group*, 48(7):777–784, 2016. ISSN 1061-4036. doi:10.1038/ng.3578. URL http://dx.doi.org/10.1038/ng.3578. → pages 2, 16, 25

Nenad B., Michael G.K., Marinus D., and Carlo D.-V. IEMBASE, a knowledgebase of inborn errors of metabolism. *Molecular Genetics and Metabolism*, 111(3):296, 2014. ISSN 1096-7192. URL http://www.embase.com/search/results?subaction=viewrecord{&}from=export{&}id=L71804976. → page 2

Kevin A Strauss, Erik G Puffenberger, and D Holmes Morton. Maple Syrup Urine Disease. In *GeneReviews*, volume 28, pages 93–97. 2013.

doi:NBK1319[bookaccession]. URL http://www.ncbi.nlm.nih.gov/books/NBK1319/. → page 3

Caroline H. Johnson, Julijana Ivanisevic, and Gary Siuzdak. Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7):451–459, 2016. ISSN 1471-0072. doi:10.1038/nrm.2016.25. URL http://www.nature.com/doifinder/10.1038/nrm.2016.25. → pages 3, 6

Exome Aggregate Consortium. ExAC Browser, 2016. URL http://exac.broadinstitute.org/variant/9-139413097-T-G. → pages 3, 27

E M Smigielski, K Sirotkin, M Ward, and S T Sherry. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research*, 28(1):352–355, 2000. ISSN 0305-1048. doi:10.1093/nar/28.1.352. → pages 3, 27

Monkol Lek, Konrad J Karczewski, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O, James S Ware, Andrew J Hill, Beryl B Cummings, Daniel P Birnbaum, Jack A Kosmicki, Laramie Duncan, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, David N Cooper, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, D Stenson, Christine Stevens, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, G Wilson, Mark J Daly, and Daniel G MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, 536(7616):030338, 2016. ISSN 0028-0836. doi:10.1101/030338. URL http://biorxiv.org/lookup/doi/10.1101/030338. → pages 3, 27

Pauline C. Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003. ISSN 03051048. doi:10.1093/nar/gkg509. → page 3

Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, (SUPPL.76), 2013. ISSN 19348266. doi:10.1002/0471142905.hg0720s76. → page 3

Maja Tarailo-Graovac, Wyeth W. Wasserman, and Clara D. M. Van Karnebeek. Impact of next-generation sequencing on diagnosis and management of neurometabolic disorders: current advances and future perspectives. *Expert Review of Molecular Diagnostics*, 17(4):307–309, 2017. ISSN 1473-7159. doi:10.1080/14737159.2017.1293527. URL https://www.tandfonline.com/doi/full/10.1080/14737159.2017.1293527. → page 4

Yaping Yang, Donna M. Muzny, Jeffrey G. Reid, Matthew N. Bainbridge, Alecia Willis, Patricia A. Ward, Alicia Braxton, Joke Beuten, Fan Xia, Zhiyv Niu, Matthew Hardison, Richard Person, Mir Reza Bekheirnia, Magalie S. Leduc, Amelia Kirby, Peter Pham, Jennifer Scull, Min Wang, Yan Ding, Sharon E. Plon, James R. Lupski, Arthur L. Beaudet, Richard A. Gibbs, and Christine M. Eng. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine*, 369(16):1502–1511, 2013. ISSN 0028-4793. doi:10.1056/NEJMoa1306555. URL http://www.nejm.org/doi/abs/10.1056/NEJMoa1306555. → page 4

Aziz Belkadi, Alexandre Bolze, Yuval Itan, Aurélie Cobat, Quentin B. Vincent, Alexander Antipenko, Lei Shang, Bertrand Boisson, Jean-Laurent Casanova, and Laurent Abel. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, 112(17):5473–5478, 2015. ISSN 0027-8424. doi:10.1073/pnas.1418631112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1418631112. → page 4

Gabrielle Bertier, Yann Joly, and Martin Hétu. Unsolved challenges of clinical whole-exome sequencing: A systematic literature review of end-users' views. Accepted 07/28/2016. *BMC medical genomics*, 9(1):doi: 10.1186/s12920–016–0213–6, 2016. ISSN 1755-8794. doi:10.1186/s12920-016-0213-6. URL http://dx.doi.org/10.1186/s12920-016-0213-6. → page 4

Aihua Zhang, Hui Sun, Ping Wang, Ying Han, and Xijun Wang. Modern analytical techniques in metabolomics analysis. *The Analyst*, 137(2):293–300, 2012. ISSN 0003-2654. doi:10.1039/C1AN15605E. URL http://xlink.rsc.org/?DOI=C1AN15605E. → page 7

Sunil U. Bajad, Wenyun Lu, Elizabeth H. Kimball, Jie Yuan, Celeste Peterson, and Joshua D. Rabinowitz. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *Journal of Chromatography A*, 1125(1):76–88, 2006. ISSN 00219673. doi:10.1016/j.chroma.2006.05.019. → page 7

Lee D. Roberts, Amanda L. Souza, Robert E. Gerszten, and Clary B. Clish. Targeted metabolomics. *Current Protocols in Molecular Biology*, 1 (SUPPL.98), 2012. ISSN 19343639. doi:10.1002/0471142727.mb3002s98. → page 7

Juntuo Zhou and Yuxin Yin. Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry. *The Analyst*, 141 (23):6362–6373, 2016. ISSN 0003-2654. doi:10.1039/C6AN01753C. URL http://xlink.rsc.org/?DOI=C6AN01753C. → page 7

Mikko Katajamaa, Jarkko Miettinen, and Matej Oresic. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics (Oxford, England)*, 22(5):634–6, 2006. ISSN 1367-4803. doi:10.1093/bioinformatics/btk039. URL http://www.ncbi.nlm.nih.gov/pubmed/16403790. → page 7

R Tautenhahn, C Bottcher, and S Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9:16, 2008. ISSN 1471-2105. doi:10.1186/1471-2105-9-504. → pages 7, 8, 28

Eugene Melamud, Livia Vastag, and Joshua D Rabinowitz. Metabolomic analysis and visualization engine for LC-MS data. *Analytical chemistry*, 82(23): 9818–9826, 2010. ISSN 1520-6882. doi:10.1021/ac1021166. → page 7

Nathaniel G. Mahieu and Gary J. Patti. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Analytical Chemistry*, 89(19):10397–10406, 2017. ISSN 15206882. doi:10.1021/acs.analchem.7b02380. → page 8

Bridgit Crews, William R. Wikoff, Gary J. Patti, Hin-Koon Woo, Ewa Kalisiak, Johanna Heideker, and Gary Siuzdak. Variability Analysis of Human Plasma and Cerebral Spinal Fluid Reveals Statistical Significance of Changes in Mass Spectrometry-Based Metabolomics Data. *Analytical Chemistry*, 81(20): 8538–8544, 2009. ISSN 0003-2700. doi:10.1021/ac9014947. URL http://pubs.acs.org/doi/abs/10.1021/ac9014947. → page 8

Leonid Brodsky, Arieh Moussaieff, Nir Shahaf, Asaph Aharoni, and Ilana Rogachev. Evaluation of peak picking quality in LC-MS metabolomics data. *Analytical Chemistry*, 82(22):9177–9187, 2010. ISSN 00032700. doi:10.1021/ac101216e. → page 8

Joanna Godzien, Vanesa Alonso-Herranz, Coral Barbas, and Emily Grace Armitage. Controlling the quality of metabolomics data: new strategies to get the best out of the QC sample. *Metabolomics*, 11(3):518–528, 2015. ISSN 15733890. doi:10.1007/s11306-014-0712-4. → pages 8, 9

Dirk Vaikenborg, Grégoire Thomas, Luc Krois, Koen Kas, and Tomasz Burzykowski. A strategy for the prior processing of high-resolution mass

spectral data obtained from high-dimensional Combined fractional diagonal chromatography. *Journal of Mass Spectrometry*, 44(4):516–529, 2009. ISSN 10765174. doi:10.1002/jms.1527. → page 9

Xiaotao Shen, Xiaoyun Gong, Yuping Cai, Yuan Guo, Jia Tu, Hao Li, Tao Zhang, Jialin Wang, Fuzhong Xue, and Zheng Jiang Zhu. Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics*, 12(5):1–12, 2016. ISSN 15733890. doi:10.1007/s11306-016-1026-5. → page 9

Alysha M. De Livera, Marko Sysi-Aho, Laurent Jacob, Johann A. Gagnon-Bartsch, Sandra Castillo, Julie A. Simpson, and Terence P. Speed. Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Analytical Chemistry*, 87(7):3606–3615, 2015. ISSN 15206882. doi:10.1021/ac502439y. → page 9

Yuliya V Karpievitch, Alan R Dabney, and Richard D Smith. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, 13(Suppl 16):S5, 2012. ISSN 1471-2105. doi:10.1186/1471-2105-13-S16-S5. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S16-S5. → page 9

Yiman Wu and Liang Li. Sample normalization methods in quantitative metabolomics, 2015. ISSN 18733778. → page 10

B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *BIOINFORMATICS*, 19(2):185–193, 2003. ISSN 1367-4803. doi:10.1093/bioinformatics/19.2.185. URL http://www.stat.berkeley.edu/bolstad/normalize/. → pages 10, 28

Robert a van den Berg, Huub C J Hoefsloot, Johan a Westerhuis, Age K Smilde, and Mariët J van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7: 142, 2006. ISSN 1471-2164. doi:10.1186/1471-2164-7-142. URL http://www.ncbi.nlm.nih.gov/pubmed/16762068. → page 10

Svati H. Shah and Christopher B. Newgard. Integrated Metabolomics and Genomics: Systems Approaches to Biomarkers and Mechanisms of Cardiovascular Disease. *Circulation: Cardiovascular Genetics*, 8(2):410–419, 2015. ISSN 19423268. doi:10.1161/CIRCGENETICS.114.000223. → page 11

Eugene P Rhee, Jennifer E Ho, Ming-Huei Chen, Dongxiao Shen, Susan Cheng, Martin G Larson, Anahita Ghorbani, Xu Shi, Iiro T Helenius, Christopher J O'Donnell, Amanda L Souza, Amy Deik, Kerry A Pierce, Kevin Bullock, Geoffrey A Walford, Ramachandran S Vasan, Jose C Florez, Clary Clish, J.-R. Joanna Yeh, Thomas J Wang, and Robert E Gerszten. A Genome-wide Association Study of the Human Metabolome in a Community-Based Cohort. *Cell Metabolism*, 18(1):130–143, 2013. ISSN 15504131. doi:10.1016/j.cmet.2013.06.013. → pages 11, 14

Tao Long, Michael Hicks, Hung-Chun Yu, William H Biggs, Ewen F Kirkness, Cristina Menni, Jonas Zierer, Kerrin S Small, Massimo Mangino, Helen Messier, Suzanne Brewerton, Yaron Turpaz, Brad A Perkins, Anne M Evans, Luke A D Miller, Lining Guo, C Thomas Caskey, Nicholas J Schork, Chad Garner, Tim D Spector, J Craig Venter, and Amalio Telenti. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature genetics*, 49(4):568–578, 2017. ISSN 1546-1718. doi:10.1038/ng.3809. URL http://www.nature.com/doifinder/10.1038/ng.3809{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/28263315. → pages 11, 14, 15

Bernd O. Keller, Jie Sui, Alex B. Young, and Randy M. Whittal. Interferences and contaminants encountered in modern mass spectrometry, 2008. ISSN 00032670. → page 11

Carsten Kuhl, Ralf Tautenhahn, Christoph Böttcher, Tony R. Larson, and Steffen Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289, 2012. ISSN 00032700. doi:10.1021/ac202450g. → pages 12, 28

Ralf Petri and Claudia Schmidt-Dannert. BioCyc—Genome and Metabolism. *Angew. Chem. Int. Ed.*, 43(15):1908, 2004. ISSN 1433-7851. doi:10.1002/anie.200483067. URL http://doi.wiley.com/10.1002/anie.200483067. → page 12

A Smith, Grace. O 'maille, Elizabeth. J. Want, Chuan. Qin, Sunia. A. Trauger, Theodore. R. Brandon, Darlene. E. Custodio, Ruben. Abagyan, and Gary. Siuzdak. METLIN A Metabolite Mass Spectral Database. *Proceedings of the 9Th International Congress of Therapeutic Drug Monitoring & Clinical Toxicology*, 27(6):747–751, 2005. ISSN 0163-4356. doi:10.1097/01.ftd.0000179845.53213.39. → page 12

David S. Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie Aude Coutouly, Ian Forsythe, Peter Tang, Savita Shrivastava, Kevin Jeroncic, Paul Stothard, Godwin Amegbey, David Block, David D. Hau, James Wagner, Jessica Miniaci, Melisa Clements, Mulu Gebremedhin, Natalie Guo, Ying Zhang, Gavin E. Duggan, Glen D. MacInnis, Alim M. Weljie, Reza Dowlatabadi, Fiona Bamforth, Derrick Clive, Russ Greiner, Liang Li, Tom Marrie, Brian D. Sykes, Hans J. Vogel, and Lori Querengesser. HMDB: The human metabolome database. *Nucleic Acids Research*, 35(SUPPL. 1), 2007. ISSN 03051048. doi:10.1093/nar/gkl923. → pages 12, 29, 30

Ines Thiele, Almut Heinken, and Ronan M T Fleming. A systems biology approach to studying the role of microbes in human health. *Current Opinion in Biotechnology*, 24(1):4–12, 2013. ISSN 09581669. doi:10.1016/j.copbio.2012.10.001. URL http://dx.doi.org/10.1016/j.copbio.2012.10.001. → page 12

S Li, Y Park, S Duraisingham, F H Strobel, N Khan, Q A Soltow, D P Jones, and B Pulendran. Predicting network activity from high throughput metabolomics. *PLoS Comput Biol*, 9(7):e1003123, 2013. ISSN 1553-7358. doi:10.1371/journal.pcbi.1003123. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701697/pdf/pcbi.1003123.pdf. → pages 12, 48

Leila Pirhaji, Pamela Milani, Mathias Leidl, Timothy Curran, Julian Avila-pacheco, Clary B Clish, Forest M White, Alan Saghatelian, and Ernest Fraenkel. Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nature biotechnology*, (October 2015), 2016. doi:10.1038/nmeth.3940. → pages 12, 48, 50

M Sysi-Aho, M Katajamaa, L Yetukuri, and M Oresic. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, 8:93, 2007. ISSN 1471-2105. doi:10.1186/1471-2105-8-93. URL http://www.ncbi.nlm.nih.gov/pubmed/17362505. → page 13

Bedilu Alamirie Ejigu, Dirk Valkenborg, Geert Baggerman, Manu Vanaerschot, Erwin Witters, Jean-Claude Dujardin, Tomasz Burzykowski, and Maya Berg. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. *Omics : a journal of integrative biology*, 17(9):473–85, 2013. ISSN 1557-8100.

doi:10.1089/omi.2013.0010. URL http://www.ncbi.nlm.nih.gov/pubmed/23808607. → page 13

Daniel Weindl, André Wegner, Christian Jäger, and Karsten Hiller. Isotopologue ratio normalization for non-targeted metabolomics. *Journal of Chromatography A*, 1389:112–119, 2015. ISSN 18733778. doi:10.1016/j.chroma.2015.02.025. → page 13

William R. Wikoff, Jon A. Gangoiti, Bruce A. Barshop, and Gary Siuzdak. Metabolomics identifies perturbations in human disorders of propionate metabolism. *Clinical Chemistry*, 53(12):2169–2176, 2007. ISSN 00099147. doi:10.1373/clinchem.2007.089011. → page 13

Marli Dercksen, Gerhard Koekemoer, Marinus Duran, Ronald J A Wanders, Lodewyk J. Mienie, and Carolus J. Reinecke. Organic acid profile of isovaleric acidemia: A comprehensive metabolomics approach. *Metabolomics*, 9(4): 765–777, 2013. ISSN 15733882. doi:10.1007/s11306-013-0501-5. → page 13

Leonie Venter, Zander Lindeque, Peet Jansen van Rensburg, Francois van der Westhuizen, Izelle Smuts, and Roan Louw. Untargeted urine metabolomics reveals a biosignature for muscle respiratory chain deficiencies. *Metabolomics*, 11(1):111–121, 2014. ISSN 15733890. doi:10.1007/s11306-014-0675-5. → page 13

Lukáš Najdekr, Alžběta Gardlo, Lucie Mádrová, David Friedecký, Hana Janečková, Elon S. Correa, Royston Goodacre, and Tomáš Adam. Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-CoA dehydrogenase deficiency. *Talanta*, 139:62–66, 2015. ISSN 00399140. doi:10.1016/j.talanta.2015.02.041. → page 13

Abela L., Steindl K., Simmons L., Joset P., Papuc M., Mathis D., Schmitt B., Wohlrab G., Klein A., Asadollahi R., Crowther L., Sass O., Hersberger M., and Rauch A. A combined metabolic-genetic approach to early-onset epileptic encephalopathies: Results from a Swiss study cohort, 2016. URL http://ovidsp.ovid.com/ovidweb.cgi?T=JS{&}PAGE=reference{&}D=emex{&}NEWS=N{&}AN=615322759. → pages 13, 15

Adam D. Kennedy, Kirk L. Pappan, Taraka R. Donti, Anne M. Evans, Jacob E. Wulff, Luke A.D. Miller, V. Reid Sutton, Qin Sun, Marcus J. Miller, and Sarah H. Elsea. Elucidation of the complex metabolic profile of cerebrospinal fluid using an untargeted biochemical profiling assay. *Molecular Genetics and Metabolism*, 121(2):83–90, 2017. ISSN 10967206. doi:10.1016/j.ymgme.2017.04.005. → pages 13, 49

Kirk L. Pappan, Adam D. Kennedy, Pilar Magoulas, Neil A. Hanchard, Qin Sun, and Sarah H. Elsea. Clinical Metabolomics to Segregate Aromatic Amino Acid Decarboxylase Deficiency From Drug-Induced Metabolite Elevations. *Pediatric Neurology*, 2017. ISSN 08878994. doi:10.1016/j.pediatrneurol.2017.06.014. URL http://linkinghub.elsevier.com/retrieve/pii/S0887899417304836. → pages 13, 15

Marcus J. Miller, Adam D. Kennedy, Andrea D. Eckhart, Lindsay C. Burrage, Jacob E. Wulff, Luke A D Miller, Michael V. Milburn, John A. Ryals, Arthur L. Beaudet, Qin Sun, V. Reid Sutton, and Sarah H. Elsea. Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. *Journal of Inherited Metabolic Disease*, 38(6):1029–1039, 2015. ISSN 15732665. doi:10.1007/s10545-015-9843-7. → page 13

Karlien L.M. Coene, Leo A.J. Kluijtmans, Ed van der Heeft, Udo F.H. Engelke, Siebolt de Boer, Brechtje Hoegen, Hanneke J.T. Kwast, Maartje van de Vorst, Marleen C.D.G. Huigen, Irene M.L.W. Keularts, Michiel F. Schreuder, Clara D.M. van Karnebeek, Saskia B. Wortmann, Maaike C. de Vries, Mirian C.H. Janssen, Christian Gilissen, Jasper Engel, and Ron A. Wevers. Next-generation metabolic screening: targeted and untargeted metabolomics for the diagnosis of inborn errors of metabolism in individual patients. *Journal of Inherited Metabolic Disease*, 2018. ISSN 15732665. doi:10.1007/s10545-017-0131-6. → pages 13, 28, 50

Christian Gieger, Ludwig Geistlinger, Elisabeth Altmaier, Martin Hrabé De Angelis, Florian Kronenberg, Thomas Meitinger, Hans Werner Mewes, H. Erich Wichmann, Klaus M. Weinberger, Jerzy Adamski, Thomas Illig, and Karsten Suhre. Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. *PLoS Genetics*, 4(11), 2008. ISSN 15537390. doi:10.1371/journal.pgen.1000282. → page 14

Andrew A. Hicks, Peter P. Pramstaller, Åsa Johansson, Veronique Vitart, Igor Rudan, Peter Ugocsai, Yurii Aulchenko, Christopher S. Franklin, Gerhard Liebisch, Jeanette Erdmann, Inger Jonasson, Irina V. Zorkoltseva, Cristian Pattaro, Caroline Hayward, Aaron Isaacs, Christian Hengstenberg, Susan Campbell, Carsten Gnewuch, A. Cecile J.W. Janssens, Anatoly V. Kirichenko, Inke R. König, Fabio Marroni, Ozren Polasek, Ayse Demirkan, Ivana Kolcic, Christine Schwienbacher, Wilmar Igl, Zrinka Biloglav, Jacqueline C.M. Witteman, Irene Pichler, Ghazal Zaboli, Tatiana I. Axenovich, Annette Peters, Stefan Schreiber, H. Erich Wichmann, Heribert Schunkert, Nick Hastie, Ben A.

Oostra, Sarah H. Wild, Thomas Meitinger, Ulf Gyllensten, Cornelia M. Van Duijn, James F. Wilson, Alan Wright, Gerd Schmitz, and Harry Campbell. Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS Genetics*, 5(10), 2009. ISSN 15537390. doi:10.1371/journal.pgen.1000672. → page 14

Thomas Illig, Christian Gieger, Guangju Zhai, Werner Römisch-Margl, Rui Wang-Sattler, Cornelia Prehn, Elisabeth Altmaier, Gabi Kastenmüller, Bernet S Kato, Hans-Werner Mewes, Thomas Meitinger, Martin Hrabé de Angelis, Florian Kronenberg, Nicole Soranzo, H Erich Wichmann, Tim D Spector, Jerzy Adamski, and Karsten Suhre. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*, 42(2):137–141, 2010. ISSN 1546-1718. doi:ng.507[pii]10.1038/ng.507. URL http://www.pubmed.org/20037589{%}5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC3773904/. → page 14

Karsten Suhre, Henri Wallaschofski, Johannes Raffler, Nele Friedrich, Robin Haring, Kathrin Michael, Christina Wasner, Alexander Krebs, Florian Kronenberg, David Chang, Christa Meisinger, H. Erich Wichmann, Wolfgang Hoffmann, Henry Völzke, Uwe Völker, Alexander Teumer, Reiner Biffar, Thomas Kocher, Stephan B. Felix, Thomas Illig, Heyo K. Kroemer, Christian Gieger, Werner Römisch-Margl, and Matthias Nauck. A genome-wide association study of metabolic traits in human urine. *Nature Genetics*, 43(6): 565–569, 2011. ISSN 10614036. doi:10.1038/ng.837. → page 14

Aye Demirkan, Cornelia M. van Duijn, Peter Ugocsai, Aaron Isaacs, Peter P. Pramstaller, Gerhard Liebisch, James F. Wilson, Åsa Johansson, Igor Rudan, Yurii S. Aulchenko, Anatoly V. Kirichenko, A. Cecile J.W. Janssens, Ritsert C. Jansen, Carsten Gnewuch, Francisco S. Domingues, Cristian Pattaro, Sarah H. Wild, Inger Jonasson, Ozren Polasek, Irina V. Zorkoltseva, Albert Hofman, Lennart C. Karssen, Maksim Struchalin, James Floyd, Wilmar Igl, Zrinka Biloglav, Linda Broer, Arne Pfeufer, Irene Pichler, Susan Campbell, Ghazal Zaboli, Ivana Kolcic, Fernando Rivadeneira, Jennifer Huffman, Nicholas D. Hastie, Andre Uitterlinden, Lude Franke, Christopher S. Franklin, Veronique Vitart, Christopher P. Nelson, Michael Preuss, Joshua C. Bis, Christopher J. O'Donnell, Nora Franceschini, Jacqueline C.M. Witteman, Tatiana Axenovich, Ben A. Oostra, Thomas Meitinger, Andrew A. Hicks, Caroline Hayward, Alan F. Wright, Ulf Gyllensten, Harry Campbell, and Gerd Schmitz. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genetics*, 8(2), 2012. ISSN 15537390. doi:10.1371/journal.pgen.1002490. → page 14

Taru Tukiainen, Johannes Kettunen, Antti J. Kangas, Leo Pekka Lyytikinen, Pasi Soininen, Antti Pekka Sarin, Emmi Tikkanen, Paul F. O'reilly, Markku J. Savolainen, Kimmo Kaski, Anneli Pouta, Antti Jula, Terho Lehtimki, Mika Knen, Jorma Viikari, Marja Riitta Taskinen, Matti Jauhiainen, Johan G. Eriksson, Olli Raitakari, Veikko Salomaa, Marjo Riitta Jrvelin, Markus Perola, Aarno Palotie, Mika Ala-korpela, and Samuli Ripatti. Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Human Molecular Genetics*, 21(6):1444–1455, 2012. ISSN 09646906. doi:10.1093/hmg/ddr581. → page 14

Johannes Kettunen, Taru Tukiainen, Antti-Pekka Sarin, Alfredo Ortega-Alonso, Emmi Tikkanen, Leo-Pekka Lyytikäinen, Antti J Kangas, Pasi Soininen, Peter Würtz, Kaisa Silander, Danielle M Dick, Richard J Rose, Markku J Savolainen, Jorma Viikari, Mika Kähönen, Terho Lehtimäki, Kirsi H Pietiläinen, Michael Inouye, Mark I McCarthy, Antti Jula, Johan Eriksson, Olli T Raitakari, Veikko Salomaa, Jaakko Kaprio, Marjo-Riitta Järvelin, Leena Peltonen, Markus Perola, Nelson B Freimer, Mika Ala-Korpela, Aarno Palotie, and Samuli Ripatti. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*, 44(3):269–76, 2012. ISSN 1546-1718. doi:10.1038/ng.1073. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605033{&}tool=pmcentrez{&}rendertype=abstract{%}5Cnhttp://dx.doi.org/10.1038/ng.1073. → page 14

So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M Valdes, Craig L Hyde, Vicky Wang, Daniel Ziemek, Phoebe Roberts, Li Xi, Elin Grundberg, Melanie Waldenberger, J Brent Richards, Robert P Mohney, Michael V Milburn, Sally L John, Jeff Trimmer, Fabian J Theis, John P Overington, Karsten Suhre, M Julia Brosnan, Christian Gieger, Gabi Kastenm?ller, Tim D Spector, and Nicole Soranzo. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550, 2014. ISSN 1061-4036. doi:10.1038/ng.2982. URL http://www.nature.com/ng/journal/v46/n6/full/ng.2982.html{%}5Cnhttp://www.nature.com/doifinder/10.1038/ng.2982. → pages 14, 15

Harmen H. M. Draisma, René Pool, Michael Kobl, Rick Jansen, Ann-Kristin Petersen, Anika A. M. Vaarhorst, Idil Yet, Toomas Haller, Aye Demirkan, Tõnu Esko, Gu Zhu, Stefan Böhringer, Marian Beekman, Jan Bert van Klinken, Werner Römisch-Margl, Cornelia Prehn, Jerzy Adamski, Anton J. M. de Craen,

Elisabeth M. van Leeuwen, Najaf Amin, Harish Dharuri, Harm-Jan Westra, Lude Franke, Eco J. C. de Geus, Jouke Jan Hottenga, Gonneke Willemsen, Anjali K. Henders, Grant W. Montgomery, Dale R. Nyholt, John B. Whitfield, Brenda W. Penninx, Tim D. Spector, Andres Metspalu, P. Eline Slagboom, Ko Willems van Dijk, Peter A. C. t Hoen, Konstantin Strauch, Nicholas G. Martin, Gert-Jan B. van Ommen, Thomas Illig, Jordana T. Bell, Massimo Mangino, Karsten Suhre, Mark I. McCarthy, Christian Gieger, Aaron Isaacs, Cornelia M. van Duijn, and Dorret I. Boomsma. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature Communications*, 6:7208, 2015. ISSN 2041-1723. doi:10.1038/ncomms8208. URL http://www.nature.com/doifinder/10.1038/ncomms8208. → page 14

Eugene P. Rhee, Qiong Yang, Bing Yu, Xuan Liu, Susan Cheng, Amy Deik, Kerry A. Pierce, Kevin Bullock, Jennifer E. Ho, Daniel Levy, Jose C. Florez, Sek Kathiresan, Martin G. Larson, Ramachandran S. Vasan, Clary B. Clish, Thomas J. Wang, Eric Boerwinkle, Christopher J. O'Donnell, and Robert E. Gerszten. An exome array study of the plasma metabolome. *Nature Communications*, 7:12360, 2016. ISSN 2041-1723. doi:10.1038/ncomms12360. URL http://www.nature.com/doifinder/10.1038/ncomms12360. → page 14

Lining Guo, Michael V Milburn, John A Ryals, Shaun C Lonergan, Matthew W Mitchell, Jacob E Wulff, Danny C Alexander, Anne M Evans, Brandi Bridgewater, Luke Miller, Manuel L. Gonzalez-Garay, and C Thomas Caskey. Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. *Proceedings of the National Academy of Sciences*, 112(35): E4901–E4910, 2015. ISSN 0027-8424. doi:10.1073/pnas.1508425112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1508425112. → pages 14, 15

Akram Yazdani, Azam Yazdani, Xiaoming Liu, and Eric Boerwinkle. Identification of Rare Variants in Metabolites of the Carnitine Pathway by Whole Genome Sequencing Analysis. *Genetic Epidemiology*, 40(6):486–491, 2016. ISSN 10982272. doi:10.1002/gepi.21980. → page 14

B. Yu, A. H. Li, G. A. Metcalf, D. M. Muzny, A. C. Morrison, S. White, T. H. Mosley, R. A. Gibbs, and E. Boerwinkle. Loss-of-function variants influence the human serum metabolome. *Science Advances*, 2(8):e1600800–e1600800, 2016. ISSN 2375-2548. doi:10.1126/sciadv.1600800. URL http://advances.sciencemag.org/cgi/doi/10.1126/sciadv.1600800. → page 14

R Gauba, T G Natarajan, L Song, K Bhuvaneshwar, S Madhavan, and Y Gusev. Metabolomic and exome sequence analysis reveal novel molecular signatures associated with colorectal cancer relapse. *BMC Proceedings*, Conference: Beyond the Genome 2012 Boston, MA United States. C, 2012. URL http://ovidsp.ovid.com/ovidweb.cgi?T=JS{&}CSC=Y{&}NEWS=N{&}PAGE=fulltext{&}D=emed12{&}AN=71478275{%}5Cnhttp://imp-primo.hosted.exlibrisgroup.com/openurl/44IMP/44IMP{_}services{_}page?sid=OVID{&}isbn={&}issn=1753-6561{&}volume=6{&}issue={&}date=2012{&}title=BMC+Proceedings{&}atitle=Met. → pages 14, 15

Edward M. Marcotte, Matteo Pellegrini, Michael J. Thompson, Todd O. Yeates, and David Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 1999. ISSN 00280836. doi:10.1038/47048. → page 18

Euan A. Adie, Richard R. Adams, Kathryn L. Evans, David J. Porteous, and Ben S. Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 2005. ISSN 14712105. doi:10.1186/1471-2105-6-55. → page 18

Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, Peter Carmeliet, and Yves Moreau. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 2006. ISSN 10870156. doi:10.1038/nbt1203. → page 18

Marc a van Driel, Koen Cuelenaere, Patrick P C W Kemmeren, Jack a M Leunissen, and Han G Brunner. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *European journal of human genetics : EJHG*, 2003. ISSN 1018-4813. doi:10.1038/sj.ejhg.5200918. → page 18

Tijl De Bie, Léon Charles Tranchevent, Liesbeth M.M. van Oeffelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. In *Bioinformatics*, 2007. ISBN 1367-4811 (Electronic)\r1367-4803 (Linking). doi:10.1093/bioinformatics/btm187. → page 18

Martin Oti, Martijn A. Huynen, and Han G. Brunner. Phenome connections, 2008. ISSN 01689525. → page 18

Ugo Ala, Rosario Michael Piro, Elena Grassi, Christian Damasco, Lorenzo Silengo, Martin Oti, Paolo Provero, and Ferdinando Di Cunto. Prediction of

human disease genes by human-mouse conserved coexpression analysis. *PLoS Computational Biology*, 2008. ISSN 1553734X. doi:10.1371/journal.pcbi.1000043. → page 18

J. Freudenberg and P. Propping. A similarity-based method for genome-wide prediction of disease-relevant human genes. In *Bioinformatics*, 2002. ISBN 1367-4803 (Print). doi:10.1093/bioinformatics/18.suppl_2.S110. → page 18

Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 2002. ISSN 10614036. doi:10.1038/ng895. → page 18

Frances S Turner, Daniel R Clutterbuck, and Colin A M Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome biology*, 2003. ISSN 1474-760X. doi:10.1186/gb-2003-4-11-r75. → page 18

Yongjin Li and Jagdish C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 2010a. ISSN 13674803. doi:10.1093/bioinformatics/btq108. → pages 18, 19

M. Oti and H. G. Brunner. The modular nature of genetic diseases, 2007. ISSN 00099163. → page 18

Hunter B. Fraser and Joshua B. Plotkin. Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biology*, 2007. ISSN 14747596. doi:10.1186/gb-2007-8-11-r252. → page 18

Shao Li, Lijiang Wu, and Zhongqi Zhang. Constructing biological networks through combined literature mining and microarray analysis: A LMMA approach. *Bioinformatics*, 2006. ISSN 13674803. doi:10.1093/bioinformatics/btl363. → page 18

Kyle J. Gaulton, Karen L. Mohlke, and Todd J. Vision. A computational system to select candidate genes for complex human traits. *Bioinformatics*, 2007. ISSN 13674803. doi:10.1093/bioinformatics/btm001. → page 18

Chad L. Myers, Drew Robson, Adam Wible, Matthew A. Hibbs, Camelia Chiriac, Chandra L. Theesfeld, Kara Dolinski, and Olga G. Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome Biology*, 2005. ISSN 1474760X. doi:10.1186/gb-2005-6-13-r114. → pages 18, 19

Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. GeneMANIA: A real-time multiple association network integration

algorithm for predicting gene function. *Genome Biology*, 9(SUPPL. 1):1–15, 2008. ISSN 14747596. doi:10.1186/gb-2008-9-s1-s4. → pages 18, 19

Koji Tsuda, HyunJung Shin, and Bernhard Schölkopf. Fast protein classification with multiple networks. *Bioinformatics (Oxford, England)*, 2005. ISSN 1367-4811. doi:10.1093/bioinformatics/bti1110. → page 18

Minghua Deng, Ting Chen, and Fengzhu Sun. An Integrated Probabilistic Model for Functional Prediction of Proteins. *Journal of Computational Biology*, 2004. ISSN 1066-5277. doi:10.1089/1066527041410346. → page 18

Sara Mostafavi and Quaid Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, 2010. ISSN 13674803. doi:10.1093/bioinformatics/btq262. → page 18

Gert R G Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 2004. ISSN 13674803. doi:10.1093/bioinformatics/bth294. → page 18

Lourdes Peña-Castillo, Murat Tasan, Chad L. Myers, Hyunju Lee, Trupti Joshi, Chao Zhang, Yuanfang Guan, Michele Leone, Andrea Pagnani, Wan Kyu Kim, Chase Krumpelman, Weidong Tian, Guillaume Obozinski, Yanjun Qi, Sara Mostafavi, Guan Ning Lin, Gabriel F. Berriz, Francis D. Gibbons, Gert Lanckriet, Jian Qiu, Charles Grant, Zafer Barutcuoglu, David P. Hill, David Warde-Farley, Chris Grouios, Debajyoti Ray, Judith A. Blake, Minghua Deng, Michael I. Jordan, William S. Noble, Quaid Morris, Judith Klein-Seetharaman, Ziv Bar-Joseph, Ting Chen, Fengzhu Sun, Olga G. Troyanskaya, Edward M. Marcotte, Dong Xu, Timothy R. Hughes, and Frederick P. Roth. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biology*, 2008. ISSN 14747596. doi:10.1186/gb-2008-9-s1-s2. → pages 18, 19

Paul Pavlidis, Jason Weston, Jinsong Cai, and William Stafford Noble. Learning Gene Functional Classi cations from Multiple Data Types. *JOURNAL OF COMPUTATIONAL BIOLOGY*, 2002. → page 18

Daniel Lancour, Adam Naj, Richard Mayeux, Jonathan L. Haines, Margaret A. Pericak-Vance, Gerard C. Schellenberg, Mark Crovella, Lindsay A. Farrer, and Simon Kasif. One for all and all for One: Improving replication of genetic studies through network diffusion. *PLoS Genetics*, 2018. ISSN 15537404. doi:10.1371/journal.pgen.1007306. → pages 18, 19, 21

R Sharan, I Ulitsky, and R Shamir. Network-based prediction of protein function. *Molecular systems biology*, 2007. ISSN 1744-4292. doi:10.1038/msb4100129. → page 18

J. Weston, A. Elisseeff, D. Zhou, C. S. Leslie, and W. S. Noble. Protein ranking: From local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences*, 2004. ISSN 0027-8424. doi:10.1073/pnas.0308067101. → pages 18, 31

Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics (Oxford, England)*, 2006. ISSN 1367-4811. doi:10.1093/bioinformatics/btl467. → page 18

Lude Franke, Harm van Bakel, Like Fokkens, Edwin D. de Jong, Michael Egmont-Petersen, and Cisca Wijmenga. Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes. *The American Journal of Human Genetics*, 2006. ISSN 00029297. doi:10.1086/504300. → page 18

Kasper Lage, E. Olof Karlberg, Zenia M. Størling, Páll Í Ólason, Anders G. Pedersen, Olga Rigina, Anders M. Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, Yves Moreau, and Søren Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 2007. ISSN 10870156. doi:10.1038/nbt1295. → page 18

Xuebing Wu, Rui Jiang, Michael Q. Zhang, and Shao Li. Network-based global inference of human disease genes. *Molecular Systems Biology*, 2008. ISSN 17444292. doi:10.1038/msb.2008.27. → pages 18, 19

Christian von Mering, Lars J Jensen, Michael Kuhn, Samuel Chaffron, Tobias Doerks, Beate Krüger, Berend Snel, and Peer Bork. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic acids research*, 2007. ISSN 1362-4962. doi:10.1093/nar/gkl825. → page 19

Insuk Lee, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 2011. ISSN 10889051. doi:10.1101/gr.118992.110. → pages 19, 21, 44

Y. Itan, S.-Y. Zhang, G. Vogt, A. Abhyankar, M. Herman, P. Nitschke, D. Fried, L. Quintana-Murci, L. Abel, and J.-L. Casanova. The human gene connectome as a map of short cuts for morbid allele discovery. *Proceedings of the National*

Academy of Sciences, 2013. ISSN 0027-8424. doi:10.1073/pnas.1218167110. → page 19

Casey S. Greene, Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, Boris M. Hartmann, Elena Zaslavsky, Stuart C. Sealfon, Daniel I. Chasman, Garret A. Fitzgerald, Kara Dolinski, Tilo Grosser, and Olga G. Troyanskaya. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 2015. ISSN 15461718. doi:10.1038/ng.3259. → page 19

Bolan Linghu, Evan S. Snitkin, Zhenjun Hu, Yu Xia, and Charles DeLisi. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biology*, 2009. ISSN 14747596. doi:10.1186/gb-2009-10-9-r91. → page 19

Yongjin Li and Jagdish C. Patra. Integration of multiple data sources to prioritize candidate genes using discounted rating system. *BMC Bioinformatics*, 2010b. ISSN 14712105. doi:10.1186/1471-2105-11-S1-S20. → page 19

Tak Lee and Insuk Lee. AraGWAB: Network-based boosting of genome-wide association studies in Arabidopsis thaliana. *Scientific Reports*, 2018. ISSN 20452322. doi:10.1038/s41598-018-21301-4. → page 19

Yu Qian, Søren Besenbacher, Thomas Mailund, and Mikkel Heide Schierup. Identifying disease associated genes by network propagation. *BMC Systems Biology*, 2014. ISSN 17520509. doi:10.1186/1752-0509-8-S1-S6. → pages 19, 21, 44

Justin K. Huang, Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo, and Trey Ideker. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell Systems*, 2018. ISSN 24054720. doi:10.1016/j.cels.2018.03.001. → page 22

Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, Luisa Castagnoli, and Gianni Cesareni. MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*, 2012. ISSN 03051048. doi:10.1093/nar/gkr930. → page 22

T S Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan,

Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C J Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y L Ramachandra, V Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic acids research*, 2009. ISSN 1362-4962. doi:10.1093/nar/gkn892. → page 22

C. Alfarano, C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F.F. Ouellette, and C. W.V. Hogue. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, 2005. ISSN 03051048. doi:10.1093/nar/gki051. → page 22

L. Salwinski. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 2004. ISSN 1362-4962. doi:10.1093/nar/gkh086. → page 22

Andrew Chatr-Aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby Joe Breitkreutz, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 2017. ISSN 13624962. doi:10.1093/nar/gkw1102. → page 22

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 2017. ISSN 13624962. doi:10.1093/nar/gkw1092. → page 22

Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter

D'Eustachio. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 2018. ISSN 13624962. doi:10.1093/nar/gkx1132. → page 22

S. Kerrien, Y. Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. IntActopen source resource for molecular interaction data. *Nucleic Acids Research*, 2007. ISSN 03051048. doi:10.1093/nar/gkl958. → page 22

Ingrid M. Keseler, Amanda Mackie, Martin Peralta-Gil, Alberto Santos-Zavaleta, Socorro Gama-Castro, César Bonavides-Martínez, Carol Fulcher, Araceli M. Huerta, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Luis Muñiz-Rascado, Quang Ong, Suzanne Paley, Imke Schröder, Alexander G. Shearer, Pallavi Subhraveti, Mike Travers, Deepika Weerasinghe, Verena Weiss, Julio Collado-Vides, Robert P. Gunsalus, Ian Paulsen, and Peter D. Karp. EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Research*, 2013. ISSN 03051048. doi:10.1093/nar/gks1027. → page 22

L J Jensen, M Kuhn, M Stark, S Chaffron, C Creevey, J Muller, T Doerks, P Julien, A Roth, M Simonovic, P Bork, and C von Mering. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 2009. ISSN 03051048. doi:10.1093/nar/gkn760. → page 22

J. Michael Cherry, Caroline Adler, Catherine Ball, Stephen A. Chervitz, Selina S. Dwight, Erich T. Hester, Yankai Jia, Gail Juvik, Taiyun Roe, Mark Schroeder, Shuai Weng, and David Botstein. SGD: Saccharomyces genome database. *Nucleic Acids Research*, 1998. ISSN 03051048. doi:10.1093/nar/26.1.73. → page 23

Joanna S. Amberger and Ada Hamosh. Searching online mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic phenotypes. *Current Protocols in Bioinformatics*, 2017. ISSN 1934340X. doi:10.1002/cpbi.27. → page 23

Han Yan, Kavitha Venkatesan, John E. Beaver, Niels Klitgord, Muhammed A. Yildirim, Tong Hao, David E. Hill, Michael E. Cusick, Norbert Perrimon, Frederick P. Roth, and Marc Vidal. A genome-wide gene function prediction resource for Drosophila melanogaster. *PLoS ONE*, 2010. ISSN 19326203. doi:10.1371/journal.pone.0012139. → page 23

Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O'roak, Gregory M. Cooper, and Jay Shendure. A general framework for estimating the relative

pathogenicity of human genetic variants. *Nature Genetics*, 2014. ISSN 15461718. doi:10.1038/ng.2892. → page 27

Gunnar Libiseller, Michaela Dvorzak, Ulrike Kleb, Edgar Gander, Tobias Eisenberg, Frank Madeo, Steffen Neumann, Gert Trausinger, Frank Sinner, Thomas Pieber, and Christoph Magnes. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics*, 16(1):118, 2015. ISSN 1471-2105. doi:10.1186/s12859-015-0562-8. URL http://www.biomedcentral.com/1471-2105/16/118. → page 28

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Sch. Learning with Local and Global Consistency. *Advances in Neural Information Processing Systems (NIPS)*, 2003. ISSN 00664162. doi:c. → page 31

Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5(1):21, 2011. ISSN 1752-0509. doi:10.1186/1752-0509-5-21. URL http://www.biomedcentral.com/1752-0509/5/21. → page 48

Jörg Martin Büscher, Dominika Czernik, Jennifer Christina Ewald, Uwe Sauer, and Nicola Zamboni. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Analytical Chemistry*, 81(6):2135–2143, 2009. ISSN 00032700. doi:10.1021/ac8022857. → page 49

Idil Yet, Cristina Menni, So Youn Shin, Massimo Mangino, Nicole Soranzo, Jerzy Adamski, Karsten Suhre, Tim D. Spector, Gabi Kastenmüller, and Jordana T. Bell. Genetic influences on metabolite levels: A comparison across metabolomic platforms. *PLoS ONE*, 11(4), 2016. ISSN 19326203. doi:10.1371/journal.pone.0153672. → page 49

Patrick Leuthold, Elke Schaeffeler, Stefan Winter, Florian Büttner, Ute Hofmann, Thomas E. Mürdter, Steffen Rausch, Denise Sonntag, Judith Wahrheit, Falko Fend, Jörg Hennenlotter, Jens Bedke, Matthias Schwab, and Mathias Haag. Comprehensive Metabolomic and Lipidomic Profiling of Human Kidney Tissue: A Platform Comparison. *Journal of Proteome Research*, 16(2): 933–944, 2017. ISSN 15353907. doi:10.1021/acs.jproteome.6b00875. → page 49

George Nicholson, Mattias Rantalainen, Anthony D. Maher, Jia V. Li, Daniel Malmodin, Kourosh R. Ahmadi, Johan H. Faber, Ingileif B. Hallgrímsdóttir,

Amy Barrett, Henrik Toft, Maria Krestyaninova, Juris Viksna, Sudeshna Guha Neogi, Marc Emmanuel Dumas, Ugis Sarkans, Bernard W. Silverman, Peter Donnelly, Jeremy K. Nicholson, Maxine Allen, Krina T. Zondervan, John C. Lindon, Tim D. Spector, Mark I. McCarthy, Elaine Holmes, Dorrit Baunsgaard, and Chris C. Holmes. Human metabolic profiles are stably controlled by genetic and environmental variation. *Molecular Systems Biology*, 7, 2011. ISSN 17444292. doi:10.1038/msb.2011.57. → page 50

Maryam Lotfi Shahreza, Nasser Ghadiri, Seyed Rasoul Mousavi, Jaleh Varshosaz, and James R. Green. Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning. *Journal of Biomedical Informatics*, 2017. ISSN 15320464. doi:10.1016/j.jbi.2017.03.006. → page 51

Erhan Kenar, Holger Franken, Sara Forcisi, Kilian Wörmann, Hans-Ulrich Häring, Rainer Lehmann, Philippe Schmitt-Kopplin, Andreas Zell, and Oliver Kohlbacher. Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Molecular & Cellular Proteomics*, 13 (1):348–359, 2014. ISSN 1535-9484. doi:10.1074/mcp.M113.031278. URL http://www.mcponline.org/content/13/1/348.full{%}5Cnpapers3://publication/doi/10.1074/mcp.M113.031278. → page 51

Appendix A

Supporting Materials

See below table for variant and clinical information for each patient included in this study.

Patient number	Causative gene	Mode of inheritance	Publication ID and variant frequency (if available)	Variant information	Clinical description
1	CPT1A	Homozygous recessive, missense	PMID: 20696606, ClinVar: 65644, gnomAD: 3.246×10^{-5}	CPT1A: g.68548130G>A (p.Pro479Leu)	Mild hypoglycemia on fasting, osteogenesis imperfecta - like phenotype (unexplained), short stature, congenital anomalies and dysmorphisms explained by methotrexate exposure during pregnancy
2	NANS	Compound heterozygous, missense	PMID:27276562, PMID:27213289, ClinVar: 235191 , gnomAD: 0	NANS: g.100843203C>T (p.Arg237Cys), g.100840588T>C (p.Tyr188His), Transcript: ENST00000210444	Skeletal dysplasia, short stature and rhixomelia, neurodevelopmental arrest, progressive epileptic encephalopathy, dysmorphisms, congenital brain abnormalities and white matter lesions
3	SCN2A	de novo, splice donor variant	PMID:27276562 and PMID:26647175, ClinVar: NA, gnomAD:0	SCN2A: g.166188079+1 G>A, Transcript: ENST00000283256	Global developmental delay, seizures, ataxia, microcephaly, autism, abnormal CSF mono-amine neurometabolite profile

Patient number	Causative gene	Mode of inheritance	Publication ID and variant frequency (if available)	Variant information	Clinical description
4	DYRK1A	de novo, missense	PMID:27276562, ClinVar: NA, gnomAD:0	DYRK1A: g.38865404C>T (p.Ser346Phe), Transcript: ENST00000398960	Neurodevelopmental delay, intractable epilepsy, absence seizures, microcephaly, mild dysmorphisms, hypoglycorrhagia
5	KIF5C	de novo, missense	ClinVar: NA, gnomAD: NA	KIF5C: g.149818513G>A (p.Val101Met, p.Val333Met, p.Val238Met , p.Val50Met), Transcript: ENST00000435030	Seizures, behavioral and psychiatric abnormalities, aggression, low CSF MTHF (folate), mild cerebral atrophy, mild ataxia, mild dysmorphism

Patient number	Causative gene	Mode of inheritance	Publication ID and variant frequency (if available)	Variant information	Clinical description
6	CNKSR2	homozygous from one heterozygous parent, missense	ClinVar: NA, gnomAD: NA	CNKSR2: g.21581499C>T (p.Phe464Ser), Transcript: ENST00000543067	Autonomic crises in infancy with hypertension, tachycardia, bladder retention, bowel dysmotility, frequent infections/sepsis responding to choline therapy, low acetylcholine levels, progressive cholinergic failure, alzheimers type memory loss (on treatment with Donepezil), central apneas and on BiPAP at night
7	ECI1	compound heterozygous, missense	PMID: 7586637 ClinVar: NA, gnomAD: 0.0071	ECI1: g.2296927G>A (p.Thr17Met) and g.2290104 (p.Thr262Met), Transcript: ENST00000566379	Spasticity and dystonia, microcephaly and cataracts, elevated methylmalonic acid, elevated malonic acid, enlargement of the ventricles, MRI signal changes in the basal ganglia and cerebellar atrophy

Patient number	**Causative gene**	**Mode of inheritance**	**Publication ID and variant frequency (if available)**	**Variant information**	**Clinical description**
8	IDS and HAL	IDS: hemizygous, missense; HAL: compound heterozygous, splice donor	IDS: ClinVar: NA, gnomAD: 0; HAL: ClinVar: NA, gnomAD: 0.000599	IDS: g. 148571971G>A (p.Arg294Trp ,p.Arg83Trp), Transcript: ENST00000340855, HAL: g. 96371767A>G (p.Trp537Arg, p.Trp329Arg, p.Trp68Arg), ENST00000261208 and g.96374333C>T, Transcript: ENST00000261208	Early onset global developmental delay, short stature, dysmorphisms, coarse facial features, severe behavioral disturbances, elevated keratan-sulphate, developmental regression, elevated glycosaminoglycans

Patient number	Causative gene	Mode of inheritance	Publication ID and variant frequency (if available)	Variant information	Clinical description
9	CHRNA1 and DHFR	CHRNA1: de novo, missense; DHFR: de novo, missense	CHRNA1: ClinVar: NA, gnomAD: NA; DHFR: ClinVar: NA, gnomAD: NA	CHRNA1: g.175619063C>T (p.Ala167Thr/ p.Ala142Thr), Transcript: ENST00000261007; DHFR: g.79950270C>G (p.Gln13His), Transcript: ENST00000439211	Progressive global developmental delay and loss of skills, microcephaly, congenital hypotonia and wheelchair bound, dysmorphic features, severe feeding difficulties, growth retardation, demyelination on brain MRI scan, elevated lactates
10	ATP8A2	Homozygous recessive, missense	NA	ATP8A2: g.26402265G>A (p.Ala897Thr), Transcript ENST00000381655	Hypotonia, ataxia since age 18 months

Patient number	**Causative gene**	**Mode of inheritance**	**Publication ID and variant frequency (if available)**	**Variant information**	**Clinical description**
11	MYO5B	Compound heterozygous, missense	ClinVar: 0.00240, gnomAD: 0.001643	MYO5B: g.47506839G>A (p.Arg344His), g.47566678G>C (p.Glu49Gln)	Intellectual Disability, hyperkinetic movement disorder, sensorineural hearing loss, myopathy, malabsorption, failure to thrive, elevated urine threonine, serine and lysine, plasma amino acids suggesting lactic acidemia, elevated lactate

Patient number	**Causative gene**	**Mode of inheritance**	**Publication ID and variant frequency (if available)**	**Variant information**	**Clinical description**
12	MAST1 and KCNQ2	MAST1: de novo, missense; KCNQ2: de novo, in frame deletion	ClinVar: NA; gnomAD: NA	MAST1: g.12975745G>A (p.Asp497Asn), Transcript: ENST00000251472; KCNQ2: g. 62038511del-GAG (p.Phe701del, p.Phe683del, p.Phe670del, p.Phe673del, p.Phe709del), Transcript: ENST00000354587	Global developmental delay, microcephaly, hypotonia, failure to thrive, epilepsy and, delayed myelination

Patient number	**Causative gene**	**Mode of inheritance**	**Publication ID and variant frequency (if available)**	**Variant information**	**Clinical description**
13	VGLL4	homozygous recessive, missense	ClinVar: NA, gnomAD: 0.001073	VGLL4: g. 11600101G>T (p.Arg268Ser, p.Arg184Ser ,p.Arg209Ser ,p.Arg273Ser, p.Arg274Ser ,p.Arg188Ser), Transcript: ENST00000430365	Progressive dystonia, spasticity, query seizures, abnormalities of the neurotransmitters, intellectual disability, cerebral atrophy, low CSF HVA, 5HIAA, MTHF