Department of Genome-Oriented Bioinformatics

Technische Universität München

Master's Thesis in Bioinformatics

Analysis of rare mitochondrial disorders by combining phenotype, genotype and metabotype

Marie-Sophie Friedl

Technische Universität München

Department of Genome-Oriented Bioinformatics

Technische Universität München

Master's Thesis in Bioinformatics

Analysis of rare mitochondrial disorders by combining phenotype, genotype and metabotype

Analyse seltener mitochondrialer Erkrankungen durch Kombination von Phänotyp-, Genotyp- und Metabotypinformationen

Author: Marie-Sophie Friedl
Supervisor: Prof. Dr. Hans-Werner Mewes
Advisors: Dr. Gabi Kastenmüller,
Dr. Andreas Ruepp
Submitted: 16.08.2016

I confirm that this master's thesis is my own work and I have documented all sources and material used.

16.08.2016 _____

Marie-Sophie Friedl

Acknowledgements

First, I am very grateful to Prof. Dr. Hans-Werner Mewes for giving me the opportunity to realize my master's thesis at the Institute of Bioinformatics and Systems Biology at Helmholtz Zentrum München.

I would like to thank my supervisors Dr. Gabi Kastenmüller and Dr. Andreas Ruepp for their valuable advice, for the inspiring, helpful discussions and for reviewing the early versions of this text.

Further on, I wish to express my gratitude to the members of the metabolomics group Dr. Matthias Arnold, Jan-Dominik Quell, Dr. Johannes Raffler, Dr. Werner Römisch-Margl and Jonas Zierer, to the annotators Barbara Brauner, Irmtraud Dunger-Kaltenbach, Dr. Gisela Fobo, Dr. Goar Frischmann, Dr. Corinna Montrone, and to Tim Jeske from the NGS group. I highly appreciate that they shared their data and their scientific experience with me.

I also wish to thank Dr. Manar Aoun, Laura Kremer and Dr. Holger Prokisch from the Institute of Human Genetics who gave me a comprehensive introduction into the biology of mitochondrial disorders and who kindly provided me with patient data from the MitoNET register. I am also grateful to the patients from MitoNET whose data are the foundation of this thesis.

Finally, special thanks to my parents and Wolfgang. Without their support this thesis would never have been possible.

Abstract

Background: There are more than 7 000 thousand known individually rare diseases, which altogether impair the health and life quality of millions of people. About 80% of the rare diseases are caused by genetic defects. The diagnosis of rare diseases is often difficult due to the lack of awareness and knowledge. Rare mitochondrial disorders are a special kind of rare diseases that arise from defects of the energy metabolism. The correct diagnosis of rare mitochondrial diseases is especially challenging because of their high genetic and phenotypic variability. The recent emergence of next-generation sequencing in combination with computational methods improved the diagnostic process and lead to the identification of the causative genes for many rare genetic disorders.

Method: This master's thesis presents a KNIME-based pipeline named PheNoBo for predicting the causal gene of a patient suffering from any kind of genetic disorder. In contrast to previous genotype-centric methods, the predictions of PheNoBo are based on the patient's phenotype, metabotype and genotype. PheNoBo uses prior knowledge for its analyses that is integrated from a variety of public databases about diseases, metabolites and genes. The performance of the method was extensively evaluated on simulated patient data, patient data from the literature and patients from the MitoNET register with mitochondrial diseases. Finally, PheNoBo was applied to make diagnostic suggestions for participants of the KORA F4 cohort and for patients from MitoNET whose causal genes were not known.

Results: Various analyses and comparisons revealed that the prior knowledge used by PheNoBo is up-to-date and of high quality. A large-scale simulation of patients succeeded in deriving the optimal settings of PheNoBo for analyzing data from real patients. The evaluation of PheNoBo on the MitoNET patients proved that all three types of information about the patient, the phenotype, metabotype and genotype, make important contributions for the prediction. PheNoBo is able to predict the causal gene of more than one fourth of the MitoNET patients correctly. Therefore, PheNoBo is even able to handle challenging data sets like the mitochondrial patients from MitoNET. Furthermore this thesis presents reasonable diagnostic suggestions for part of the KORA F4 cohort and the MitoNET patients. The predictions of PheNoBo for the MitoNET patients have the potential to assist in the elucidation of the genetic causes of rare mitochondrial diseases.

Zusammenfassung

Hintergrund: Es sind mehr als 7 000 einzeln sehr seltene Krankheiten bekannt, die zusammen genommen die Gesundheit und Lebensqualität mehrerer Millionen Menschen beeinträchtigen. 80% der seltenen Erkrankungen werden durch genetische Defekte verursacht. Die Diagnose von seltenen Erkrankungen wird oft durch einen Mangel an Bewusstsein und Wissen erschwert. Seltene mitochondriale Erkrankungen sind spezielle seltene Krankheiten, die durch Defekte im Energie-Metabolismus verursacht werden. Die Diagnose von mitochondrial bedingten Krankheiten ist besonders schwierig wegen der hohen genetischen und phänotypischen Variabilität dieser Erkrankungen. Der Einsatz von Next-Generation Sequencing zusammen mit Computer-basierten Datenanalysen konnte den Diagnoseprozess verbessern und ermöglichte es, die kausalen Gene für viele der seltenen Erkrankungen zu identifizieren.

Methode: Diese Masterarbeit stellt die Pipeline PheNoBo vor. PheNoBo versucht, das kausale Gen eines Patienten mit einer genetisch bedingten Erkrankung vorherzusagen. Im Gegensatz zu den bisherigen Genotyp-basierten Methoden verwendet PheNoBo den Phänotyp, Metabotyp und Genotyp des Patienten für die Vorhersage. Dazu greift die Methode auf bestehendes Wissen aus öffentlich zugänglichen Datenbanken über Krankheiten, Metabolite und Gene zurück. Die Qualität der Vorhersagen von PheNoBo wurde ausführlich auf simulierten Patientendaten, Patient aus wissenschaftlichen Publikationen und Patienten mit mitochondrialen Erkrankungen aus dem MitoNET Register validiert. Schließlich wurde PheNoBo auf Teilnehmer der KORA F4 Kohorte und auf Patienten aus dem MitoNET Register mit unbekanntem kausalen Gen angewandt.

Ergebnisse: Mehrere verschiedene Analysen und Vergleiche ergaben, dass das die PheNoBo zugrunde liegenden Datensätze aktuell und von hoher Qualität sind. Die Simulation einer großen Patientenkohorte erlaubte die Festlegung der optimalen Parameter von PheNoBo für die Analyse realer Patientendaten. Die Anwendung von PheNoBo auf die Patientendaten von MitoNET zeigte, dass alle drei Arten von Informationen (Phänotyp, Metabotyp und Genotyp) einen bedeutenden Beitrag zur Vorhersage des kausalen Gens leisten. PheNoBo ermittelt das korrekte kausale Gen für mehr als ein Viertel aller MitoNET Patient. Dies impliziert, dass PheNoBo auch mit schwierigen Datensätzen wie den Patienten aus MitoNET umgehen kann. Des Weiteren beschreibt diese Arbeit die Ergebnisse von PheNoBo für einen Teil der KORA F4 Kohorte und der MitoNET Patienten. Die Vorhersagen von PheNoBo für die Patienten aus MitoNET sind nachvollziehbar und plausibel und können somit potentiell zur Klärung der genetischen Ursachen von seltenen mitochondrialen Erkrankungen beitragen.

Contents

1 Introduction

“Although individually rare, the cumulative burden of rare diseases is significant” [33]. This quote points out the importance of the medical research in rare diseases, which had been neglected until the 1980s [89]. Therefore, this introduction first gives an overview of the current challenges faced when diagnosing and treating rare diseases (section 1.1). These challenges can be partly met by including computational methods into the diagnostic process. The second part of the introduction presents three existing methods to find causal genes for patients suffering from rare diseases (section 1.2). The analysis methods implemented by these tools rely on different combinations of genotype, phenotype and metabotype data of the patient. However, there is currently no prediction method that considers all three types of patient data together [44]. The aim of this thesis is to develop and evaluate a new method for predicting the causal gene of a patient that uses any combination of phenotype, genotype and metabotype data (section 1.3).

1.1 Rare Diseases

The European Union classifies a disease as rare disease if it affects less than one in 2 000 individuals [33]. Altogether there are more than 7 000 known rare diseases [21] that affect about 30 million Europeans, including 4 million people in Germany [33].

80% of the rare diseases are caused by genetic defects [33]. Due to the genetic origin of rare diseases these diseases often manifest early in life [33]. The affected infants require a fast diagnosis to reduce the harmful effects of their rare diseases. A timely diagnosis can considerable reduce the morbidity and mortality of the affected children [76]. However, the large number of rare diseases and their phenotypic variability represent a big challenge for doctors when diagnosing rare diseases. [66].

In particular rare mitochondrial disorders are difficult to diagnose [45]. Mitochondrial disorders are caused by defects impairing the energy metabolism of human cells. As these defects are harmful for all energy-consuming tissues, mitochondriopathies are often severe, multi-systemic diseases. Mitochondrial diseases show only a poor correlation between the genotype and phenotype [25]. For example, the phenotype caused by a mutation within the mitochondrial genome depends on the number of mitochondria carrying the mutation [100]. Inversely, diseases with different underlying genetic defects cannot be distinguished by their phenotype [25]. Therefore, a classical diagnostic approach based on the patient's phenotype is often insufficient to deduce reliably the cause of the patient's mitochondriopathy.

The emergence of next-generation sequencing, especially whole-exome sequencing, facilitates the detection of the causal genes and consequently the diagnosis of rare diseases. Whole-exome sequencing is a special sequencing technique to determine the sequence of the coding region of a genome. This technique makes it possible to test for genetic defects in all human genes with a single experiment [21]. The decreasing costs for applying the new sequencing technology enable large scale projects investigating the causal genes for specific classes of rare diseases [76]. For example, the project Deciphering Developmental Disorders (DDD) analyzed the exomes of more than 4 000 families affected by rare developmental disorders [3]. Furthermore, the Canadian consortium FORGE succeeded in discovering causal genes for more than 100 rare diseases within two years [12]. Both projects successfully applied computational methods to find the causal gene within the patient's genotype data. These projects proved that computational methods can considerably speed up and improve the diagnostic process of rare diseases.

1.2 Prediction of Causal Genes for Rare Diseases

The success of rare disease projects like DDD and FORGE has fueled the development of increasingly accurate computational methods to predict the causal gene of a patient [95].

Miller *et al.* [76] developed a diagnostic pipeline called STATseq. STATseq is able to identify the causal gene of a patient within 26 hours after taking a blood sample from the patient. The routine includes whole-genome sequencing of the patient as well as the computational analysis of the sequencing data. The computational part of the pipeline identifies putative harmful variants within the sequencing data with a sensitivity of more than 99.5%. These variants are annotated and filtered according to their annotations. The final result of the pipeline is a list of potentially causal genes that is interpreted manually.

Genotype-centric prediction methods with a setup similar to STATseq have a success rate of about 25% in identifying the causal gene of a patient [95]. This low success rate is due to the fact that even exomes of healthy individuals contain about 10 000 non-synonymous variants [44]. Furthermore, 100 of these variants are deleterious loss-of-function variants (LOFs) that completely inactivate the gene that they lie in [73]. Various recently developed tools try to prioritize those putative harmful sequence variants by taking into account additional omics data of the patient [85]. These tools rely on genotype data produced by external pipelines like STATseq (without the final variant annotation and filtering).

Phen-Gen [56] is another computational method, that combines phenotype and genotype data to predict the causal gene of a patient. The phenotype data of a patient is a set of symptoms describing the current state and the disease history of the patient. Phen-Gen compares the phenotype of the patient against the symptoms annotated to all known diseases. This comparison allows Phen-Gen to identify the diseases that best explain the patient's symptoms. Phen-Gen translates the similar diseases into a ranking of potentially causal genes using knowledge about causal disease-gene relations and genetic interaction. In the second step, Phen-Gen analyzes the genotype data of the patient. The genotype data is a set of variants identified by whole-exome or whole-genome sequening of the patient's genome. Phen-Gen predicts the effect of each variant on protein coding genes and summarizes the predictions into a ranking of genes. Phen-Gen produces an independent ranking of genes for each patient data type. The Bayesian framework of Phen-Gen finally combines the two rankings into a final prediction. By combining phenotype and genotype, Phen-Gen was able to predict the causal genes for 88% of the patients from a simulated cohort and for 8 of 11 real patients with severe intellectual disability.

The approach of Guo *et al.* [44] combines metabotype and genotype data to identify disease-causing variants. The metabotype data used in this approach arises from non-targeted metabolomics experiments. Such experiments are high-throughput methods for measuring several hundred metabolites within a blood sample from the patient. The method of Guo *et al.* compares the metabotype measurements of an individual against reference metabolite levels from a healthy population. The method identifies and scores all metabolites of the patient under consideration that deviate significantly from the reference values. The genotype data is processed separately from the metabotype data. The genotype data comprises variants found by whole-exome sequencing. Guo *et al.* built a pipeline for annotating and scoring the variants according to their effect on the genes they lie in. The prediction procedure is completed by combining the metabotype and genotype data. The scored variants are searched for those variants that affect genes related to metabolic pathways of the deviating metabolites. This method succeeded in detecting variants related to fructose intolerance, xanthinuria and carnitine deficiency using metabotype and genotype data of 80 volunteers

with normal health.

Phen-Gen and the method of Guo *et al.* provide complementary and efficient methods to predict causal genes for patients suffering from genetic diseases. However, there is still considerable potential to further improve these methods by combining all three types of patient data, the phenotype, genotype and metabotype data.

1.3 Aim of this Thesis

The aim of this master's thesis is to develop a new computational method for predicting the causal gene of a patient suffering from a genetic disease. The predictions of the method are based on combining all three types of patient data introduced above: phenotype, genotype and metabotype data. Finally, the method was evaluated on data from patients affected by rare mitochondrial disorders.

1.3.1 Combination of Phenotype, Genotype and Metabotype (PheNoBo)

The method presented in this thesis unites the approaches of Phen-Gen and Guo *et al.* (figure 1). The method is called PheNoBo which stands for the Combination of **Phe**notype, Ge**No**type and Meta**Bo**type. PheNoBo predicts the causal gene of a patient suffering from a genetic disease. For this purpose, PheNoBo combines phenotype, metabotype and genotype data into a comprehensive ranking of scored genes. The genes with the highest scores are most likely to cause the patient's disease.

The workflow of PheNoBo processes each data type separately yielding three independent analyses: the genotype analysis, the metabotype analysis and the phenotype analysis. The genotype analysis works on the genotype data of the patient, which is a set of genetic variants identified from whole-exome sequencing of the patient's genome. These variants are scored according to their effect on the genes which they lie in. The individual variant scores are summarized into genes scores.
The metabotype analysis uses metabolite profiles generated by non-targeted metabolomics measurements. The levels of each metabolite of the patient's metabotype are scored by comparing them against reference values from a healthy control group. The metabolite scores are translated into gene scores using metabolite-gene and gene-gene interactions.

The phenotype analysis is similar to the metabotype analysis. The patient's phenotype is described by a collection of symptoms. These symptoms are used to score known diseases according to their similarity to the patient's phenotype. The resulting disease scores are transformed into gene scores using disease-gene and gene-gene interactions.
Each of the three analyses yields a set of scored genes. The score of a gene reflects the probability that the gene is causal given the patient's genotype, metabotype or phenotype. In a final analysis step PheNoBo combines the three scores for each gene into a single score. The gene with the highest final score is most likely to be the patient's causal gene.

1.3.2 Application of PheNoBo on Rare Mitochondrial Diseases

This master's thesis is realized in collaboration with the Institute of Human Genetics at Helmholtz Zentrum München (IHG). The IHG is part of MitoNET [11], the German network for mitochondrial disorders. The MitoNET is a network of researchers and clinicians as well as a patient register for patients suffering from rare mitochondrial disorders. The aim of MitoNET is to find the causal genes and effective treatments for the patients of the register.

The research of the IHG focuses on MitoNET patients for which standard diagnostic methods and treatments are insufficient. The IHG kindly provided an extract from the MitoNET patient register providing phenotype, genotype and metabotype data for 112 patients with mitochondrial diseases. For a subset of the patients the causal gene of their disease is known. This thesis covers the evaluation of PheNoBo with this subset as well as the application of PheNoBo on the remaining patients to learn more about the causes of their diseases.

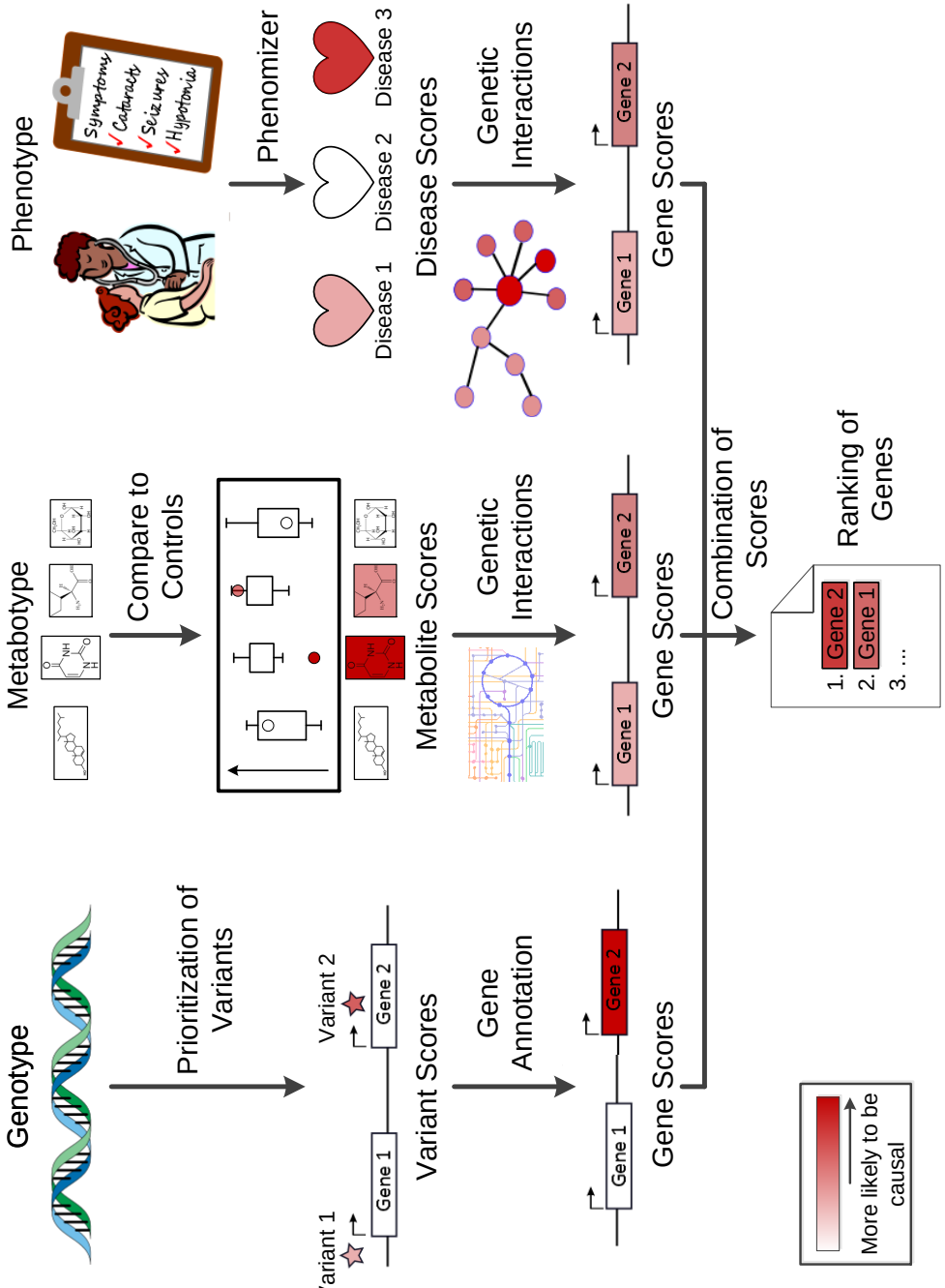


Figure 1: Workflow for predicting causal genes by combining phenotype, genotype and metabotype. The image gives an outline of the method PheNoBo which is presented in this thesis. The aim of the method is to predict the causal gene of a patient suffering from a genetic disease.

2 Materials and Methods

The workflow described in the introduction was realized as a set of tools called PheNoBo (Combination of **Phe**notype, Ge**No**type and Meta**Bo**type).

Figure 2 gives a schematic overview of the setup of PheNoBo. Altogether there are 6 different tools in place, which form seven distinct steps in the PheNoBo workflow. The tools of PheNoBo require additional data from different databases about genes, metabolites and diseases. The input patient data of the pipeline result from measurements of the patient's phenotype, genotype and metabotype. PheNoBo also requires metabolite profiles from a control group as a reference. The final output of PheNoBo is a list of ranked genes.

PheNoBo mainly covers the analysis of the phenotype data and the metabotype data. There are already several in-house pipelines in use at Helmholtz Zentrum for searching sequencing data for putative harmful genomic variants (e.g. [5], [38], [50] and [57]). The output of such a pipeline is passed as genotype data to PheNoBo.

This chapter gives more detailed information about the components of the PheNoBo pipeline. Section 2.1 describes the external databases on which PheNoBo relies. The second section (section 2.2) gives an overview of the patient data used for PheNoBo. The remaining sections present the algorithms of PheNoBo (section 2.3) and outline how PheNoBo was applied and evaluated on the patient data (section 2.4).

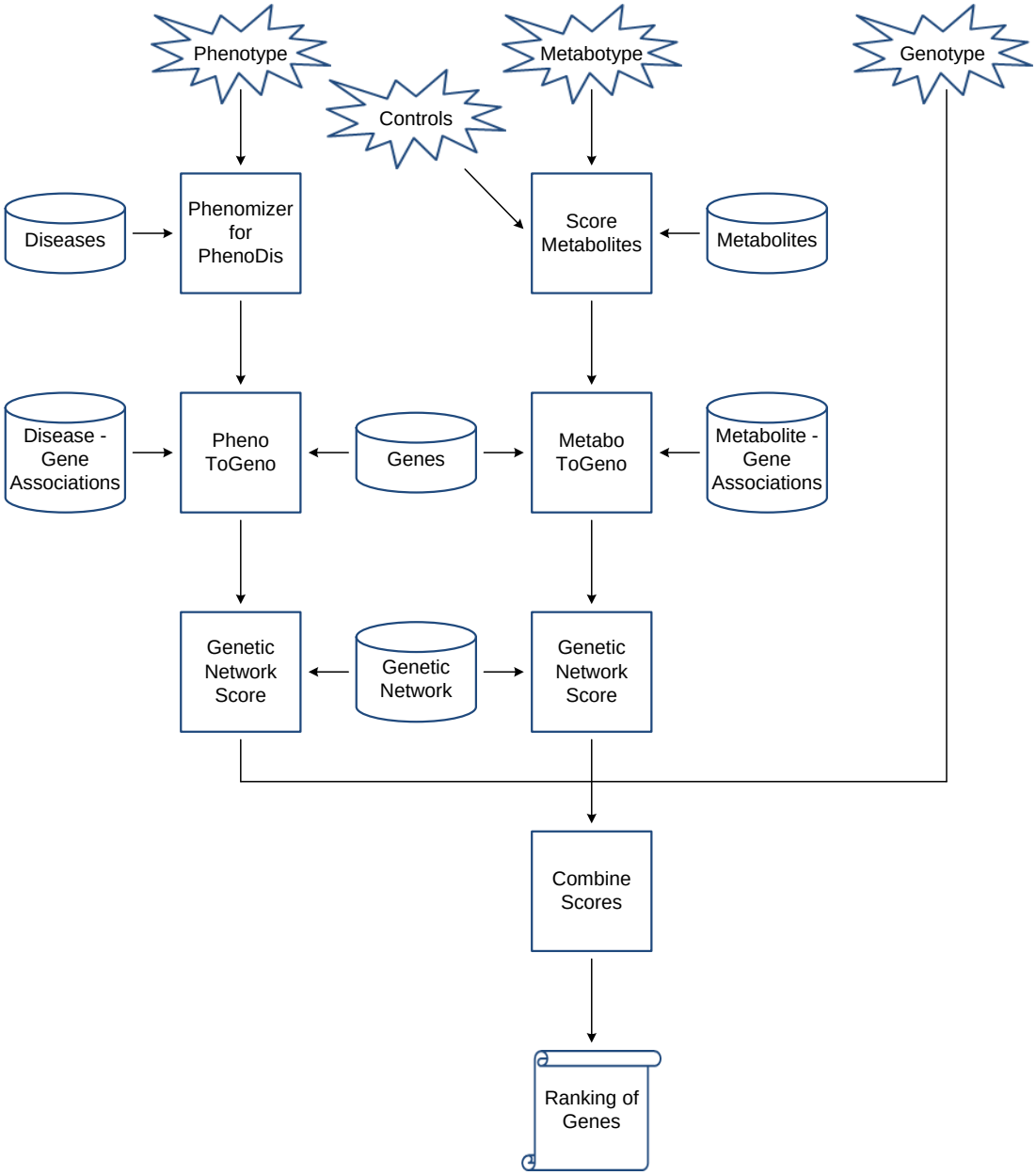


Figure 2: Overview of the PheNoBo pipeline. The squares represent the different tools of PheNoBo that process different types of data. The patient data are shown as jagged ovals. The cylinders depict the public resources (i.e. external databases) providing information for the tools. The resulting prediction, a list of ranked genes, is drawn as a scroll. An arrow between any of the entities indicates a directed data transfer between them.

2.1 Public Resources Underlying PheNoBo

The predictions of PheNoBo are based on prior knowledge about genes, metabolites, diseases and their relations among each other. These data sets were extracted from several bioinformatics databases and combined into a uniform representation. The data processing was implemented with KNIME 3.1.0 [15], Python 2.7.6 [37], R 3.0.2 [81] and MATLAB 2016a [74].

The following sections describe the underlying data sets of PheNoBo (shown as cylinders in figure 2).

2.1.1 Disease Database

The information about diseases is taken from the PhenoDis database (version of February 2016). PhenoDis is an in-house database developed and maintained at Helmholtz Zentrum. The database unites the resources of OMIM (Online Mendelian Inheritance in Man) [46], Orphanet [83] and the HPO (Human Phenotype Ontology) [86].

PhenoDis stores a comprehensive collection of diseases and their phenotypes. The phenotype of a disease in PhenoDis is characterized by a set of symptoms. This knowledge of PhenoDis is consolidated from OMIM and Orphanet. The HPO provides a controlled terminology for describing the symptoms of human diseases [86]. Therefore, all symptoms of the diseases in PhenoDis are mapped to the standardized vocabulary of the HPO.

PhenoDis provides four tables for this work:

- a table of 8 263 rare and common diseases
- a table of 11 573 symptoms.
- a table of 138 912 symptom-disease pairs
- a table of 14 941 symptom-symptom pairs

The first two tables contain some basic entities used in PheNoBo whereas the last two tables represent relationships of algorithmic importance for PheNoBo.

The symptom-disease associations describe the phenotypes of the diseases in PhenoDis.

The associations indicate which symptoms occur with a particular disease. Additionally some symptom-disease pairs are annotated with frequency values. The frequencies describe how often the symptom is observed in patients suffering from the respective disease. For example, the entry "strabismus – Leigh syndrome – very frequent" means that most of the patients with Leigh Syndrome suffer from strabismus. The interpretation of the term "very frequent" in terms of exact frequency values depends on the source and time of annotation. Currently, the annotators of PhenoDis use "very frequent" for symptoms that occur in at least 90% of all patients with the disease. Note that PheNoBo uses a different interpretation from [39] that is based on the Orphanet annotations from August 2015. For PheNoBo the entry "strabismus – Leigh syndrome – very frequent" means that more than 75% of the patients with Leigh Syndrome suffer from strabismus.

The table of symptom-symptom pairs stores the Human Phenotype Ontology (HPO) [86]. An ontology is a feature to represent well-defined entities and the relationship between those entities [41]. In this case the ontology models is-a relationships between the symptoms [86].

Figure 3 provides an excerpt from the HPO. The ontology is displayed as a directed graph. Each node of the graph (oval) corresponds to a symptom. Each directed edge (arrow) corresponds to an is-a relation between two symptoms. For example, the edge from "Abnormal iron deposition in mitochondria" to "Abnormality of mitochondrial metabolism" means that the aberrant deposition of iron in the mitochondria is a specific abnormality of the mitochondrial metabolism. "Abnormality of mitochondrial metabolism" is called parent term of "Abnormal iron deposition in mitochondria" [86]. The HPO is no strict hierarchy as symptoms can have more than one parent term, e.g. "Decreased plasma carnitine" is an "Abnormality of mitochondrial metabolism" as well as an "Abnormality of carnitine metabolism". However, the HPO does not allow any directed circles within the graph of the ontology, i.e. the HPO is represented by a directed acyclic graph [86].

Each symptom of the HPO is associated with a defined number of diseases in PhenoDis (squares connected to the symptoms in figure 3). The associated diseases are inherited along the edges of the ontology. For example, "Decreased plasma total carnitine" is a symptom of Carnitine palmitoyl transferase II deficiency. As "Decreased plasma total carnitine" is also an "Abnormality of carnitine metabolism", Carnitine palmitoyl transferase II deficiency is automatically annotated to "Abnormality of carnitine metabolism". This transfer of symptoms is also included in the number of associated diseases. The eleven diseases with the symptom "Decreased plasma carnitine" are composed of

- one disease annotated to the symptom "Decreased plasma total carnitine"

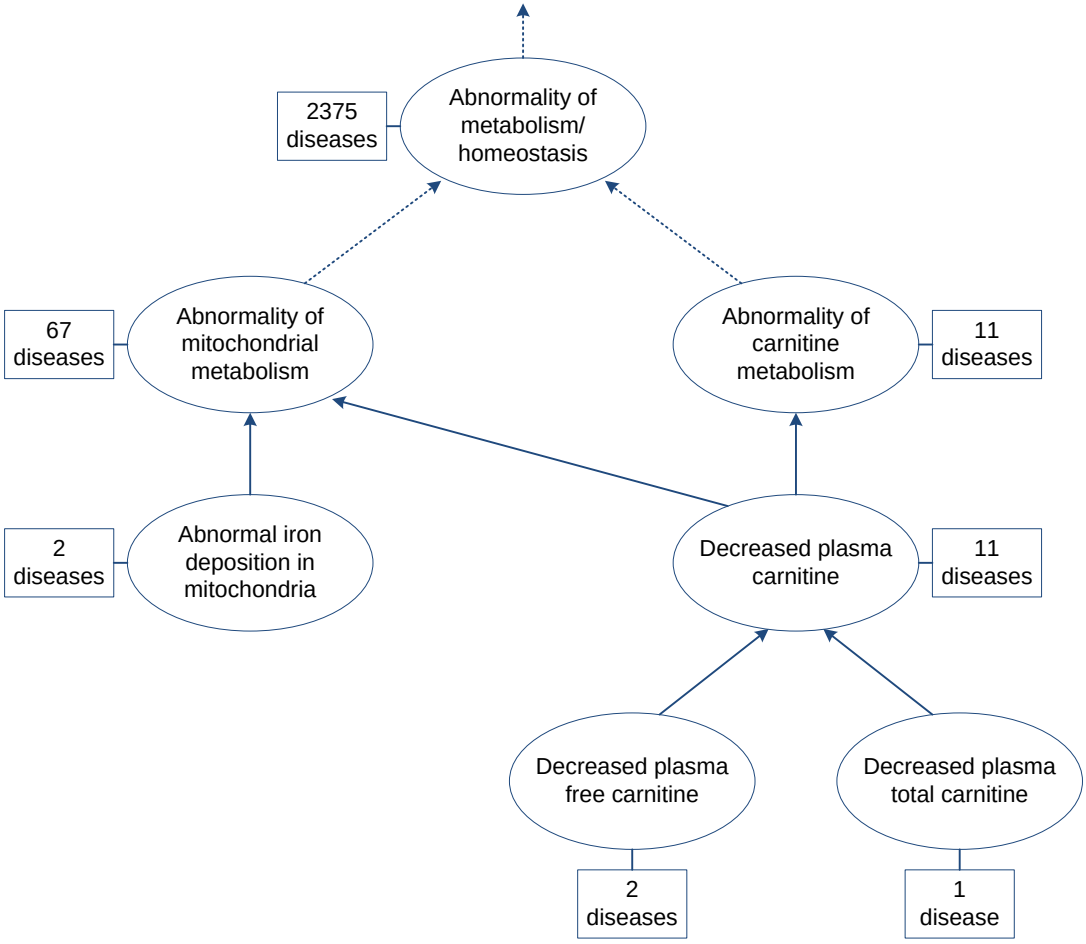


Figure 3: Sample from the HPO. Each oval represents a symptom. An arrow represents an is-a relationship. The dashed arrows indicate that some intermediate symptoms are not shown for reason of clarity. The squares connected to the symptoms annotate the numbers of diseases associated with the symptoms.

- two diseases annotated to the symptom “Decreased plasma free carnitine”
- eight diseases directly associated with “Decreased plasma carnitine”

This feature of the HPO is essential for the Phenomizer algorithm of PheNoBo [39].

2.1.2 Metabolite Database

PheNoBo mainly gathers information about metabolites from the Human Metabolome Database (HMDB) [114] (version 3.6, downloaded in May 2016). The HMDB aims to provide a comprehensive collection of the components of the human metabolome. The database stores the current knowledge about the metabolites and metabolic reactions in the human body. The HMDB includes classical metabolites such as lipids or amino acids as well as small molecules taken up from the environment like toxins and food additives [114]. The database currently manages information about 41 993 metabolites.

PheNoBo requires additional reference values for each metabolite (section 2.3.2). The reference values give the expected amount of the metabolite in the blood of health individuals. These values are determined from the MitoNET control group which is part of the patient data of PheNoBo (section 2.2.4). The control group provides reference values for 602 metabolites.

Therefore, the set of the 41 993 human metabolites is reduced to 276 metabolites with reference values. The 326 metabolites from the reference set which are not part of the HMDB are added to the set of metabolites.

Hence, the final data set of metabolites embraces 602 metabolites. Each metabolite in PheNoBo is addressed by its unique Metabolon identifier which is also part of the reference data from MitoNET. PheNoBo calculates a score for every metabolite in this data set.

2.1.3 Gene Database

The data about genes are based on the gene annotations from Ensembl [118] (downloaded in March 2016). Ensembl comprises in total 66 203 genes for the human genome. This set was restricted to all genes with at least one known protein product, yielding a collection of 22 784 protein coding genes.

As PheNoBo also uses disease-gene associations (section 2.1.4), metabolite-gene associations (section 2.1.5) and genetic interactions (section 2.1.6), all non-coding genes from these data sets were added to the collection of genes.

- The disease-gene data includes 3 711 genes. 24 of those gene do not encode proteins. Most of the 24 genes are mitochondrial tRNA genes, which are prone to mutations leading to mitochondrial diseases [98].
- The metabolite-gene data comprises 2 530 genes. The metabolite data contains one pseudogene.
- The genetic interactions provide data for 18 872 genes. There are 512 non-coding genes in the set. The 512 genes are made up of pseudogenes, antisense RNA genes, long intergenic non-protein coding RNA genes and open reading frames.

The origin and content of the three enumerated data sets is further explained in the following sections.

After the addition of the non-coding genes the final set of genes consists of 23 330 genes. PheNoBo makes a prediction for every gene in this data set.

2.1.4 Disease-Gene Associations

A disease-gene relation is defined as an etiologic relationship. A gene is associated with a disease if mutations in the gene give rise to the disease.

All relations between the genes of section 2.1.3 and the diseases of section 2.1.1 are taken from the SNiPA [8] resource (version 3.1). As SNiPA is derived from various databases including Ensembl [118], OMIM [46] and Orphanet [83], it provides all required identifiers for associating the disease entities and the gene entities of PheNoBo. The resulting data set consists of 8 585 associations covering 3 711 genes and 4 915 diseases.

2.1.5 Metabolite-Gene Associations

The information about the metabolite-gene relations was collected from six different resources. Consequently, this data set comprises relationships of different meaning and quality. The following paragraphs explain each of the six databases and the provided types of data.

HMDB: The Human Metabolome Database (HMDB) [114] stores information about the biochemical reactions and the enzymes of the human metabolism (section 2.1.2). All associations between metabolites and genes are based on reaction mechanisms. A gene is annotated to a metabolite in HMDB if the gene encodes an enzyme that forms or consumes the metabolite.

Recon: Recon [101] (version 2.04, downloaded in May 2016) is a systems biology project with the aim of building a computational model for simulations of the human metabolism. The metabolite-gene relations represent direct biochemical reactions and are equivalent in meaning to those of the HMDB.

SMPDB: The Small Molecule Pathway Database (SMPDB) [58] (version 2.0, downloaded in May 2016) collects the metabolic pathways for the metabolites and enzymes of the HMDB. A pathway is a set of linked reactions where the product of a reaction is the substrate of the next one. The pathways of SMPDB cover essential cellular processes, the metabolism of drugs and metabolic diseases [58]. The SMPDB enables a less strict definition of a metabolite-gene pair. A gene and a metabolite have to participate in the same metabolic pathway in order to become associated.

GWAS Server: The Metabolomics GWAS Server [91] manages the currently most comprehensive data set of genome-wide association studies (GWAS) on metabolites. A metabolite-gene relation of the GWAS Server implies a statistical association at genome-wide significance between the concentration of a metabolite and a genomic variant within or near a gene.

Manually Curated Data Set Based on HPO Symptoms: The vocabulary of the Human Phenotype Ontology (HPO) (section 2.1.1) also includes terms for disturbances in some metabolite concentrations like “Abnormality of carnitine metabolism” (figure 3). These metabolic symptoms are descendants of the parent terms “Abnormality of metabolism/homeostasis” (HPO:0001939) and “Abnormality of the endocrine system” (HPO:0000818). The annotators of the IBIS kindly provided manual annotations of metabolite names for 561 metabolic symptoms. The annotated symptoms were transformed into metabolite-gene pairs by joining the disease-symptom pairs of the disease database (section 2.1.1) and the disease-gene associations (section 2.1.4). Consequently, an association between a metabolite and a gene based on the HPO means that abnormal concentrations of the metabolite occur in diseases that are caused by the gene.

Manually Curated Data Set Based on GO Terms: The Gene Ontology (GO) [40] is an ontology providing a controlled terminology for gene product functions. The manual

annotations of the 561 metabolic symptoms introduced in the previous paragraph also included GO terms related to the symptoms. There are four GO terms for almost every symptom referring to the transport, metabolic process, the homeostasis and the triggered response of the affected metabolite. Addition of the GO annotations of Ensembl [118] yields associations of metabolites and genes. The resulting metabolite-gene pairs imply that the function of the gene is related to the metabolite.

As the final data set should match the entities of the gene set (section 2.1.3) and the metabolite set (section 2.1.2), all database identifiers were mapped to Ensembl gene identifiers and to Metabolon metabolite ids respectively. The metabolites of HMDB, SMPDB and Recon are identified by HMDB accession numbers. The mapping between the HMDB metabolite identifiers and the Metabolon ids was provided by an in-house metabolomics database at the IBIS. The gene symbols of the enzymes in HMDB, SMPDB and GWAS Server were mapped via Ensembl [118]. The Metabolon ids belonging to the metabolites of the GO terms and the HPO symptoms were found by exact string matching of the metabolite names. All other databases already contained the necessary identifiers to create the set of metabolite-gene associations. The consolidation of the six data sources results in 12 676 associations covering 252 metabolites and 2 530 genes.

2.1.6 Genetic Network

The genetic network represents the interactions among the genes of section 2.1.3. The interaction data was extracted from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) version 10 [99], which collects interaction data from a wide range of sources. The network of STRING is modeled as an undirected, weighted graph. Each node of the network corresponds to a gene. An edge between two genes indicates a physical or functional interaction [99]. The weight of an edge is a number between 0 and 1 which gives the confidence of the interaction. A weight of at least 0.7 corresponds to a high confidence score and a weight between 0.4 and 0.7 is considered as medium confidence [110].

Figure 4 illustrates the concept of the STRING network for four exemplary genes and their interactions among each other. The genes for Carnitine/Acylcarnitine Translocase (CACT), Carnitine Palmitoyltransferase 1A (CPT1A) and Carnitine Palmitoyltransferase 2 (CPT2) are linked together by edges of high confidence as the three proteins catalyze well-studied, subsequent steps of the import of acylcarnitine into the mitochondria. The interaction between Estrogen Receptor 1 (ESR1) and CACT is only of medium confidence because the data stems from a less reliable high-throughput interaction screening.

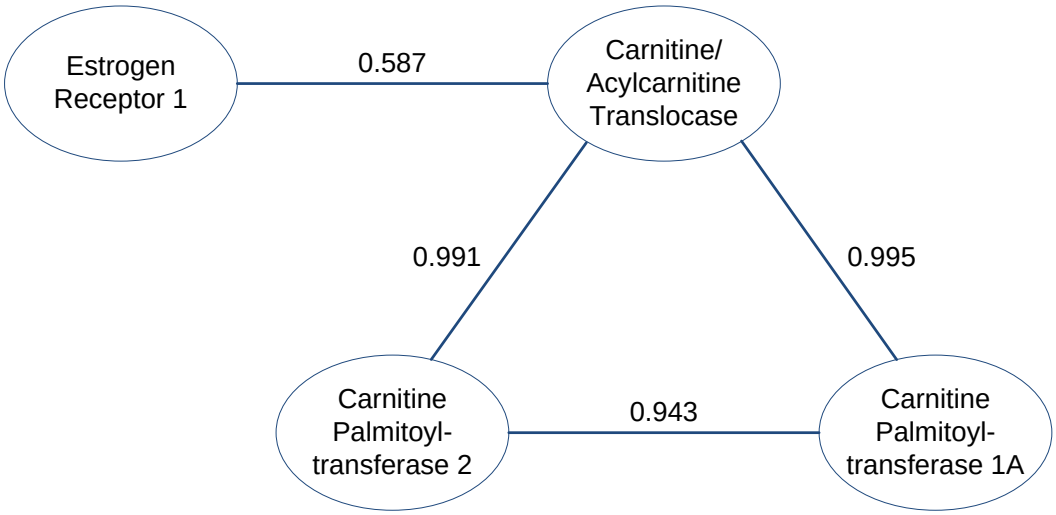


Figure 4: Example from the STRING network for the genes CACT, CPT1A, CPT2 and ESR1. An edge between two genes indicates an interaction and is annotated with a confidence score between 0 (low confidence) and 1 (high confidence).

The network of STRING was processed in a filtering step before applying it as data set for PheNoBo. The STRING network also contains false positive interactions with a low confidence score, e.g. from text mining. As such interactions could affect the predictions of PheNoBo, all edges with a confidence score of less than 0.4 were removed [65]. The resulting network consists of 18 872 genes and 740 415 interactions.

2.2 Patient Data

In addition to the data sets from public resources, PheNoBo requires two types of patient data (shown as jagged ovals in figure 2): a control group and a description of the patient for whom the causal gene should be predicted.

The control group is necessary for the analysis of the patient's metabotype. The metabotype analysis depends on a reference data set giving the expected metabolite levels for every metabolite (further explained in section 2.3.2.1). This information is based on metabolomics measurements from blood samples of the members of the control group. All participants of the control group are expected to be healthy. This master's thesis uses two control groups: a control group from the MitoNET register [11] and the KORA F4 cohort [112].

The description of the patient is composed of three different types of information:

- **Phenotype Data:** The phenotype data is simply a list of symptoms. The symptoms should give a precise and comprehensive description of the current state and the disease history of the patient. PheNoBo requires that these symptoms match the thesaurus of symptom terms provided by the HPO (section 2.1.1).

- **Metabotype Data:** The metabotype data is a set of metabolite levels detected in a blood sample from the patient. PheNoBo compares the metabotype data of the patient against reference values derived from a control group (above). As a consequence PheNoBo has two requirements to the patient's metabotype data. First, the blood sample from the patient has to be analyzed with the same experimental setup as the samples from the control group. Ideally, one should measure the sample from the patient together with these control samples. The second requirement concerns the normalization of the metabolite concentrations of the patient. The normalization applied to the patient data has to follow the steps for constructing the reference data set (described in section 2.3.2.1). Both requirements ensure comparability between the reference values and the metabotype data of the patient. Finally, the normalized concentrations should be annotated with the phenotype group of the patient (table 4). The information about the phenotype group is a prerequisite for running the ScoreMetabolites algorithm (further explained in section 2.3.2).

- **Genotype Data:** The genotype data is a list of scored genes. The score should reflect the probability that the gene is causal given the patient's genomic variants. The patient's genome is determined by next-generation sequencing experiments. PheNoBo relies on external pipelines for identifying and scoring putative harmful genomic

variants from sequencing data. This thesis focuses on loss-of-function variants (LOF), a special kind of harmful sequence mutations. LOFs are genomic mutations that completely disrupt the function of protein-coding genes [73]. Therefore, each LOF gene g (i.e. a gene with at least one LOF) in the genome of a patient p gets a score of

$$\text{score}(g) = \frac{1}{\text{LOFs}(p)} \tag{2.1}$$

LOFs(p) denotes the total number of LOF genes of the current patient p. This score is chosen such that all LOF genes are equally likely to cause the patient's condition.

PheNoBo runs on any combination of these data. There is no specific requirement regarding the source of this data: the input of PheNoBo can stem from simulated patients, from patients described in the literature or from a patient register.

Table 1 gives an overview of the patient data that was used in this master's thesis. Altogether there are 8 patient data sets comprising phenotype, metabotype data and genotype data for several patients. Most patient data sets provide phenotype data. Due to the strict requirements on the metabotype data, only two patient data sets with metabotype information are used in this thesis. The genotype data is even more sparse. PheNoBo was applied on two patient data sets with genotype data, one of real and one of simulated genotype data.

This section aims to describe the patient data of table 1 and the corresponding control groups. The description is structured according to the source of the patient data (column source of table 1).

Data Set	Source	Size	Phenotype	Metabotype	Genotype
1	Simulation	1 400	Very Frequent Symptoms	–	–
2	Simulation	7 890	Mitochondrial Symptoms	–	–
3	Literature	21	Case Studies Rare Diseases	–	–
4	Literature	15	Case Studies CACTD and PCD	–	–
5	KORA	106	–	KORA Metabolite Samples	KORA LOFs
6	MitoNET	82	–	Metabolite Samples Fasting Adults	–
7	MitoNET	19	Symptom Table Known	Metabolite Samples Known	Simulated LOFs
8	MitoNET	93	Symptom Table Unknown	Metabolite Samples Unknown	–

Table 1: Patient data sets for the application of PheNoBo. Each patient data set is a combination of phenotype data, metabotype data and the genotype data from a group of patients. The column size gives the number of patients covered by the respective data set. The column source indicates the origin of the patient data set.

2.2.1 Simulated Patients

The simulated patient cohorts comprise only phenotype data describing the individual patients. These data are used for the patient data sets 1 and 2 (table 1). The following paragraphs present the simulation procedures for generating the patient data.

Simulated Patients – Very Frequent Symptoms: The phenotype data set called *Very Frequent Symptoms* consists of 1 400 simulated patients. Each virtual patient is defined to suffer from a particular disease of PhenoDis (section 2.1.1). There is one virtual patient per disease. The symptoms of the virtual patient are all very frequent symptoms known for the patient's disease. The simulation procedure classifies a symptom of a disease as very frequent if it occurs with a frequency of more than 75% together with the disease (table 2). This frequency information is also provided by PhenoDis (section 2.1.1). The frequency information is interpreted based on the definitions used in [39] and Orphanet [83] (from August 2015).

Simulated Patients – Mitochondrial Symptoms: *Mitochondrial Symptoms* is another set of simulated patients with phenotype data. The data set comprises 7 890 patients with mitochondrial diseases. In this context a mitochondrial disease is defined as a disease that is classified as "mitochondrial" in PhenoDis or that is caused by a gene with a GO term [40] containing the word "mitochondrion" or "mitochondrial". According to this definition the PhenoDis database contains 789 mitochondrial diseases. The set *Mitochondrial Symptoms* has ten patients per mitochondrial disease.
The lists of symptoms for the patients result from a stochastic sampling procedure. The procedure includes each symptom of a disease with a certain probability. The probability distribution for the simulation is derived from a grouping of symptom-disease pairs according to their frequency. The group definition is adopted from [39] and given in table 2.
The simulation procedure iterates over all patients. The simulation draws a random number between 0 and 1 from a uniform distribution for each symptom of the patient's disease. If the random number is at most as large as the simulation threshold of the current symptom-disease pair (table 2), the symptom is added to the symptom list of the current patient.
The simulation procedure is repeated until the simulated phenotype data for a patient contains at least 5 symptoms.

Group	Frequency Range	Simulation Threshold
Occasional	]0, 0.25]	0.125
Frequent	]0.25, 0.75]	0.5
Very Frequent	]0.75, 1]	0.875

Table 2: Grouping of symptom-disease pairs according to frequency. The column Group gives the name of the group. The column Frequency Range contains intervals of frequency values. If the frequency of a symptom-disease pair lies within the interval of a group, it is part of this group. The definitions of the groups and the frequency ranges are taken from [39]. The column Simulation Threshold shows the probability distribution used for the simulation of the data set *Mitochondrial Symptoms*.

2.2.2 Patients from the Literature

Two groups of patients were derived from scientific publications describing individual patients. The patient data sets 3 and 4 are made up of phenotype data from these groups (table 1). The subsequent paragraphs describe the specific sources of the phenotype data for each group of patients from the literature.

Literature – Case Studies Rare Diseases: The phenotype data set *Case Studies Rare Diseases* is based on the 21 publications that were already used in [39]. These medical articles describe individual patients suffering from rare diseases. Table 3 lists the publications and groups them according to the disease of the presented patient. The symptoms of the resulting 21 patients were extracted manually from the texts.

Literature – Case Studies CACTD and PCD: The set *Case Studies CACTD and PCD* is derived from eight case studies characterizing the phenotype of 15 patients affected by carnitine-related diseases. Three publications [2, 23, 116] present patients with Primary Carnitine Deficiency (PCD). The remaining manuscripts [53, 72, 87, 97, 109] are about patients suffering from Carnitine/Acylcarnitine Translocase Deficiency (CACTD). The symptoms of the patients were manually assembled from the case studies.

Disease	Publications
Acromelanosis	[43], [7], [61], [93]
Blue diaper syndrome	[34], [71]
CADASIL	[22], [104], [10]
Hawkinsinuria	[78], [103], [19], [18]
Leigh syndrome	[52], [27], [70]
Merosin-deficient congenital muscular dystrophy	[49], [59], [54]
Neonatal hemochromatosis	[63]
Systemic lupus erythematosus	[60]

Table 3: Overview of the case studies about rare diseases. The table shows the 21 case studies and the corresponding 8 rare diseases of the phenotype data set *Case Studies Rare Diseases*.

2.2.3 KORA

KORA (Kooperative Gesundheitsforschung in der Region Augsburg) is a population-based cohort started in 1996. KORA provides a platform for research in epidemiology, health economics and health care [112]. The patient data for PheNoBo comprises data from 1 768 members of the KORA F4 cohort. KORA F4 is a follow up study of KORA that was carried out from 2006 to 2008. All 1 768 subjects are adults aged between 32 and 77 [91]. The IBIS kindly provided metabotype data for the 1 768 individuals and additional genotype data for 106 participants from a non-diabetic subgroup.

Control group: The control group derived from KORA comprises 1 768 members from KORA F4. Blood samples were taken from the 1 768 fasting subjects for the determination of metabolite profiles. The samples were analyzed using a non-targeted metabolomics platform established by Metabolon [35, 29]. The platform of Metabolon applies ultra high performance liquid chromatography with subsequent tandem mass spectrometry and gas chromatography with subsequent mass spectrometry. The experimental details of the metabolomics measurements are given in the supplementary material of [91]. The measurements resulted in metabolite levels for 517 metabolites. The summarization of the metabolite levels into a reference data set for the metabotype analysis is covered in section 2.3.2.1.

Metabotype Data – KORA Metabolite Samples: The metabotype data of the set *KORA Metabolite Samples* is a subset of the metabotype data of the control group. *KORA Metabolite Samples* collects the metabolite profiles of 106 samples from a non-diabetic subgroup

of KORA F4. These 106 samples were chosen for the application of PheNoBo because the donors of the samples were also sequenced and analyzed for loss-of-function variants [57] (see genotype data below).

Genotype Data – KORA LOFs: The set *KORA LOFs* is based on the exome sequencing of 106 participants from the KORA F4 cohort. The sequencing data is described and analyzed in [57]. This external analysis yielded a list of LOF genes per subject. For enabling the processing by PheNoBo a score was assigned to each gene (equation 2.1). The resulting data set consists of 106 lists of scored genes. Each list contains on average 164 LOF genes.

2.2.4 MitoNET

MitoNET [11] is a German patient register for patients suffering from mitochondrial diseases. MitoNET is run by an interdisciplinary network of clinicians and researchers. The register collects data from patients with mitochondrial disorders, e.g. it manages biobanks with plasma and DNA samples from the patients. The aim of MitoNET is the improvement of the medical treatment of the patients and the promotion of research in mitochondrial diseases.

The Institute of Human Genetics at Helmholtz Zentrum München (IHG) kindly provided an extract of the MitoNET register. The extract comprises information about 143 patients and 97 controls (presented in detail in [90]). There is metabotype data available for all cases and controls. In addition the MitoNET register provides the phenotype data for 112 patients. For 19 of those 112 patients the causal gene of the patient's disease is known. Note that there is also genotype data derived from exome sequencing for most of the patients of MitoNET. However, this data was not available in the time of this thesis. Instead, simulated LOF data were used as genotype data for analyzing the MitoNET patients with PheNoBo.

Control group: The control group comprises 97 blood samples from healthy individuals. The composition of the control group is given in table 4. The blood samples from the controls were analyzed in a similar manner as the samples from the KORA cohort. A non-targeted metabolomics of Metabolon was applied to gather metabolite profiles from the samples. The platform uses liquid chromatography with subsequent mass spectrometry. The details of these measurements are described in [90]. The application of the MetaP server [62] in [90] proved that the measurements are of high quality. The metabolite profiles from the control group cover 602 metabolites. However, the chemical structure of 227 of those metabolites is unknown. The analyses in [90] revealed that some of the unknown metabolites are associated with mitochondrial diseases. For that reason all 602 measured metabolites

were included in the construction of reference metabolite data set. This construction is further explained in section 2.3.2.1.

Phenotype Data – Symptom Table: The phenotypic data set for the MitoNET patients is called *Symptom Table*. The phenotype data from MitoNET is organized as a table collecting the results of medical examinations. Each row of the table corresponds to a patient. Each column represents a symptom. The names and encodings of the columns of the table were manually translated into symptoms. The table covers 112 patients and provides information about more than 300 symptoms. Each patient has on average 29 symptoms.
The resulting data set was divided into two subsets

- Symptom Table Known: phenotype data of 19 patients with known causal gene
- Symptom Table Unknown: phenotype data of 93 patients without known causal gene

Metabotype Data – Metabolite Samples: The set *Metabolite Samples* comprises metabolite concentrations for the 143 MitoNET patients. The donors of these samples are suffering from diseases with mitochondrial involvement. Table 4 gives an overview of these donors. The samples were measured together with the 97 control samples described above. This metabotype data is presented and analyzed in more detail in [90]. Three not necessarily disjoint subsets were chosen from the 143 samples and analyzed with PheNoBo:

- Metabolite Samples Fasting Adults: metabotype data of 82 fasting adults
- Metabolite Samples Known: metabotype data of 19 patients with known causal gene
- Metabolite Samples Unknown: metabotype data of 93 patients without known causal gene

The metabolite levels of all samples were normalized according to the procedure given in section 2.3.2.1.

Genotype Data – Simulated LOFs: The *Simulated LOFs* are derived from two sources: the *KORA LOFs* and the MitoNET register. The first source, *KORA LOFs*, is explained in section 2.2.3. The second source provides the validated causal genes for 19 patients from the MitoNET register. The two resources were combined into $19 \cdot 106 = 2\,014$ lists of genes. There is a gene list for every combination of KORA proband and MitoNET patient containing the LOF genes of the proband and the causal gene of the patient. The genotype scores for the 2 014 lists were determined by applying equation 2.1.

Group	Age	Fasting State	Number of Controls	Number of Cases
1	0 - 1 year	unknown	14	5
2	2 - 12 years	unknown	18	35
3	>12 years	fasting	44	83
4	>12 years	no fasting	21	20

Table 4: Phenotype groups of the control and patient samples. The 97 control samples (controls) and 143 patient samples (cases) are grouped according to age and fasting state.

2.3 Implementation of PheNoBo

The tools of PheNoBo are implemented as KNIME [15] nodes for KNIME version 3.1.0. These nodes are combined into a KNIME workflow to realize the functionality of PheNoBo (figure 2).

The PheNoBo pipeline is composed of six different tools: Phenomizer, ScoreMetabolites, PhenoToGeno, MetaboToGeno, GeneticNetworkScore and CombineScores. The phenotype is evaluated by Phenomizer, PhenoToGeno and GeneticNetworkScore. The metabotype analysis consists of the tools ScoreMetabolites, MetaboToGeno and GeneticNetworkScore. PheNoBo does not cover the analysis of the genotype but there is e.g. the collection of KNIME nodes called KNIME4NGS which provides this functionality [48]. Finally, CombineScores produces the output of PheNoBo.

The listing below gives a brief description of the purpose of each node.

- **Phenomizer for PhenoDis** calculates a score for each disease based on the patient's phenotype.
- **ScoreMetabolites** scores metabolites by evaluating the metabotype of the patient.
- **PhenoToGeno** transforms disease scores into gene scores.
- **MetaboToGeno** transforms metabolite scores into gene scores.
- **GeneticNetworkScore** distributes the gene scores within a genetic network.
- **CombineScores** combines the gene scores of the phenotype, metabotype and genotype analysis into an overall result.

The next sections present and explain the algorithms and settings of these tools in more detail.

2.3.1 Phenomizer for PhenoDis

The phenotype analysis of PheNoBo starts with Phenomizer for PhenoDis [39]. The tool requires two inputs: the disease data set (section 2.1.1) and a set of symptoms describing the phenotype of the patient of interest (section 2.2). The main task of Phenomizer for PhenoDis

is to compare the patient's symptoms with every disease in the disease data set. For this purpose Phenomizer for PhenoDis implements and extends the Phenomizer algorithm [66] of Köhler *et al*.

Basically, the Phenomizer algorithm compares symptoms pairwise and evaluates their similarity by applying the following formula [66]:

$$\mathrm{sim}(s_1, s_2) = \max_{a \in \mathrm{ancestors}(s_1, s_2)} \left\{ -\log \left(\frac{\#\text{diseases annotated to } a}{\#\text{all diseases}} \right) \right\} \tag{2.2}$$

s_1 and s_2 denote the symptoms to compare. # is an abbreviation for 'number of'. The term ancestors(s_1, s_2) refers to the set of common parent terms of s_1 and s_2. The final result of the formula sim(s_1, s_2) is a similarity score. The similarity score is a positive number with high values indicating a high similarity between the symptoms. For the exemplary terms depicted in figure 3, the set ancestors('Abnormal iron deposition in mitochondria', 'Decreased plasma free carnitine') contains the terms 'Abnormality of mitochondrial metabolism' and 'Abnormality of metabolism/ homeostasis'. Consequently, the similarity of 'Abnormal iron deposition in mitochondria' and 'Decreased plasma free carnitine' (figure 3) is equal to $\max\{-\log(67/8\,263), -\log(2\,375/8\,263)\} = -\log(67/8\,263) \approx 4.8$

Phenomizer for PhenoDis uses equation 2.2 to compare the set of the patient's symptoms P with the set of symptoms of a disease D yielding a final similarity score sim(P, D):

$$\mathrm{sim}\,(P \to D) = \frac{1}{|P|} \left(\sum_{s_1 \in P} \mathrm{weight}(D, s) \max_{s_2 \in D} \mathrm{sim}(s_1, s_2) \right) \tag{2.3}$$

$$\mathrm{sim}\,(D \to P) = \frac{1}{|D|} \left(\sum_{s_1 \in D} \max_{s_2 \in P} \mathrm{sim}(s_1, s_2) \right) \tag{2.4}$$

$$\mathrm{sim}\,(P, D) = \frac{\mathrm{sim}\,(P \to D) + \mathrm{sim}\,(D \to P)}{2} \tag{2.5}$$

The symbol s in equation 2.3 is an abbreviation for $s = \mathrm{argmax}_{s_2 \in \mathrm{D}} \mathrm{sim}(s_1, s_2)$. Hence, s denotes the symptom of the disease that maximizes the similarity score to the current symptom s_1 of the patient in the summation of equation 2.3. The term weight(D, s) reflects the frequency of the symptom s in patients with the disease D. The exact translation of frequency annotations into weights weight(D, s) is given in table 5. The expression $|D|$ describes the number of symptoms annotated to the disease D and $|P|$ gives the total number of the patient's symptoms.

Frequency Annotation	Frequency Range	Weight
Occasional	]0, 0.25]	0.5
Frequent	]0.25, 0.75]	1.0
Very Frequent	]0.75, 1]	1.5

Table 5: Translation of symptom-disease frequencies into weights. The symptom-disease pairs are grouped according to their annotated frequencies into three groups (column Frequency Annotation). The choice of groups is based on the Orphanet frequency annotations (Orphanet release from August 2015) [39]. The column Weights gives a weight for every group. The weight is plugged in for the expression weight(D, s) in equation 2.3. The data of the table is directly taken from [39].

Equations 2.4 and 2.5 are directly adopted from [66]. Equation 2.3 was derived from the original Phenomizer algorithm [66] by adding the term weight(D, s).

Phenomizer for PhenoDis calculates a similarity score sim(P, D) for every disease in the disease set (section 2.1.1). These similarity scores are translated into p values giving the probability of observing at random a score at least as large as the actual score [66]. The p values are estimated by accessing an empiric score distribution sim(R, D) that was sampled from 100 000 random sets of symptoms R for every disease D and every size of R between one and ten [39]. The resulting p values are corrected for multiple testing with the Benjamini-Hochberg procedure [13].

In summary, Phenomizer for PhenoDis outputs a similarity score and a p value for every disease of the disease data set (section 2.1.1). The similarity score indicates the similarity between the disease and the phenotype of the patient and the p value represents the significance of this score. This result is passed on the PhenoToGeno procedure.

2.3.2 ScoreMetabolites

The metabotype analysis of PheNoBo starts with the tool ScoreMetabolites. ScoreMetabolites evaluates the metabolite measurements of a patient with respect to the reference metabolite levels provided by a control group (section 2.2).

The algorithm of ScoreMetabolites consists of two parts: the construction of a reference set and the calculation of metabolite scores. The construction of a reference set is a

preprocessing step that is done once per control group. The resulting reference set is then used repeatedly to score the metabolites levels measured for the patient under consideration.

2.3.2.1 Construction of the Reference Set

The reference set is a collection of metabolites and their expected levels in plasma samples from a healthy population. These expected levels are derived from metabolite measurements of a control group. The procedure to construct the reference set has to account for the experimental setup that was used for the measurements. Therefore, the construction of the reference set is not part of the KNIME implementation of ScoreMetabolites. The procedure is realized in R [81] and requires specific adaption for each control group.

PheNoBo was applied with two reference sets. Both sets are determined from non-targeted metabolomics measurements done with a platform of Metabolon (section 2.2.3 and 2.2.4). The algorithm presented below is specifically designed for such non-targeted metabolomics measurements.

Four steps are necessary to build the reference sets for ScoreMetabolites:

- Normalization of the metabolite concentrations
- Introduction of phenotype groups
- Handling of missing values
- Summary of the metabolites

These procedures are explained in the following paragraphs.

Normalization of the Concentrations: The samples are normalized so that the median of each metabolite in all samples of the same run day is equal to one. Furthermore, all concentrations are transformed by taking the decadic logarithm. The application of the logarithm ensures that the concentrations of a metabolite follow approximately a normal distribution [90].

Introduction of Phenotype Groups: Recent studies demonstrated that metabolite concentrations are dependent on factors like sex [77], age [75] and fasting state [68]. To account for this in PheNoBo, the samples of the control group are divided into phenotype groups. Table 4 defines the phenotype groups that are used in this work. Note that the definition

of the phenotype group is dependent on the composition of the control group. The current grouping considers the age and fasting state of the controls. Unfortunately, the number of samples used in this thesis is not sufficient to include in addition the gender of the controls.

Handling of Missing Values: Not all metabolites were measured in each of the samples. The missing measurements are due to the fact that the affected metabolites are not present in the sample or due to technical effects [32]. The aim of this work is to gain as much usable information from the control samples as possible. Consequently, all metabolites with missing concentrations in less than 70% of the samples were imputed with the R package MICE [108]. MICE was run ten times with default options and the imputed values were chosen as the average over all runs. The work of [32] demonstrated that the MICE (Multivariate Imputation by Chained Equations) can handle metabolites with up to 70% missingness without introducing artificial correlations. The remaining metabolites with more than 70% missing concentrations are treated as binary information, i.e. missing values are replaced by 0 and measured concentrations by 1. The binary information is also valuable because the pattern of missing concentrations potentially differs between healthy people and people with mitochondrial diseases [90].

Summary of the Metabolites: The last step of data preparation is to summarize the metabolite concentrations into a single table. The summary table serves as reference set for ScoreMetabolites. Table 6 shows five sample entries of the final table for the metabolites pyruvate and carnosine. There are two types of entries in the summary table: entries of type *concentration* and of type *binary*.
The entries of type *concentration* belong to metabolites whose concentrations were imputed in the previous step (e.g. pyruvate in table 6). There are four entries for such metabolites, one for each phenotype group. Each entry provides three values: the mean, the standard deviation and the missingness. The mean and the standard deviation are calculated from the metabolite concentrations across all samples of the respective phenotype group. The column missingness gives the fraction of missing values in all samples before the imputation.
A binary-encoded metabolite (e.g. carnosine in table 6) is transformed into a summary entry of type *binary*. A binary entry reports only the metabolite's missingness across all samples of the control group.

The four steps were applied to the control group of MitoNET (section 2.2.4) and to the KORA F4 cohort (section 2.2.3).

The steps for constructing the MitoNET reference set followed exactly the procedure explained above. The MitoNET reference set comprises expected metabolite levels for 456 metabolites of type concentration and 146 metabolites of type binary.

Metabolon ID	Name	Type	Group	Mean	Standard Deviation	Missingness (in %)
M42582	pyruvate	concentration	1	0.47	0.51	1.0
M42582	pyruvate	concentration	2	0.22	0.47	1.0
M42582	pyruvate	concentration	3	−0.22	0.36	1.0
M42582	pyruvate	concentration	4	−0.06	0.35	1.0
M01768	carnosine	binary	–	–	–	94.8

Table 6: Extract from the reference set of ScoreMetabolites. The table contains the information resulting from summarizing concentrations of 97 metabolomics samples from the MitoNET control group (section 2.2.4). The table shows two types of entries. Rows of type concentration provide information about metabolites with less than 70% missing values. Rows of type binary provide information about metabolites with more than 70% missing values.

The determination of the KORA reference set required adapting the four steps. The following changes in the algorithm are due to the differences in size and scope between the KORA and MitoNET data.

- Normalization of the Concentrations: The measurements from the KORA F4 samples provided for this work were already normalized. The normalization procedures applied on the samples of KORA F4 are described in [91]. The metabolite concentrations were normalized according to run day and logarithmized as explained above. There was an additional processing step that removed outliers (i.e. concentrations that differ more than 4 standard deviations from the mean of the metabolite) from the measurements.

- Introduction of Phenotype Groups: All KORA samples are taken from fasting adults and are assigned to the corresponding phenotype group of fasting adults (group 3).

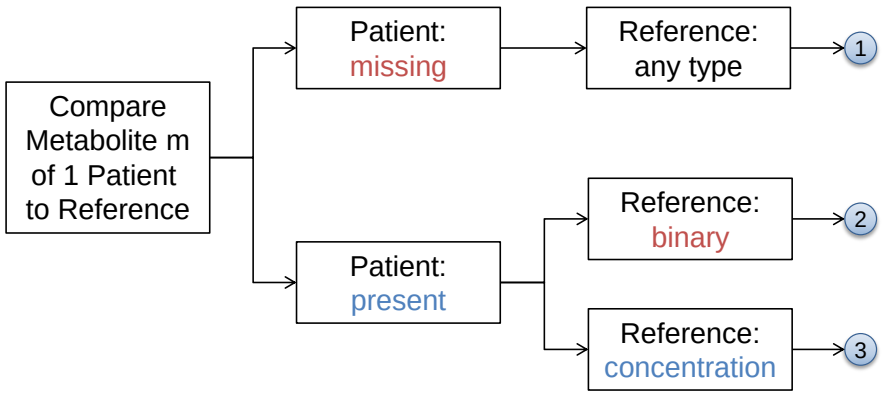
- Handling of Missing Values: The application of an imputation tool on metabolites with less than 70% missingness was left out because of the large number of samples. If there are 70% missing values (corresponding to 1 238 samples) for a metabolite, there are still 530 measured concentrations. These 530 values provide a reasonable basis to derive the expected concentrations of the metabolite.

- Summary of the Metabolites: The summary step was not changed.

The KORA reference set consists of 51 metabolites of type binary and 466 metabolites of type concentration.

2.3.2.2 Calculation of Metabolite Scores

The core algorithm of ScoreMetabolites is concerned with the comparison of the metabotype of the patient and the reference set. The algorithm is derived from [44] and is outlined in figure 5.



Case	Type of Score	Score	Probability
1	binary	0	missingness(m)
2	binary	1	1-missingness(m)
3	z score	z(m)	$2 \cdot P(Z \geq \lvert z(m) \rvert)$

1

Figure 5: Overview of the algorithm of ScoreMetabolites. ScoreMetabolites calculates a score and a probability for each metabolite of the patient's metabotype. The algorithm is based on three cases shown as circles with numbers. The upper part of the figure displays a decision tree for choosing the appropriate case for the current metabolite. The lower part of the figure gives the resulting metabolite score and probability for every case. The term $z(m)$ in table (case 3) denotes the z score of metabolite m which results from the application of formula 2.6.

ScoreMetabolite determines a score and a probability for every metabolite of the metabolite

set (section 2.1.2). The algorithm implements three different methods to calculate these values, henceforth named case 1, case 2 and case 3.

The choice of the case depends on the missing values of the current metabolite m in the reference data set and in the measurement of the patient (top of figure 5). The metabolites of the reference data are divided into two types: metabolites of type binary have a high missingness (>70%) and metabolites of type concentration have a low missingness (≤70%) (section 2.3.2.1). Furthermore the concentration of a metabolite can be present or missing in the measurement of the patient.

ScoreMetabolites is able to calculate different types of scores and probabilities depending on the current case (bottom of figure 5). There are two types of scores: binary scores and z scores. A binary score is computed if it is not possible to directly compare the measurement and reference values because of the missing values. Otherwise ScoreMetabolite calculates a z score (equation 2.6 below). The probability depends on the score and indicates the likelihood of observing the actual score or a more extreme score at random. The paragraphs below give a description of the three cases.

Case 1: The metabotype of the patient contains a missing value for the current metabolite m. It is not possible to compare a missing value to the reference metabolite concentrations. Therefore ScoreMetabolite applies a binary scoring scheme: the score of the current metabolite is set to 0 indicating a missing value. The expected probability of observing a missing value for m in the patient's measurement (and hence of getting a score of 0) is approximated by missingness(m), i.e. the missingness of the metabolite m in the reference data set.

Case 2: The current metabolite m was measured for the patient (i.e no missing value) and the reference values of m are of type binary. In this case there are not sufficient measurements in the reference data set to make a reliable comparison based on concentration values. The metabolite gets a binary score of 1 indicating that there is no missing value in the measurement for the patient. The probability of observing no missing value for m in the patient (and hence of getting a score of 1) is estimated by $1 - \text{missingness}(m)$.

Case 3: The current metabolite m was measured in the patient's sample and the reference values of m are of type concentration. This case allows a direct comparison between the patient and the reference data. ScoreMetabolite calculates a z score $z(m)$ for the current metabolite m [44]:

$$z(m) = \frac{\text{concentration}(m) - \text{mean}(m)}{\text{standard deviation}(m)} \tag{2.6}$$

concentration(m) denotes the measured concentration of m for the patient. mean(m) and

standard deviation(m) describe the expected concentration of m in the reference data. The resulting z score $z(m)$ follows a normal distribution with mean 0 and standard deviation 1 as the original metabolite levels are approximately normally distributed (after logarithmization). [90, 44]. This fact allows the direct calculation of the probability of obtaining a more extreme z score Z than the observed score $z(m)$.

$$\text{probability}(m) = 2 \cdot P(Z \geq |z(m)|) = 2 \cdot (1 - \text{pnorm}\,(|z(m)|, 0, 1)) \tag{2.7}$$

The function pnorm(x, a, b) gives the probability of drawing a value lower than x from a normal distribution with mean a and standard deviation b. PheNoBo uses the pre-implemented pnorm function of the Java library Apache Commons Math version 3.6.1 [28].

The results for the patient's metabolites can be ranked according to the reported probabilities. A low probability indicates that the metabolite differs from the reference values and might be involved in the patient's disease. These probabilities are passed on to the MetaboToGeno tool.

2.3.3 PhenoToGeno and MetaboToGeno

The tool PhenoToGeno is the second step of the phenotype analysis and works on the results of Phenomizer for PhenoDis (section 2.3.1). MetaboToGeno is the successor of ScoreMetabolites (section 2.3.2) in the metabotype analysis and processes the results of ScoreMetabolites. PhenoToGeno and MetaboToGeno are grouped together in this section as they apply the same algorithm on different entities. PhenoToGeno applies the procedure to diseases and MetaboToGeno uses it on metabolites. Consequently the term "entity" refers to "disease or metabolite" in the following explanation of the algorithm.

The algorithm of PhenoToGeno and MetaboToGeno transforms the results of the respective predecessor tool into gene scores. Three types of information are required to derive the gene scores:

- the result of the predecessor tool in PheNoBo that provides scored entities
- a mapping of entities to genes which is covered by the data sets of section 2.1.4 and section 2.1.5.
- a set of all genes to be scored (section 2.1.3)

This information is processed in two steps, which are adapted from Phen-Gen [56]. The first step, conversion of probabilities, transforms the output of Phenomizer for PhenoDis or ScoreMetabolites. The scoring of genes is the second step. It distributes the results of the previous step among all genes. The calculations used in these steps are given below.

Conversion of Probabilities: The conversion of probabilities works on the probability p_e associated with an entity e. In the metabotype analysis p_e stands for the probability of metabolite e reported by ScoreMetabolites. If the conversion of probabilities is applied in the phenotype analysis, p_e is the p value of disease e in the result of Phenomizer for PhenoDis. p_e is transformed into an intermediate score t_e for entity e with the following formula (derived from [56]):

$$t_e = \frac{1}{1 + \#\text{entities} \cdot p_e} \tag{2.8}$$

#entities means the total number of entities. If the entities are diseases, #entities is equal to the number of diseases in the disease data set (section 2.1.1). If #entities is used in the metabotype analysis, the expression refers to the number of metabolites in the metabolite data set (section 2.1.2). The transformed score t_e of a disease e is proportional to the conditional probability of the disease being causal given that the disease is similar to

the patient's symptoms [56]. Similarly t_e of a metabolite e is related to the conditional probability of the metabolite being causal given the patient's metabotype.

Scoring of Genes: The scoring of genes is concerned with the transfer of t_e to genes using the entity-gene pairs from the input. A gene g gets an annotation of t_e if it is associated with entity e. If an entity e^* is not associated with any gene, its score t_{e^*} is distributed among all genes of the gene set (section 2.1.3) [56]. Every gene gets an annotation of $\frac{t_{e^*}}{n}$ where n is the total number of genes in the genetic data set.
After the application of these rules every gene g ends up with annotated scores from more than one entity. The algorithm of PhenoToGeno and MetaboToGeno provides two modes for resolving the annotations into a single score per gene which is denoted by score(g):

- The Multiple Annotation Mode combines all annotated scores t_i of gene g (abbreviated as set annotation(g)) [56].

$$\text{score}(g) = 1 - \prod_{t_i \in \text{ annotation}(g)} (1 - t_i) \tag{2.9}$$

- The Maximum Annotation Mode keeps only the maximum score of all annotated scores t_i.

$$\text{score}(g) = \max_{t_i \in \text{ annotation}(g)} t_i \tag{2.10}$$

The final result of the algorithm is a score for each gene. This score ranges from 0 to 1 and represents a relative and unnormalized probability that the gene is causal for the patient's phenotype or metabotype. These scores are passed to the GeneticNetworkScore tool of PheNoBo, which is the subject of the next paragraphs.

2.3.4 GeneticNetworkScore

The third and final part of the metabotype and phenotype analysis of PheNoBo is covered by the tool GeneticNetworkScore. The aim of the program is to refine the gene scores of MetaboToGeno and PhenoToGeno by spreading them in a genetic network. This step increases the score of genes that interact with many high-scoring genes and therefore promotes the detection of novel disease-gene relations. GeneticNetworkScore implements a random walk with restart on the genetic network presented in section 2.1.6. The implementation follows the procedures described in [65] and [94].

A random walk with restart is a widely used model to evaluate the importance of nodes within a network [64]. Figure 6 gives an example of a random walk with restart. A random walk is an arbitrarily chosen path through the network [88]. The path starts at a randomly selected node. In each step of the walk one moves along an edge from the current node to another node. The inclusion of a restart into the random walk allows to return to the start node in every move. The restart is possible from every node even though the current node is not connected to the start node.

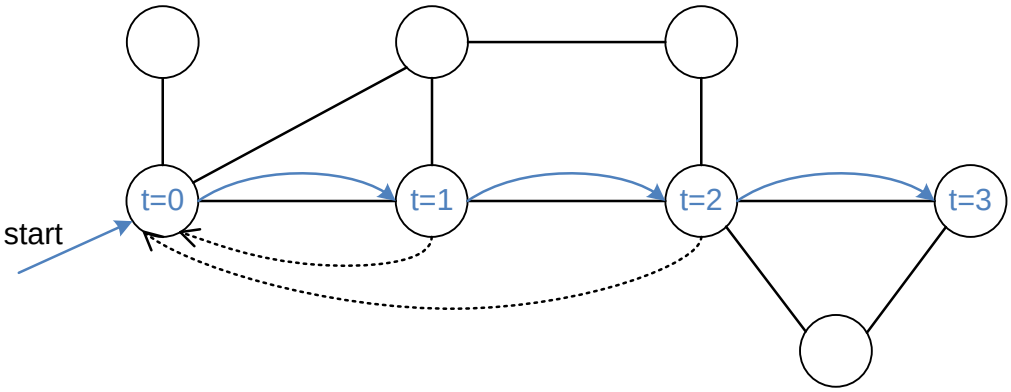


Figure 6: Example of a random walk with restart. The figure shows a small network consisting of eight nodes (circles) and ten edges (solid lines). The random walk with restart is depicted by the blue arrows. A random walk with restart is a randomly chosen path through a network. The steps of the walk are numbered with $t = 0$ to $t = 3$. $t = 0$ is the starting point of the walk. At $t = 0$ a node from the network is chosen as start which is indicated by the blue arrow labeled with "start". At $t = 1$ to $t = 3$ the walk moves from one node to another by following a randomly selected edge. The dashed arrows indicate the possibility of a restart. A restart allows to return to the start node in every step of the walk.

There is an iterative formula for calculating the probabilities of arriving at a certain node after $t + 1$ random steps from node to node [65].

$$s_{t+1} = (1 - r)Ms_t + rs_0 \tag{2.11}$$

s_t is a vector with an entry for every node n containing the probability of being at node n after t steps. s_0 is an initial probability distribution for every node n giving the probability of choosing n as start of the path. r models the probability of a restart, i.e. to jump back to the first node of the path. M is a square matrix describing the structure of the network. There is an entry m_{ij} in M for every pair of nodes i and j that gives the probability of moving from j to i. If one applies the formula until convergence (i.e. s_t and s_{t+1} do not differ significantly

any more), one obtains a final probability for every node. A node with a high probability is an important frequently-visited node of the network.

GeneticNetworkScore uses this model to re-evaluate the gene scores of MetaboToGeno and PhenoToGeno. It applies formula 2.11 iteratively. The subsequent paragraphs explain how GeneticNetworkScore models the variables of the random walk with restart.

Initial Probabilities s_0: GeneticNetworkScore uses s_0 as a vector with an entry for every gene g. The entries of s_0 represent the normalized gene scores of MetaboToGeno or PhenoToGeno. The normalization is done by the following formula

$$s_0(g) = \frac{\text{score}(g)}{\sum_{\text{all genes } g'} \text{score}(g')} \tag{2.12}$$

$s_0(g)$ denotes the entry of s_0 for gene g and score(g) refers to the original score of g in the output of MetaboToGeno or PhenoToGeno. Formula 2.12 is required to obtain a valid probability distribution such that all entries of s_0 sum up to 1.

Restart Probability r: The parameter r is used to control the fraction of the gene scores that is distributed within the network. This means that each gene keeps a base score of $r \cdot s_0(g)$ while the remainder $(r-1) \cdot s_0(g)$ is distributed [56].

Transition Matrix M: M is derived from the adjacency matrix of the genetic network of section 2.1.6. An entry m_{ij} of M gives the fraction of the score that is passed from gene j to gene i in every iteration of formula 2.11. M is a stochastic matrix. All entries of such a matrix have values between 0 and 1 and all columns sum up to 1. GeneticNetworkScore provides two ways to calculate M from the genetic network.
If one does not consider the edge weights of the network, m_{ij} is derived by

$$m_{ij} = \begin{cases} \dfrac{1}{\left|\text{neighbors}(j)\right|} & \text{if there is an edge between } i \text{ and } j \\ 0 & \text{else} \end{cases} \tag{2.13}$$

neighbors (j) refers to the set of neighbors of j, i.e the set of nodes which are connected to j via an edge. $\left|\text{neighbors}(j)\right|$ denotes the number of neighbors of j.
If one includes the edge weights of the network, GeneticNetworkScore applies

$$m_{ij} = \begin{cases} \dfrac{\text{weight}(i,j)}{\sum_{k \in \text{ neighbors}(j)} \text{weight}(k,j)} & \text{if there is an edge between } i \text{ and } j \\ 0 & \text{else} \end{cases} \tag{2.14}$$

weight(i, j) denotes the weight of the edge between gene i and j as described in section 2.1.6.
GeneticNetworkScore does not calculate and store M directly as two-dimensional array. Instead GeneticNetworkScore uses a data structure named *coordinate form* which stores all non-zero entries of M in an array of triples (i, j, m_{ij}) [96]. This data structure reduces the computation time and main memory usage of GeneticNetworkScore.

Resulting Probabilities s_t: s_t is a vector of scores for each gene. $s_t(g)$ gives the resulting score of gene g after t iterations of formula 2.11. Parameter t controls how far the score is spread within the network. For example, if $t = 2$ the score of a node is passed on to all nodes within a distance of two edges. GeneticNetwork score is able to calculate s_t for any number of steps t. It is also possible to approximate s_t for $t \to \infty$ by repeating formula 2.11 until s_t and s_{t+1} differ by less than 10^{-12} (using the maximum norm [20] of $s_{t+1} - s_t$ for comparison).

The output of GeneticNetworkScore translates $s_t(g)$ into an enrichment score $e(g)$ which facilitates the interpretation of the results from GeneticNetworkScore. The enrichment score is calculated by

$$e(g) = -\log_{10}\left(s_t\left(g\right) \cdot n\right) \tag{2.15}$$

n denotes the total number of genes. The enrichment score compares $s_t(g)$ against a uniform distribution of scores where each gene gets a score of $\frac{1}{n}$. If $e(g)$ is positive, the score of gene g is higher than expected for a (random) uniform distribution of scores. If $e(g)$ is negative, the score of g is lower than expected at random.

GeneticNetworkScore outputs these enrichment scores $e(g)$ together with the adjusted gene scores $s_t(g)$ for every gene. PheNoBo uses two instances of GeneticNetworkScore (figure 2). The first instance produces a set of scores based on the analysis of the phenotype whereas the second instance is applied to evaluate the metabotype. The next section will cover how the results of the different instances of GeneticNetworkScore are joined into a final prediction.

2.3.5 CombineScores

CombineScores is the last tool in the workflow of PheNoBo. The purpose of CombineScores is to merge the individual predictions of PheNoBo based on the phenotype and metabotype of the patient. CombineScores is also able to integrate genotype-based results from an external pipeline.

CombineScores applies the following formula to generate a single score combi_raw(g) for every gene g from these individual predictions:

$$\text{combi\_raw}(g) = \frac{\text{g}(g) \cdot \text{p}(g) \cdot \text{m}(g)}{\text{g}(g) \cdot \text{p}(g) \cdot \text{m}(g) + (1 - \text{g}(g)) \cdot (1 - \text{p}(g)) \cdot (1 - \text{m}(g))} \tag{2.16}$$

g(g) is the genotype-based score of g. p(g) denotes the score of g that is derived from the phenotype. m(g) refers to the result of the metabotype analysis. p(g) and m(g) are directly taken from the output of the respective instance of GeneticNetworkScore. g(g), p(g), m(g) are normalized probabilities, i.e. all gene scores of a prediction type sum up to one.

Formula 2.16 extends the combination procedure of Phen-Gen [56] to include the results of the metabotype analysis. The formula is derived from Bayes' formula [88] by assuming that g(g), p(g) and m(g) are independent probabilities [56]. Hence, the result combi_raw(g) gives the probability that g is causal for the patient's condition given the observed phenotype, metabotype and genotype.

CombineScores adapts formula 2.16 automatically depending on the input as one may join two instead of three predictions. For example, if there is no genotype-based prediction, formula 2.16 is simplified to

$$\text{combi\_raw}(g) = \frac{\text{p}(g) \cdot \text{m}(g)}{\text{p}(g) \cdot \text{m}(g) + (1 - \text{p}(g)) \cdot (1 - \text{m}(g))} \tag{2.17}$$

The result combi_raw(g) is normalized analogous to equation 2.12:

$$\text{combi}\,(g) = \frac{\text{combi}\,(g)}{\sum_{\text{all genes } g'} \text{combi}\,(g')} \tag{2.18}$$

CombineScores translates combi(g) into an enrichment score by applying formula 2.15 on combi(g) for $s_t(g)$. The enrichment score derived from combi(g) can be interpreted the same way as explained in section 2.3.4.

The final output of PheNoBo is a ranking of genes. CombineScores produces this ranking by sorting the genes according to combi(g) in descending order. Genes with low ranks and consequently high scores are most likely causative of the patient's disease.

2.4 Evaluation and Application of PheNoBo

The modules of PheNoBo were evaluated and applied on the patient data introduced in section 2.2 (summarized in table 1). The evaluation of PheNoBo aims to find the best setup of PheNoBo and to estimate the accuracy of the predictions. The application of PheNoBo on patient data without known causal genes tries to make diagnostic suggestions.

First, section 2.4.1 introduces the options of PheNoBo that were included into the optimization of PheNoBo. The second part (section 2.4.2) presents the evaluation criteria used to assess the predictions of PheNoBo. The last two sections explain the purpose and the analyses of the individual patient data set listed in table 1. The patient data sets 1, 2, 3, 4, 6 and 7 were used to evaluate and optimize the individual steps of PheNoBo (section 2.4.3). PheNoBo was applied on the remaining data sets 5 and 8 to learn more about the patients of those sets (section 2.4.4).

2.4.1 Options of PheNoBo

The algorithms of PheNoBo (except for CombineScores) provide several options. The following paragraphs explain the options and the specific values tried for them in the evaluation (section 2.4.3).

Output of Phenomizer for PhenoDis – Filtering of diseases: This option controls the interface of Phenomizer for PhenoDis and PhenoToGeno. Phenomizer for PhenoDis yields a score and a p value for every disease in PhenoDis (section 2.1.1). Only the diseases that are selected by the filter are passed from Phenomizer for PhenoDis to PhenoToGeno. PheNoBo offers 3 kinds of filters:

- Top20 selects the 20 diseases that are most similar to the patient's phenotype.
- SignificantP selects all diseases with a p value of less than 0.05.
- TopP selects the diseases with the best (i.e. lowest) p value.

The filter option is realized with the Row Filter node of KNIME.

ScoreMetabolites – Reference Set: ScoreMetabolites requires a set of reference metabolite values for calculating metabolite scores (section 2.3.2). Currently, there are two reference sets available: the KORA reference set derived from the KORA F4 cohort (section 2.2.3)

and the MitoNET reference set based on the MitoNET control group (section 2.2.4).

PhenoToGeno and MetaboToGeno – Gene annotation mode: As explained in section 2.3.3, PhenoToGeno and MetaboToGeno provide two modes to resolve annotations from multiple diseases or metabolites per gene. Multiple Annotation Mode combines the scores of all diseases or metabolites that are associated with a gene (equation 2.9). Maximum Annotation Mode considers only the highest score of all diseases or metabolites annotated to a gene (equation 2.10).

GeneticNetworkScore: GeneticNetworkScore has three parameters to adjust.

- **Edge weights:** This option is a binary option. It decides whether GeneticNetworkScore includes the edge weights of the network of section 2.3.4 into the random walk with restart. This option influences the calculation of the transition matrix M of equation 2.11. If the edge weights are taken into account, M is calculated using equation 2.14. Otherwise, formula 2.13 is applied.

- **Restart Probability:** The restart probability controls the fraction of gene scores that is distributed in the random walk with restart of GeneticNetworkScore (section 2.3.4). Two different values were tested for this option: 0.5 (50% of the score is distributed) and 0.9 (10% of the score is distributed).

- **Iterations:** The option iterations controls the number of steps of the random walk with restart in GeneticNetworkScore (section 2.3.4). GeneticNetworkScore was either run for two iterations or for an unlimited number of iterations (i.e. the random walk continues until the scores do not change considerably anymore).

Finally, table 7 lists the default parameters of PheNoBo that were used for the application of PheNoBo on real patient data (section 2.4.4). The choice of these options is established and explained in sections 3.2 and 3.3.

Data Type	Tool	Option	Value
Phenotype	Phenomizer & PhenoToGeno	Filtering of Diseases	No Filter
	PhenoToGeno	Annotation Mode	Multiple
	GeneticNetworkScore	Edge Weights Restart Probability Iterations	Yes 0.9 Unlimited
Metabotype	ScoreMetabolites	Reference Set	MitoNET
	MetaboToGeno	Annotation Mode	Maximum
	GeneticNetworkScore	Edge Weights Restart Probability Iterations	Yes 0.9 Unlimited

Table 7: Default options of PheNoBo. The table provides the default options for every tool of PheNoBo. The tools are group according to the data type they analyze (column Data Type). Note that each option of GeneticNetworkScore is listed twice as the tool is applied once during the phenotype analysis and once during the metabotype analysis. Annotation Mode = Gene Annotation Mode, Multiple = Multiple Annotation Mode, Maximum = Maximum Annotation Mode.

2.4.2 Evaluation Criteria

The evaluation of PheNoBo basically used three criteria to assess the predictive performance of the different steps of PheNoBo. Each module of PheNoBo produces a sorted list of scored objects (diseases, metabolites or genes). The list is sorted according to scores such that objects with a high probability of being causal are placed on top of the list. All three evaluation criteria are applied on these sorted lists. The subsequent paragraphs present these criteria in more detail.

Rank: The quality of most predictions is evaluated by considering the rank of the causal gene in a list of scored genes. The rank of a gene g in this list is calculated by the following formula:

$$\text{rank}(g) = \frac{\text{first}(g) + \text{last}(g)}{2} \tag{2.19}$$

first(g) is the lowest position of a gene with a score equal to the score of g. Similarly, last(g) denotes the highest position of a gene with a score equal to the score of g. The rank defined in formula 2.19 accounts for the fact that there can be several genes with identical scores. All genes with identical scores get the same rank, which is equal to the average of their positions in the list.
Some of the patients (in the patient data set 1 and 2) have more than one causal gene. The rank of each causal gene is determined separately with the formula above. The resulting ranks for all causal genes of a patient are summarized into a single value by taking the average rank of all causal genes.
This procedure can be applied analogously on a list of sorted diseases or metabolites.

Sensitivity: The sensitivity is a metric for summarizing the individual predictions of a patient data set into a single value. For calculating the sensitivity one first has to determine the rank of the causal gene for every patient in the patient data set (see above). The sensitivity is then defined as the fraction of patients whose causal gene has a rank of at most a certain value x:

$$\text{top}x\text{\_sensitivity} = \frac{|\text{rank} \leq x|}{|\text{Patients}|} \tag{2.20}$$

|Patients| denotes the total number of patients in the patient data set under consideration. $|\text{rank} \leq x|$ corresponds to the number of patients whose causal gene has a rank of at most x. x is an arbitrarily chosen threshold. The evaluation of PheNoBo is based on the top10, top20, top30 and top50 sensitivity.
Again one can also calculate the sensitivity from the ranks of causal diseases or metabolites.

Set of Deviating Metabolites: This evaluation criterion is a special criterion for evaluating and comparing the results of ScoreMetabolites. The result of ScoreMetabolites is a list of scored metabolites sorted according to the probabilities of the metabolites (section 2.3.2.2). Low probabilities hint at metabolites whose measured concentrations strongly deviate from the expected values. Therefore, a deviating metabolite is defined as a metabolite with a probability of less than 0.05. The set of deviating metabolites for a patient is the set of all metabolites with a probability of less than 0.05 in the results of ScoreMetabolites for the patient's metabotype.

2.4.3 Evaluation of PheNoBo

This section describes the steps to evaluate and optimize PheNoBo on the patient data sets 1, 2, 3, 4 and 6. It outlines the combination of options (section 2.4.1) and the evaluation

criteria (section 2.4.2) used on each data set.

The section is structured according to the part of the PheNoBo under consideration. Phenomizer for PhenoDis is assessed on data set 3 (section 2.4.3.1). The evaluation of the phenotype analysis relies on the data sets 1, 2 and 4 (section 2.4.3.2). Data set 6 is used to compare ScoreMetabolites on different metabolite reference sets (section 2.4.3.3). Finally, PheNoBo is evaluated on different combinations of the data types provided by the patient data set 7 (section 2.4.3.4).

2.4.3.1 Evaluation of Phenomizer for PhenoDis

The aim of the evaluation of Phenomizer for PhenoDis is to compare different implementations of the Phenomizer algorithm: the original Phenomizer of [66], Phenomizer for PhenoDis from August 2015 (presented in [39]) and Phenomizer for PhenoDis from February 2016 (section 2.3.1). The implementations are assessed using the patient data set 3 derived from the literature (section 2.2.2).

Data Set 3: The patient data of set 3 had been used previously to test the performance of Phenomizer for PhenoDis [39]. Phenomizer for PhenoDis, which now runs on a new release of PhenoDis, was re-evaluated on this set. For this re-evaluation Phenomizer for PhenoDis from 2015 and from 2016 as well as the original Phenomizer were applied on the patients of the data set. Then the rank of the disease of the patient (given in table 3) in the results of the respective Phenomizer algorithm was determined. The analysis of patient data set 3 yielded three ranks per patient, one rank for each variant of Phenomizer. The ranks obtained from each tool were summarized using the top10, top20 and top30 sensitivity (section 2.4.2). These sensitivities are the basis for comparing the different implementations of the Phenomizer algorithm.

2.4.3.2 Evaluation of the Phenotype Analysis

The phenotype analysis consists of the tools Phenomizer for PhenoDis (section 2.3.1), PhenoToGeno (section 2.3.3) and GeneticNetworkScore (section 2.3.4). The evaluation of the phenotype analysis is based on the data sets 1, 2 and 4. The simulated patient data (set 1 and 2) are used to test and to compare different options of the algorithms applied in the phenotype analysis. As the patient data set 4 is derived from real patients, the set can help to assess the diagnostic performance of the phenotype analysis.

Patient Data Set 1: Data set 1 comprises the phenotype data of 1 400 simulated patients suffering from defined diseases (section 2.2.1). The virtual patients were analyzed 16 times with PheNoBo, each time with a different combination of the options of PhenoToGeno and GeneticNetworkScore (section 2.4.1). The evaluation is based on the rank of the causal gene in the list of scored genes produced by GeneticNetworkScore (section 2.4.2). The causal genes of a simulated patient are defined as all genes that are associated with the patient's disease in the disease-gene data set (section 2.1.4). The evaluation resulted in one rank for every patient in every run of PheNoBo. The ranks are used to find the optimal parameters for the phenotype analysis.

Patient Data Set 2: The set consists of the phenotype data of 7 890 simulated patients affected by mitochondrial diseases (section 2.2.1). The purpose of data set 2 is to evaluate a subset of the parameters of PheNoBo more thoroughly on mitochondrial diseases. Patient data set 2 was analyzed in a similar way as data set 1. PheNoBo was run with five different combinations of options on the simulated phenotype data. Data Set 2 produced two ranks per patient and run of PheNoBo: the rank of the causal gene in the predictions of PhenoToGeno and GeneticNetworkScore (section 2.4.2). The ranks of a run were summed up by calculating the top10, top20 and top50 sensitivity for PhenoToGeno and GeneticNetworkScore (section 2.4.2). These results allow a comparison of PhenoToGeno and of GeneticNetworkScore in the context of mitochondrial diseases. Furthermore, the results indicate which one of the five combinations of options leads to the best predictions.

Patient Data Set 4: Data set 4 is based on the symptoms of 15 patients affected by the carnitine-related diseases CACTD and PCD (section 2.2.2). The whole phenotype analysis of PheNoBo was applied on these patients. The analysis was repeated several times with the five combinations of options that were also applied on data set 2. The resulting predictions were evaluated by considering the rank of the causal gene after the application of PhenoToGeno and GeneticNetworkScore. The causal gene for the patients with CACTD is Carnitine/Acylcarnitine Translocase. PCD is caused by a defect in the high-affinity sodium-dependent carnitine cotransporter encoded by the gene SLC22A5. The results consist of two ranks for every patient and every run of the phenotype analysis. Like in data set 2, these ranks were used to determined the top10, top20 and top50 sensitivity of PhenoToGeno and GeneticNetworkScore (section 2.4.2). The aim of patient data set 4 is to estimate the sensitivity of the phenotype analysis and to optimize the options of PheNoBo on real patients.

2.4.3.3 Evaluation of the Metabolite Reference Sets

This section is dedicated to the evaluation of the first step of the metabotype analysis, ScoreMetabolites (section 2.3.2). ScoreMetabolites can be used with two reference sets providing expected metabolite levels: the KORA reference set and the MitoNET reference set (section 2.4.1). The following analysis investigates the exchangeability of the reference sets on patient data set 6.

Patient Data Set 6: Set 6 is a purely metabolic data set. It consists of the metabotype data of the set *MitoNET Metabolite Samples Group 3* (section 2.2.4). The set is made up of samples from fasting adults who suffer from mitochondrial diseases. Score Metabolites is applied on these data with two different reference sets (section 2.4.1): the set derived from the MitoNET controls (section 2.2.4) and the set constructed from participants of the KORA F4 cohort (section 2.2.3). As the KORA F4 cohort consists of fasting adults only, *MitoNET Metabolite Samples Group 3* is used for this evaluation instead of the whole set *MitoNET Metabolite Samples*. Furthermore, the reference sets cover different metabolites. For this reason, the metabotype data of *MitoNET Metabolite Samples Group 3* was reduced to the concentrations of the 273 metabolites that are part of both reference sets.
ScoreMetabolites was run twice for every patient of the set, one run per reference data set. The predictions of ScoreMetabolites were evaluated by considering the sets of deviating metabolites (section 2.4.2). In a final step the deviating metabolites predicted for a patient were divided into three subsets: MitoNET and KORA, KORA only, MitoNET only. MitoNET and KORA contains all metabolites which are deviating metabolites for both reference sets. KORA only comprises the metabolites which are deviating metabolites for the KORA reference set but not for the MitoNET reference set. MitoNET only consists of the deviating metabolites for the MitoNET reference set that are not deviating metabolites for the KORA reference set. These subsets are used to study the effect of the reference set on the predictions of ScoreMetabolites.

2.4.3.4 Evaluation of PheNoBo on Different Combinations of Patient Data

In an ideal setting PheNoBo runs on the genotype, metabotype and phenotype data of the patient of interest. However, it is not alway possible to obtain all three types of data for a patient. Therefore, patient data set 7 is used to evaluate the performance of PheNoBo on different combinations of phenotype, metabotype and genotype data (section 2.2.4).

Patient Data Set 7: This set is the most complete patient data set used in this work. It deals

with 19 patients from the MitoNET register with known etiology, i.e. one has identified and verified the genetic defects that cause the condition of the patient. The data set comprises real phenotype and metabotype data. The genotype data of the set was constructed from the KORA LOF data. There are 106 genotype data sets per patient. Each the genotype data set is analyzed together with the phenotype and metabotype data of the corresponding patient from MitoNET. In total, patient data set 6 requires $106 \cdot 19 = 2\,014$ runs of PheNoBo. Each run of PheNoBo applies the default options given in table 7. The rank of the causal gene of the patient is recorded for every run (section 2.4.2). For a comprehensive analysis the ranks resulting from different steps of PheNoBo were taken into account: phenotype analysis, metabotype analysis and all combinations of phenotype, metabotype and genotype data. As there are 106 runs per patient, a final rank per patient is determined as the mean rank of the 106 runs. This analysis provides an overview of the strengths and weaknesses of all parts of PheNoBo.

2.4.4 Application of PheNoBo

The evaluation of PheNoBo allows to derive the optimal setup of PheNoBo for analyzing and interpreting real patient data. After the evaluation PheNoBo is ready to make diagnostic predictions for real patients. PheNoBo was used to examine patient data from KORA (section 2.4.4.1) and MitoNET (section 2.4.4.2)

2.4.4.1 Application on KORA

The patient data set 5 deals with the data from the subjects of the population-based KORA F4 cohort. As all participants are not expected to suffer from any severe genetic disease, there is no phenotype data for them available. The purpose of the KORA data is to investigate how PheNoBo behaves on data from supposedly healthy people.

Patient Data Set 5: This set consists of 106 participants from the KORA F4 cohort (section 2.2.3). PheNoBo was applied on the LOFs and the metabolite samples of these 106 individuals. The options of PheNoBo were set to the default values for the pipeline (table 7). The analysis of the predictions focuses on the top scoring gene predicted by CombineScores and on the top scoring metabolite of ScoreMetabolites for every patient. These results were inspected manually.

2.4.4.2 Application on MitoNET

The final patient data set is derived from the MitoNET patients without known causal gene (section 2.2.4). The task of PheNoBo is to gain insight into potential causes of the patient's disease.

Patient Data Set 8: Data set 8 comprises the patients from MitoNET for whom the cause of their disease is unknown. PheNoBo was run on the phenotype and metabotype data of these patients to suggest a diagnosis for them. For this purpose the raw results of every step of PheNoBo were recorded for every patient. The results were checked for plausibility and interpreted manually.

3 Results and Discussion

The major result of this thesis is the development of the pipeline PheNoBo. PheNoBo combines phenotype, metabotype and genotype data of a patient to predict the causal gene for the patient's disease. PheNoBo is realized as a workflow in KNIME (figure 7). The pipeline is based on six KNIME nodes:

- **Phenomizer for PhenoDis** (section 2.3.1) calculates a score for each disease based on the patient's phenotype.
- **ScoreMetabolites** (section 2.3.2) scores metabolites by evaluating the metabotype of the patient.
- **PhenoToGeno** (section 2.3.3) transforms disease scores into gene scores.
- **MetaboToGeno** (section 2.3.3) transforms metabolite scores into gene scores.
- **GeneticNetworkScore** (section 2.3.4) distributes the gene scores within a genetic network.
- **CombineScores** (section 2.3.5) combines the gene scores of the phenotype, metabotype and genotype analysis into an overall result.

The KNIME nodes, their source code and the KNIME workflow of PheNoBo are available on the attached DVD (section DVD Content) and at `https://github.com/marie-sophie/mapra`.

This implementation was applied and evaluated on various patient data sets from simulations, from the literature and from real patient cohorts (section 2.2 and table 1). In a first step, I analyzed three different parts of PheNoBo independently from each other: Phenomizer for PhenoDis (section 3.1), the phenotype analysis pipeline consisting of Phenomizer for PhenoDis, PhenoToGeno and GeneticNetworkScore (section 3.2) and ScoreMetabolites

Figure 7: Screenshot of the PheNoBo pipeline in KNIME. The figure shows how the PheNoBo workflow (figure 2) is realized with KNIME nodes.
Yellow nodes = tools of PheNoBo (section 2.3), orange nodes = file readers providing the data sets to run PheNoBo (section 2.1 and section 2.2).

(section 3.3). The aim of these analyses is to find the optimal settings for each step in PheNoBo and to assess the impact of the underlying data sets on the predictions. Further on, I tested the capability of PheNoBo for analyzing real patient data by applying the whole PheNoBo pipeline on a set of MitoNET patients with known causal genes (section 3.4). After deriving and verifying the optimal setup of PheNoBo, I applied the method on the individuals from the KORA cohort and the MitoNET patients with unknown causal gene. The last part of this chapter reports and interprets the resulting predictions of PheNoBo (section 3.5).

3.1 Evaluation of Phenomizer for PhenoDis

The Phenomizer algorithm compares the symptoms of a patient against a set of diseases using the HPO (section 2.3.1). There are three implementations of the Phenomizer algorithm using different disease databases and weighting schemes for formula 2.3 of the algorithm.

- The Original Phenomizer [66] uses disease annotations derived from the OMIM database [46]. The tool applies the algorithm described in section 2.3.1 without considering the weights in equation 2.3 (hence weight$(D, s) = 1$ for all diseases D and symptoms s).
- Phenomizer for PhenoDis 2015 [39] relies on the PhenoDis database which combines the diseases of OMIM [46] and Orphanet [83]. The tool uses the release of PhenoDis from August 2015 and the algorithm described in section 2.3.1.
- Phenomizer for PhenoDis 2016 is identical to Phenomizer for PhenoDis 2015 but uses the recent release of PhenoDis from February 2016.

This evaluation compares the three different implementations of the Phenomizer algorithm (section 2.4.3.2). The aim of the comparison is to find the variant of Phenomizer that is best suited for the application in PheNoBo.

The three implementations of the Phenomizer algorithm were applied on a patient data set derived from the literature that was already used for the evaluation of Phenomizer for PhenoDis in [39]. The data set comprises phenotype data of 21 patients suffering from rare diseases (section 2.2.2) and is henceforth referred to as patient data set 3 (table 1). Each patient of the set is described by a manually curated set of symptoms. The application of each variant of Phenomizer on the symptoms of a patient yielded a sorted list of scored

diseases. The predictions are assessed in terms of the rank of the patient's causal disease (i.e. the position of the disease within the list of scored diseases). The ranks are summarized into three sensitivities for each variant of the Phenomizer algorithm: the top10 sensitivity, the top20 sensitivity and the top30 sensitivity (i.e. the fraction of causal diseases with a rank of at most 10, 20 and 30).

Phenomizer for PhenoDis 2016 performs best on patient data set 3 (figure 8). The causal diseases of 17 patients (corresponding to 81% of all patients) got a rank of 20 or less in the predictions of Phenomizer for PhenoDis 2016. Both variants of Phenomizer for PhenoDis have a higher sensitivity than the original Phenomizer when considering the fraction of predictions with a rank of at most 20 (81% and 71% for Phenomizer for PhenoDis vs. 62% for the original Phenomizer).

The comparison shows that Phenomizer for PhenoDis 2016 is the most sensitive variant for patient data set 3. This variant of the Phenomizer algorithm produces the most accurate disease scores for PheNoBo. There might two reasons for the increased performance of Phenomizer for PhenoDis 2016 compared to the other implementations: the usage of weights (section 3.1.1) and the new PhenoDis release from 2016 (section 3.1.2).

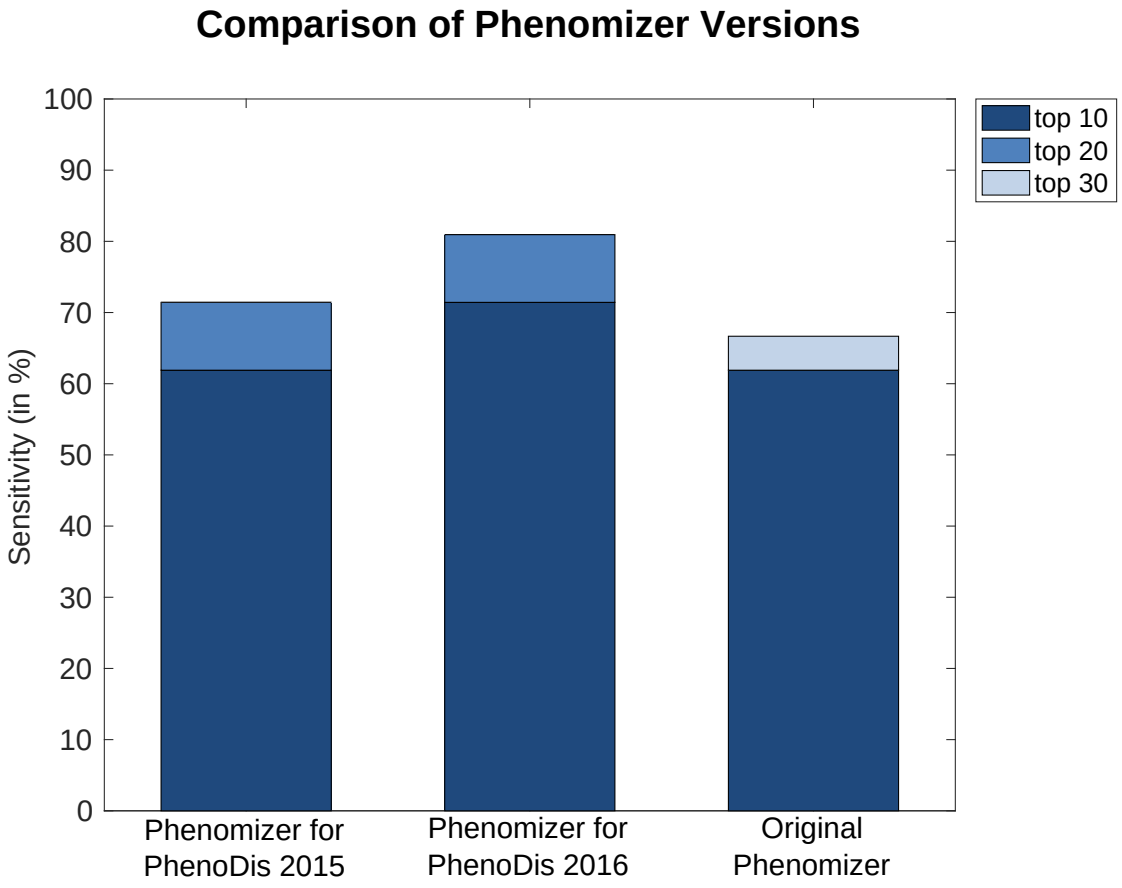


Figure 8: Comparison of different variants of Phenomizer. The bar plot shows the sensitivities for three different implementations of Phenomizer: Phenomizer for PhenoDis 2015, Phenomizer for PhenoDis 2016 and the original Phenomizer. Every tool was applied on the symptoms of 21 patients suffering from different rare diseases (patient data set 3, section 2.2.2 and 2.4.3.1). The height of the bars indicates different sensitivity values: dark blue bars = fraction of causal diseases with a rank of at most 10 (top10 sensitivity), dark + medium blue bars = fraction of causal diseases with at rank of at most 20 (top20 sensitivity), dark + medium + light blue bars = fraction of causal diseases with a rank of at most 30 (top30 sensitivity).

3.1.1 Impact of the Weighting

Phenomizer for PhenoDis 2015 and 2016 have a higher top20 sensitivity than the original Phenomizer (figure 8). This indicates that the weighting of symptom-disease pairs according to their frequencies produces a more accurate comparison between the patient and the diseases.

To further investigate the impact of the weights, Phenomizer for PhenoDis 2016 was evaluated on patient data set 1, a set of 1 400 simulated patients (section 2.2.1). The patients of the set suffer from defined diseases. The symptoms of the patients are chosen as all very frequent symptoms (i.e. all symptoms that occur with a frequency of more than 75% with the disease, table 5). Phenomizer for PhenoDis 2016 was run twice on the patient data set, once without weights and once with the weights given in table 5. The resulting predictions made without and with weights are compared in terms of the rank of the causal disease (figure 9).

The weighted variant of Phenomizer for PhenoDis 2016 produces lower ranks of the causal diseases than the unweighted variants. For example, the third quartile of the rank distribution for the unweighted variant is equal to 6 whereas the third quartile of the weighted variant is equal to 1. This indicates that the weighted similarity scores yield more accurate predictions than the unweighted scores. The application of weights allows to include the frequencies of symptom-disease pairs into the disease score (equation 2.3). Due to the weighting the very frequent symptoms of a disease make a higher contribution to the score of the disease than the rare symptoms. As a result the causal disease of a patient with the very frequent symptoms of the disease gets a higher score when using the weighted variant of Phenomizer for PhenoDis 2016.

These observations are confirmed by [39]. This work describes the evaluation of Phenomizer for PhenoDis 2015 with and without weights on patient data set 3 and reports slightly elevated sensitivities for the weighted variant of the algorithm. The publication [80] also analyzed the effect of weighting disease-symptom pairs according to their frequency. The authors discovered that comparing diseases among each other using these weights leads to a more biologically coherent clustering of diseases than without weights.

The inclusion of weights into the Phenomizer algorithm indeed increases the sensitivity of the method. Hence, the application of weights is also beneficial for the overall performance of PheNoBo.

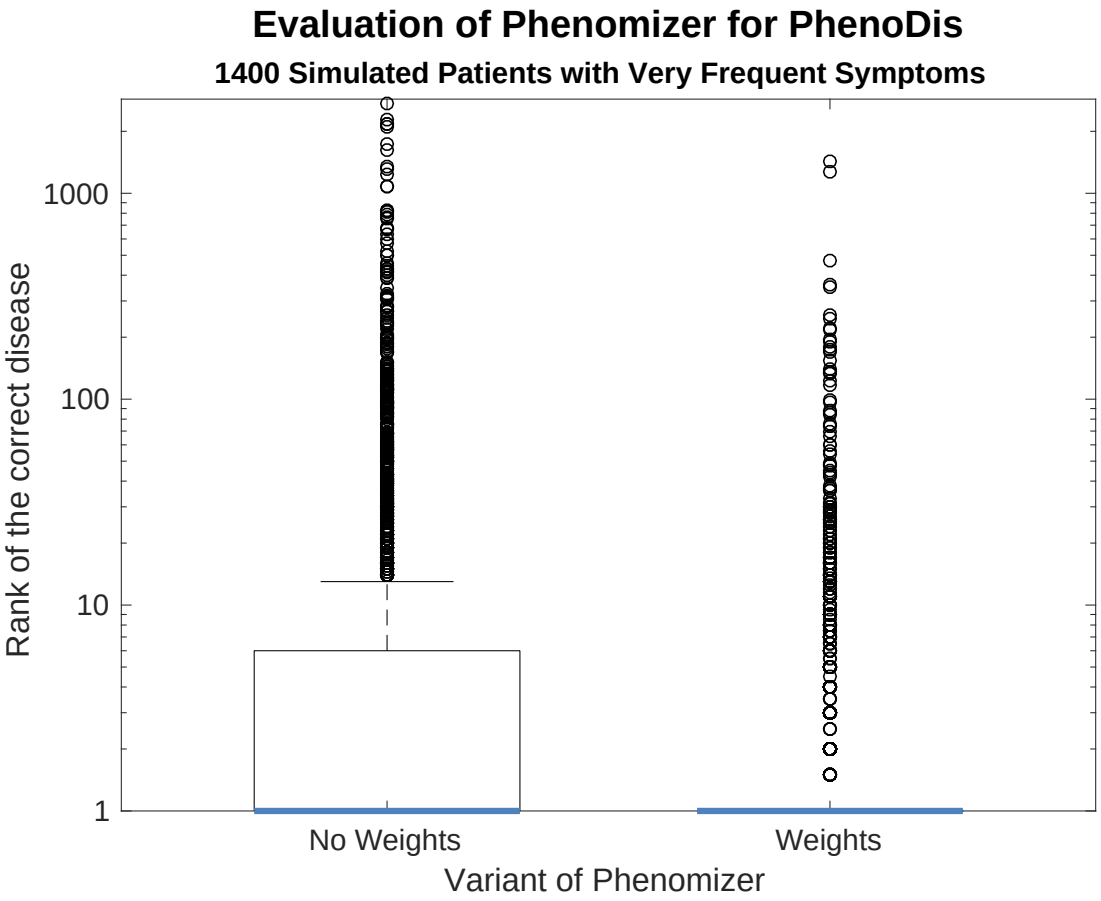


Figure 9: Impact of weights on the predictions of Phenomizer for PhenoDis. The box plot shows the distribution of the rank of the causal disease obtained for Phenomizer for PhenoDis 2016 without weights (left box) and with weights (right box) in equation 2.3. Both variants were applied on 1 400 simulated patients with very frequent symptoms (patient data set 1, section 2.2.1). The position of the box indicates the interquartile range (between the first and the third quartile) of the ranks. The thick, blue line in the box gives the median rank.

3.1.2 Impact of the PhenoDis Version

Phenomizer for PhenoDis was originally developed and evaluated on the release of PhenoDis as of August 2015 [39] (Phenomizer for PhenoDis 2015). Phenomizer for PhenoDis 2016 runs on a newer version of the PhenoDis database from February 2016. Phenomizer for PhenoDis 2016 has a higher top10 sensitivity than Phenomizer for PhenoDis 2015 (figure 8). This section investigates the differences between the two releases of PhenoDis and suggests how these changes might explain the higher sensitivity of Phenomizer for PhenoDis 2016.

The comparison of PhenoDis from 2015 and from 2016 focuses the symptom-disease annotations (figure 10). There are two major changes in the version of 2016: an increase in size of the PhenoDis database (i.e. increased number of diseases and symptoms) and an improvement of the disease-symptom annotations.

Increase in Size: There are more diseases with two to 19 annotated symptoms in the version of 2016 than in the version of 2015 (5 868 diseases in PhenoDis 2016 vs. 4 248 diseases in PhenoDis 2015). Inversely, the latest release of PhenoDis has fewer diseases with one annotated symptom (317 diseases in 2016 vs. 834 diseases in 2015). This demonstrates an increase in the amount of disease-symptom data provided by the new release of PhenoDis. The increase in information in the PhenoDis database is also reflected by the total number of diseases and symptoms. The release of PhenoDis from 2015 stored 7 554 diseases and 10 355 symptoms. The new version of PhenoDis of 2016 manages 8 263 diseases and 11 573 symptoms.
This change of PhenoDis is due to an increasing knowledge about rare diseases. Rare diseases are an active field of research with constantly evolving and improving knowledge. Ongoing studies like the project DDD (Deciphering Developmental Disorders) [115] discovered new rare diseases [3]. PhenoDis is updated along with the databases OMIM and Orphanet to keep track of new insights about rare diseases.

Improvement of the Disease-Symptom Annotations: Furthermore, the distribution of symptoms per disease indicates that the quality of disease-symptom annotations has been improved in the current release of PhenoDis. The number of diseases that are annotated with only one symptom decreased considerably (834 diseases in 2015 vs. 317 diseases in 2016, figure 10).
These enhancements of PhenoDis can be illustrated by the annotations of the disease “Cardiomyopathy, familial hypertrophic, 3”. This rare cardiac disease has only one annotated symptom, “Hypertrophic cardiomyopathy”, in the release of 2015. The new release of PhenoDis describes the disease more accurately with eleven additional symptoms including “Dyspnea”, “Right ventricular hypertrophy” and “Hypertension”.

This improvement is due to updates in OMIM and Orphanet and to the annotators of the IBIS, who verify and complete the content of PhenoDis manually.

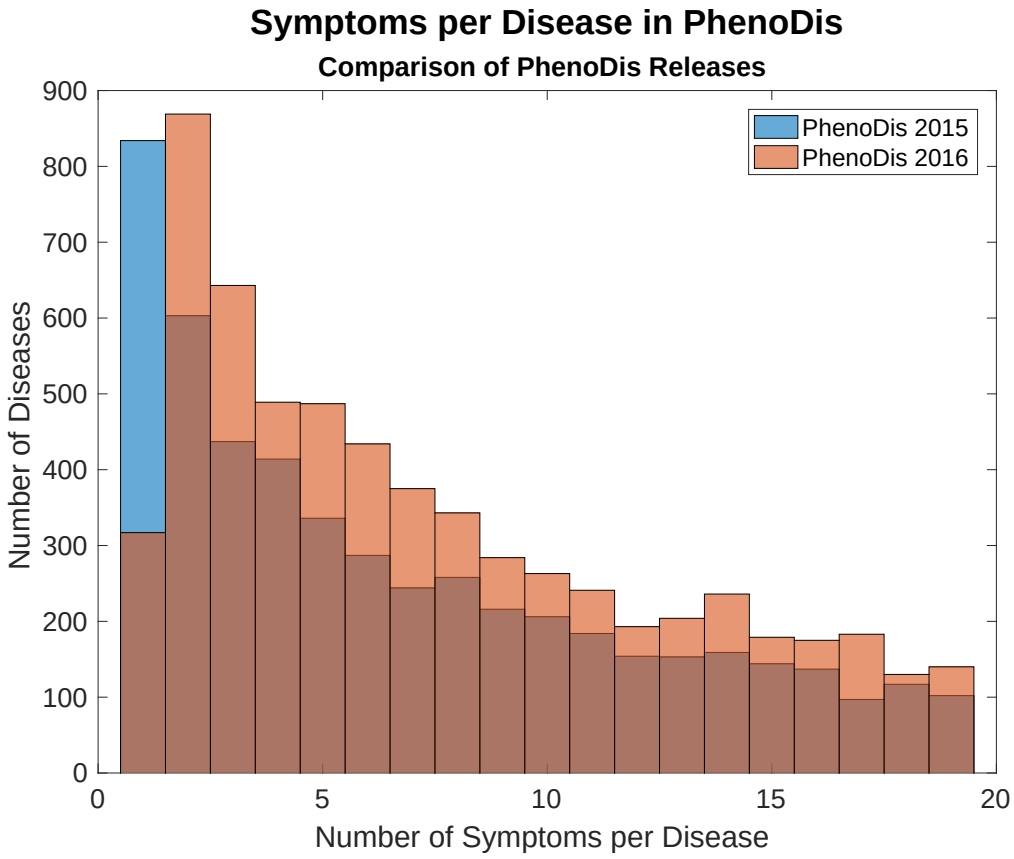


Figure 10: Symptoms per disease in PhenoDis. This figure visualizes the symptom-disease annotations of the releases of PhenoDis from August 2015 (blue bars) and February 2016 (red bars). The height of a bar gives the number of diseases annotated with a defined number of symptoms. The bars are drawn as overlapping histograms. Consequently, the dark red parts of the bars result from the overlay of a blue and a red bar.

The increased number of diseases and symptoms as well as the improvements in the disease-symptom annotations provide a more detailed basis for the comparisons between the diseases and the phenotype of the patient. If the causal disease of the patient is not part of PhenoDis, Phenomizer for PhenoDis is not able to predict the disease. The larger number of diseases allows to predict more causal diseases correctly. The improvement of the symptom-disease annotations enables a more accurate comparison of the phenotype of the patient and the disease.

This argumentation can also be applied when comparing Phenomizer for PhenoDis and the original Phenomizer. The original Phenomizer uses data derived from OMIM. The research of [80] demonstrates that OMIM is underannotated. PhenoDis unites disease annotations from OMIM and Orphanet and therefore provides more complete disease-symptom annotations than OMIM alone.

Therefore, PhenoDis 2016 provides a solid data basis for the predictions of PheNoBo. The beneficial impact of weights and PhenoDis 2016 shows that Phenomizer for PhenoDis 2016 is currently the optimal choice for the first step of the phenotype analysis in PheNoBo.

3.2 Evaluation of the Phenotype Analysis Pipeline

The phenotype analysis pipeline of PheNoBo consists of three steps. First Phenomizer for PhenoDis compares the patient's phenotype with the diseases in PhenoDis to calculate disease scores. PhenoToGeno transforms the disease scores into gene scores using known disease-gene associations. Finally, GeneticNetworkScore calculates new gene scores using genetic interactions in order to detect new disease genes. The phenotype analysis pipeline predicts the causal gene that best explains the patient's phenotype.

The phenotype analysis pipeline runs independently from the metabotype analysis pipeline and the genotype analysis pipeline. The final step of PheNoBo combines the results of the three pipelines. Therefore, PheNoBo equally depends on the prediction quality of the three separate analyses. This modular setup enables to evaluate and optimize each analysis pipeline independently.

The focus on the phenotype analysis directly allows to optimize the settings of the phenotype analysis pipeline to achieve good phenotype-base predictions. Section 3.2.1 and section 3.2.2 derive the optimal settings for PhenoToGeno and GeneticNetworkScore. The determination of the best settings also includes an estimate of the sensitivity of the pipeline. As the phenotype and the metabotype analysis pipeline rely on similar algorithms, the results of this evaluation are also used to infer the optimal setup of the metabotype analysis pipeline (section 3.2.3).

3.2.1 Settings of PhenoToGeno

PhenoToGeno is the successor tool of Phenomizer for PhenoDis in the phenotype analysis. PhenoToGeno translates the disease scores of Phenomizer into gene scores (section 2.3.3). This section examines the best parameters for the algorithm. There are two different options controlling the calculations of PhenoToGeno: the filtering and the gene annotation mode (section 2.4.1). The filtering determines which diseases are passed from Phenomizer on to PhenoToGeno. The gene annotation mode defines how PhenoToGeno computes a score for genes that are associated with multiple scored diseases.

The optimal settings of PhenoToGeno are inferred by applying the phenotype analysis pipeline with different combinations of the options on patients from simulations and from the literature. The options are compared using rank distributions (section 3.2.1.1) and sensitivities (section 3.2.1.2). When optimizing the parameters of PhenoToGeno one also needs to consider the underlying disease-gene data of PhenoToGeno. Section 3.2.1.3 presents these data as a disease-gene network and discusses the impact of the data set on the settings of PhenoToGeno.

3.2.1.1 Evaluation of Rank Distributions

Different combinations of filtering and gene annotation mode were applied on a set of 1 400 simulated patients (section 2.2.1 and 2.4.3.2). This set comprises phenotype data and is named patient data set 1. The phenotype data of a patient in set 1 are all very frequent symptoms of the patient's disease. These patient data were analyzed with the phenotype analysis of PheNoBo using different combinations of filtering and gene annotation mode. GeneticNetworkScore was applied with the default settings (table 7). The options are evaluated by considering the distribution of the ranks of the causal genes in the predictions of GeneticNetworkScore. The prediction of GeneticNetworkScore is a sorted list of scored genes, which is also the final result of the phenotype analysis pipeline. The rank of the causal gene is defined as the position of the gene within this final result (section 2.4.2).

Filter: PhenoToGeno was tested with three different filters. The following listing indicates the set of scored diseases that is used by PhenoToGeno if the respective filter is applied:

- SignificantP: all diseases with p value less than 0.05
- Top20: the 20 diseases with the lowest p values

- TopP: all diseases with the best (i.e. lowest) p value

The application of a filter results in a median rank that is at most as high as the median rank that is obtained without filter (median 3.5 for no filter, median 3.5 for SignP, median 2.5 for Top20 and median 3 for topP in multiple annotation mode). The rank distribution of the predictions with the filter Top20 has the lowest median and third quartile of all filters. The filter Top20 yields also better ranks than the unfiltered predictions for both annotation modes (figure 11).
However, the filtering of the diseases leads to more outliers in the rank distribution. There are some patients for whom the causal gene gets a rank of more than 10 000. These outliers arise when the patient's disease is removed from the predictions of Phenomizer before applying PhenoToGeno.
Hence, the application of a filter (especially the Top20 filter) improves the rank of the causal gene if the causal disease is predicted correctly by Phenomizer. Otherwise, the usage of a filter can decrease the quality of the predictions of PheNoBo.

Gene Annotation Mode: PhenoToGeno provides two gene annotation modes for calculating gene scores:

- Multiple annotation mode (abbreviated mult) combines the scores of all diseases annotated to a gene.

- Maximum annotation mode (abbreviated max) uses the maximum score of all diseases annotated to a gene.

The multiple annotation mode performs better than the maximum annotation mode if a filter is applied (distributions of mult have a higher first and third quartile than max). If no filter is used, the third quartile of max is lower than for mult and the first quartile of max is higher than for mult (figure 11). This indicates that the maximum annotation mode results in a smaller number of causal genes with a high rank. However, the maximum annotation mode without filter places the causal gene less frequently at rank 1 than multiple annotation mode without filter.

It is difficult to deduce the optimal combination of options for PhenoToGeno from these results. None of the combinations clearly performs best for all patients of data set 1. Nevertheless, the simultaneous application of a filter and maximum annotation mode proved to be less performant than the other options.
Furthermore, this analysis was done with virtual patients generated by a simple procedure (selection of very frequent symptoms). The options might perform differently on real

patients. For example, some diseases are pleiotropic and have a variable expression [66]. This means that two patients with the same disease might manifest different symptoms. Therefore, Phenomizer for PhenoDis might not always predict the causal disease correctly. In such a case a filter might remove the causal disease and impact negatively on the prediction.
For this reasons these options are further evaluated on more realistic data.

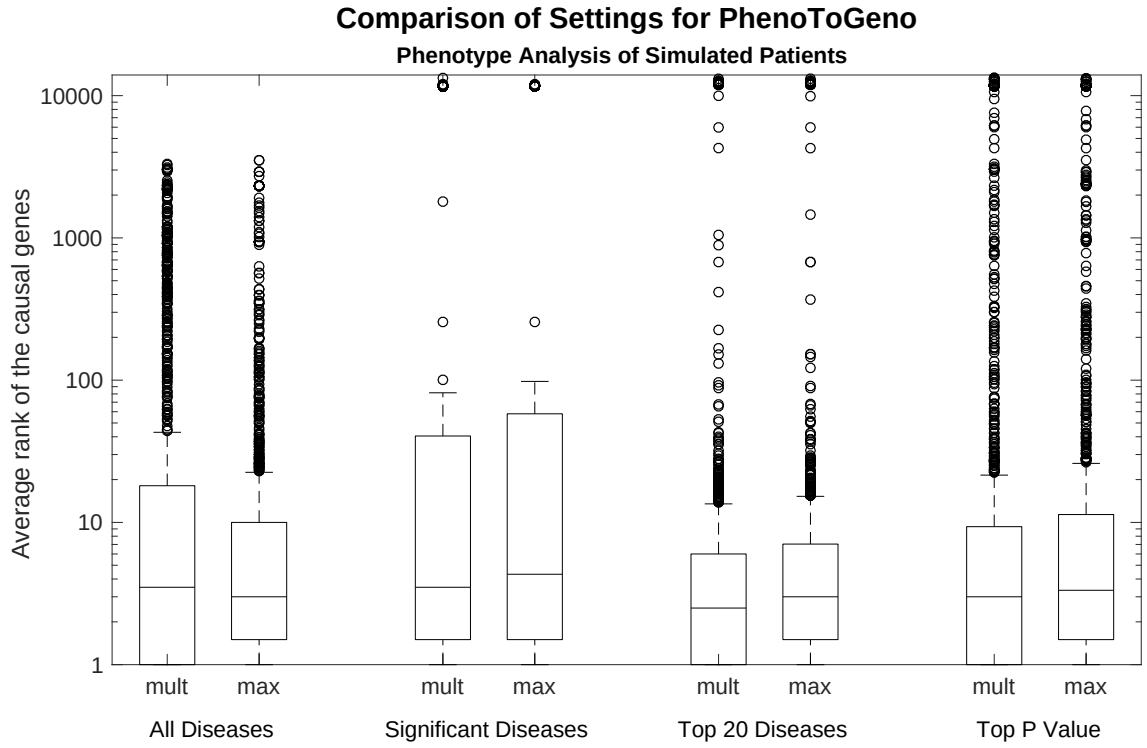


Figure 11: Comparison of the settings for PhenoToGeno. The box plot shows eight rank distributions resulting from the application of the phenotype analysis on patient data set 1 (section 2.4.3.2). For each patient and each combination of options a rank was obtained. If the disease of the patient has more than one causal gene, the average rank of the genes is plotted. The patients were analyzed with different options of PhenoToGeno (section 2.4.1): mult = multiple annotation mode, max = maximum annotation mode, All diseases = no filter, Significant Diseases = filter SignificantP, Top20 Diseases = filter Top20, Top P Value = filter TopP. The resulting distributions are shown as boxes. The position of the box indicates the interquartile range (between the first and the third quartile) of the ranks. The line in the box gives the median rank.

3.2.1.2 Evaluation of Sensitivities

As the previous comparison of rank distributions suggested, the options of PhenoToGeno were further evaluated on two more realistic patient data sets: patient data set 2 and patient data set 4. Patient data set 2 comprises simulated phenotype data of 7 890 patients with mitochondrial disorders (section 2.2.1). Patient data set 4 is based on 15 patients from the literature that suffer from rare carnitine-related disorders (section 2.2.2). The phenotype analysis was applied several times on the patient data with different options for PhenoToGeno: multiple annotation mode with different filters and maximum annotation mode without filter. The usage of a filter together with maximum annotation mode proved less advantageous (section 3.2.1.1) and therefore was not considered in this analysis.

The application of the phenotype analysis pipeline resulted in two ranks of the causal gene for each patient: the rank in the predictions of PhenoToGeno and of GeneticNetworkScore (section 2.4.3.2). The predictions of PhenoToGeno are an intermediate result of the phenotype analysis pipeline whereas the predictions of GeneticNetworkScore are the final result of the pipeline. The ranks for each combination of options and each set were translated into sensitivity values (figure 12 and figure 13). The analysis provides three different sensitivities: the top10 sensitivity, the top20 sensitivity and the top50 sensitivity (the fraction of causal genes with a rank of at most 10, 20 and 50, section 2.4.2). The calculated sensitivities enable a comparison of PhenoToGeno and GeneticNetworkScore and a comparison of the different options for PhenoToGeno.

PhenoToGeno vs. GeneticNetworkScore: The predictions of GeneticNetworkScore have a higher top10 sensitivity than the predictions of PhenoToGeno for both data sets (figure 12 and 13). However, the top50 sensitivity of GenenticNetworkScore is lower than the top50 sensitivity of PhenoToGeno for some combinations of the options. These changes are more pronounced in set 4 because the set consists only of 15 patients. For example, if the number of patients with a rank below 10 changes by one, the top10 sensitivity for data set 4 will change by more than 6%.
This observation demonstrates that the application of GeneticNetworkScore, which aims to detect new disease genes, does not impair the prediction of known disease genes. The patient data of set 4 also provide several cases in which GeneticNetworkScore is able to promote the prediction of new disease genes. For example, the disease of the second patient of the case study [109] gets a rank of 42 in the predictions of Phenomizer for PhenoDis. If one applies the top20 filter, the disease of the patient is removed from the predictions and the causal gene ends up with a rank of 11 681.5 in the result of PhenoToGeno. The high rank results from the fact that PhenoToGeno does not use the known link between the

patient's disease and the causal gene. GeneticNetworkScore is able to improve the rank of the causal gene to 68. Therefore, GeneticNetworkScore supports the prediction of known disease genes and still allows the detection of new disease genes.

Options of PhenoToGeno: For patient data set 2 (figure 12), the application of multiple annotation mode results in higher sensitivity values than the application of any other combination of options (up to 80% top10 sensitivity in the results of GeneticNetworkScore). The other setups using multiple annotation mode and a filter do not differ in their sensitivities (top10 sensitivity of 73% to 75%). The maximum annotation mode without filter does not perform as good as the other combinations of options (top 10 sensitivity 66%).
These observations go along with the results from section 3.2.1.1. However, section 3.2.1.1 provides more detailed rank distributions and reveals more differences between the filters.
The behavior of the options of PhenoToGeno in patient data set 4 is similar to set 2 (figure 13). There is one difference to the previous analyses. Maximum annotation mode without filter performs equally well as multiple annotation mode without filter. Maximum annotation mode even has higher top50 sensitivity than multiple annotation mode. This difference might result from the fact that maximum annotation mode works especially well on this particular set of patients.

Sensitivity of the Phenotype Analysis Pipeline: Overall the application of the phenotype analysis pipeline on patient data set 2 yields a top10 sensitivity of more than 65%. These sensitivities are derived from simulated patient data. This data is not entirely realistic e.g. the simulation does not consider symptoms that are not related to the patient's disease. The results from patient data set 4 suggest a sensitivity of about 30% for the phenotype analysis. Set 4 comprises only 15 patients and is restricted to the diseases CACTD and PCD. The sensitivity of the phenotype analysis is likely to be higher for patients suffering from diseases other than CACTD and PCD. Therefore, the true sensitivity of the pipeline might be lower than 70% as suggested for the simulated patients, but higher than 30% as reported for patient data set 4.

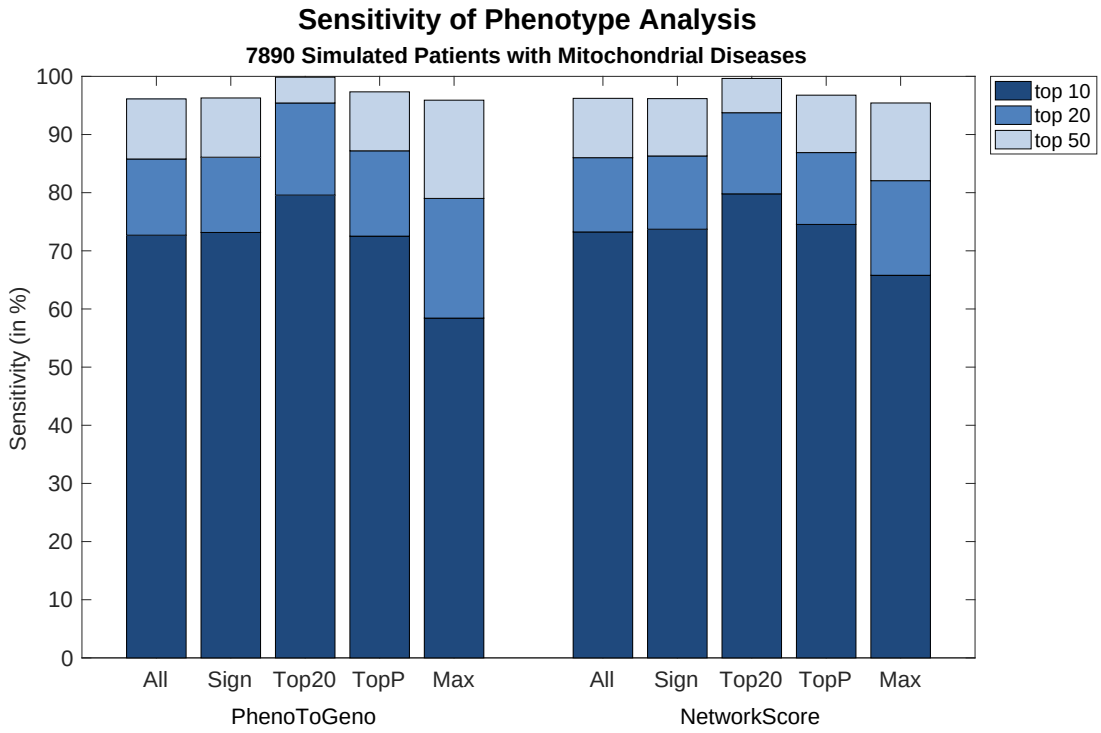


Figure 12: Sensitivity of the phenotype analysis based on simulated patients (patient data set 2). The plot shows sensitivity values for the predictions of PhenoToGeno and GeneticNetworkScore on set 2 (section 2.4.3.2). There are three different sensitivity value for each analysis: top10/top20/top50 sensitivity = fraction of patients whose causal genes got an average rank of at most 10 (dark blue)/ 20 (dark+medium blue)/ 50 (dark+medium+light blue). The phenotype analysis was run with different options. There is a stacked bar for every option: all = multiple annotation mode without filter, Sign = multiple annotation mode with SignificantP filter, Top20 = multiple annotation mode with Top20 filter, TopP = multiple annotation mode with TopP filter, max = maximum annotation mode without filter.

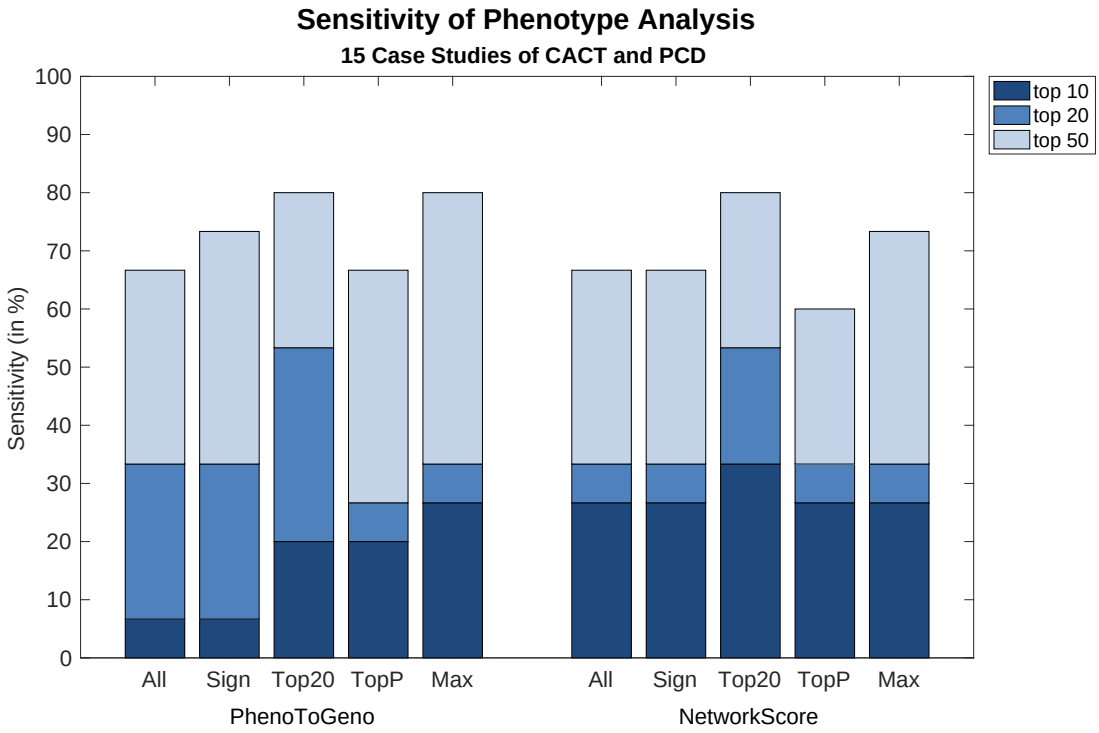


Figure 13: Sensitivity of the phenotype analysis based on patients from the literature (patient data set 4). The plot shows sensitivity values for the predictions of PhenoToGeno and GeneticNetworkScore on set 4 (section 2.4.3.2). There are three different sensitivity value for each analysis: top10/top20/top50 sensitivity = fraction of patients whose causal genes got an average rank of at most 10 (dark blue)/ 20 (dark+medium blue)/ 50 (dark+medium+light blue). The phenotype analysis was run with different options. There is a stacked bar for every option: all = multiple annotation mode without filter, Sign = multiple annotation mode with SignificantP filter, Top20 = multiple annotation mode with Top20 filter, TopP = multiple annotation mode with TopP filter, max = maximum annotation mode without filter.

The three different patient data sets, set 1, 2 and 4, were analyzed with different combinations of gene annotation mode and filtering. Comparing the results of all three sets allows to derive the settings of PhenoToGeno that optimize the overall predictions of PheNoBo.

Gene Annotation Mode: For the patients with CACTD and PCD, maximum annotation mode shows a higher sensitivity than multiple annotation mode. For the simulated patients, however, multiple annotation mode performs more favorable. In contrast, the 20 simulated patients of data set 2 with CACTD and PCD get lower ranks if PhenoToGeno is run with maximum annotation mode instead of multiple annotation mode. This implies that there might be special subsets of patients for which maximum annotation mode makes better predictions than multiple annotation mode.
Patient data set 2 is analyzed again to investigate this assumption (figure 14). For this analysis the patients of the data set are divided into two groups: patients with singleton diseases (1 640 patients) and patients with connected diseases (6 250 patients). Singleton diseases are diseases with one causal gene that is not associated with any other disease. All other diseases are classified as connected diseases. Indeed, maximum annotation mode works better than multiple annotation mode for singleton diseases (top 10 sensitivity of 51% for multiple and 61% for maximum annotation mode). The causal genes of connected diseases get lower ranks when using multiple annotation mode.
This result demonstrates that multiple annotation mode favors the prediction of genes associated with several diseases. Such genes tend to get a higher gene score than genes with only one known disease. This observation is also confirmed by the structure of equation 2.9. As the set of connected diseases is much larger than the set of singleton diseases, multiple annotation mode yields overall a higher sensitivity. Therefore, multiple annotation mode is chosen as default option for PhenoToGeno.

Filtering: PheNoBo does not use any filter in its default setup. The application of a filter is problematic. If the filter removes the causal disease from the output of Phenomizer for PhenoDis, the phenotype analysis produces inaccurate predictions. The removal of the causal disease is likely during the application of PheNoBo on real patients. The phenotype of real patients often does not fit perfectly the expected symptoms of the underlying disease [66]. For example, the causal diseases of 20% of the patients of data set 4 get a rank above 20 in the results of Phenomizer and are removed by the Top20 filter. Furthermore the sensitivity obtained with the filters SignificantP and TopP is comparable to the sensitivity of multiple annotation mode without any filter. Therefore, the inclusion of a filter is unlikely to have a positive effect on the predictions of PheNoBo for real patients.

In summary the optimal setup for the phenotype analysis pipeline of PheNoBo uses multiple annotation mode without any filtering (table 7).

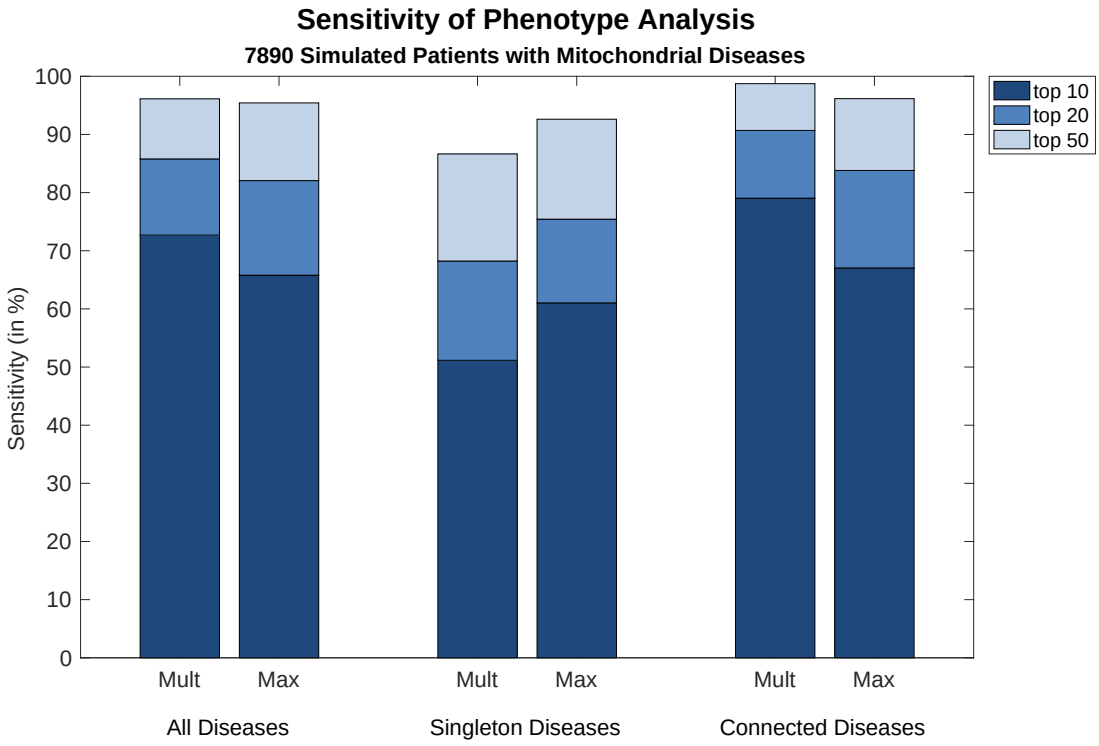


Figure 14: Sensitivity of multiple and maximum annotation mode for different subgroups of patient data set 2. The subgroups of data set 2 are analyzed with the phenotype analysis using maximum and multiple annotation mode (section 2.4.3.2): mult = multiple annotation mode without filter, max = maximum annotation mode without filter. Each bar gives three sensitivity values: top10/top20/top50 sensitivity = fraction of patients whose causal genes got an average rank of at most 10 (dark blue)/ 20 (dark+medium blue)/ 50 (dark+medium+light blue). The pair of bars at the left-hand side shows the sensitivity values for all 7 890 patients. The pair of bars in the middle gives the sensitivities for 1 640 patients with singleton diseases = diseases that have only one causal gene which is associated with only one disease. The remaining two bars at the right-hand side represent the sensitivities for 6 250 patients with connected diseases = all diseases that are not singleton diseases.

3.2.1.3 Impact of the Disease-Gene Data

The decision to use multiple annotation mode for PhenoToGeno is based on the assumption that there are more connected diseases than singleton diseases. The ratio of connected and singleton diseases depends on the underlying disease-gene data set of PhenoToGeno (section 2.1.4). A disease-gene pair is part of the data set if defects in the gene are known to cause the disease.

This section presents a disease-gene network that was constructed to visualize the disease-gene data set. The resulting network is compared to the disease network of [42] (henceforth called network of Barabási *et al.*) and to the disease-gene associations used in Phen-Gen [56] (henceforth referred to as network of Phen-Gen). The network of Barabási *et al.* was built from the diseases and the known causal genes provided by the OMIM database in December 2005 [42]. The network of Phen-Gen was derived from diseases and causal genes collected from OMIM and Orphanet in 2014.

Figure 15 visualizes the disease-gene data set of PheNoBo as bipartite network. The network consists of two type of nodes: nodes corresponding to diseases (red nodes) and nodes representing genes (blue nodes). An edge between a disease and a gene indicates that mutations in the gene are known to cause the disease. Hence, there is an edge for every disease-gene pair in the disease-gene data set.

The following analyses and comparisons focus on the size and the topology of the network.

Size of the Network: Figure 16 provides information about the number of diseases, genes and disease-gene pairs in the network of PheNoBo (light blue bars), of Barabási *et al.* (dark blue bars) and of Phen-Gen (medium blue bars). The network of PheNoBo is made up of more nodes and edges than the other networks.
This result indicates a considerable increase in knowledge over time. The network of Barabási *et al.* was published in 2007. The data for the network of Phen-Gen were collected in 2014. The network of PheNoBo is the most recent data set (from 2016). This fact goes along with the observation from section 3.1.2 that there is ongoing research about rare diseases. Research consortia like BRIDGE [111] and FORGE [12] frequently report the identification of new diseases and new causal genes.
Consequently PheNoBo is endowed with a comprehensive up-to-date data set about diseases and their causal genes. However, the constantly increasing knowledge also creates the need of regular updates of the disease-gene data of PheNoBo. As PheNoBo relies on prior knowledge, the inclusions of new disease-gene data will enhance the predictions of PheNoBo.

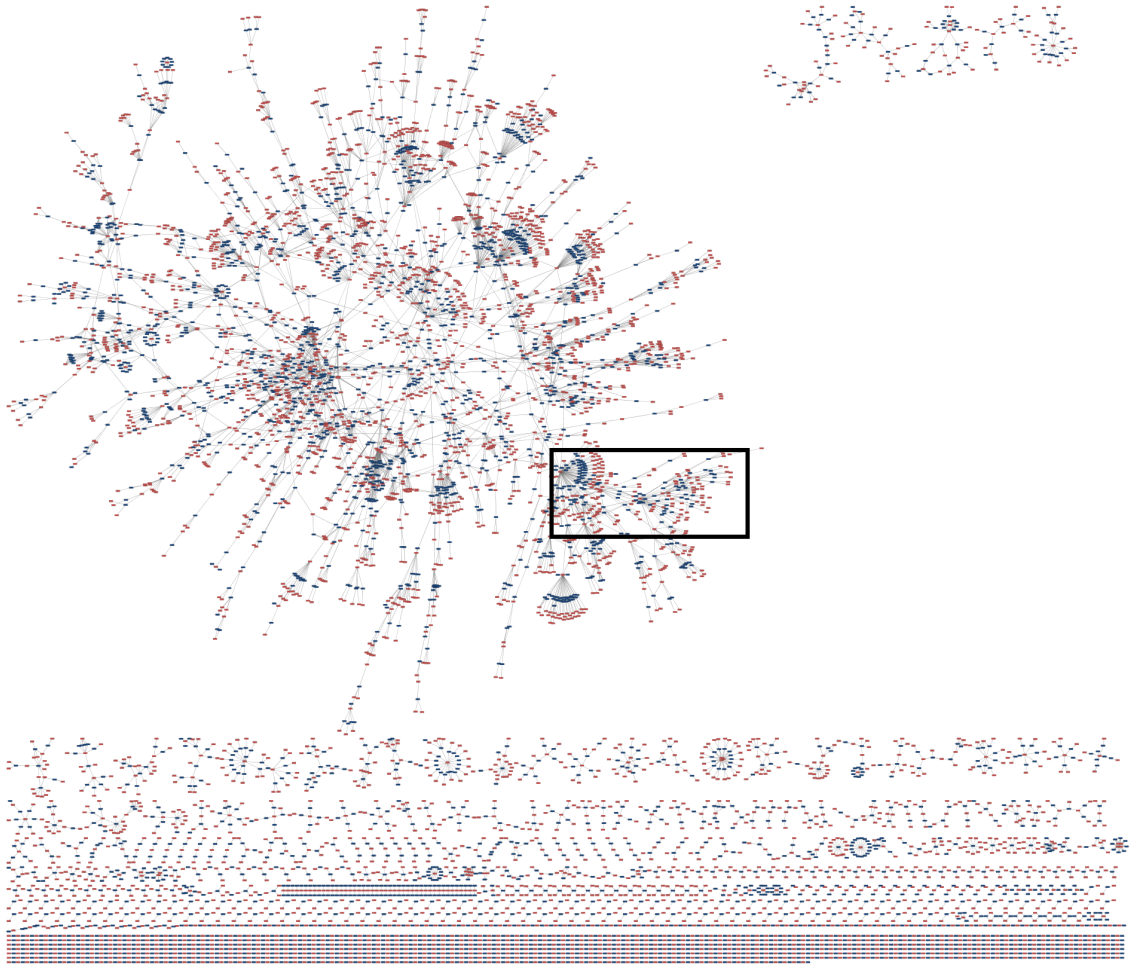


Figure 15: Disease-gene network of PheNoBo. The network was built from the disease-gene data set (section 2.1.4). Diseases are shown as red nodes and genes as blue nodes. Each edge of the network corresponds to a disease-gene pair of the disease-gene data set. The black rectangle marks the part of the network that is shown in figure 17.

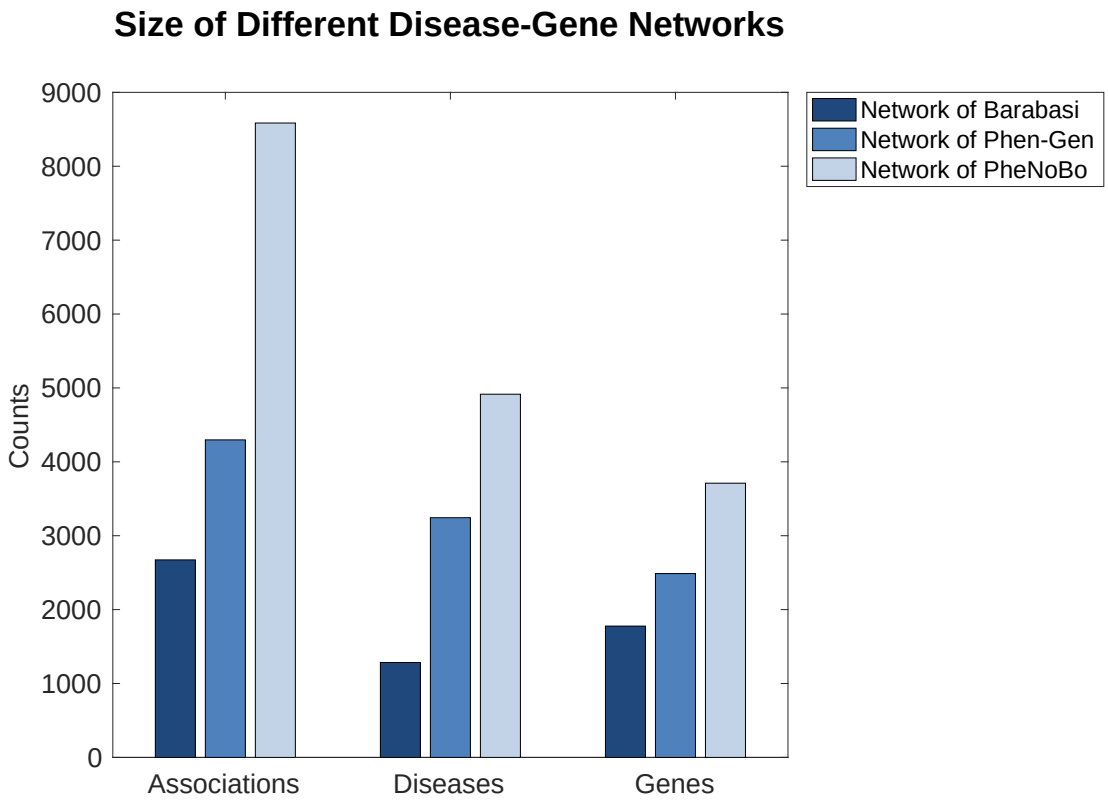


Figure 16: Size of different disease-gene networks. The bar plot shows the number of disease-gene associations, diseases and genes of the networks of Barabási *et al.* [42], of Phen-Gen [56] and of PheNoBo (section 2.1.4).

Topology of the Network: The network of Barabási *et al.* was constructed the same way as the network of PheNoBo. Both networks represent causal relationships between genes and diseases. Although the network of Barabási *et al.* was set up almost ten years ago, the topology of the network is similar to the topology of the network of PheNoBo: The majority of genes and diseases are organized in a single large connected component (i.e. the connected diseases). Both networks have several small components and some singleton disease-gene pairs (i.e. the singleton diseases).
Furthermore, related diseases are clustered together in the network of Barabási *et al.* and the network of PheNoBo (the clustering in the network of PheNoBo is not apparent from figure 15 but a digital version of the network is provided in this thesis that allows an in-depth inspection of the network of PheNoBo, DVD Content). Nevertheless, the clustering of diseases differs slightly from network to network. These differences are most likely due to new findings about disease genes. The changes in the clustering are illustrated with two examples:

- Charcot-Marie-Tooth Disease (CMT): CMT is a hereditary neuropathy which is classified into a variety of subtypes [16]. The network of Barabási *et al.* has a single cluster for CMT. The network of PheNoBo contains also a cluster of different forms of CMT. Additionally some subtypes of the disease are distributed within various small components of the network of PheNoBo. These subtypes are not part of the network of Barabási *et al.* as their causal genes were discovered after 2005 [16]. The fact that the subtypes are not linked together in the network of PheNoBo reflects the genetic heterogeneity of CMT. The causal genes of CMT comprise for example genes encoding myelin-related proteins and aminoacyl-tRNA synthetases [16].

- Bardet-Biedl Syndrome (BBS): BBS is a rare ciliopathy [36]. In the network of Barabási *et al.* BBS is located in a small connected component. This component comprises one additional disease, McKusick-Kaufman Syndrome. The component of BBS and McKusick-Kaufman Syndrome is part of the main connected component in the network of PheNoBo (rectangle in figure 15). BBS is linked to the main component via a gene that is known to cause retinitis pigmentosa and BBS (figure 17). The connection between BBS and retinitis pigmentosa makes sense as retinitis pigmentosa is a characteristic symptom of BBS [36]. This difference between the network of Barabási *et al.* and PheNoBo is due to an increased understanding of the pathogenesis of ciliopathies [80].

The similarity of the two networks implies that there is some fundamental and stable knowledge about diseases and their causal genes. The identification of new disease genes rather completed the network than changed the underlying structure. In particular, the

number of connected diseases is much larger than the number of singleton diseases for both, the network of Barabási *et al.* and of PheNoBo. Therefore, assumption for choosing multiple annotation mode is expected to be stable even if new causal relations between genes and diseases will be added to the data set of PheNoBo.

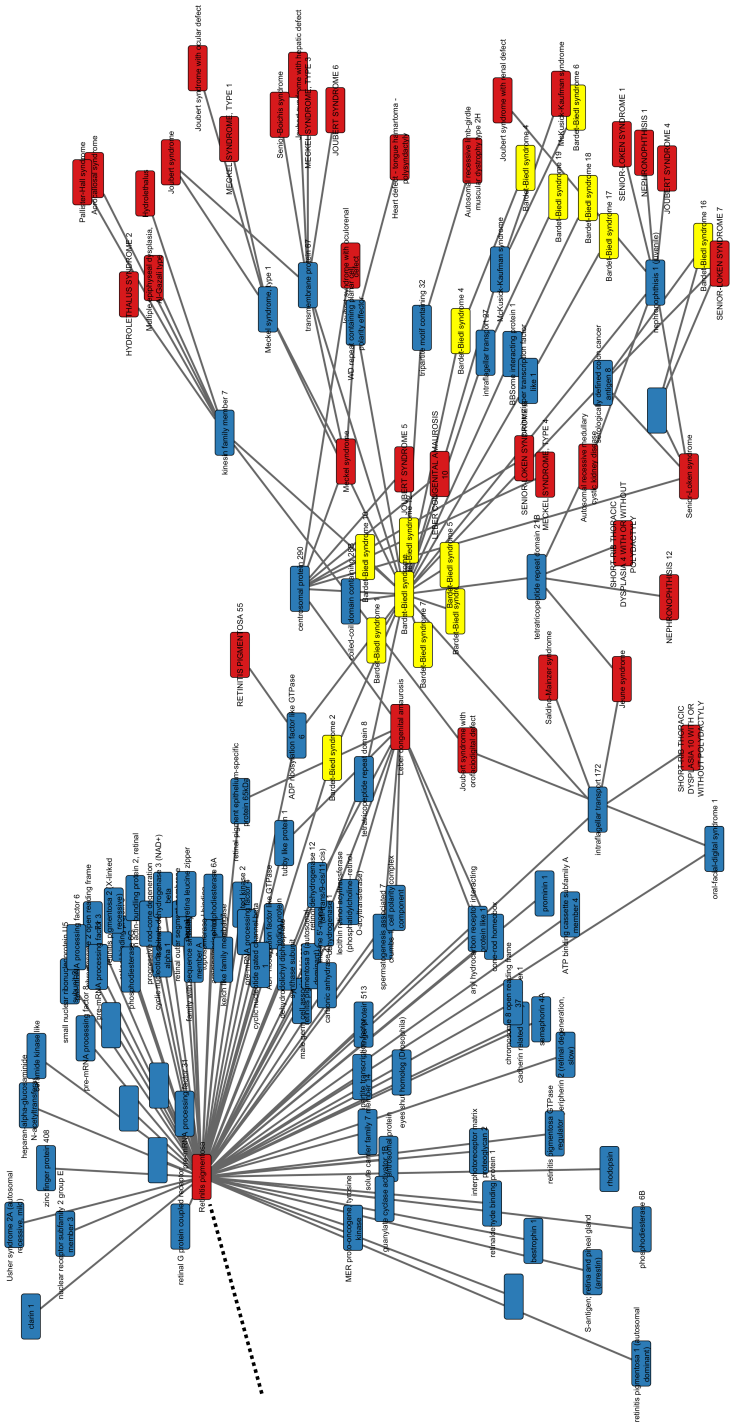


Figure 17: Cluster of Bardet-Biedl Syndromes in the disease-gene network of PheNoBo. blue nodes = genes, red nodes = diseases, yellow nodes = different types of the disease BBS, dotted line = connection of the diseases retinitis pigmentosa to the main connected component of the network.

3.2.2 Settings of GeneticNetworkScore

This section is dedicated to the optimal setup of GeneticNetworkScore. GeneticNetworkScore applies a random walk with restart on a genetic network to detect novel disease/metabolite-gene relationships (section 2.3.4). The random walk is used to distribute the initial gene scores of PhenoToGeno within the network. The underlying algorithm has three parameters to optimize (equation 2.11 and section 2.4.1). The transition matrix (abbreviated as M) can be derived from the edge weights or alternatively from the adjacency matrix of the network. The restart probability (abbreviated as r) regulates the fraction of the original scores that is distributed in the network. The number of steps (abbreviated as t) indicates how far a gene score is spread within the network.

The analysis done for GeneticNetworkScore is similar to the analysis of PhenoToGeno in section 3.2.1.1. The phenotype analysis pipeline was run on patient data set 1 (i.e. the 1 400 simulated patients with very frequent symptoms already used in section 3.2.1.1) with different combinations of the aforementioned options of GeneticNetworkScore. The options of PhenoToGeno were not changed (default options given in table 7). Altogether eight combinations of the options were tested on the patient data of set 1. The different setups were evaluated by considering the rank of the causal gene of each patient (section 2.4.3.2). The rank is the position of the causal gene in the sorted predictions of GeneticNetworkScore, the final result of the phenotype analysis pipeline. The results for the patients were summarized into a rank distribution for each combination of options (figure 18). The following paragraphs discuss the results and best choices for the three options of GeneticNetworkScore.

Inclusion of Edge Weights: GeneticNetworkScore was tested on two transition matrices. The unweighted matrix is derived from the adjacency matrix of the network (equation 2.13). The weighted matrix is calculated from the confidence scores of the genetic interactions represented by the edge weights of the network (formula 2.14). The application of the different matrices in the phenotype analysis does not cause noticeable changes in the rank distributions (figure 18).
Nevertheless, the weighted matrix is a more detailed model as it includes the confidence scores of the genetic interactions.

Restart Probability: The restart probability can be set to any value between 0 and 1. Two values were chosen for the evaluation of this option:

- $r = 0.9$: A restart probability of 90% means that 10% of the initial gene scores are distributed within the network. Choosing a larger value for r would limit the ability

to detect new disease genes as the initial scores would not change considerably.

- $r = 0.5$: Setting the restart probability to 50% results in spreading 50% of the scores within the network. If one uses a smaller restart probability, one might not be able to predict known disease genes with few interactions because their scores will decrease considerably. Furthermore, one risks to bias the predictions towards the hub nodes (genes with many interactions) of the network.

The predictions of PheNoBo show that PheNoBo performs better with $r = 0.9$ than with $r = 0.5$ (figure 18). The ranks of most causal genes are increased if a restart probability of 50% is chosen (distributions for $r = 0.5$ are slightly shifted towards higher ranks).
Therefore, the more conservative option $r = 0.9$ might be appropriate for analyzing real patients without any prior knowledge about the cause of their diseases. Phen-Gen [56], the template of PheNoBo, also makes predictions with a restart probability of 90%.
However, a lower restart probability might be appropriate if one directly aims at discovering new disease genes (e.g. like GeneWanderer in [65]). The evaluation of PheNoBo in such a use case is beyond the scope of this master's thesis.

Number of Steps: GeneticNetworkScore can be applied with any number of steps. To investigate the effect of this option on the predictions two extreme values were tested for this option:

- $t = 2$: If the random walk consists two steps, the scores are spread locally in the network. All nodes within a distance of two edges from the original node receive a fraction of the score from the original node.

- $t \to \infty$: A random walk with an unlimited number of steps is continued until the gene scores do not change anymore considerably. As a consequence, the scores of the nodes are distributed globally throughout the whole network.

The predictions for the simulated patients indicate that the predictive performance does not depend on the number of steps if a restart probability of 90% is used. The rank distributions obtained for a restart probability of 50% with different numbers of steps vary slightly. The ranks are lower for $t = 2$ than for $t \to \infty$ (first quartile of 1.25 for $t = 2$ vs. 1.5 for $t \to \infty$). These differences result from the fact that 50% of the scores is spread. Spreading a larger amount of scores notably increases the scores of genes with a distance of more than 2 edges from the original node.
As the previous paragraph suggested the choice of t should have little effect on the predictions of PheNoBo for a restart probability of 90%. Phen-Gen applies a random walk

with restart that spreads the scores globally within all nodes of the network [56]. Therefore, PheNoBo is applied with $t \to \infty$ on real patient data.

This analysis demonstrates that the options of GeneticNetworkScore lead to rather small changes in the rank distributions. The effect of these options is less strong than for the options of PhenoToGeno (section 3.2.1). The options of GeneticNetworkScore might have a stronger impact if a different underlying network is used. Therefore, the default options of GeneticNetworkScore are chosen conservatively according to the options that have been shown to work for the tool Phen-Gen [56]. The resulting optimal setup of GeneticNetworkScore is a random walk with restart using a weighted transition matrix, $r = 0.9$ and $t \to \infty$ (table 7).

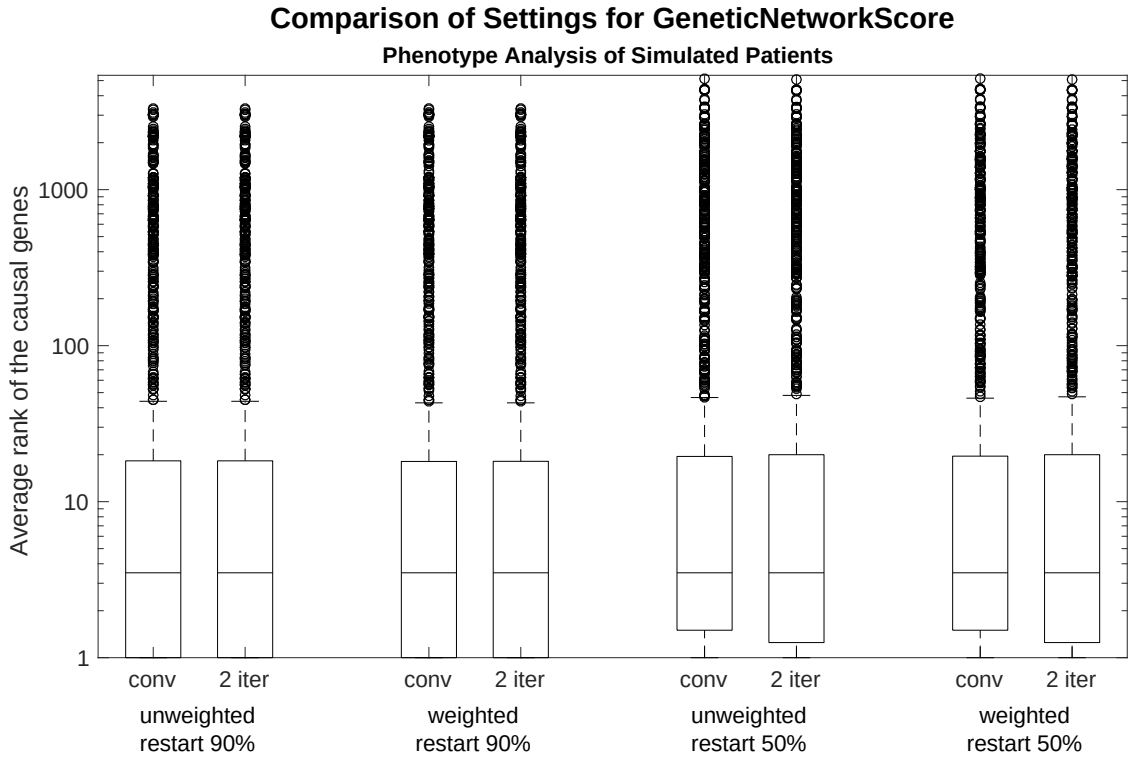


Figure 18: Comparison of the settings for GeneticNetworkScore. The box plot shows eight rank distributions resulting from the application of the phenotype analysis on patient data set 1 (section 2.4.3.2). For each patient and each combination of options a rank was obtained. If the disease of the patient has more than one causal gene, the average rank of the genes is plotted. The patients were analyzed with different options of GeneticNetworkScore (equation 2.11 and section 2.4.1). Restart 90% = restart probability $r = 0.9$, restart 50% = restart probability $r = 0.5$, weighted = weighted transition matrix M (equation 2.14), unweighted = unweighted transition matrix M (equation 2.13), conv = unlimited number of steps $t \to \infty$, 2 iter = two steps $t = 2$. The resulting distributions are shown as boxes. The position of the box indicates the interquartile range (between the first and the third quartile) of the ranks. The line in the box gives the median rank.

3.2.3 Consequences for the Metabotype Analysis Pipeline

The preceding analyses suggest the best parameters for PhenoToGeno and GeneticNetworkScore based on the application of the phenotype analysis pipeline of PheNoBo. The metabotype analysis pipeline is similar to the phenotype analysis pipeline. The metabotype analysis comprises three tools: ScoreMetabolites computes scores for the metabolites in the metabolite profile of the patient. MetaboToGeno transforms the metabolite scores into gene scores. MetaboToGeno basically uses the same algorithm as PhenoToGeno, but applies the algorithm on metabolite-gene associations instead of disease-gene associations (section 2.3.3). The last tool is GeneticNetworkScore which is also the last step of the phenotype analysis pipeline. The gene scores of GeneticNetworkScore are the final prediction of the metabotype pipeline and represent the probabilities that the genes are causal given the metabotype of the patient.

The similarities between the pipelines allow to transfer the settings derived for PhenoToGeno and GeneticNetworkScore on the metabotype analysis pipeline. However, settings like the gene annotation mode for resolving multiple disease/metabolite scores per gene depend on the underlying data sets of the analysis pipeline. Therefore, this section first presents the composition metabolite-gene associations (section 3.2.3.1) and derives in a second step the optimal settings for the metabotype analysis pipeline (section 3.2.3.2).

3.2.3.1 Composition of the Metabolite-Gene Associations

This analysis deals with the composition of the metabolite-gene associations presented in section 2.1.5 and compares them to the disease-gene associations used by PhenoToGeno (section 2.1.4). The metabolite-gene associations are composed of data from six different sources. The metabolite-gene pairs from HMDB and Recon are derived from biochemical reactions and involve physical interactions of genes and metabolites. The resources SMPDB and GWAS Server and the curated data sets based on GO and on symptoms from HPO provide indirect metabolite-gene interactions. The following paragraphs compare the number of metabolites, genes and associations contributed from the different resources (figure 19 and table 8).

Metabolites: The metabolite-gene associations cover 252 metabolites. However, the HMDB manages information about approximately 42 000 metabolites. The small number of metabolites in this data set results from the intersection of all metabolites from HMDB and the reference set (section 2.3.2.1). The algorithms of PheNoBo are not able to score

metabolites without any reference data.
All sources of the metabolite-gene data except the GWAS server cover similar metabolites (Venn diagram of metabolites in figure 19). The GWAS Server provides metabolite-gene pairs from statistical associations. These associations are based on data from non-targeted metabolomics experiments and include unknown metabolites: 28 of the 53 metabolites unique to the GWAS Server are metabolites without known chemical structure.
Notably, 350 metabolites from the reference set are not part of the metabolite-gene associations. 189 of those metabolites are unknown. The chemical structure of the remaining 161 metabolites is known. The 161 known metabolites are mainly specific lipids and oligopeptides which are not listed in the HMDB. The reason for this might be that their origin and their role in the human metabolism is not yet fully understood.
As a consequence the number of metabolites in the metabolite-gene data is about 20 times smaller than the number of diseases in the disease-gene data (table 8). The number of metabolites is limited by two factors. The first limiting factor is the aforementioned incomplete mapping between the metabolites of the reference set and metabolites of HMDB. The second reason for the small amount of metabolites is the limited metabolite coverage of non-targeted metabolomics measurements underlying the reference set and the metabotype data of the patient. This limitation results from the need of extracting and separating the metabolites of a sample before measuring them [117].
The small number of metabolites in the set is an important difference to consider when deriving the settings of the metabotype analysis from the phenotype-based results.

Genes: The metabolite-gene data set comprises associations for 2 530 genes. Each subset covers genes that are unique to the respective resource (Venn diagram of genes in figure 19). However, there is a considerable overlap between the genes of the three subsets from Recon, the HMDB and the SMPDB. This is due to the fact that the three resources focus on biochemical reactions and the enzymes catalyzing these reactions.
The genes provided by the set derived from GO terms and from HPO symptoms are especially valuable for PheNoBo. 53% of the genes in this subset are not found in any of the other resources. More than 40% of the associations from these sources involve mitochondrial genes (defined as genes with GO terms containing “mitochondrion” or “mitochondrial”). Therefore, this data is useful for the analysis of mitochondrial diseases.
Altogether, the number of genes in the metabolite-gene set is in the same order of magnitude as the number of genes in the disease-gene set.

Associations: The whole data set includes 12 676 metabolite-gene pairs. The majority of the metabolite-gene pairs originates from the SMPDB. 8 357 associations are unique to the SMPDB (Venn diagram of associations in figure 19). The other 5 sources also make important contributions. More than 50% of the associations from each source are unique to

the respective source (percentage of unique associations: HMDB 63%, Recon 54%, GWAS Server 80%, GO+HPO symptoms 73%). The small overlap between the subsets is due to the different definitions of a metabolite-gene pair depending on the data source. The data set includes metabolite-gene pairs from biochemical reactions (HMDB and Recon) and indirectly linked metabolite-gene pairs involving statistical associations and pathways.
Furthermore, the density of the metabolite-gene associations varies with the source under consideration: The SMPDB has on average 6.4 associations per gene due to the indirect nature of the associations. The HMDB and Recon provide on average 2.1 and 2.5 associations per gene. The biochemical reactions of the metabolism usually involve at least one metabolite as substrate and one metabolite as product. The data based on the GWAS Server, GO terms and HPO symptoms are even less dense with 1.9, 1.2 and 1.3 associations per gene.
Altogether the metabolite-gene data is more dense than disease-gene data set (5 metabolites per gene vs. 2.3 diseases per gene). If one excludes the metabolite-gene associations from HMDB, the density of the reduced metabolite-gene set is similar to the density of the disease-gene data (table 8).

This analysis allows to derive the composition of the metabolite-gene data set that is finally used for the predictions of PheNoBo in metabotype analysis pipeline:

- HMDB and Recon: Both resources provide verified data from biochemical reactions.
- GO and HPO Symptoms: The two manually curated data sets are enriched for mitochondrial genes.
- GWAS Server: The database includes associations for unknown metabolites.

The data from SMPDB are excluded from the application of PheNoBo because of two shortcomings:

- The density of metabolite-gene links from Small Molecule Pathway Database (SMPDB) is remarkably higher than the density of the other data sets providing indirect associations. This suggests that there are uninformative and even misleading associations resulting from the pathway data of the SMPDB. For example the metabolite phosphate is linked to 1 052 genes via SMPDB pathways. Phosphate is a universal metabolite arising in all ATP-consuming reactions. As almost every pathway uses phosphate, the metabolite is associated with more than 70% of all genes in the SMPDB subset. The link between the protein Methionyl-tRNA formyltransferase, mitochondrial (MTFMT) and the metabolite betaine is another instance for an uninformative as-

sociation. Betaine is required for the synthesis of methionine from homocysteine and MTFMT formylates methionyl-tRNA in the mitochondrion [114]. Defects in MTFMT are unlikely to have an impact on the concentration of betaine. These spurious associations might lead to false positive predictions of PheNoBo.

- The set of indirect associations is incomplete. For example the link between the protein MTFMT and the metabolite N-Formyl-L-methionine is missing. MTFMT is the only known source of N-Formyl-L-methionine in the human metabolism [114]. Defects of MTFMT lead to a shortage of N-Formyl-L-methionine and cause the disease Leigh Syndrome [105]. The inclusion of this association in the data set could improve the predictions of PheNoBo.

Furthermore, the removal the SMPDB data creates a set of associations that is more similar to the disease-gene data set than the original metbaolite-gene set. This similarity is important for deriving settings for MetaboToGeno, which uses the metabolite-gene data, from PhenoToGeno that relies on the disease-gene data.

The reduced data set consists of 4 319 associations covering 1 853 genes and 248 metabolites. The collection of reliable and interpretable metabolite-gene associations is more difficult than the assembly of disease-gene pairs. This is due to the fact that metabolomics is a relatively new field of study. The systematic and large-scale investigation of the human metabolome was enabled recently by modern high-throughput techniques [17]. Furthermore, the HMDB was established in 2007 whereas OMIM was started 50 years ago [46]. Recent studies like [69] aim to resolve the structure of unknown metabolites and to gain deeper insights in the human metabolism. Therefore, the data about metabolite-gene association is likely to improve with the ongoing research. The improvement of the metabolite-gene data would have a beneficial effect on the predictions of PheNoBo.

Data Set	Entities	Genes	Associations per Entity	Associations per Gene
Disease-Gene	4 915 Diseases	3 711	1.7	2.3
Metabolite-Gene	252 Metabolites	2 530	50	5
Metabolite-Gene without SMPDB	248 Metabolites	1 853	17.4	2.3

Table 8: Comparison of the disease-gene and metabolite-gene data. The table provides some basic parameters of the disease-gene associations and metabolite-gene associations of PheNoBo. The table compares three sets: the disease-gene data set (section 2.1.4), the original metabolite-gene data set (section 2.1.5) and the reduced metabolite-gene set without data from SMPDB. The term "entity" denotes "disease or metabolite".

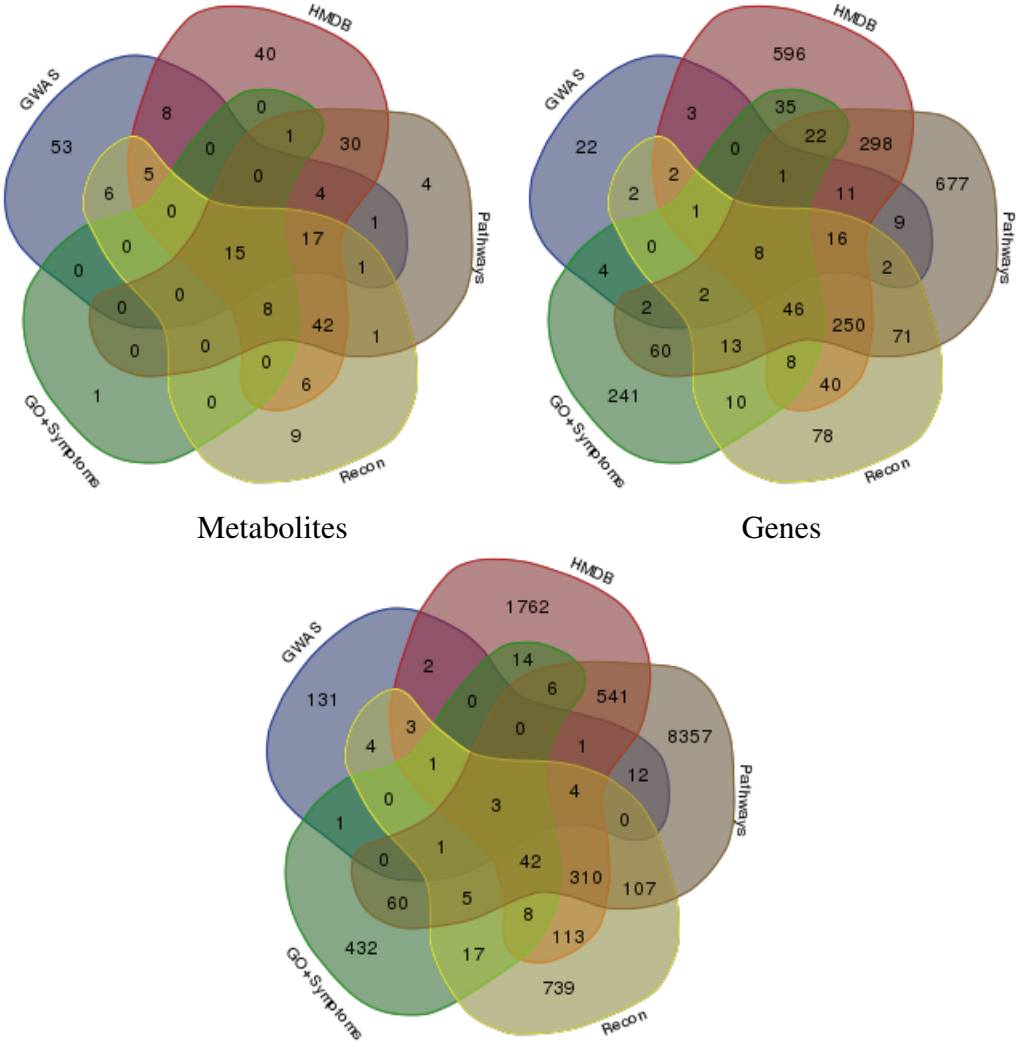


Figure 19: Composition of the metabolite-gene associations. The figure shows three Venn diagrams for the number of metabolites, the number of genes and the number of associations in the metabolite-gene data (section 2.1.5). Each set depicted in the Venn diagrams corresponds to a source used to construct the metabolite-gene data set. There are six sources: HMDB, SMPDB (named Pathways in the figure), Recon, GWAS Server, HPO Symptoms and GO. The two manually curated data sets based on the GO and on the symptoms from HPO were grouped into a single set (named GO+Symptoms) for reasons of clarity. The plots were drawn by the Venn diagram tool available at `http://bioinformatics.psb.ugent.be/webtools/Venn/`.

3.2.3.2 Settings of the Metabolite Analysis Pipeline

There is not enough metabotype patient data available to derive the settings of the metabotype analysis pipeline in a similar way than in section 3.2.1 and section 3.2.2. The integration of metabolomics data into prediction tools for causal genes is not yet common practice [44]. Methods and data sets for systematic evaluation are not yet established. Nevertheless, one can make recommendations for the options of the metabotype analysis pipeline by considering the phenotype-based results.

MetaboToGeno: The previous evaluation of the metabolite-gene data assured that the metabolite-gene data of metabotype analysis is similar to the disease-gene data of the phenotype analysis. Therefore, it is possible to infer the optimal settings of MetaboToGeno from the best settings of PhenoToGeno.
Section 3.2.1.2 shows that multiple annotation mode favors the prediction of genes that are associated with multiple diseases. The usage of multiple annotation mode in PhenoToGeno accounts for the fact that genes with multiple disease might be more prone to disease-causing mutations. This argumentation cannot be applied on metabolites. The number of metabolites that are associated with a gene is limited by the metabolites that were measured in the patient and in the reference set (section 3.2.3.1). It is not appropriate to give a higher priority to the genes that are associated with multiple metabolites. For this reason, MetaboToGeno is applied by default with maximum annotation mode (table 7). As section 3.2.1.1 revealed a decreased sensitivity when maximum annotation mode is applied in combination with a filter, the results of ScoreMetabolites are not filtered.

GeneticNetworkScore: PhenoToGeno and MetaboToGeno produce similar results. Both tools provide a prediction of genes that are most likely causal for the patient's phenotype or metabotype. Consequently there is no need to use options that differ from the default options proposed in section 3.2.2. GeneticNetworkScore of the metabotype analysis is run with a weighted similarity matrix, a restart probability of 90% and an unlimited number of steps.

3.3 Evaluation of ScoreMetabolites

The evaluation of the phenotype analysis pipeline allowed to analyze parts of the metabotype analysis pipeline (section 3.2.3). The previous analysis focused on MetaboToGeno and GeneticNetworkScore, which perform the second and third step of the metabotype analysis

pipeline. This section is dedicated to the first step of the metabotype analysis pipeline, ScoreMetabolites. The basic concept of ScoreMetabolites is to compare the metabolite levels of a patient with a set of reference values. The comparison is translated into a score for each metabolite.

The development of the respective algorithm is based on the MitoNET reference set. The MitoNET reference set is derived from 97 blood samples from healthy individuals that were analyzed with a non-targeted metabolomics platform (section 2.2.4). The MitoNET control samples are taken from individuals of different sex, age and fasting state. This fact was included in ScoreMetabolites by using phenotype groups. The optimal choice of these groups is presented and discussed in section 3.3.1. Further on, the non-targeted metabolomics measurements of the control groups contain missing values. Section 3.3.2 analyzes the amount of missing values obtained for the MitoNET controls and suggests how to handle the missing values. However, the design of the ScoreMetabolites algorithm allows to apply the tool on any kind of reference set derived from non-targeted metabolomics experiments. Section 3.3.3 presents a comparison of the predictions of ScoreMetabolites using the standard reference set from MitoNET and the reference set derived from the KORA F4 cohort (section 2.2.3).

3.3.1 Introduction of the Phenotype Groups

The measured metabolite levels of a patient considerably depend on factors like sex, age and fasting state [77, 75, 68]. Therefore, ScoreMetabolites introduces phenotype groups, which classify patients and controls according to sex, age and fasting state. A patient of a certain phenotype group is compared against the reference values derived from the controls of the respective phenotype group.

The reference set of MitoNET is derived from 97 controls. Including all three factors, sex, age and fasting state, would yield extremely small phenotype groups with inaccurate reference values. As a compromise only two of the three factors were considered and the following four phenotype groups were chosen for the MitoNET reference set: infants, children, fasting adults and non-fasting adults (table 6).

The reference concentrations for the metabolite caffeine (figure 20) provide an an extreme example for studying two features of the phenotype groups: the differences in metabolite levels depending on the phenotype group and the composition of the phenotype groups.

Differences of Metabolite Levels depending on the Phenotype Group: Figure 20 re-

veals differences in the concentration of caffeine depending on the phenotype group. The metabolite caffeine is taken up from coffee, tea and soft drinks [102]. The children (group 2) have the lowest median concentration of caffeine. Children usually consume few caffeine-containing drinks. The non-fasting adults (group 4) have the highest median concentration of caffeine because many of them had drunken tea or coffee before their blood samples were taken. This example demonstrates that the concentrations of some metabolites are dependent on the age and the fasting state.

Composition of the Phenotype Groups: The data presented in figure 20 also describe the composition of the phenotype groups.

- Group 1 (infants) and group 2 (children) are small sets of samples. The fasting state of the children and the infants is unknown. However, the individuals of group 1 and 2 are most likely non-fasting as it is difficult to control or restrict the fasting state of an infant without impact on its well-being. The small size and the uncontrolled fasting state complicate the comparison of a patient and the reference values implemented in ScoreMetabolites (section 2.3.2). The comparisons done with group 1 and 2 might be less reliable and accurate than for group 3.

- Group 3 (fasting adults) is the largest phenotype group.

- Group 4 (non fasting adults) comprises individuals with high and low caffeine concentrations. The concentration of the caffeine depends on whether the individual consumed caffeine or not before the sample was taken. Consequently it is difficult to interpret a comparison with the reference values of group 4 when dealing with the concentrations of metabolites that are taken up from food.

The example of caffeine demonstrates that the introduction of phenotype groups is necessary to account in PheNoBo for the dependence of metabolite levels on factors like age and fasting state. The composition of the phenotype groups suggests that there is still potential to improve the reference metabolite data set. The inclusion of more metabolite samples would allow a more accurate interpretation of metabolic profiles in PheNoBo. The usage of more reference samples from children is essential as many rare diseases manifest early in life [33].
However, given the MitoNET control group, the current choice of control groups provides and optimal trade-off between group size and number of factors used for the grouping.

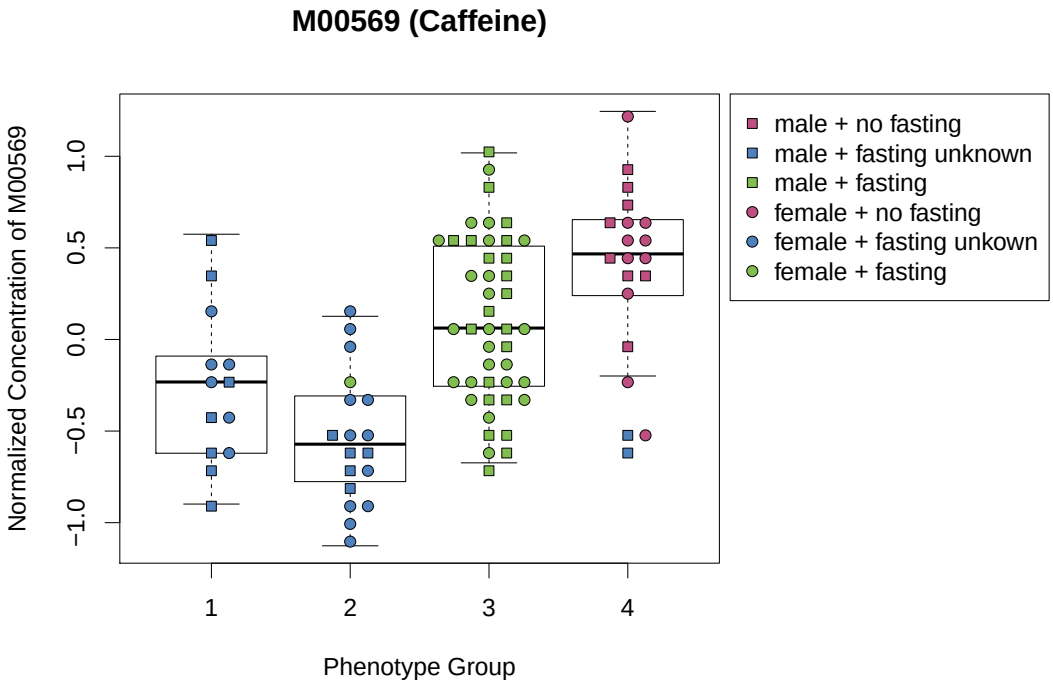


Figure 20: Reference concentrations of the metabolite caffeine. The figure shows a box plot giving the distribution of the concentration of caffeine in the metabolite reference set (section 2.3.2.1). There is one box for each of the four different phenotype groups (group 1 = infants, group 2 = children, group 3 = fasting adults, group 4 = non-fasting adults). The position of the box indicates the interquartile range (between the first and the third quartile) of the concentration. The thick line in the box gives the median concentration. The squares and circles drawn in the boxes represent the 97 samples from which the reference set was created. The x-coordinate of a sample corresponds to its phenotype group. The y-coordinate shows the concentration of caffeine in the sample.

3.3.2 Handling of the Missing Values

As ScoreMetabolites was specifically designed for the application on metabotype data from non-targeted metabolomics measurements, ScoreMetabolites needs to be able to handle missing values. Every measurement with non-targeted metabolomics yields missing values for some metabolites. There are basically two reasons for missing values: Either the metabolite is not present in the samples or the metabolite was not detected due to technical issues [32]. This section analyzes the amount of missing values in the samples that were used to derived the MitoNET reference (figure 21) and proposes how to handle them.

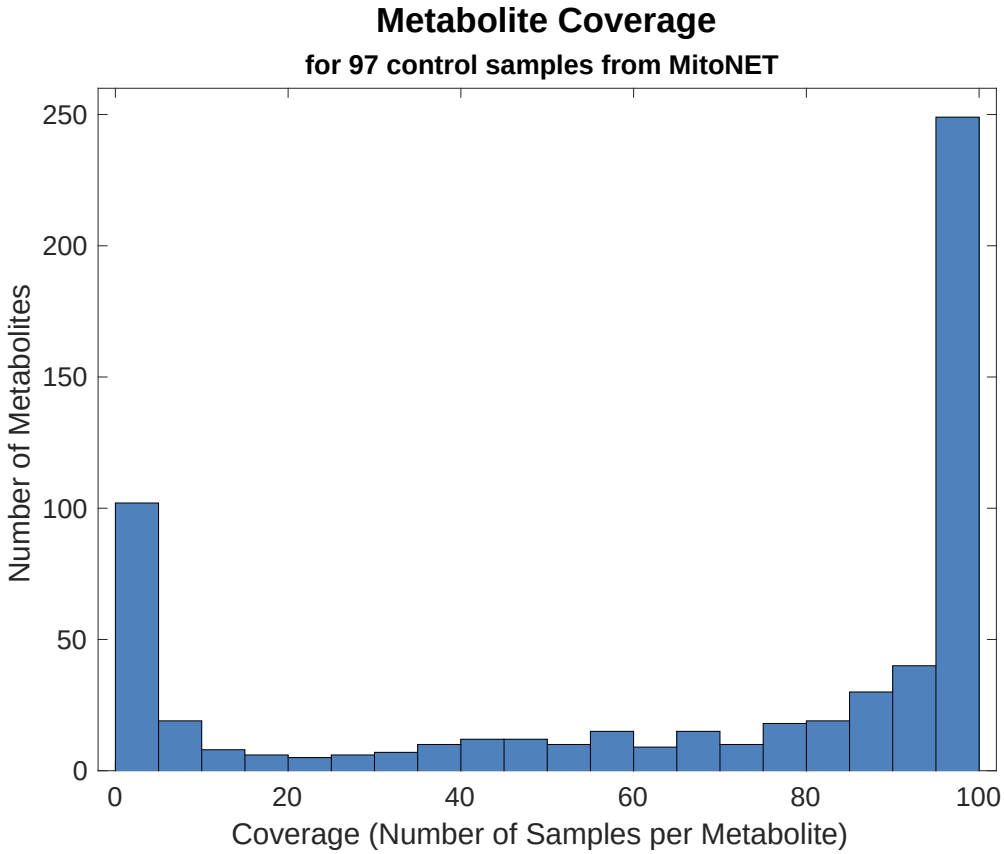


Figure 21: Metabolite coverage for the 602 metabolites of the MitoNET reference set. The coverage of a metabolite is equal to the number of samples that do not have a missing value for the metabolite. The plot shows a histogram of metabolite coverage for the 97 control samples from MitoNET.

The coverage of a metabolite is defined as the number of samples in which the metabolite was detected (i.e. the number of samples that do not have a missing value for the metabolite). About 250 metabolites were quantified in almost all samples (i.e. 95 samples or more).

Around 100 metabolites have a coverage of less than 5 samples and consequently have a missingness of more than 95%. This fact confirms that there are two distinct types of metabolites in the reference data set: metabolites with a low missingness (type binary) and metabolites with a high missing (type concentration).

Metabolites of Type Binary: All metabolites with a coverage in the range from 1-29 samples are classified as binary. This definition applies to about one fourth of all metabolites (146 metabolites). The metabolites of type binary do not provide sufficient data points to derive reliable reference values for their concentrations. However, the high amount of metabolites in this group does not necessarily imply that the measurements are of bad quality. The group includes 42 drugs and 68 unknown metabolites. The drug metabolites are not present in the samples from the MitoNET control group as this cohort is composed of healthy participants. The unknown metabolites might also result from the metabolism of drugs or they might be some other rarely observed metabolites.

Metabolites of Type Concentration: All metabolites with a coverage between 30 and 97 samples are of type concentration. This group comprises 456 metabolites. These metabolites provide sufficient information for the subsequent calculation of reference values.

The different properties of the two types of metabolites highlight the necessity to handle these types separately in ScoreMetabolites (section 2.3.2). The missing values of metabolites of type concentration can be replaced by statistically derived values using imputation without introducing artificial correlations [32]. Discarding of all binary metabolites with a low coverage would result in a considerable loss of information. As shown in [90] the pattern of missing values can differ significantly for some metabolites between healthy and diseased subjects. Therefore, the introduction of binary encoding is an appropriate means to handle these metabolites.

3.3.3 Exhangeability of the Reference Set

The previous analyses derived some basic features of the algorithm of Score metabolites from the MitoNET control group. Nevertheless, ScoreMetabolites can be run on any reference set derived from non-targeted metabolomics measurements.

As a consequence, the tool ScoreMetabolites has one parameter to optimize: the reference values giving the expected metabolite levels for calculating the metabolite scores (section 2.4.1). The selection of the reference values is critical as non-targeted metabolomics measurements do not provide absolute concentration values. The measurements of controls for the reference set and of patients should be ideally done together or at least measured within the same experimental setup. Therefore, this section is dedicated to analyze the exchangeability of the reference set.

The following evaluation is based on metabotype data from 82 MitoNET patients (patient data set 6 , section 2.4.3.3). ScoreMetabolites was applied with two different reference sets on the patient data: the MitoNET reference set and the KORA reference set. The samples underlying these reference sets were measured with similar non-targeted metabolomics platforms (section 2.2.3 and 2.2.4). The results for the two reference set are compared in terms of the metabolites with a probability of less than 0.05 (henceforth named deviating metabolites) in the output of ScoreMetabolites (figure 22).

There is a large overlap of deviating metabolites in the results based on the MitoNET and the KORA reference. More than 50% of the deviating metabolites per patient that are derived from the MitoNET reference are also present when using ScoreMetabolites with the KORA reference. Summarizing over the predictions for all 82 patients of data set 6, one obtains 2 047 deviating metabolites with the MitoNET reference and 2 334 with the KORA reference. 1 368 deviating metabolites (67% of the results from the MitoNET and 59% of the results from the KORA reference) are found with both reference sets. Consequently, both reference sets result in similar sets of deviating metabolites.

Nevertheless, the resulting set of deviating metabolites still varies to some extent depending on the reference. The differences in deviating metabolites are due to small differences in the experimental setup and the preparation of the reference data. For example, the measurements with gas chromatography and the outlier removal were only done for the samples form KORA.
Furthermore, this analysis is restricted to 273 metabolites that were covered in both reference sets. There are 329 metabolites left that were detected in the MitoNET patients but that are not part of the KORA reference set. In addition all patients of the data set belong to

the phenotype group 3 (fasting adults) as the KORA reference set does not comprise any samples from children or non-fasting subjects. The application of the KORA reference is restricted to the analysis of fasting adults. However, children are more often affected by rare diseases than adults [33].
These results suggest that it is possible to use ScoreMetabolites with reference values from a control group that was not measured together with the patients of interest. Nevertheless, the results of ScoreMetabolites are more accurate and include more metabolites if the controls and patients are measured in the same experiment. As this master's thesis focuses on the prediction of causal genes for the MitoNET patients, the MitoNET reference values are used in the default setup of PheNoBo (table 7).

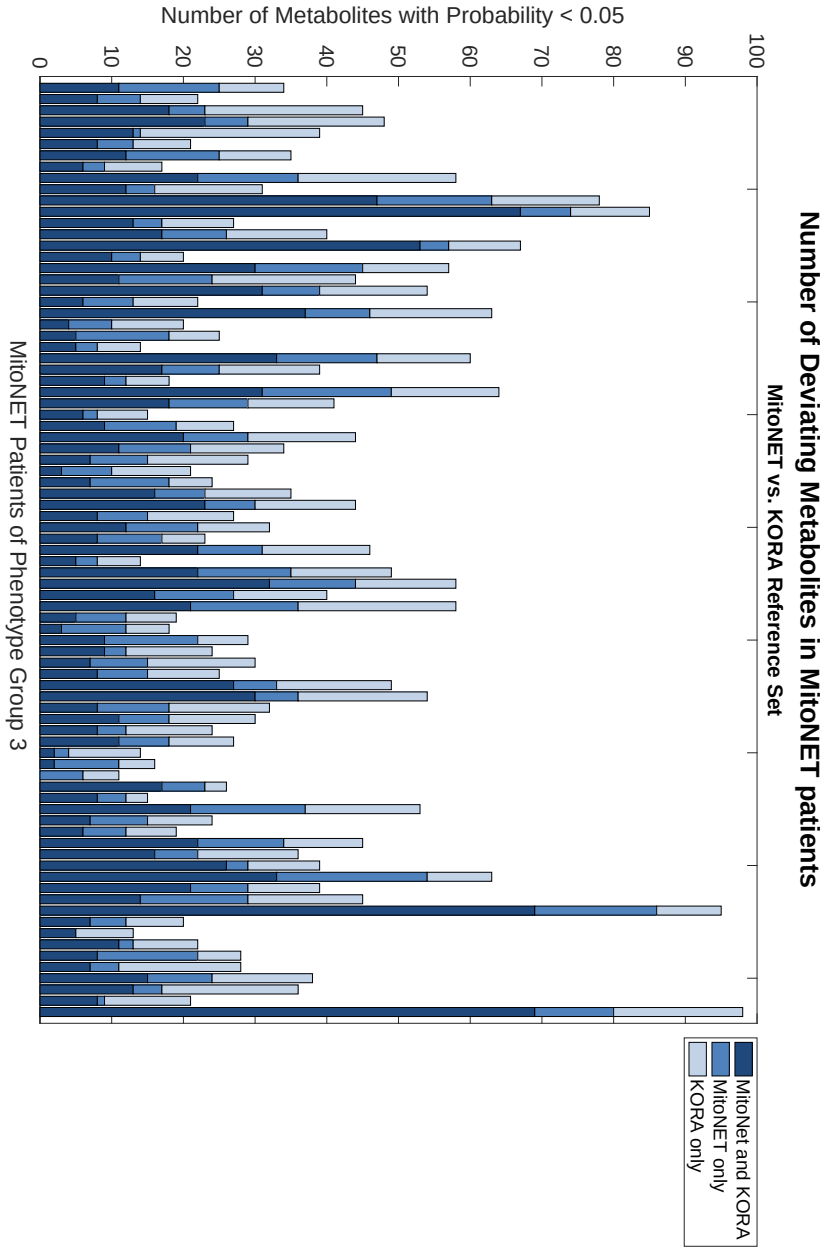


Figure 22: Comparison of the metabolite reference sets built from the MitoNET controls and the KORA cohort. The comparison is based on the number of deviating metabolites in the output of ScoreMetabolites, i.e. the number of metabolites with a probability < 0.05. Each stacked bar corresponds to a patient of data set 6 (section 2.4.3.3). Height of dark blue bar = number of deviating metabolites in KORA and MitoNET reference, height of medium blue bar = number of deviating metabolites in MitoNET reference only, height of light blue bar = number of deviating metabolites in KORA reference only.

3.4 Overall Evaluation of PheNoBo on Different Types of Patient Data

The patient data from MitoNET (patient data set 7, table 1) are the only data set providing all three types of data for PheNoBo: phenotype, metabotype and genotype data. The set comprises 19 patients from MitoNET with known causal genes. Therefore, this set was used to evaluate the performance of the whole PheNoBo pipeline on different combinations of patient data types. Figure 23 gives an overview of the prediction results of PheNoBo for this data set.

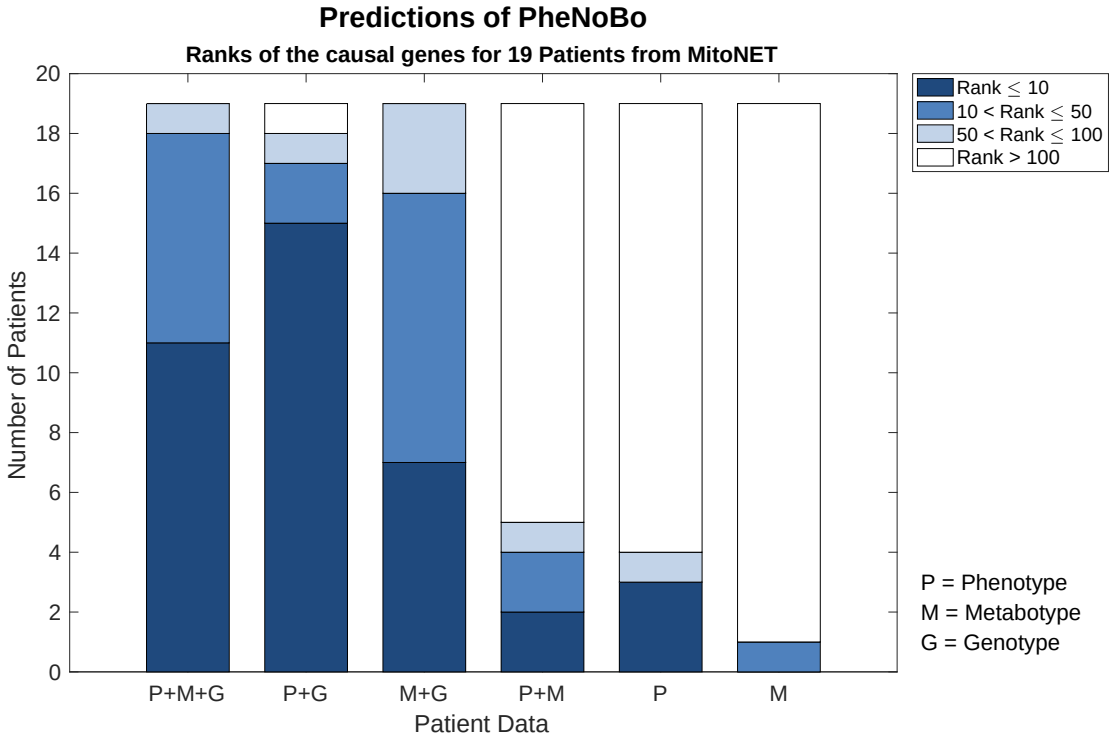


Figure 23: Predictions of PheNoBo for different types of patient data from MitoNET (patient data set 7). The figure shows the number of patients whose causal gene gets a rank of at most 10 (dark blue bars), at most 50 (dark and medium blue bars) and at most 100 (dark, medium and light blue bars) in the predictions of PheNoBo. The height of the white part of the bar gives the number of patients with a causal gene that has a rank of more than 100. Each bar corresponds to a different combination of patient data types. P = phenotype data, M = metabotype data, G = genotype data.

The results for patient data set 7 were evaluated in terms of the rank of the known causal gene in the final predictions of PheNoBo. These results show a general trend that the

sensitivity of PheNoBo increases with the number of patient data types used. The phenotype or metabotype data alone are often not sufficient to predict the known causal gene with a rank below 100. The combination of phenotype and metabotype improves the predictions from one causal gene with a rank below 100 (metabotype) to five genes with a rank below 100. The number of causal genes with a rank below 100 increases considerably when integrating genotype data into the predictions of PheNoBo. The predictions of PheNoBo are most accurate if phenotype, metabotype and genotype data are combined: five causal genes get a rank of 1, 11 causal genes get a rank of at most 10 and all known causal genes have a rank below 100.

The following subsections provide more detailed results and analyses of the phenotype-based predictions (section 3.4.1), metabotype-based predictions (section 3.4.2) and the predictions based on different combinations of data (section 3.4.3).

3.4.1 Evaluation on Phenotype Data

The phenotype data of the 19 MitoNET patients was analyzed by the application of Phenomizer for PhenoDis (section 2.3.1), PhenoToGeno (section 2.3.3) and GeneticNetworkScore (section 2.3.4). Table 9 gives an overview of the resulting ranks of the causal genes and of the likely causal diseases. In this context the likely causal disease refers to the disease that is associated with the causal gene and that is most similar to the patient's symptoms. The results show that the rank of the causal gene in the predictions of GeneticNetworkScore correlates with the rank of the corresponding disease in the results of Phenomizer for PhenoDis. This is illustrated for three patients.

Patient 300087: The causal gene of the patient, MRPL44, has only one known associated disease: Infantile hypertrophic cardiomyopathy due to MRPL44 deficiency (ORPHA352563). Both, the disease and the gene get a high rank in PheNoBo. This cardiomyopathy was discovered recently [24] and is characterized by cardiac symptoms and liver problems in PhenoDis. However, the only symptom of patient 300087 related to the disease is "Abnormality of cardiovascular system physiology" (HP:0000822) [86]. This is an unspecific symptom that is not further described in the patient data provided by MitoNET. The case study [31] gives a more complete description of the patient's cardiac problems that fits the known symptoms of the disease. The current annotation of symptoms in PhenoDis for this type of cardiomyopathy might be incomplete. There are only four known cases of the disease in the scientific literature (including this patient and another patient from MitoNET) [24, 31]. It is difficult to judge if the additional symptoms of the patient 300087 like brain le-

sions and migraine are also due to the defect in MRPL44 [31]. Therefore, the disagreement between the patient's phenotype and the current knowledge about the phenotype caused by MRPL44 mutations leads to a high rank in the predictions of PheNoBo.

Patient 300014: PheNoBo fails completely to predict the causal gene TRIM46 for patient 300014. This is due to the fact that the function of TRIM46 is not yet fully understood. A recent study demonstrated that TRIM46 is required for establishing the polarity of neurons [107]. The disease-gene data set (section 2.1.4) underlying PheNoBo does not list any diseases for TRIM46. Consequently, there is no rank for the causal disease of patient 300014 and TRIM46 gets a rank above 10 000.

Patient 300075: Patient 300075 provides an example where the likely causal disease and the causal gene for patient get a low rank. The patient most likely suffers from Leigh syndrome which can be caused by defects in the gene MTFMT [105]. The patient has many typical symptoms of Leigh syndrome like elevated lactate levels, neurological symptoms affecting the movement and muscle weakness including cardiomyopathy [82]. The high similarity between Leigh Syndrome and the patient's phenotype as well as the known link between Leigh syndrom and MTFMT lead to low ranks in the predictions of PheNoBo. Patient 300075 provides an example with low ranks for the causal gene and the associated disease.

The examples above show that the performance of PheNoBo is dependent on the prior knowledge provided about the causal gene. The phenotype data of MitoNET provides challenging data for PheNoBo. The patients were taken up into the MitoNET register because standard diagnostic methods failed to find the cause of their disease. As illustrated by the examples of patient 300087 and 300014, most patients of the data set suffer from an atypical representation of their disease or a previously unknown disease. Nevertheless, the phenotype analysis alone can successfully predict the causal gene for three patients with a rank below 10. Furthermore, including genotype and metabotype data into the predictions improves the prediction results considerably, e.g. MRPL44 of patient 300087 ends up with a rank below 20 in the final predictions (table 11).

Patient	Causal Gene	Rank Phenotype Analysis	Enrichment Score	Rank Phenomizer
300003	ADAR	295	1.005	328
300004	SUCLA2	296	0.731	266
300007	SURF1	61	1.44	41
300013	IARS2	1 175	0.043	3 243
300014	TRIM46	12 131	-0.832	–
300017	ACAD9	694	0.752	259
300030	SLC6A8	738	0.403	642
300040	CYP4X1	19 926	-0.714	–
300049	ECHS1	122	1.601	29
300073	TTC19	644	0.431	272
300075	MTFMT	7	2.429	1
300076	C1QBP	4 542	-0.512	–
300087	MRPL44	3 685	0.196	5 597
300093	ECHS1	387	0.986	360
300106	PRPS1	3	2.417	4
300151	OPA1	261	0.657	409
300152	ANO5	314	0.953	201
300219	C10orf2	2	2.892	1
300230	SACS	158	1.451	112

Table 9: Predictions of PheNoBo on the phenotype data of patient data set 7 from MitoNET.
Rank Phenotype Analysis = rank of the causal gene in the result of GeneticNetworkScore (section 2.4.2), these ranks are summarized in figure 23 (bar labeled *P*),
Enrichment Score = score calculated by GeneticNetworkScore (equation 2.15), a score higher than 0 indicates that the score of the gene is higher than expected by random, a score lower than 0 indicates that the score is lower than expected (section 2.3.4),
Rank Phenomizer = lowest rank in Phenomizer for PhenoDis for a diseases associated with the causal gene in the disease-gene data (section 2.1.4).

3.4.2 Evaluation on Metabotype Data

The analysis of the metabotype data of the 19 MitoNET patients was done with ScoreMetabolites (section 2.3.2), MetaboToGeno (section 2.3.3) and GeneticNetworkScore (section 2.3.4). The results of the analysis focus on the rank of the causal gene and the lowest rank of a metabolite associated with the gene (table 10). Overall the ranks obtained by the analysis of the metabotype are higher than the ranks of the phenotype-based predictions. The metabotype-based predictions of PheNoBo show a dependence between the rank of the causal gene and the rank of the associated metabolite. To support and explain this observation, the results for three patients are introduced below.

Patient 300004: The causal gene of the patient SUCLA2 and the associated metabolite phosphate get both a rank above 100 in the predictions. Phosphate is an unspecific metabolite that is produced when the enzyme encoded by SUCLA2 transforms succinic acid into succinyl-CoA [14]. The concentrations of succinic acids and succinyl-CoA are not covered by the reference set provided from MitoNET and therefore are not taken into account in the analysis. However, the metabolite profile of patient 300004 shows a strong increase in succinylcarnitine. The publication [55] reported elevated succinylcarnitine for two cases with a mutation in SUCLA2. The indirect link between SUCLA2 and succinylcarnitine is not part of the metabolite-gene data set underlying PheNoBo (section 2.1.5). Therefore, the rank of SUCLA2 is likely to improve when completing the metabolite-gene data used by PheNoBo.

Patient 300151: Patient 300151 suffers from a defect in the gene OPA1. OPA1 has a rank of more than 1 000 in the prediction results as it does not have any associated metabolites. The lack of associated metabolites for OPA1 is due to the function of OPA1. OPA1 does not encode a classical metabolic enzyme but a GTPase that mediates the fusion of mitochondria [79]. A defect in OPA1 does not cause specific changes in the metabolism that could be interpreted by PheNoBo.

Patient 300007: PheNoBo yields a rank of 49 for SURF1 which is the causal gene of patient 300007. SURF1 is required for the assembly of complex IV of the respiratory chain [67]. Although SURF1 is not a classical metabolic enzyme, it is associated with lactate in the metabolite-gene data set used by PheNoBo (section 2.1.5). Lactate gets a rank of 8 in the results of ScoreMetabolites as the metabolite profile of patient 300007 is characterized by an abnormal lactate concentration. Elevated lactate is an unspecific indicator for mitochondrial diseases [45]. Thus, the link between SURF1 and lactate is reasonable and leads to a low rank of SURF1 in the predictions. Most other patients of the patient data set, who also suffer from mitochondrial diseases, do not have elevated lactate levels. Elevated lactate levels have

serious effects on the health and should be treated immediately [6]. For this reasons, other causal genes that are associated with elevated lactate levels, like C10orf2, get high ranks in the predictions.

The results show again that the patients from MitoNET are a challenging test set for PheNoBo. The evaluation of the metabotype-based predictions on this patient data set is limited by the high number of patients, who do not suffer from a classical metabolic disease such as patient 300151. Therefore, the performance of PheNoBo is expected to be better on patients with classical metabolic defects. This assumption is also supported by the results of section 3.5.2 showing that the metabotype analysis is able to interpret metabolite profiles in terms of causal genes. Furthermore, the results for patient 300004 also demonstrate the importance of indirect metabolite-gene associations (i.e. that do not originate from direct biochemical reactions) for the predictions. As already stated in section 3.2.3.1 the set of indirect metabolite-gene associations of PheNoBo is incomplete.
Despite these shortcomings, the signal of the metabotype analysis is sufficient to predict the causal gene when using additional genotype data. All 19 causal genes of the data set get a rank below 100 and seven genes even get a rank below 10, when running PheNoBo with genotype and metabotype data (table 11).

Patient	Causal Gene	Rank Metabotype Analysis	Enrichment Score	Rank Score-Metabolites
300003	ADAR	6 155	-1.321	–
300004	SUCLA2	904	0.519	120
300007	SURF1	49	1.853	8
300013	IARS2	313	0.942	71
300014	TRIM46	12 084	-1.292	–
300017	ACAD9	1 767	-0.217	–
300030	SLC6A8	939	0.343	153
300040	CYP4X1	1 909	-0.265	209
300049	ECHS1	306	0.929	38
300073	TTC19	296	0.923	34
300075	MTFMT	1 154	0.351	224
300076	C1QBP	4 279	-0.802	–
300087	MRPL44	2 039	-0.322	361
300093	ECHS1	1 187	0.215	132
300106	PRPS1	506	0.61	78
300151	OPA1	3 906	-0.824	–
300152	ANO5	13 810	-1.408	–
300219	C10orf2	1 890	0.006	329
300230	SACS	10 753	-1.341	–

Table 10: Predictions of PheNoBo on the metabotype data of patient data set 7 from MitoNET.
Rank Metabotype Analysis = rank of the causal gene in the result of GeneticNetworkScore (section 2.4.2), these ranks are summarized in figure 23 (bar labeled *M*),
Enrichment Score = score calculated by GeneticNetworkScore (equation 2.15), a score higher than 0 indicates that the score of the gene is higher than expected by random, a score lower than 0 indicates that the score is lower than expected (section 2.3.4),
Rank ScoreMetabolites = lowest rank in ScoreMetabolites for a metabolite associated with the causal gene in the metabolite-gene data (section 2.1.5).

3.4.3 Combination of Data

The previous analyses showed that the predictions based only on one type of patient data are often insufficient for predicting the causal gene with PheNoBo. The main idea of PheNoBo is not to use a single type of data but to combine evidence from several patient data types into a single prediction. As patient data set 7 from MitoNET provides phenotype, metabotype and genotype data, PheNoBo was tried on all combinations of these types. The prediction results can be used to investigate the contributions of the individual data types to the predictions. The results are evaluated and compared by considering the rank of the causal gene for each patient of the data set for all possible combination of data types (table 11). The ranks are determined from the position of the causal gene in the results of CombineScores (section 2.3.5). CombineScores is the last step of PheNoBo and combines the individual predictions of the phenotype, metabotype and genotype analysis.

The ranks obtained for combining metabotype and phenotype data are ten times higher than the ranks which result from combining metabotype and genotype or phenotype and genotype (except for patients 300106 and 300219, who already got low ranks in the phenotype analysis). This observation shows that the genotype data considerably improves the predictions of PheNoBo. There are usually many genes which could explain the patient's phenotype or metabotype, e.g a disease is on average associated with 2 genes and a metabolite is associated with 17 genes (table 8). However, if the patient does not have a harmful sequence variant within a gene that fits his phenotype or metabotype, this variant gets a high rank in the prediction of CombineScores. The inclusion of genotype data allows to filter the prediction results based on the phenotype or metabotype data. This filtering results in a considerable decrease of the rank of the causal gene.

The 19 MitoNET patients can be divided into three subgroups according to their rank in the predictions for the combination of all three types of data. The first group comprises all patients whose causal gene has a rank below 10 (eleven patients). All patients of the group except for 300017 have a causal gene that is associated with at least one disease and one metabolite in the underlying data sets of PheNoBo. The second group consists of the patients with a causal gene that has a rank in the range from 10 to 25 (six patients). The causal genes for the group except for MRPL44 do have either no associated metabolites or no associated diseases in the data sets of PheNoBo. These patients show decreased ranks in the predictions of PheNoBo if one patient data type is excluded from PheNoBo. The causal genes without associated diseases get a lower rank for the combination of metabotype and genotype data. Inversely, the causal genes without related metabolites get a lower rank for the combination of phenotype and genotype data. The last group includes two patients whose

causal genes have a rank of more than 25 (but less than 100) in the combined predictions. These causal genes do not have any associated diseases and metabolites in the data sets used by PheNoBo. This shows that the prediction results of PheNoBo are dependent on the underlying disease-gene and metabolite-gene data (section 2.1.4 and 2.1.5). If the causal gene is part of both data sets, PheNoBo is able to make good predictions for the gene (the causal gene is within the 10 top scoring genes). However, if there is no information (no disease and no metabolites) about the causal genes in the data, the predictions of PheNoBo are less accurate. This implies that one requires special approaches for specifically detecting causal genes without any information in the data sets. One possibility would be to exclude all genes from the final ranking of CombineScores with known disease- and metabolite-gene associations.

In summary the results show a good overall performance for the combination of phenotype, metabotype and genotype data. The majority of causal genes has a rank below 10. The predictions considerably improved compared to the phenotype-based and metabotype-based predictions (table 9 and table 10). For example, the causal gene of patient 300049, ECHS1, initially got a rank above 100 in the phenotype analysis and in the metabotype analysis. The combination of all three types of data predicts the gene correctly at rank 1.

Patient	Causal Gene	Disease	Metabolite	Rank all	Rank P+G	Rank M+G	Rank P+M
300003	ADAR	✓	✗	11.2	3.0	39.7	1 660
300004	SUCLA2	✓	✓	3.1	2.5	8.5	267
300007	SURF1	✓	✓	1.0	1.2	1.0	13
300013	IARS2	✓	✓	3.8	8.1	3.9	405
300014	TRIM46	✗	✗	82.3	72.1	77.3	13 085
300017	ACAD9	✓	✗	6.4	5.9	15.8	738
300030	SLC6A8	✓	✓	5.8	5.6	9.8	465
300040	CYP4X1	✗	✓	23.8	120.9	14.8	3 682
300049	ECHS1	✓	✓	1.0	2.4	3.1	67
300073	TTC19	✓	✓	2.2	5.1	3.7	236
300075	MTFMT	✓	✓	1.0	1.0	10.0	29
300076	C1QBP	✗	✗	27.9	24.2	29.5	4 083
300087	MRPL44	✓	✓	15.5	19.6	18.8	1 748
300093	ECHS1	✓	✓	4.7	3.2	10.8	271
300106	PRPS1	✓	✓	1.0	1.0	7.0	5
300151	OPA1	✓	✗	10.3	1.6	26.3	1 396
300152	ANO5	✓	✗	23.9	4.4	94.9	3 355
300219	C10orf2	✓	✓	1.0	1.0	17.7	5
300230	SACS	✓	✗	11.1	1.8	68.7	1 470

Table 11: Predictions of PheNoBo on patient data set 7 from MitoNET. The set comprises phenotype, metabotype and genotype data for 19 MitoNET patients suffering from rare mitochondrial disorders. The predictions of PheNoBo on the patients were assessed using the rank (i.e. position within the sorted results, section 2.4.2) of the causal gene in the output of CombineScores. Note that there were 106 simulated genotype data sets per patient. The ranks reported for predictions using the genotype data are the average ranks of all 106 genotype data sets per patient. checkmark = causal gene of the patient has an associated disease/metabolite in the disease-/metabolite-gene data set (section 2.1.4 and 2.1.5), cross = causal gene of the patient does not have an associated disease/metabolite in the disease-/metabolite-gene data set, all= Combination of phenotype, metabotype and genotype (bar *P+M+G* in figure 23), P+G = Combination of phenotype and genotype (bar *P+G* in figure 23), M+G = Combination of metabotype and genotype (bar *M+G* in figure 23), P+M = Combination of phenotype and metabotype (bar *P+M* in figure 23).

3.5 Application of PheNoBo on Real Patient Data

This section illustrates the results of PheNoBo for the participants of KORA F4 (patient data set 5) and for the MitoNET patients without known causal genes (patient data set 8). These patient data sets differ from the previously presented sets: the causal genes for the members of the sets are unknown. PheNoBo was applied on these data to learn more about the causal genes of the subjects. The first part of the section (section 3.5.1) analyzes the predictions of PheNoBo for the members of KORA F4. The second part (section 3.5.2) deals with the predictions of PheNoBo for the patient data from MitoNET.

3.5.1 Predictions for Participants of KORA F4

The following paragraphs present the results of PheNoBo for patient data set 5. The set consists of metabotype and genotype data from 106 participants of the population-based KORA F4 cohort (section 2.2.3). These participants are not known to suffer from rare genetic diseases. Therefore, this patient data set allows to find out how PheNoBo interprets data from supposedly healthy subjects. The analysis starts with interpreting the top scoring metabolites of the patient data set and finally presents the predicted causal genes.

Top Scoring Metabolites: ScoreMetabolites (section 2.3.2) was applied on the patient data set from KORA using the MitoNET reference data set. The KORA reference set was not available at the time of the analysis. Section 3.3.3 already demonstrated that the reference sets of KORA and MitoNET are to some extent exchangeable. Nevertheless, the inspection of the top scoring metabolites of ScoreMetabolites is necessary to learn about artifacts due to the MitoNET reference set. Table 12 lists the three most frequently occurring top scoring metabolites for patient data set 5.
The high frequencies of the top scoring metabolites are most likely caused by differences in the preparation of the metabolite samples from data set 5 and the MitoNET reference. Biliverdin and indolelactate have a missingness of 0% in the MitoNET reference set but a missingness of more than 15% in the KORA samples. All samples from data set 5 with a missing value for biliverdin or indolelactate will get a high score when comparing them with the MitoNET reference. The increased missingness of the metabolites in KORA might result from the outlier removal that was done for the KORA samples but not for the MitoNET reference. This argumentation does not hold for the third metabolite, the dipeptide phenylalanylphenylalanine (Phe-Phe). The metabolite has a lower missingness in the KORA samples than in the MitoNET reference.

The high frequency of Phe-Phe as top scoring metabolite results from the distribution of the measured Phe-Phe metabolite levels in the MitoNET control group (figure 24). The measured concentrations of the individual control samples from MitoNET are widely spread around the median. Hence, the distribution is expected to have a large standard deviation. The standard deviation of Phe-Phe is estimated seperately from each phenotype group. As the phenotype groups are composed of less than 50 samples per group, the standard deviation of the MitoNET reference for Phe-Phe might be inaccurate. The actual standard deviation observed in a population for Phe-Phe might be higher than the standard deviation in the reference set. The inaccurate standard deviation results in an overestimation of the score of Phe-Phe by ScoreMetabolites.
The scores of the metabolites discussed above are the most strongly biased metabolite scores obtained for this patient data set. However, the reference values for most metabolites in the KORA and the MitoNET samples are similar (section 3.3.3).

Metabolite	Number of Subjects	Missingness KORA in %	Missingness MitoNET in %
Biliverdin	18	35.3	0.0
Phenylalanylphenylalanine	11	0.3	21.6
Indolelactate	7	17.3	0.0

Table 12: Most frequent top scoring metabolites of patient data set 5 from KORA. The table shows the three metabolites which get most frequently a rank of 1 in the predictions of ScoreMetabolites. Number of Subjects = number of subjects in patient data set 5 for which the metabolite is the top scoring metabolite in ScoreMetabolites, Missingness KORA = missingness of the metabolite in the 1 768 samples from KORA F4 (section 2.2.3). Missingness MitoNET = missingness of the metabolite in the 97 control samples from MitoNET (section 2.2.4). Note that patient data set 5 is a subset of the 1 768 samples from KORA F4.

Predicted Causal Genes: The application of PheNoBo on the genotype and metabotype data from KORA yielded a predicted causal gene for every member of the patient data set. Some genes are predicted frequently as causal genes for the data set. Altogether there are four putative causal genes that are predicted for 70 of the 106 samples under consideration (table 13).
These four genes are not known to be associated with a disease (i.e they are not listed in the disease-gene data presented in section 2.1.4). Furthermore, all variants found within the genes have a minor allele frequency of more than 20%. A genetic variant with a high frequency is unlikely to cause a rare genetic disorder. Hence, these observations are in line

with the assumption that the subjects of KORA F4 do not suffer from rare genetic diseases. The technical bias shown above has only a minor influence on the predictions of PheNoBo. The predictions of Cysteine conjugate-beta lyase, cytoplasmic (CCBL1) might be partly due to the bias. CCBL1 is associated with indolelactate in the metabolite-gene set (section 2.1.5). However, the gene is known to affect the levels of five additional metabolites, including methionine and glutamine. It is likely that not all predictions of CCBL1 result from spuriously elevated scores of indolelactate.

Finally, three of the four genes in table 13 are part of the metabolite-gene associations provided by the GWAS Server (section 2.1.5). GWAS Server collects information about common polymorphisms that influence the human metabolism [91]. This finding indicates that the variants in the predicted causal genes indeed have a visible effect on the metabolite profiles of their carriers. In summary, the four genes predicted frequently by PheNoBo are able to explain the subject's metabotype without implying a strong impact on the subject's health.

There are interesting exceptions to this statement. The predicted genes for four participants of KORA are associated with a disease and contain rare LOF mutations with a minor allele frequency below 1%. A loss-of-function variant (LOF) is a deleterious mutation that completely inactivates the affected gene. All four subjects are heterozygous for the rare LOF, i.e. they have one inactivated and one still functioning copy of the gene. Therefore, the carriers of the LOF are likely to experience no or at most a very mild effect on their health. Gathering phenotype data for these four interesting cases would allow a more complete interpretation of these predictions. The phenotype data would further provide an opportunity to re-analyze the patients with PheNoBo and the reference set from KORA.

Gene Name	Number of Subjects with Prediction	Number of Subjects with LOFs	Minor Allele Frequency
Cysteine conjugate-beta lyase, cytoplasmic	22	90	0.48
Organic anion transporter 5	17	73	0.45
Organic cation transporter 1	16	77	0.42
Cytochrome P450 family 2 subfamily C member 8	15	44	0.23

Table 13: Most frequent top scoring genes of patient data set 5 from KORA. The table shows the four genes which are most frequently predicted as causal genes by PheNoBo. Number of Subjects with Prediction = number of subjects in patient data set 5 for which the gene is predicted as causal, Number of Subjects with LOF = number of subjects in patient data set 5 that have a loss-of-function variant (LOF) in the gene, Minor Allele Frequency = lowest minor allele frequency of a LOF in the gene.

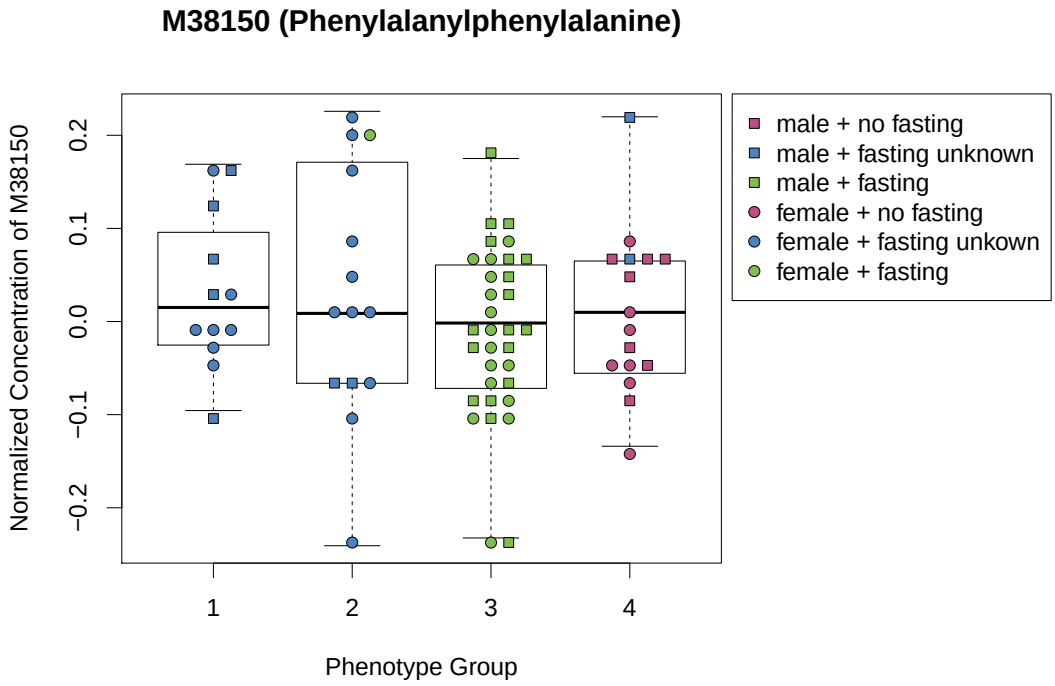


Figure 24: Reference concentrations of the metabolite phenylalanylphenylalanine (Phe-Phe). The figure shows a box plot giving the distribution of the concentration of Phe-Phe in the metabolite reference set derived from MitoNET (section 2.3.2.1). There is one box for each of the four different phenotype groups (group 1 = infants, group 2 = children, group 3 = fasting adults, group 4 = non-fasting adults). The position of the box indicates the interquartile range (between the first and the third quartile) of the concentration. The thick line in the box gives the median concentration. The squares and circles drawn in the boxes represent the 97 MitoNET control samples from which the reference set was created (section 2.2.4). The x-coordinate of a sample corresponds to its phenotype group. The y-coordinate shows the concentration of Phe-Phe in the sample.

3.5.2 Predictions for MitoNET Patients without Known Causal Gene

The last patient data set, set 8, consists of phenotype and metabotype data from 93 patients of the MitoNET register (section 2.2.4). The causal genes of these patients are unknown. Patient data set 8 represents a realistic use case for PheNoBo as the patients suffer from rare mitochondrial disorders. The predictions of PheNoBo for the set provide ranked suggestions of potentially causal genes. These suggestions could be ruled out or verified by additional, targeted examinations of the patients. This section explains and discusses the predictions for three patients of the set. The patients were chosen to present different types of mitochondrial diseases.

Patient 300079: PheNoBo predicts DNA polymerase gamma, catalytic subunit (POLG) as causal gene for the disease of patient 300079. POLG is a subunit of DNA polymerase γ which replicates the mitochondrial genome. POLG is not part of the mitochondrial genome itself. The subunit is encoded in the nuclear genome and imported into the mitochondrion after translation [51].
The prediction of POLG as causal gene is consistent with the patient's phenotype. The symptoms of patient 300079 describe mainly neurological and motor problems. In addition the patient suffers from seizures. These symptoms are clinical features of the disease Mitochondrial spinocerebellar ataxia and epilepsy (MSCAE) [106]. MSCAE is predicted by Phenomizer for PhenoDis with a p value below 0.001 (rank 7). As a consequence the causal gene of MSCAE, POLG, [106] gets a rank of 2 in the phenotype analysis. The most likely causal gene predicted for the patient's phenotype is C10orf2. C10orf2 encodes Twinkle, the mitochondrial DNA helicase. Twinkle is part of the replication fork of the mitochondrial replication and interacts with polymerase y [51]. Hence, the disease of patient 300079 is likely due to a defect in the maintenance and replication system of the mitochondrial genome.
The metabotype of patient 300079 supports the results of the phenotype analysis. Most of the deviating metabolites (high scoring metabolites identified by ScoreMetabolites) of the patient are unknown. The most strongly deviating known metabolite is citrulline. The concentration of the amino acid citrulline is strongly reduced in the patient's sample. The patient's hypocitrullinemia is indicative of decreased mitochondrial function [9]. A defect of POLG affects all genes of the mitochondrial genome. As the mitochondrial genome contains the most essential components of the mitochondrion, a defective polymerase γ reduces all functionality of the mitochondrion. However, citrulline is a relatively unspecific marker for many types of mitochondriopathies. More than 50 genes are associated with citrulline in the metabolite-gene data underlying PheNoBo. For this reason POLG gets a

rank of 46 in the predictions of the metabotype analysis.
Overall, the metabolite profile of patient 300079 hints at a disease with mitochondrial involvement. The phenotype analysis narrows down the diagnostic suggestion to MSCAE which is caused most likely by POLG.

Patient 300163: The top scoring gene for patient 300163 is Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 6 (MT-ND6). MT-ND6 is part of the mitochondrial genome and encodes a subunit of complex I. Complex I is located in the inner mitochondrial membrane and participates in the respiratory chain. Therefore, MT-ND6 is an essential component for the production of chemical energy (in form of ATP) [4].
The phenotype of the patient supports this prediction. Patient 300163 suffers from visual problems, neuropathy and muscle weakness. These symptoms fit quite well the annotations of the disease Leber optic atrophy and dystonia (LDYT) in PhenoDis. LDYT gets a rank of 18 and a p value below 0.05 in the predictions of Phenomizer for PhenoDis. Defects of MT-ND6 are known to cause LDYT [92]. The prediction of MT-ND6 is further supported by the patient's symptom "Decreased activity of mitochondrial complex I". Finally, the phenotype analysis yields a rank of 16 for MT-ND6. Most top-scoring genes of the phenotype analysis are components of the respiratory chain. The phenotype of patient 300163 hints at an impairment of the electron transport in the mitochondria.
The interpretation of the patient's metabolite profile is less straightforward. The set of deviating metabolites comprises two classes of metabolites: metabolites with unknown chemical structure and steroid hormones like cortisol, cortisone, androsterone and pregnenolone [14]. The levels of the steroid hormones are strongly reduced in the patient's sample. Altogether this metabotype is indicative of adrenal insufficiency (HPO:0000846). The HPO [86] defines this symptom as "Insufficient production of steroid hormones (primarily cortisol) by the adrenal glands as a result of a primary defect in the glands themselves". This symptom is missing in the patient's phenotype. The phenotype data of the MitoNET patients does not provide any information about abnormalities of the adrenal gland. However, adrenal insufficiency was recently recognized as symptom of MELAS syndrome [1]. The link between adrenal insufficiency and MELAS syndrome is supported by PhenoDis. The metabotype goes along with the prediction of MT-ND6 as mutations in the gene are also known to cause MELAS [30]. The metabolite-gene data of MetaboToGeno comprises the link between MT-ND6 and cortisol, but there are no links to other steroid hormones. Therefore, MT-ND6 ends up with a rank of 92 in the metabotype-based prediction. The top scoring genes of the metabotype analysis are mitochondrial t-RNA genes which can also cause MELAS syndrome [30].
Combining phenotype and metabotype analysis into a single prediction yields MT-ND6 as predcited causal gene. Defects in the gene can cause LDYT as well as MELAS syndrome.

Patient 300163 might suffer from a combination of both diseases.

Patient 300224: Patient 300224 is predicted to suffer from a defect of Carnitine Palmitoyltransferase 2 (CPT2). CPT2 is part of the pathway for importing long-chain fatty acids into mitochondria (figure 25). The import pathways enables the β-oxidation that breaks down fatty acids and generates energy from them.
The phenotype of patient 300224 provides little information for the predictions of PheNoBo. MitoNET reports only five symptoms for the patient that mainly refer to abnormalities of the musculature. The major symptom of the patient is myoglobinuria, i.e. the "presence of myoglobin in the urine" (HPO:0002913) [86]. As a consequence Phenomizer for PhenoDis suggests myoglobinuria as causal disease for patient 300224. Myoglobinuria is stored in PhenoDis as a symptom and as a disease. LPIN1 is known to be the causal gene of the disease myoglobinuria. Hence, the phenotype analysis of PheNoBo predicts LPIN1 as causal gene. LPIN1 encodes a phosphatidic acid phosphohydrolase and regulates the fatty acid metabolism [84].
There is also a disease caused by CPT2 that fits the symptoms of the patient: the myopathic form of CPT2 deficiency. The disease, which is a relative mild subtype of CPT2 deficiency, is associated with muscle weakness and myoglobinuria [113]. The myopathic form of CPT2 deficiency gets a rank of 562 in predictions of Phenomizer for PhenoDis for patient 300224. The reason for this is that PhenoDis annotates the myopathic form of CPT2 deficiency to rhabdomyolysis rather than to myoglobinuria. Rhabdomyolysis is the "breakdown of muscle fibers that leads to the release of muscle fiber contents (myoglobin) into the bloodstream" (HPO:0003201) [86]. The release of myoglobin into the blood finally leads also to the presence of myoglobin in the urine. Despite the causal relationship between rhabdomyolysis and myoglobinuria, these symptoms are not related in the HPO. The HPO classifies rhabdomyolysis as abnormality of the musculature and myoglobinuria as abnormality of the metabolism. Therefore, the disease CPT2 deficiency gets a rank of 562 and the corresponding gene CPT2 a rank of 286 in the phenotype analysis.
The metabotype of patient 300224 provides evidence for a defect in CPT2. The process of rhabdomyolysis is not detectable in the patient's sample as myoglobin is not part of the reference set. However, the metabotype shows an interesting pattern hinting at the underlying molecular mechanism that leads to the breakdown of muscle fibers. The metabolite profile of the patient shows a strong elevation of the acylcarnitine levels, e.g. for oleoylcarnitine, palmitoylcarnitine and stearoylcarnitine. The concentration of free carnitine is remarkably decreased. This metabolic pattern is characteristic for a defect in the import of fatty acids into the mitochondrion (figure 25) [97]. In accordance with the pattern, the metabotype analysis predicts three top scoring genes: CPT1A, CACT and CPT2. CPT1A links carnitine and a fatty acids for forming an acylcarnitine. However, a loss-of-function

of CPT1A would result in opposite changes in the metabolite profile: increase of carnitine and decrease of acylcarnitine. PheNoBo currently does not consider the direction of change in the metabolite profile. The second predicted gene, CACT, transfers the acylcarnitine across the inner mitochondrial membrane. CACT is able to explain the metabotype of the patient. The phenotype of the corresponding disease CACTD is usually more severe than the patient's phenotype [97]. CPT2 removes carnitine from the acylcarnitine to release the fatty acid into the mitochondrial lumen for β-oxidation.
Therefore, the final combined prediction of PheNoBo suggests CPT2 as potential cause of the patient's condition.

The sample results from patient data set 8 confirm again that the predictions of PheNoBo profit from combining phenotype and metabotype data. For patient 300224 the metabotype data was decisive for the prediction whereas the phenotype data was helpful to analyze patient 300163.
The descriptions given above focus on the top scoring gene of each patient. The top scoring genes usually are those genes which have known associations with the patient's phenotype and metabotype. Nevertheless, it is reasonable to consider also alternative causal genes that might result from atypical or unknown forms of rare genetic diseases. For example, CACT could be an alternative causal gene for patient 300224. The disease of patient 300079 might result from a different mitochondrially-encoded subunit of complex I. These considerations could be supported by using PheNoBo with additional genotype data of the patients.
Unfortunately, the genotype data were not available within the scope of this master's thesis.
In summary, the predictions of PheNoBo can facilitate and speed up the diagnostic process for the MitoNET patients. The predictions of PheNoBo can be used e.g. to prioritize the genetic variants of the patients for targeted verification experiments.

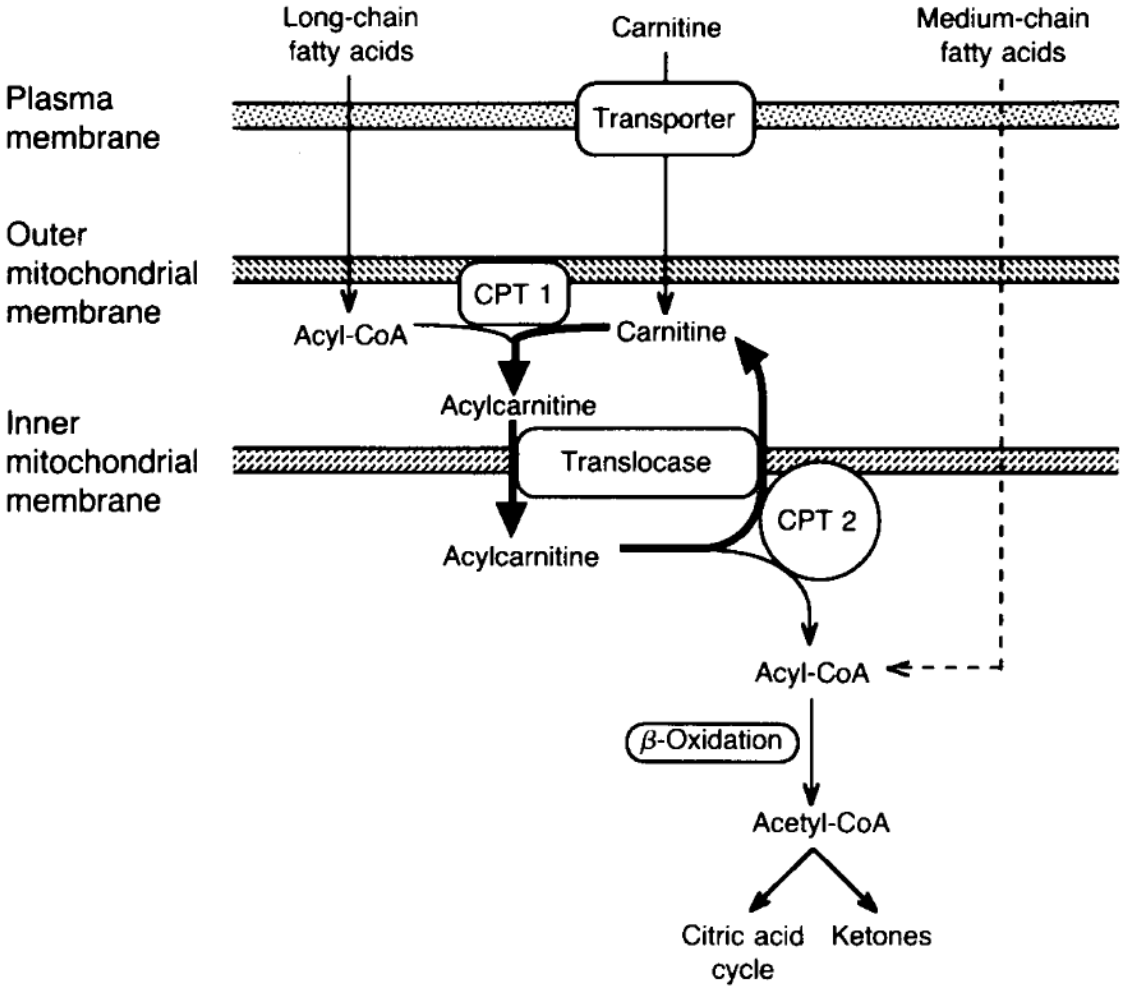


Figure 25: Import of long-chain fatty acids into the mitochondrion. Long-chain fatty acids cannot directly enter the mitochondrion for β-oxidation. The import of long-chain fatty acids into the mitochondrion requires carnitine and 3 proteins: CPT1= Carnitine Palmitoyltransferase 1, Translocase = Carnitine/Acylcarnitine Translocase (CACT) and CPT2=Carnitine Palmitoyltransferase 2. The image is taken from [97].

4 Conclusion and Outlook

This master's thesis introduces PheNoBo, a new tool to predict the causal gene from omics-data of an individual patient suffering from a genetic disorder. Most computational methods for identifying causal genes consider only genotype data from sequencing the patient's genome. A few tools include two types of omics-data into the prediction procedure: Phen-Gen uses genotype and phenotype data and the approach of Guo *et al.* relies on genotype and metabotype data. PheNoBo, which is short for the Combination of **Phe**notype, Ge**No**type and Meta**Bo**type, successfully combines the procedures of Phen-Gen and of Guo *et al.* to use all three different data types. PheNoBo is realized as modular pipeline in KNIME. Therefore, PheNoBo is able to produce separate predictions for each type of data that are finally united into a single diagnostic suggestion. The modular setup is a valuable advantage of PheNoBo, as it allows to understand and interpret the prediction process step-by-step.

The second main result of the thesis is the application of PheNoBo on patients from the MitoNET register who suffer from rare mitochondrial diseases. These patients are challenging test cases for PheNoBo because they are affected by known diseases with atypical presentations or even by previously unknown diseases. Nevertheless, PheNoBo performs well on these patient data. The procedure predicted the causal gene correctly for more than one fourth of the patients with known etiology. The causal gene of more than half of these patients was found within the top 10 scoring genes predicted by PheNoBo. In addition PheNoBo produced testable hypotheses about the causal genes of the remaining patients from MitoNET with unclear diagnosis. These predictions facilitate the interpretation of the genotype data from these patients and help to finally uncover the genetic cause of their disorder.

These results provide a proof-of-principle for the general concept of PheNoBo to integrate phenotype, genotype and metabotype data. The PheNoBo pipeline is a solid foundation for the development of a computational tool that can be applied in clinics and research projects on patients with rare diseases. Such a tool requires basically three extensions.

The first extension deals with the data basis and the algorithms of the metabotype analysis. The research about the human metabolome started recently. Therefore, there are only few standardized methods for interpreting metabolite profiles in terms of the causal genes. Trying different data sets and approaches for the metabotype analysis has the potential to further improve the predictive quality of PheNoBo.
The quality of the predictions could profit from the inclusion of more data about indirect associations between metabolites and genes, i.e. data that is not directly derived from biochemical reactions. The metabolite-gene data could be even further enhanced by adding relations between chemical entities. The Chemical Entities of Biological Interest (ChEBI) database [47] offers an ontology for this purpose.
The construction of a biochemical network is another promising approach for the metabotype analysis. The network represents metabolic reactions and the enzymes catalyzing and regulating the reactions. Executing a random walk with restart on the metabolic network is an alternative way to derive gene scores from the patient's metabolite profile. A potential benefit of such a method is the integration of metabolites that are not part of the reference set required for the metabotype analysis.

Another valuable extension of PheNoBo is the integration of additional types of patient data. The modular architecture of PheNoBo allows adding new algorithms and analysis pipelines for making predictions based on these new data. For example, MitoNET collects also transcriptomics data from RNA-sequencing experiments and proteomics measurements for many patients. The integration of the metabotype data already proved that using different types of patient data increases the predictive success.

Finally, an extension of PheNoBo is needed to enhance the usability and accessibility of the method. PheNoBo is already freely available at `https://github.com/marie-sophie/mapra`. However, it requires a bioinformatician to collect the biological data for running PheNoBo and to set up KNIME. The transfer of PheNoBo to a different platform would make the tool available to a larger public. The application area of PheNoBo in research and clinics suggests two possible platforms for PheNoBo. Designing PheNoBo as web application would grant an easy access to the algorithms but creates also the need of a secure way to upload and store patient data on the underlying web server. A platform-independent stand-alone implementation with an intuitive graphical user interface could be run locally and allows for more data privacy. To keep PheNoBo up-to-date and to improve its predictions continuously, PheNoBo should be equipped with automatic update procedures for collecting new knowledge from the databases underlying PheNoBo.
This extension is the major step for enabling the application of PheNoBo in clinics.

Bibliography

[1] Bushra Afroze, Nida Amjad, Shahnaz H. Ibrahim, Khadija Nuzhat Humayun, and Yusnita Yakob. Adrenal insufficiency in a child with MELAS syndrome. *Brain & Development*, 36(10):924–927, 2014.

[2] Aldo Agnetti, Lee Bitton, Bertrand Tchana, Akamin Raymond, and Nicola Carano. Primary carnitine deficiency dilated cardiomyopathy: 28 years follow-up. *International Journal of Cardiology*, 162(2):e34–e35, 2013.

[3] Nadia Akawi, Jeremy McRae, Morad Ansari, Meena Balasubramanian, Moira Blyth, Angela F Brady, Stephen Clayton, Trevor Cole, Charu Deshpande, Tomas W Fitzgerald, Nicola Foulds, Richard Francis, George Gabriel, Sebastian S Gerety, Judith Goodship, Emma Hobson, Wendy D Jones, Shelagh Joss, Daniel King, Nikolai Klena, Ajith Kumar, Melissa Lees, Chris Lelliott, Jenny Lord, Dominic McMullan, Mary O'Regan, Deborah Osio, Virginia Piombo, Elena Prigmore, Diana Rajan, Elisabeth Rosser, Alejandro Sifrim, Audrey Smith, Ganesh J Swaminathan, Peter Turnpenny, James Whitworth, Caroline F Wright, Helen V Firth, Jeffrey C Barrett, Cecilia W Lo, David R FitzPatrick, and Matthew E Hurles for the DDD study. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature Genetics*, 47(11):1363–1372, 2015.

[4] Bruce Albers, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molekularbiologie der Zelle*. Ulrich Schäfer, Wiley-VHC, 5.Auflage 2011.

[5] Uwe Altermann. Identification and assignment of disease-related Loss Of Function mutations in NGS (Next Generation Sequencing) data. Master's thesis, Institute of Bioinformatics and Systems Biology (Helmholtz Zentrum München), 2013.

[6] Lars W. Andersen, Julie Mackenhauer, Jonathan C. Roberts, Katherine M. Berg, Michael N. Cocchi, and Michael W. Donnino. Etiology and therapeutic approach to elevated lactate. *Mayo Clinic Proceedings*, 88(10):1127–1140, 2013.

[7] A.W. Arnold, J.S. Kern, P.H. Itin, M. Pigors, R. Happle, and C. Has. Acromelanosis Albo-Punctata: A Distinct Inherited Dermatosis with Acral Spotty Dyspigmentation without Systemic Involvement. *Dermatology*, 224:331–339, 2012.

[8] Matthias Arnold, Johannes Raffler, Arne Pfeufer, Karsten Suhre, and Gabi Kastenmüller. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics*, 31(8):1334–1336, 2015. Available at `http://www.snipa.org`.

[9] Kondala R. Atkuri, Tina M. Cowan, Tony Kwan, Angelina Ng, Leonard A. Herzenberg, Leonore A. Herzenberg, and Gregory M. Enns. Inherited disorders affecting mitochondrial function are associated with glutathione deficiency and hypocitrullinemia. *Proceedings of the National Academy of Sciences of the United States of America*, 106(10):3941–3945, 2009.

[10] Marielle Baudrimont, Frédéric Dubas, Anne Joutel, Elizabeth Tournier-Lasserve, and Marie-Germaine Bousser. Autosomal Dominant Leukoencephalopathy and Subcortical Ischemic Stroke. *Stroke*, 24(1):122–125, 1993.

[11] B. Büchner, C. Gallenmüller, R. Lautenschläger, K. Kuhn, I. Wittig, L. Schöls, D. Rapaport, D. Seelow, P. Freisinger, H. Prokisch, W. Sperl, T. Wenz, C. Behl, M. Deschauer, C. Kornblum, P. Schneiderat, A. Abicht, M. Schuelke, T. Meitinger, T. Klopstock, and the mitoNET Consortium. Das Deutsche Netzwerk für mitochondriale Erkrankungen (mitoNET). *Medizinische Genetik*, 3:193–199, 2012.

[12] Chandree L. Beaulieu, Jacek Majewski, Jeremy Schwartzentruber, Mark E. Samuels, Bridget A. Fernandez, Francois P. Bernier, Michael Brudno, Bartha Knoppers, Janet Marcadier, David Dyment, Shelin Adam, Dennis E. Bulman, Steve J.M. Jones, Denise Avard, Minh Thu Nguyen, Francois Rousseau, Christian Marshall, Richard F. Wintle, Yaoqing Shen, Stephen W. Scherer, FORGE Canada Consortium, Jan M. Friedman, Jacques L. Michaud, and Kym M. Boycott. FORGE Canada Consortium: Outcomes of a 2-Year National Rare-Disease Gene-Discovery Project. *American Journal of Human Genetics*, 94:809–817, 2014.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[14] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. *Biochemie*. Spektrum Akademischer Verlag, 6. Auflage 2007.

[15] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, 2007.

[16] Thomas D Bird. *Charcot-Marie-Tooth Hereditary Neuropathy Overview*. In: Pagon RA, Adam MP, Ardinger HH, et al., editors. GeneReviews® [Internet], 1998 Sep 28 [Updated 2015 May 7]. Available from: `http://www.ncbi.nlm.nih.gov/books/NBK1358/`.

[17] Nathan Blow. Biochemistry's new look. *Nature*, 455:697–700, 2008.

[18] Hazel R. Bloxam, M. G. Day, Nancy K. Gibbs, and L. I. Woolf. An Inborn Defect in the Metabolism of Tyrosine in Infants on a Normal Diet. *Biochemical Journal*, 77:320–326, 1969.

[19] Margaret Borden, Jan Holm, Jack Leslie, Lawrence Sweetman, William L. Nyhan, Lynn Fleisher, Henry Nadler, David Lewis, and C. Ronald Scott. Hawkinsinuria in Two Families. *American Journal of Medical Genetics*, 44:52–56, 1992.

[20] Folkmar Bornemann. *Konkrete Analysis für Studierende der Informatik*. Springer-Verlag, 2008.

[21] Kym M. Boycott, Megan R. Vanstone, Dennis E. Bulman, and Alex E. MacKenzie. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14:681–691, 2013.

[22] Steven D. Brass, Eric E. Smith, Joseph F. Arboleda-Velasquez, William A. Copen, and Matthew P. Frosch. Case 12-2009: A 46-Year-Old Man with Migraine, Aphasia, and Hemiparesis and Similarly Affected Family Members. *New England Journal of Medicine*, 360:1656–1665, 2009.

[23] Aline Cano, Caroline Ovaert, Christine Vianey-Saban, and Brigitte Chabrol. Carnitine Membrane Transporter Deficiency: A Rare Treatable Cause of Cardiomyopathy and Anemia. *Pediatric Cardiology*, 29:163–165, 2008.

[24] Christopher J Carroll, Pirjo Isohanni, Rosanna Pöyhönen, Liliya Euro, Uwe Richter,

Virginia Brilhante, Alexandra Götz, Taina Lahtinen, Anders Paetau, Helena Pihko, Brendan J Battersby, Henna Tyynismaa, and Anu Suomalainen. Whole-exome sequencing identifies a mutation in the mitochondrial ribosome protein MRPL44 to underlie mitochondrial infantile cardiomyopathy. *Journal of Medical Genetics*, 50:151–159, 2013.

[25] Patrick F Chinnery. *Mitochondrial Disorders Overview*. In: Pagon RA, Adam MP, Ardinger HH, et al., editors. GeneReviews® [Internet], 2000 Jun 8 [Updated 2014 Aug 14].

[26] Kyle C Chipman and Ambuj K Singh. Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics*, 10:17, 2009.

[27] Barbara E. Clayton, R. H. Dobbs, and A. D. Patrick. Leigh's Subacute Necrotizing Encephalopathy: Clinical and Biochemical Study, with Special Reference to Therapy with Lipoate. *Archives of Disease in Childhood*, 42:467–478, 1967.

[28] Commons Math Developers. *Apache Commons Math, Release 3.6.1*. The Apache Software Foundation, 2016. Available at `http://commons.apache.org/math/download_math.cgi`.

[29] Corey D DeHaven, Anne M Evans, Hongping Dai, and Kay A Lawton. Organization of GC/MS and LC/MS metabolomics data into chemical libraries. *Journal of Cheminformatics*, 2, 2010.

[30] Salvatore DiMauro and Michio Hirano. *MELAS*. In: Pagon RA, Adam MP, Ardinger HH, et al., editors. GeneReviews® [Internet], 2001 Feb 27 [Updated 2013 Nov 21].

[31] Felix Distelmaier, Tobias B. Haack, Claudia B. Catarino, Constanze Gallenmüller, Richard J. Rodenburg, Tim M. Strom, Fabian Baertling, Thomas Meitinger, Ertan Mayatepek, Holger Prokisch, and Thomas Klopstock. MRPL44 mutations cause a slowly progressive multisystem disease with childhood-onset hypertrophic cardiomyopathy. *Neurogenetics*, 16:319–323, 2015.

[32] Kieu Trinh Do, Simone Wahl, Johannes Raffler, Sophie Molnos, Jerzy Adamski, Karsten Suhre, Konstantin Strauch, Christian Gieger, Fabian Theis, Harald Grallert, Jan Krumsiek, and Gabi Kastenmüller. Characterization of missingness in untargeted MS-based metabolomics data sets and evaluation of missing data handling strategies. publication in preparation.

[33] John A Dodge, Tamara Chigladze, Jean Donadieu, Zachi Grossman, Feliciano Ramos, Angelo Serlicorni, Liesbeth Siderius, Constantinos J Stefanidis, Velibor Tasic, Arunas Valiulis, and Jola Wierzba. The importance of rare diseases: from the gene to society. *Archives of Disease in Childhood*, 96(9):791–792, 2011.

[34] Keith N. Drummond, Alfred F. Michael, Robert A. Ulstrom, and Robert A . Good. The Blue Diaper Syndrome: Familial Hypercalcemia with Nephrocalcinosis and Indicanuria. *American Journal of Medicine*, 37:928–948, 1964.

[35] Anne M. Evans, Corey D. DeHaven, Tom Barrett, Matt Mitchell, and Eric Milgram. Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. *Analytical Chemistry*, 81(16):6656–6667, 2009.

[36] Elizabeth Forsythe and Philip L Beales. *Bardet-Biedl Syndrome*. In: Pagon RA, Adam MP, Ardinger HH, et al., editors. GeneReviews® [Internet], 2003 Jul 14 [Updated 2015 Apr 23]. Available from: `http://www.ncbi.nlm.nih.gov/books/NBK1363/`.

[37] Python Software Foundation. *Python Language Reference, version 2.7.6*. Python Software Foundation. Available at `https://www.python.org/`.

[38] Marie-Sophie Friedl. A KNIME Toolkit for Identification and Annotation of Loss-of-Function Mutations. Bachelor's thesis, Institute of Bioinformatics and Systems Biology (Helmholtz Zentrum München), 2014.

[39] Marie-Sophie Friedl, Carolin Prexler, Maria Schelling, Gabi Kastenmüller, Andreas Ruepp, and Jan-Dominik Quell. Decision support for the diagnosis of rare diseases based on symptoms. Master pratical, Institute of Bioinformatics and Systems Biology (Helmholtz Zentrum München), 2015.

[40] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.

[41] The Gene Ontology Consortium, Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for

the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[42] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690, 2007.

[43] José R. González and Miguel Vázquez Botet. Acromelanosis: A case report. *Journal of the American Academy of Dermatology*, 2:128–131, 1980.

[44] 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 of the United States of America*, 112(35):E4901–E4910, 2015.

[45] Richard H. Haas, Sumit Parikh, Marni J. Falk, Russell P. Saneto, Nicole I. Wolf, Niklas Darin, Lee-Jun Wong, Bruce H. Cohen, and Robert K. Naviaux. The In-Depth Evaluation of Suspected Mitochondrial Disease: The Mitochondrial Medicine Society's Committee on Diagnosis. *Molecular Genetics and Metabolism*, 94(1):16–37, 2008.

[46] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(D1):D514–D517, 2005.

[47] Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, and Christoph Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41(D1):D456–D463, 2013.

[48] Maximilian Hastreiter, Tim Jeske, Jonathan Hoser, Michael Kluge, Kaarin Ahomaa, Marie-Sophie Friedl, Sebastian J. Kopetzky, Jan-Dominik Quell, H.-Werner Mewes, and Robert Küffner. KNIME4NGS: a comprehensive toolbox for high-throughput Next Generation Sequencing analysis. Available at `http://ibisngs.github.io/knime4ngs/`.

[49] Zhanwen He, Xiangyang Luo, Liyang Liang, Pinggan Li, Dongfang Li, and Meng

Zhe. Merosin-deficient congenital muscular dystrophy type 1A: A case report. *Experimental and therapeutic medicine*, 6:1233–1236, 2013.

[50] Michael Heinzinger and Isabel Reinhardt. Validation of Loss-of-function Variants in human protein coding genes using RNA-Sequencing data. Master pratical, Institute of Bioinformatics and Systems Biology (Helmholtz Zentrum München), 2015.

[51] Ian J. Holt and Aurelio Reyes. Human Mitochondrial DNA Replication. *Cold Spring Harbor Perspectives in Biology*, 4:a012971, 2012.

[52] F. A. Hommes, H. A. Polman, and J. D. Reerink. Leigh's Encephalomyelopathy: an Inborn Error of Gluconeogenesis. *Archives of Disease in Childhood*, 43:423–426, 1968.

[53] Vito Iacobazzi, Federica Invernizzi, Silvia Baratta, Roser Pons, Wendy Chung, Barbara Garavaglia, Carlo Dionisi-Vici, Antonia Ribes, Rossella Parini, Maria Dolores Huertas, Susana Roldan, Graziantonio Lauria, Ferdinando Palmieri, and Franco Taroni. Molecular and Functional Analysis of SLC25A20 Mutations Causing Carnitine-Acylcarnitine Translocase Deficiency. *Human Mutation*, 24:312–320, 2004.

[54] Janice JK Ip, Peter KT Hui, MT Chau, and Wendy WM Lam. Merosin-Deficient Congenital Muscular Dystrophy (MDCMD): A Case Report with MRI, MRS and DTI Findings. *Journal of Radiology Case Reports*, 6(8):1–7, 2012.

[55] Elham Jaberi, Fereshteh Chitsazian, Gholam Ali Shahidi, Mohammad Rohani, Farzad Sina, Iman Safari, Maryam Malakouti Nejad, Masoud Houshmand, Brandy Klotzle, and Elahe Elahi. The novel mutation p.Asp251Asn in the β-subunit of succinate-CoA ligase causes encephalomyopathy and elevated succinylcarnitine. *Journal of Human Genetics*, 58:526–530, 2013.

[56] Asif Javed, Saloni Agrawal, and Pauline C Ng. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nature Methods*, 11(9):935–941, 2014.

[57] Tim Jeske. Systematic Analysis of LOF Alleles in Exome Data. Master's thesis, Institute of Bioinformatics and Systems Biology (Helmholtz Zentrum München), 2015.

[58] Timothy Jewison, Yilu Su, Fatemeh Miri Disfany, Yongjie Liang, Craig Knox, Adam Maciejewski, Jenna Poelzer, Jessica Huynh, You Zhou, David Arndt, Yannick Djoumbou, Yifeng Liu, Lu Deng, An Chi Guo, Beomsoo Han, Allison Pon, Michael

Wilson, Shahrzad Rafatnia, Philip Liu, and David S. Wishart. SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. *Nucleic Acids Research*, 42(D1):D478–D484, 2014.

[59] Kristi J Jones, Graeme Morgan, Heather Johnston, Vivienne Tobias, Robert A Ouvrier, Ian Wilkinson, and Kathryn N North. The expanding phenotype of laminin $\alpha 2$ chain (merosin) abnormalities: case series and review. *Journal of Medical Genetics*, 38:649–657, 2001.

[60] Sona S. Kamat, Peri H. Pepmueller, and Terry L. Moore. Triplets With Systemic Lupus Erythematosus. *Arthritis and Rheumatism*, 48(11):3176–3180, 2003.

[61] A.J. Kanwar, R. Jaswal, G.P. Thami, and G.K. Bedi. Acquired Acromelanosis due to Phenytoin. *Dermatology*, 194:373–374, 1997.

[62] Gabi Kastenmüller, Werner Römisch-Margl, Brigitte Wägele, Elisabeth Altmaier, and Karsten Suhre. metaP-Server : A Web-Based Metabolomics Data Analysis Tool. *Journal of Biomedicine and Biotechnology*, 2011.

[63] Alison L Kelly, Peter W Lunt, Fernanda Rodrigues, P J Berry, Diana M Flynn, Patrick J McKiernan, Deirdre A Kelly, Giorgina Mieli-Vergani, and Timothy M Cox. Classification and genetic features of neonatal haemochromatosis: a study of 27 affected pedigrees and molecular analysis of genes implicated in iron metabolism. *Journal of Medical Genetics*, 38:599–610, 2001.

[64] Alfons Kemper and André Eickler. *Datenbanksysteme – Eine Einführung*. Oldenbourg Verlag, München, 8. Auflage, 2011.

[65] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N. Robinson. Walking the Interactome for Prioritization of Candidate Disease Genes. *American Journal of Human Genetics*, 82:949–958, 2008.

[66] Sebastian Köhler, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *American Journal of Human Genetics*, 85:457–464, 2009.

[67] Nikola Kovářová, Alena Čížková Vrbacká, Petr Pecina, Viktor Stránecký, Ewa Pronicka, Stanislav Kmoch, and Josef Houštěk. Adaptation of respiratory chain biogenesis to cytochrome c oxidase deficiency caused by SURF1 gene mutations.

Biochimica et Biophysica Acta, 1822:1114–1124, 2012.

[68] Susanne Krug, Gabi Kastenmüller, Ferdinand Stückler, Manuela J. Rist, Thomas Skurk, Manuela Sailer, Johannes Raffler, Werner Römisch-Margl, Jerzy Adamski, Cornelia Prehn, Thomas Frank, Karl-Heinz Engel, Thomas Hofmann, Burkhard Luy, Ralf Zimmermann, Franco Moritz, Philippe Schmitt-Kopplin, Jan Krumsiek, Werner Kremer, Fritz Huber, Uwe Oeh, Fabian J. Theis, Wilfried Szymczak, Hans Hauner, Karsten Suhre, and Hannelore Daniel. The dynamic range of the human metabolome revealed by challenges. *The FASEB Journal*, 26:2607–2619, 2012.

[69] Jan Krumsiek, Karsten Suhre, Anne M. Evans, Matthew W. Mitchell, Robert P. Mohney, Michael V. Milburn, Brigitte Wägele, Werner Römisch-Margl, Thomas Illig, Jerzy Adamski, Christian Gieger, Fabian J. Theis, and Gabi Kastenmüller. Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genetics*, 8(10), 2012.

[70] Denis Leigh. Subacute necrotizing encephalomyelopathy in an infant. *Journal of Neurology, Neurosurgery and Psychiatry*, 14:216–221, 1951.

[71] Sherwood A. Libit, Robert A. Ulstrom, and Doris Doeden. Fecal Pseudomonas aeruginosa as a cause of the blue diaper syndrome. *Journal of Pediatrics*, 81(3):546–547, 1972.

[72] Enrico Lopriore, Reinoud J. B. J. Gemke, Nanda M. Verhoeven, Cornelis Jakobs, Ronald J. A. Wanders, Angelique B. C. Roeleveld-Versteeg, and Bwee Tien Poll-The. Carnitine-acylcarnitine translocase deficiency: phenotype, residual enzyme activity and outcome. *European Journal of Pediatrics*, 160:101–104, 2001.

[73] Daniel G. MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K. Pickrell, Stephen B. Montgomery, Cornelis A. Albers, Zhengdong D. Zhang, Donald F. Conrad, Gerton Lunter, Hancheng Zheng, Qasim Ayub, Mark A. DePristo, Eric Banks, Min Hu, Robert E. Handsaker, Jeffrey A. Rosenfeld, Menachem Fromer, Mike Jin, Xinmeng Jasmine Mu, Ekta Khurana, Kai Ye, Mike Kay, Gary Ian Saunders, Marie-Marthe Suner, Toby Hunt, If H. A. Barnes, Clara Amid, Denise R. Carvalho-Silva, Alexandra H. Bignell, Catherine Snow, Bryndis Yngvadottir, Suzannah Bumpstead, David N. Cooper, Yali Xue, Irene Gallego Romero, 1000 Genomes Project Consortium, Jun Wang, Yingrui Li, Richard A. Gibbs, Steven A. McCarroll, Emmanouil T. Dermitzakis, Jonathan K. Pritchard, Jeffrey C. Barrett, Jennifer Harrow, Matthew E.

Hurles, Mark B. Gerstein, and Chris Tyler-Smith. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, 335:823–828, 2012.

[74] The MathWorks, Inc. *MATLAB and Statistics Toolbox Release 2016a*. The MathWorks, Inc., Natick, Massachusetts, United States. Available at `https://www.mathworks.com`.

[75] Cristina Menni, Gabi Kastenmüller, Ann Kristin Petersen, Jordana T Bell, Maria Psatha, Pei-Chien Tsai, Christian Gieger, Holger Schulz, Idil Erte, Sally John, M Julia Brosnan, Scott G Wilson, Loukia Tsaprouni, Ee Mun Lim, Bronwyn Stuckey, Panos Deloukas, Robert Mohney, Karsten Suhre, Tim D Spector, and Ana M Valdes. Metabolomic markers reveal novel pathways of ageing and early development in human populations. *International Journal of Epidemiology*, 42:1111–1119, 2013.

[76] Neil A. Miller, Emily G. Farrow, Margaret Gibson, Laurel K. Willig, Greyson Twist, Byunggil Yoo, Tyler Marrs, Shane Corder, Lisa Krivohlavek, Adam Walter, Josh E. Petrikin, Carol J. Saunders, Isabelle Thiffault, Sarah E. Soden, Laurie D. Smith, Darrell L. Dinwiddie, Suzanne Herd, Julie A. Cakici, Severine Catreux, Mike Ruehle, and Stephen F. Kingsmore. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Medicine*, 7(100), 2015.

[77] Kirstin Mittelstrass, Janina S. Ried, Zhonghao Yu, Jan Krumsiek, Christian Gieger, Cornelia Prehn, Werner Roemisch-Margl, Alexey Polonikov, Annette Peters, Fabian J. Theis, Thomas Meitinger, Florian Kronenberg, Stephan Weidinger, Heinz Erich Wichmann, Karsten Suhre, Rui Wang-Sattler, Jerzy Adamski, and Thomas Illig. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genetics*, 7(8), 2011.

[78] A. Niederwieser, Ana Matasovic, Patricia Tippett, and D.M. Danks. A new sulfur amino acid, named hawkinsin, identified in a baby with transient tyrosinemia and her mother. *Clinica Chimica Acta*, 76:345–356, 1977.

[79] Jodi Nunnari and Anu Suomalainen. Mitochondria: In Sickness and in Health. *Cell*, 148:1145–1159, 2012.

[80] Martin Oti, Martijn A. Huynen, and Han G. Brunner. The Biological Coherence of Human Phenome Databases. *American Journal of Human Genetics*, 85:801–808, 2009.

[81] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. Available at `https://www.R-project.org`.

[82] Shamima Rahman and David Thorburn. *Nuclear Gene-Encoded Leigh Syndrome*. In: Pagon RA, Adam MP, Ardinger HH, et al., editors. GeneReviews® [Internet], 2015 Oct 1.

[83] Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Segolene Ayme. Representation of Rare Diseases in Health Information Systems: The Orphanet Approach to Serve a Wide Range of End Users. *Human Mutation*, 33(5):803–808, 2012.

[84] Karen Reue. The Lipin Family: Mutations and Metabolism. *Current Opinion in Lipidology*, 20(3):165–170, 2009.

[85] Marylyn D. Ritchie, Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16:85–97, 2015.

[86] Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *American Journal of Human Genetics*, 83:610–615, 2008.

[87] ME Rubio-Gozalbo, P Vos, PPh Forget, SB Van Der Meer, RJA Wanders, HR Waterham, and JA Bakker. Carnitine-acylcarnitine translocase deficiency: case report and review of the literature. *Acta Paediatrica*, 92:501–504, 2003.

[88] Thomas Schickinger and Angelika Steger. *Diskrete Strukturen 2: Wahrscheinlichkeitstheorie und Statistik*. Springer-Verlag, 2002.

[89] Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. Why rare diseases are an important medical and social issue. *Lancet*, 371:2039–2041, 2008.

[90] Rene Christoph Schoeffel. Prioritizing Pathways and Genes linked to rare Mitochondriopathies based on metabolic Extremes in individual Patients. Bachelor's thesis, Institute of Bioinformatics and Systems Biology (Helmholtz Zentrum München), 2014.

[91] 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, The Multiple Tissue Human Expression Resource (MuTHER) Consortium, 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 Kastenmüller, Tim D Spector, and Nicole Soranzo. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–553, 2014.

[92] John M. Shoffner, Michael D. Brown, Carol Stugard, Albert S. Jun, Stephen Pollock, Richard H. Haas, Allan Kaufman, Deborah Koontz, Yoon Kim, Jennifer R. Graham, Edwin Smith, John Dixon, and Douglas C. Wallace. Leber's Hereditary Optic Neuropathy Plus Dystonia Is Caused by a Mitochondrial DNA Point Mutation. *Annals of Neurology*, 38(2):163–169, 1995.

[93] H. W. Siemens. Acromelanosis albo-punctata. *Dermatologica*, 128:86–87, 1964.

[94] Damian Smedley, Sebastian Köhler, Johanna Christina Czeschik, Joanna Amberger, Carol Bocchini, Ada Hamosh, Julian Veldboer, Tomasz Zemojtel, and Peter N. Robinson. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, 30(22):3215–3222, 2014.

[95] Damian Smedley and Peter N. Robinson. Phenotype-driven strategies for exome prioritization of human mendelian disease genes. *Genome Medicine*, 7(81), 2015.

[96] Ivan P. Stanimirović and Milan B. Tasić. Performance comparison of storage formats for sparse matrices. *FACTA UNIVERSITATIS Series Mathematics and Informatics*, 24:39–51, 2009.

[97] Charles A. Stanley, Daniel E. Hale, Gerard T. Berry, Susan Deleeuw, Jay Boxer, and Jean-Paul Bonnefont. Brief report: a deficiency of carnitine-acylcarnitine translocase in the inner mitochondrial membrane. *The New England Journal of Medicine*, 327(1):19–23, 1992.

[98] Tsutomu Suzuki, Asuteka Nagao, and Takeo Suzuki. Human Mitochondrial tRNAs: Biogenesis, Function, Structural Aspects, and Diseases. *Annual Review of Genetics*, 45:299–329, 2011.

[99] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide

Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2015.

[100] Robert W. Taylor and Doug M. Turnbull. Mitochondrial DNA Mutations in Human Disease. *Nature Reviews Genetics*, 6:389–402, 2005.

[101] Ines Thiele, Neil Swainston, Ronan M T Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, Stefan G Thorleifsson, Rasmus Agren, Christian Bölling, Sergio Bordel, Arvind K Chavali, Paul Dobson, Warwick B Dunn, Lukas Endler, David Hala, Michael Hucka, Duncan Hull, Daniel Jameson, Neema Jamshidi, Jon J Jonsson, Nick Juty, Sarah Keating, Intawat Nookaew, Nicolas Le Novère, Naglis Malys, Alexander Mazein, Jason A Papin, Nathan D Price, Evgeni Selkov, Sr, Martin I Sigurdsson, Evangelos Simeonidis, Nikolaus Sonnenschein, Kieran Smallbone, Anatoly Sorokin, Johannes H G M van Beek, Dieter Weichart, Igor Goryanin, Jens Nielsen, Hans V Westerhoff, Douglas B Kell, Pedro Mendes, and Bernhard Ø Palsson. A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, 31(5):419–427, 2013.

[102] The Metabolomics Innovation Centre (TMIC). FooDB Version 1.0. Available at `http://foodb.ca/`.

[103] Kaede Tomoeda, Hisataka Awata, Toshinobu Matsuura, Ichiro Matsuda, Engelbert Ploechl, Tom Milovac, Avihu Boneh, C. Ronald Scott, David M. Danks, and Fumio Endo. Mutations in the 4-Hydroxyphenylpyruvic Acid Dioxygenase Gene Are Responsible for Tyrosinemia Type III and Hawkinsinuria. *Molecular Genetics and Metabolism*, 71:506–510, 2000.

[104] Elizabeth Tournier-Lasserve, Marie-Thérèse Iba-Zizen, Norma Romero, and Marie-Germaine Bousser. Autosomal Dominant Syndrome With Strokelike Episodes and Leukoencephalopathy. *Stroke*, 22(10):1297–1302, 1991.

[105] Elena J. Tucker, Steven G. Hershman, Caroline Köhrer, Casey A. Belcher-Timme, Jinal Patel, Olga A. Goldberger, John Christodoulou, Jonathon M. Silberstein, Matthew McKenzie, Michael T. Ryan, Alison G. Compton, Jacob D. Jaffe, Steven A. Carr, Sarah E. Calvo, Uttam L. RajBhandary, David R. Thorburn, and Vamsi K. Mootha. Mutations in MTFMT underlie a human disorder of formylation causing impaired

mitochondrial translation. *Cell Metabolism*, 14(3):428–434, 2011.

[106] Charalampos Tzoulis and Laurence A Bindoff. The Syndrome of Mitochondrial Spinocerebellar Ataxia and Epilepsy caused by POLG mutations. *Advances in Clinical Neuroscience and Rehabilitation*, 9(3):13–16, 2009.

[107] Sam F.B. van Beuningen, Lena Will, Martin Harterink, Anaël Chazeau, Eljo Y. van Battum, Cátia P. Frias, Mariella A.M. Franker, Eugene A. Katrukha, Riccardo Stucchi, Karin Vocking, Ana T. Antunes, Lotte Slenders, Sofia Doulkeridou, Peter Sillevis Smitt, A.F. Maarten Altelaar, Jan A. Post, Anna Akhmanova, R. Jeroen Pasterkamp, Lukas C. Kapitein, Esther de Graaff, and Casper C. Hoogenraad. TRIM46 Controls Neuronal Polarity and Axon Specification by Driving the Formation of Parallel Microtubule Arrays. *Neuron*, 88(6):1208–1226, 2015.

[108] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 2011.

[109] Isidro Vitoria, Elena Martín-Hernández, Luis Peña-Quintana, María Bueno, Pilar Quijada-Fraile, Jaime Dalmau, Sofia Molina-Marrero, Belén Pérez, and Begoña Merinero. Carnitine-Acylcarnitine Translocase Deficiency: Experience with Four Cases in Spain and Review of the Literature. *Journal of Inherited Metabolic Disease*, 20:11–20, 2014.

[110] Christian von Mering, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A. Huynen, and Peer Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(D1):D433–D437, 2005.

[111] Sarah K Westbury, Ernest Turro, Daniel Greene, Claire Lentaigne, Anne M Kelly, Tadbir K Bariana, Ilenia Simeoni, Xavier Pillois, Antony Attwood, Steve Austin, Sjoert BG Jansen, Tamam Bakchoul, Abi Crisp-Hihn, Wendy N Erber, Rémi Favier, Nicola Foad, Michael Gattens, Jennifer D Jolley, Ri Liesner, Stuart Meacham, Carolyn M Millar, Alan T Nurden, Kathelijne Peerlinck, David J Perry, Pawan Poudel, Sol Schulman, Harald Schulze, Jonathan C Stephens, Bruce Furie, Peter N Robinson, Chris van Geet, Augusto Rendon, Keith Gomez, Michael A Laffan, Michele P Lambert, Paquita Nurden, Willem H Ouwehand, Sylvia Richardson, Andrew D Mumford, Kathleen Freson, and on behalf of the BRIDGE-BPD Consortium. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Medicine*, 7(36), 2015.

[112] H.-E. Wichmann, C. Gieger, and T. Illig. KORA-gen – Resource for Population Genetics, Controls and a Broad Spectrum of Disease Phenotypes. *Gesundheitswesen*, 67(Sonderheft 1):S26–S30, 2005.

[113] Thomas Wieser. *Carnitine Palmitoyltransferase II Deficiency*. In: Pagon RA, Adam MP, Ardinger HH, et al., editors. GeneReviews® [Internet], 2004 Aug 27 [Updated 2014 May 15].

[114] David S. Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, Souhaila Bouatra, Igor Sinelnikov, David Arndt, Jianguo Xia, Philip Liu, Faizath Yallou, Trent Bjorndahl, Rolando Perez-Pineiro, Roman Eisner, Felicity Allen, Vanessa Neveu, Russ Greiner, and Augustin Scalbert. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807, 2013. Available at `http://www.hmdb.ca/` (accessed in 2016 May 9).

[115] Caroline F Wright, Tomas W Fitzgerald, Wendy D Jones, Stephen Clayton, Jeremy F McRae, Margriet van Kogelenberg, Daniel A King, Kirsty Ambridge, Daniel M Barrett, Tanya Bayzetinova, A Paul Bevan, Eugene Bragin, Eleni A Chatzimichali, Susan Gribble, Philip Jones, Netravathi Krishnappa, Laura E Mason, Ray Miller, Katherine I Morley, Vijaya Parthiban, Elena Prigmore, Diana Rajan, Alejandro Sifrim, G Jawahar Swaminathan, Adrian R Tivey, Anna Middleton, Michael Parker, Nigel P Carter, Jeffrey C Barrett, Matthew E Hurles, David R FitzPatrick, and Helen V Firth on behalf of the DDD study. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*, 385:1305–1314, 2015.

[116] AA Yamak, F Bitar, P Karam, and G Nemer. Exclusive cardiac dysfunction in familial primary carnitine deficiency cases: a genotype-phenotype correlation. *Clinical Genetics*, 72:59–62, 2007.

[117] Oscar Yanes, Ralf Tautenhahn, Gary J. Patti, and Gary Siuzdak. Expanding Coverage of the Metabolome for Global Metabolite Profiling. *Analytical Chemistry*, 83:2152–2161, 2011.

[118] Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavi-

das, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, and Paul Flicek. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, 2016.

List of Figures

List of Tables

List of Abbreviations

ACAD9	Acyl-CoA dehydrogenase family member 9
ADAR	Adenosine Deaminase, RNA-Specific
ANO5	Anoctamin 5
ATP	adenosine triphosphate
BBS	Bardet-Biedl Syndrome
BRIDGE	Biomedical Research Centres/Units Inherited Disorders and Genetic Evaluation
C1QBP	Complement C1q binding protein
C10orf2	Chromosome 10 open reading frame 2
CACT	Carnitine/Acylcarnitine Translocase
CACTD	Carnitine/Acylcarnitine Translocase Deficiency
CADASIL	Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy
CCBL1	Cysteine conjugate-beta lyase, cytoplasmic
ChEBI	Chemical Entities of Biological Interest
CMT	Charcot-Marie-Tooth Disease
CPT1	Carnitine Palmitoyltransferase 1

CPT1A	Carnitine Palmitoyltransferase 1A
CPT2	Carnitine Palmitoyltransferase 2
CYP4X1	Cytochrome P450 family 4 subfamily X member 1
DDD	Deciphering Developmental Disorders
DNA	deoxyribonucleic acid
ECHS1	Enoyl-CoA hydratase, short chain 1
ESR1	Estrogen Receptor 1
FORGE	Finding of Rare Disease Genes
GO	Gene Ontology
GWAS	genome-wide association study
HMDB	Human Metabolome Database
HPO	Human Phenotype Ontology
IARS2	Isoleucine-tRNA synthetase 2, mitochondrial
IBIS	Institute of Bioinformatics and Systems Biology at Helmholtz Zentrum München
IHG	Institute of Human Genetics at Helmholtz Zentrum München
KORA	Kooperative Gesundheitsforschung in der Region Augsburg
KNIME	Konstanz Information Miner
LDYT	Leber optic atrophy and dystonia
LOF	loss-of-function variant
LPIN1	lipin-1
MELAS	Mitochondrial encephalomyopathy, lactic acidosis and stroke-like

	episodes
MICE	Multivariate Imputation by Chained Equations
MRPL44	Mitochondrial ribosomal protein L44
MSCAE	Mitochondrial spinocerebellar ataxia and epilepsy
MTFMT	Methionyl-tRNA formyltransferase, mitochondrial
MT-ND6	Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 6
OMIM	Online Mendelian Inheritance in Man
OPA1	OPA1, mitochondrial dynamin like GTPase
PCD	Primary Carnitine Deficiency
PheNoBo	Combination of **Phe**notype, Ge**No**type and Meta**Bo**type
Phe-Phe	phenylalanylphenylalanine
POLG	DNA polymerase gamma, catalytic subunit
PRPS1	Phosphoribosyl pyrophosphate synthetase 1
RNA	ribonucleic acid
SACS	Sacsin molecular chaperone
SLC22A5	Solute carrier family 22 member 5
SLC6A8	Solute carrier family 6 member 8
SMPDB	Small Molecule Pathway Database
SNiPA	Single Nucleotide Polymorphism Annotator
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SUCLA2	Succinate-CoA ligase ADP-forming beta subunit

SURF1	Surfeit 1
TRIM46	Tripartite motif containing 46
TTC19	Tetratricopeptide repeat domain 19

DVD Content

The attached DVD contains an electronic copy of this thesis and three folders with the following data:

DataSets: The folder provides all data sets for running PheNoBo. The content of the folder covers the data presented in section 2.1, the pre-calculated score distributions for Phenomizer for PhenoDis (section 2.3.1) and the metabolic reference sets described in section 2.3.2.1.

PheNoBo: The folder provides an executable of the KNIME nodes of PheNoBo version 2.1.6 (section 2.3) and the corresponding KNIME workflow for KNIME 3.1 shown in figure 7. In addition the source code of PheNoBo is included as Eclipse project in the folder. These data are also available from github at `https://github.com/marie-sophie/mapra`.

Results: The folder contains the results from applying PheNoBo on the eight patient data sets summarized in table 1. Note that, for reasons of medical data protection, the patient data sets 5 to 8 (section 2.2.3 and 2.2.4) are excluded from the DVD. The folder includes the input data used on PheNoBo for each data set (except for sets 5 to 8) (section 2.2) and a summary of the predictions from PheNoBo (section 3).