UNIVERSITÉ DU LUXEMBOURG

PhD-FSTC-2018-19
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 13/03/2018 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Alberto SILVA DE NORONHA
Born on 04 April 1988 in Póvoa de Lanhoso (Portugal)

DEVELOPMENT OF A COMPUTATIONAL RESOURCE FOR PERSONALIZED DIETARY RECOMMENDATIONS

Dissertation defence committee

Dr. Ines Thiele, dissertation supervisor
Associate Professor, Université du Luxembourg

Dr Lorraine Brennan
Professor, University College Dublin

Dr Rejko Krüger, Chairman
Professor, Université du Luxembourg

Dr Elmar Heinzle
Professor, Universität des Saarlandes

Dr Reinhard Schneider, Vice-Chairman
Head of bioinformatics core facility, Université du Luxembourg

Molecular Systems Physiology

Luxembourg Centre for Systems Biomedicine

Faculty of Life Sciences, Technology and Communication

Doctoral School in Systems and Molecular Biomedicine

Disseration Defence Committee:

Committee members:	Prof. Rejko Krüger
	Dr. Reinhard Schneider
	Prof. Lorraine Brennan
	Prof. Elmar Heinzle
Supervisor:	Prof. Ines Thiele

I hereby confirm that the PhD thesis entitled “DEVELOPMENT OF A COMPUTATIONAL RESOURCE FOR PERSONALIZED DIETARY RECOMMENDATIONS” has been written independently and without any other sources than cited.

Luxembourg, _____ _____

Author Name

If a man knows not to which port
he sails, no wind is favorable.
Seneca

Acknowledgments

First and foremost, I would like to express my gratitude to Prof. Ines Thiele for giving me the opportunity to do this project under her supervision and for the interesting discussions we had during these last four years. My sincere thanks to all present and past members of the MSP group, with whom I had the pleasure to collaborate. I am also thankful for all the help they, and other collaborators, have provided to make this document possible. A word for everyone at the LCSB who make this institute an amazing place to do science but more importantly, a great working place.

When I moved from Portugal I left family and friends to evolve academically and professionally. I did not expect this to be easy but I find consolation in the many great people I met here. To all my friends, from the "Happy Wednesday" crew to my brothers in arms, the Bítores, my deepest and sincerest thanks. I am truly blessed and thankful for having crossed paths with you and wherever the future takes you, know that you will find a friend in me. You are my second family and the reason I can call Luxembourg my new "home". I would also like to have a special word for Anne-Catherine, for showing me how even in the grayest of days the sun can shine the brightest. Thank you for being that sunshine, Annie.

Finally, I want to thank my family, especially my parents and my brother. You are my references, the people I look up to. Despite the distance and the pains, your unfaltering support keeps me going forward. I am forever indebted and everything I accomplish is thanks to you. Obrigado!

Contents

List of Figures

List of Tables

List of Abbreviations

ATP	Adenine triphosphate
AGORA	Assembly of gut organisms through reconstruction and analysis
API	Application programming interface
ATP	Adenosine triphosphate
BMI	Body Mass Index
CHD	Coronary heart disease
COBRA	Constraint Based Reconstruction and Analysis
CSV	Comma separated values
DNA	Deoxyribonucleic acid
DNS	Domain Name System
EWAS	Epigenome-wide association studies
FBA	Flux Balance Analysis
FFQ	Food Frequency Questionnaire
FVA	Flux Variability Analysis
GDP	Guanosine diphosphate
GEMs	Genome-Scale Metabolic Models
GENREs	Genome-Scale Metabolic Reconstructions
GMD	Golm Metabolome Database
GPR	Gene-Protein-Rule
GWAS	Genome wide association studies
HGNC	HUGO Gene Nomenclature Committee

HMDB	Human Metabolome Database
HPO	Human Phenotypic Ontology
HTML	Hypertext Markup Language
IBD	Inflammatory bowel disease
IMDs	Inherited metabolic diseases
JSON	Javascript Object Notation
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC/MS	Liquid chromatography/mass spectrometry
MD	Maltodextrin
MINERVA	Molecular Interaction Network Visualization
NCBI	National Center for Biotechnology Information
NCDs	Non-communicable diseases
NMC	Netherlands Metabolomics Centre
OMIM	Online Mendelian Inheritance of Man
PBPK	Physiologically based pharmacokinetic
PKU	Phenylketonuria
RCTs	Randomized Controlled Trials
REST	Representational state transfer
RNA	Ribonucleic acid
SBML	Systems Biology Markup Language
SFCAs	Short-chain fatty acids
SQL	Structured Query Language
tSNE	t-Distributed Stochastic Neighbor Embedding
UDP	Uridine triphosphate
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VM	Virtual Machine
VMH	Virtual Metabolic Human
WES	Whole exome sequencing

Summary

There is a global increase in the incidence of non-communicable diseases associated with unhealthy food intakes. Conditions such as diabetes, heart disease, high blood pressure, and strokes represent a high societal impact and an economic burden for health-care systems around the world. To understand these diseases, one needs to account the several factors that influence how the human body processes food, some of which are determined by the genome and patterns of gene expression that translate to the ability - or lack of - to degrade and absorb certain nutrients. Other factors, like the gut microbiota, are more volatile because its composition is highly moldable by diet and lifestyle.

Multi-omics technologies can support the comprehensive collection of dietary intake data and monitoring of the health status of individuals. Also, a correct analysis of this data could lead to new insights about the complex processes involved in the digestion of dietary components and their involvement in the prevention or the appearance of health problems, but its integration and interpretation is still problematic.

Thus, in this thesis, we propose the utilization of Constraint-Based Reconstruction and Analysis (COBRA) methods as a framework for the integration of this complex data. To achieve this goal, we have created a knowledge-base, the Virtual Metabolic Human (VMH), that combines information from large-scale models of metabolism from the human organism and typical gut microbes, with food composition information, and a disease compendium. VMH's unique combination of resources leverages the exploration of metabolic pathways from different organisms, the inclusion of dietary information into *in-silico* experiments through its own diet designer tool, visualization and analysis of experimental and simulation data, and exploring disease mechanisms and potential treatment strategies.

VMH is a step forward in providing the necessary tools to investigate the mechanisms behind the influence of diet in health and disease. Tools such as the diet designer can be

used as a basis for diet optimization by predicting combinations of foods that can contribute to specific metabolic outcomes, which has the potential to be integrated and translated into treatment development and dietary recommendations in the foreseeable future.

Chapter 1

Introduction

Abstract

Non-communicable diseases (NCDs) have a high societal impact and represent significant costs for the healthcare systems around the world. These diseases result from a combination of factors but are closely related to unhealthy lifestyle and nutrition. Understanding the mechanisms behind the effect of nutritional patterns in health is not trivial and there are limitations associated with dietary assessment tools and studies of nutrition that further complicate this task. For this purpose, novel technologies, such as metabolomic or metagenomic sequencing are being used in an attempt at better characterizing the effect of different diets, foods, and nutrients. Due to the high complexity of these data, more and more, a systems biology approach becomes necessary for the study of nutrition. Constraint-Based Reconstruction and Analysis (COBRA) uses Genome-Scale Metabolic Models (GEMs) to study the metabolism of human and microbial species. We propose that GEMs and the COBRA approach as a suitable framework to integrate the complex data generated in nutrition studies and provide the simulation tools that will allow formulating hypothesis to explain the mechanisms behind the effect of different dietary patterns in health. Achieving this will pave the way for personalized dietary recommendations.

1.1 Nutrition

Societies are facing an increase in non-communicable diseases (NCDs), also known as chronic diseases. These conditions are known to be a result of a combination of genetic, physiological and environmental factors. They are often associated with older populations but affect people in all age groups. The main risk factors associated with NCDs are very closely related to lifestyle, consisting of unhealthy diets, physical inactivity, exposure to tobacco smoke or the harmful use of alcohol [87]. Diet-associated diseases and risk factors are widespread across the population worldwide. According to the Global Nutrition Report of 2017, more than half of the European population is overweight [71]. The statistics of different european countries reveal a trend of high incidence of risk factors for diet-related non-communicable diseases, such as raised blood pressure, blood glucose, and blood cholesterol (Figure 1.1). Particularly in Luxembourg, the ORISCAV-LUX study (2007-2008) reported that 85% of the population displayed one or more risk factors for cardiovascular disease: notably 35% of the population has hypertension, 70% increased lipid levels in blood, and 54% of the population is overweight (BMI above 25) with 31% of these considered to be obese [6]. These numbers demonstrate the high societal impact and an associated increase in costs for the health care system resulting from unhealthy lifestyle and nutrition. For these reasons, there is great interest in promoting the understanding of how health is influenced by different diet compositions and how the complex systems involved in food digestion interact with each other. Nutrition is a subject that undoubtedly attracts a lot of attention from the general public when compared with other fields of science. It is common to come across contradictory information and passionate discussions about the efficacy of specific diets. Often, nutritional studies receive broad media coverage in the form of misleading headlines and very little detail on the used methodologies and their limitations. In fact, it is extremely difficult to derive knowledge from results obtained from nutritional studies due to the involvement of a great many confounding factors. To understand these limitations we will start by covering the main features of dietary assessment tools and types of nutrition studies.

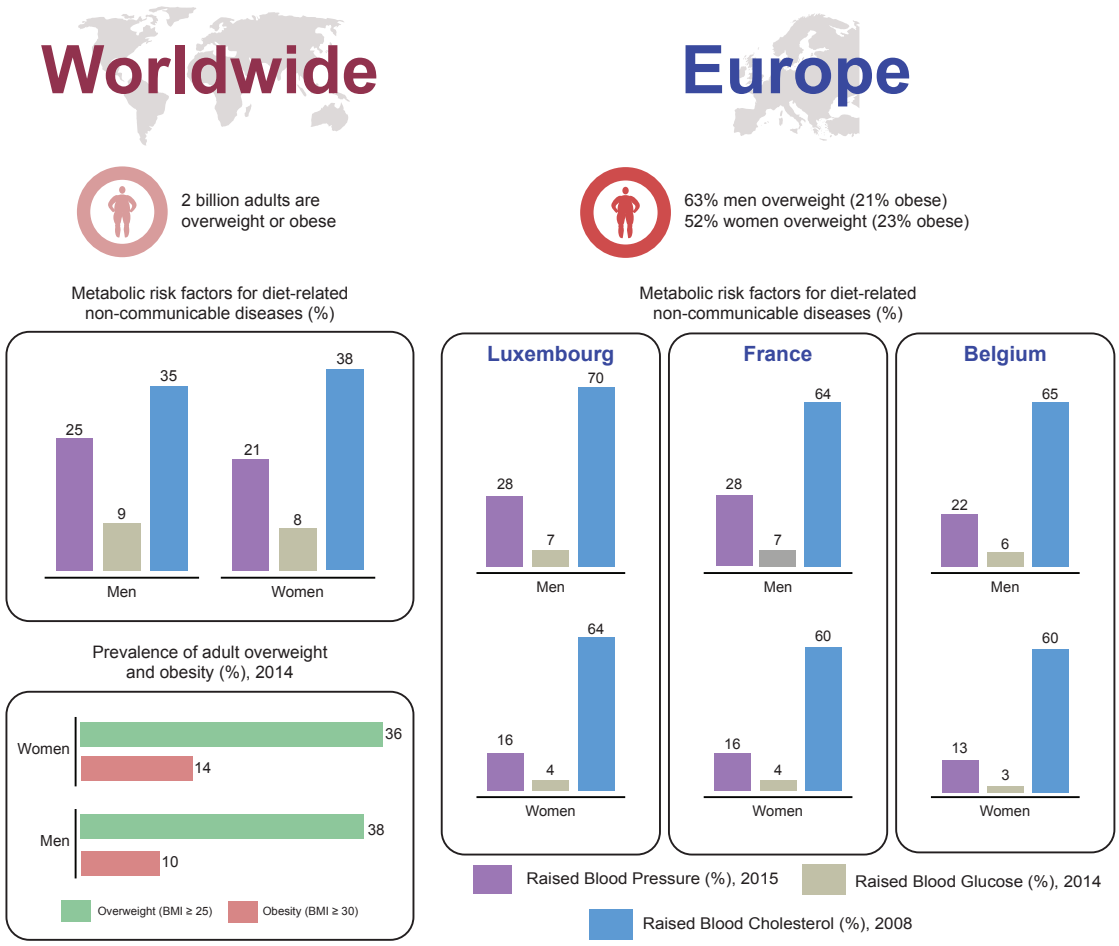

Figure 1.1: Incidence of obesity and risk factors for NCDs worldwide and in Europe. Adapted from the Global Nutrition Report 2017 [71].

1.1.1 Dietary assessment tools

Acquiring reliable and accurate information on food intake from free-living individuals is a known limitation of nutrition studies. Nutrition studies often require that volunteers remember what they have eaten over certain periods of time. Some studies go to the extent of collecting information on dietary habits for periods of years [198]. But how reliable and accurate is this information? Typically, the three most common dietary assessment tools used are: (i) food records (or journals/diaries), (ii) 24-hour recall, and the most popular of these, (iii) food frequency questionnaires (FFQ) [308, 309].

Food journals/records are a collection of all food and beverages consumed over a period of time, typically in the range of a few days. Food records tend to be more accurate if participants weigh food portions. The amounts consumed can be recorded using a scale, household

measures, or with the aid of pictures [38]. This assessment tool carries a high respondent burden and misreporting issues tend to occur when periods of consecutive recording go beyond 4 days [101].

A 24-hour recall, as the name indicates, involves remembering and recording all foods consumed during a whole day. This tool has a low respondent burden and is suitable for large-scale studies. The recall can be conducted by an interviewer [41, 44] or be self-administered [12, 11, 321]. 24h recalls have some limitations regarding the accuracy of portion size estimation and misreports. Additionally, they are single observations and do not portray the typical diet of the respondent. For this reason, it is necessary to collect several data points. To address some of these limitations automated self-administered instruments were developed, such as the ASA24 [297] developed at the National Cancer Institute [298, 295]. These systems collect food composition data from established databases and in some cases use food photographs to help users indicate portion size [295]. In this fashion, the collection of high-quality dietary data in large-scale is more effective [321, 167].

An FFQ is a collection of the relative frequency of consumption of a list of foods or food groups. Many FFQs also integrate portion size information. They are suitable for large-scale surveys with a low respondent burden and are designed to capture the overall dietary habits of the respondents. For this reason, the food list of the FFQ must be carefully designed according to the target population. These lists can have different lengths, and research suggests that longer lists perform better [209]. At the same time, it is not clear if portion size questions in FFQs are useful [209]. Food intake frequency has a bigger impact than the size of serving in the intake variance of most foods [121, 84] but some studies report slight increases in performance by having portion size information [61, 34]. Similarly to other retrospective assessment tools, FFQs are susceptible to misreports, estimation of portion sizes and possible over-representation of healthy foods. In fact, validation studies of FFQs using recovery biomarkers have shown that they significantly underestimate energy [123, 296, 248, 267, 218] and protein intake [33, 245, 227, 246, 32] of respondents. Due to this large measurement error, the NCI Dietary Assessment Primer suggests that other assessment tools should be prioritized [38] and that FFQs might be more useful if used in combination with other tools [43, 119, 165].

As previously stated, some of the limitations of dietary assessment tools are being ad-

dressed with the use of technology. Electronic dietary assessment tools can ease the respondent burden and increase adherence [88]. More importantly, the ubiquitous nature of mobile technologies provides the means to assist dietary assessment or the creation of novel methods that can address the pitfalls of the classical approaches. Electronic dietary assessment methods have been used for several years but more recently, tools to address the portion size estimation and respondent burden problems using image recognition technology have started emerging [268]. This ongoing effort uses novel technologies such as deep learning to develop such applications [207].

Dietary assessment tools allow collection of data that can be used to capture dietary habits of study participants. These tools have, as shown, some limitations that make the task of deriving knowledge from study results more difficult. Additionally, when interpreting results from nutrition studies, one also has to consider the study design and associated benefits and shortcomings. We will discuss this in the next section.

1.1.2 Observational Studies and Randomized Controlled Trials

The search for optimal diets is an ongoing effort in society and the research community. Most nutrition studies that try to address how specific foods or dietary regimes are affecting health are observational. These studies are by definition subject to many confounding factors and this led to several producing contradictory results and controversy [91, 292, 197]. Data from observational studies, in particular from prospective cohorts, is critical for research in the nutritional field but it has limitations that should be considered. Measurement errors are a common issue with limitations innately associated with the already discussed dietary assessment tools. Another known issue relates to the fact that study participants show a tendency for under-reporting their energy intake [266, 20, 33]. This tendency also seems to correlate with social desirability and social approval traits (e.g. BMI)[271, 122, 126, 125, 127]. Additionally, this issue is also more closely associated with foods that are perceived as bad nutritional choices, such as sugary foods and alcohol [181]. Perceived physical activity, on the other hand, shows the opposite trend with a tendency for over-reporting [3].

Isolating the effects of specific nutrients is a challenge as nutrients can be highly correlated as they are often present in the same types of food. Food nutrient composition is

heterogeneous and depends on the type and cost of the product (organic/local farm products versus industrialized). Furthermore, food databases often lack complete information about nutrient content. Substitution effects might also occur: when testing for a specific nutrient the volunteers will change this diet. This change can cause unforeseen effects, confusing the obtained results. Another important point to bear in mind is the individual response variation to dietary exposure, depending on factors that might not be taken into account, such as lifestyle. Cardiorespiratory fitness is often not assessed and replaced with questionnaire-based physical activity assessment which often provides imprecise estimations [3, 277]. The composition of the gut-microbiota of each individual has a role in health and disease [55] and is often disregarded in nutrition studies. Some of these problems, however, are not limited to observational studies.

Randomized Controlled Trials (RCTs) are considered the most reliable studies for evidence-based medicine with some researchers advocating for their wider use in nutrition [35, 146]. These trials reduce the probability bias by randomly allocating treatments or dietary interventions. RCTs have, in some cases contradicted observational studies. A meta-analysis of several observational studies showed that B-vitamin could potentially decrease coronary heart disease risk [143]. However, 8 large RCTs failed to demonstrate this association [54]. RCTs have also failed to demonstrate that higher consumption of antioxidant vitamins reduces CHD risk [252, 343] as previous observational studies had proposed [192, 291].

RCTs are also not free of limitations [124] or failures. The vitamin supplement efficacy is still an ongoing discussion [213]. A double-blind randomized controlled trial involving pregnant woman showed that choline supplementation did not boost fetal brain supplementation [48], but it is estimated that 44% of women have increased dietary choline requirements due to genetic variants [346]. The Women's Health Initiative calcium and vitamin D controversial results showed that calcium and vitamin D supplementation did not have a positive effect on the risk of bone fractures [148]. However, this result seems to have been caused by an erroneous estimation of calcium intake by members of the control group [213].

RCTs minimize some of the effects of confounding factors when compared with observational studies. This does not mean, however, that observational data is not useful. There are situations where performing a RCT is simply not possible. Additionally, there are cases where data from both studies are concordant. One of these cases was shown for the relation between

the Mediterranean diet and reduced cardiovascular risk. Results from the PREDIMED study [78] were aligned with previous observational data [66, 273, 287].

While classical drug studies use placebo groups that allow the verification of a significant and measurable effect of the tested drug, such is impossible for nutrition studies. The effects of nutrition interventions are, by design, more subtle and nutrients cannot be completely removed from the diet of the participants. Additionally, there are various confounding factors and limitations in the dietary assessment of habitual diet of free-living populations. Biomarkers of nutrient intake can support this effort objectively assessing dietary consumption, avoiding the bias and errors of self-reporting. The rise of multiple omics technologies is an opportunity to objectively quantify the effects of nutritional patterns in the organism.

1.2 Foodomics

In previous sections, we discussed the subjective nature of self-reported dietary assessment methods and how that poses challenges in the interpretation of results obtained from nutrition studies. In addition, other factors increase the difficulty of result interpretation in nutrition. Nutrient-nutrient interactions caused by the ingestion of different combinations of foods can influence nutrient absorption. Food-composition databases are also prone to imprecision due to the natural variation of nutrient content in the same food item. For instance, selenium amounts in cereals and grains is determined by the amount of selenium in the cultivated soil [211], which implies that location and distribution plays a role in the validity of data [254]. Another good example is Vitamin E, coming from different fat and oils [247, 200], that is affected by the processing procedure, shelf life, and whether antioxidants were added to the product, restoring oxidized vitamin E [310, 281, 297].

Biomarkers of food intake have the potential to address some of these limitations by assessing intake and exposure to foods more accurately and on different time-scales (short, medium, and long-term [247]). To avoid the effect of confounding factors such as lifestyle and genetic variability a more "complete picture" is necessary. *Omics* technologies can play an important part in this as they are used to collectively quantify and characterize pools of biological molecules. In this section, we will discuss several of these technologies and how they are being used for the study of nutrition.

1.2.1 Nutritional Genomics

The study of the complexity, diversity, and influence of genomes started with the discovery of the DNA structure but the origin of genomic technologies can be traced back to the 1970's with the development of DNA cloning [96]. Since then, genomic technologies went through incredible developments. In the early 2000s, the establishment of the first reference human genome promoted the development of high-throughput sequencing platforms which caused an abrupt decrease in cost for sequencing [107]. These high-throughput technologies, known as Next-Generation Sequencing (NGS), are today important research and clinical tools.

Thanks to these technological developments, the field of Nutritional Genomics arose to study gene-nutrient interactions, with the potential for the development of personalized nutrition approaches based on the genetic make-up of each individual [155]. Despite this, research and applications are sensitive to the complexity of food and variable mechanisms that cause diseases [158, 157]. Other questions arise, how will gene expression change in response to different exposures or interactions with nutrients? Will these changes affect the health of an individual? Nutritional genomics covers not only the analysis of the genome but also the epigenetic and transcriptomic modifications and interactions caused by the intake of food [242, 263].

There are known conditions caused by single gene defects and associated nutrient interactions. For instance, Phenylketonuria (PKU), an inborn error of metabolism, is caused by a PAH faulty gene that codes for an enzyme that degrades the amino acid phenylalanine. If left untreated the disease can cause intellectual disability [67]. Despite not being curable, the disease is treatable thanks to a combination of a specific diet low in phenylalanine and medication. Another example is the genetic variant of the lactase gene that causes lactose intolerance. The main treatment consists of cutting out foods with lactose but alternatives exist, such as enzyme supplementation.

Not all gene-diet interactions involve a single gene. Gene-diet interactions exist for the FTO and MC4R genes, consistently associated with obesity risk and with type-2 diabetes. In a case-control study, it was suggested that the association of specific polymorphisms of these genes with type 2 diabetes depends on the dietary pattern. Adherence to a Mediterranean diet counteracts this genetic predisposition [233]. The same diet was also found to counteract

the predisposition for cardiovascular disease caused by a polymorphism in the TCF7L2 gene [60].

Several genetic variants were identified for their association with diabetes and obesity [205]. Some of these variants are closely related to obesogenic dietary exposure [156, 154, 159]. Genome-wide association studies (GWAS) have helped improve the understanding of these diseases pathophysiology but fail to explain why, for instance why Asians tend to develop diabetes at a younger age with a lower prevalence of obesity when compared to European populations [340]. Other GWAS have identified gene variants associated with diseases related to energy metabolism and aging. These associations usually are indicatives of small increments in risk suggesting that there are other mechanisms playing a role, such as epigenetic changes. Epigenome-wide association studies (EWAS) have reported that epigenetic changes caused by diet and other factors complement genetic predispositions and contribute to the development of metabolic and other types of disorders [208, 323, 105, 240]. In the effort to understand the association between genetic variation and diseases it is also worth mentioning the work by the NIH Roadmap Epigenomics Consortium that has generated the largest collection of reference human epigenomes for the study of the molecular basis of human disease [178]. In this front, there are currently efforts in managing inflammatory disease and general health through dietary factors [40, 147, 166, 69, 188, 120, 79, 319]. The search for these nutritional targets and promotion of healthy aging will also encompass the determination of optimal doses and exposure windows during different phases of development [320].

Despite all the good examples previously discussed, there is still a lack of clear associations between specific genes and dietary intake or nutrient-related diseases. In fact, a recent meta-analysis of commercially available nutrigenomic tests failed to find any statistically significant association [241]. It seems that the genetic predisposition by itself fails to explain the effect of dietary patterns on health. As nutrition research moves to the identification of the physiological role of minor dietary components and monitoring of dietary interventions [265], more systematic broad ranged techniques become relevant. Such is the case of proteomics and metabolomics technologies.

1.2.2 Proteomics

Nutritional proteins and peptides can have beneficial or adverse effects. Proteins are the only source of essential amino acids and nitrogen for humans and can be obtained from animal and plant sources [265]. Food allergies are generally caused by proteins and affect around 250 million people worldwide [327]. The continuing increase in the occurrence of these allergies without an apparent reason is a theme of increased interest in nutritional research [52]. The World Health Organization and the International Union of Immunological Societies created a resource for the systematic nomenclature of allergens [133] in November of 2017 listed 882 allergens with 310 food allergens. While food legislation demands detailed allergen content in food, detection and quantification remain challenging due to the high complexity of food composition and food matrices. Emerging mass spectrometry methods might be able to address some of these shortcomings [169]. This research might increase food safety but also possibly support the discovery of the mechanisms underlying these allergies, still poorly understood.

Proteomic analyses are commonly used to study bioactive peptides, specific fragments of a protein that can potentially influence health [168]. Various studies sought to identify bioactive protein and peptides in milk from human and other mammalian sources [272, 285, 201, 13]. These studies identified various functions for these proteins such as beneficial effects on host immune response [80, 238], antimicrobial and anti-amnestic activities, antioxidant effects among others [26]. Attempts to mechanistically identify endogenously produced peptides in human milk were also pursued. More than 700 naturally occurring peptides were found to derive from 30 human milk proteins [112].

Another focus of nutriproteomics is in studying these bioactive proteins and peptides in plant sources. These proteins are a valuable replacement for animal protein and have a smaller ecological impact. Typically, plant-based protein sources include soy, wheat, and legumes. The nutritional value of these has been thoroughly analyzed in terms of their bioactive protein and peptide content [180]. In the case of wheat, these analyses are important as bread quality depends on the protein content of the seeds [98]. Soybean, traditionally used in Asia, offers the complete set of essential amino acids and proteomic analysis support some of the health claims made in their favor [92, 180]. Characterization of protein content in peas, for instance,

was also studied to evaluate plasticity of protein composition [37].

The increasing volume of information generated by proteomic studies specifically on the bioactive proteins and peptides content, their characteristics, and how they can affect health can promote the design of strategies to tailor nutritional interventions or the development of products that promote beneficial physiological effects.

1.2.3 Metabolomics

Analogous to other *omics*, metabolomics is the field that aims at characterizing the full set of small molecules (metabolites) that are the substrates and products of metabolism [306, 219]. In the human, metabolites can be produced by the body, by colonizing microorganisms [318, 113], but also from exogenous sources such as drugs, diet, and other exposures such as toxins from the pollutants [150] (Figure 1.2). These small molecules are involved in key cellular functions such as energy production and storage, apoptosis, and signal transduction [186]. They are also indirectly involved in other processes such as regulating epigenetic mechanisms [290, 160, 316] or modifying protein activity [328, 216].

Metabolomics technologies allow researchers to measure hundreds of metabolites simultaneously, mainly through the use of mass-spectrometry, an analytical tool that measures mass-to-charge ratio of ionized chemical species to determine their identity. The typical approaches consist in measuring metabolites in a targeted or untargeted manner. Targeted metabolomics, as the name indicates, is used to target metabolites and metabolic pathways of interest and thus, requires *a priori* knowledge. When compared with untargeted approaches, these methods offer more sensitivity and selectivity. However, untargeted approaches, especially liquid chromatography/mass spectrometry (LC/MS) the most common method used in metabolomics studies, measures thousands of signals offering a more global overview of metabolism[239]. Despite this, hundreds of unknown signals that might correspond to unknown metabolites are found in metabolomics datasets [344]. The challenges of translating these signals into metabolite identities and analysis of complex metabolomics datasets are being addressed with the development of numerous analytical tools and databases to store metabolite identity information. Methods for the analysis of these type of data are available for researchers in tools such as MetaboAnalyst [338, 339], XCMS Online [139], Bayesil

[253], Workflow4Metabolomics [102], MetFrag [334], and MyCompoundID [140]. To complement these methods, several databases of metabolite identifiers are available, notably, HMDB [333, 331, 330], METLIN [286], MassBank [137], and GMD [172, 142].

This capacity to measure and analyze this many chemicals led to several projects to identify the metabolome of human [333, 331, 330], plants [18], and microbes [318]. Initially, the promise of metabolomics was to address the discovery of biomarkers correlated with various diseases but researchers started finding that the metabolome shows noticeable differences that are related with gender, age, health status, genetics, and most importantly, diet [136, 284, 175, 129]. In this context, different studies also characterized responses to the intake of whole foods or food constituents [289, 191, 312].

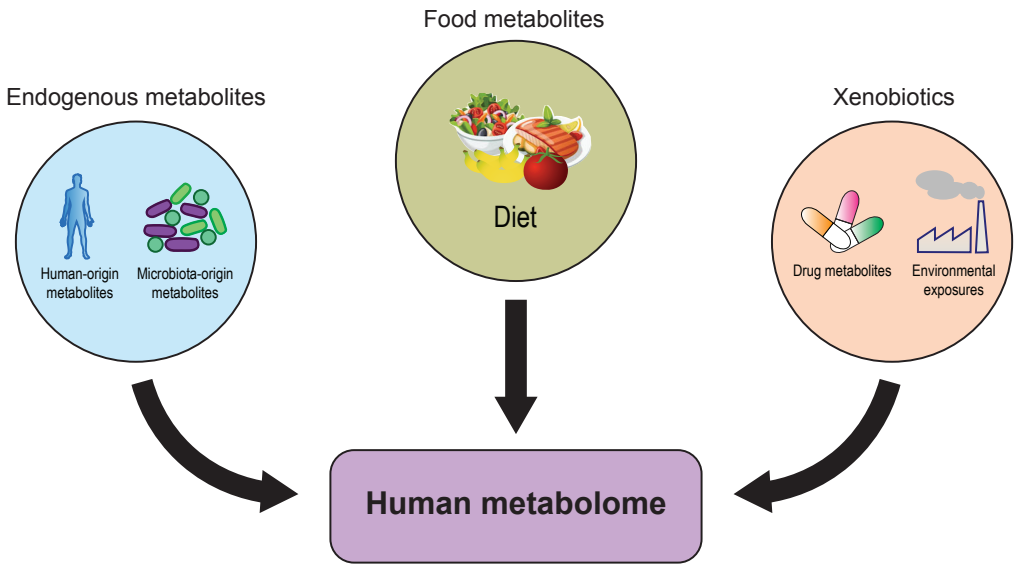

Figure 1.2: Overview of the different origin of metabolites that compose the human metabolome.

As such, metabolomics in nutrition has the potential to address some of the limitations of current dietary assessment tools if objective biomarkers of food intake are characterized. Traditionally, biomarkers of food intake were identified/measured in epidemiological studies from samples such as serum, red blood cells, and urine. These biomarkers can be used to assess specific food or food group intake. A systematic review of intervention studies has identified carotenoids and vitamin C as biomarkers for fruit and vegetable intake [19], urinary and phenolic acids can serve as biomarkers for polyphenol-rich food intake [206]

such as spices and dried herbs, darkly coloured berries, cocoa products, seeds, nuts and some vegetables such as olives and globe artichoke heads [244] but also tea and wine [132]. Alkylresorcinol concentrations seem to correlate with whole-grain food intake [10, 257], plasma concentrations of daidzein and genistein (isoflavones) can assess the intake of soy [322] and fatty acids are related with the intake of meat, dairy products, and fish [7, 215, 15].

Dietary biomarkers are also susceptible to other influences besides diet. In this context, the Food Biomarkers Alliance (FoodBall) project was created, spanning 11 countries. This project aims at the development of clear strategies for food intake biomarker discovery and validation [foo]. Dietary biomarkers have typically short half-lives, and are most suitable for identification of frequently consumed foods or food groups. While classical approaches remain relevant, *omics* technologies, in particular, metabolomics, give a more global snapshot of the biological phenotype and thus, enable the analysis of combinations of dietary components and diet-induced metabolic changes [103]. Taken together these technologies carry the potential for the development of personalized nutrition approaches [228].

1.3 The Gut Microbiota

Symbiosis is an essential aspect of nature. In fact, mitochondria and organelles in human cells are remnants of prokaryotes and are essential to life [179]. Symbiotic interactions with microbial communities are present throughout nature and the human bodies alike. In particular, the human gut microbiota is composed of thousands of different microbial species [348]. In human adults, 60% of the bacteria belong to the *Bacteroidetes* and *Firmicutes* phyla [17]. Common found genera of bacteria are *Bacteroides*, *Bifidobacterium*, *Lactobacillus*, *Clostridium*, *Escherichia*, *Ruminococcus*, and *Streptococcus*. While the community composition is highly variable between individuals their metabolic capabilities are well conserved [58].

Several studies have reported that imbalances in gut microbial populations can be associated with diseases such as obesity [347], type 2 diabetes [249], and inflammatory bowel disease (IBD) [90]. For this reason, there is an increased interest in understanding how the dysbiosis of these microbial communities contributes to the mechanism of disease and how strategies can be devised to avoid or treat these imbalances and maintain health. These studies investigate correlations between composition and a disease state and in line with this, treat-

ments for gut-associated diseases are designed to revert the composition of a dysbiotic gut microbiota community to mimic one of a healthy individual. These treatments are typically ingestion of probiotics [99], changes in diet to include prebiotics [42], and the most radical form of gut microbiota modulation, fecal microbiota transplantation [258].

The composition of the gut microbiota is highly influenced by age, genetics, and diet [104]. These microbial communities indeed play an important role in nutrition. They digest dietary fibers [74] providing fermentation products such as short-chain fatty acids (SFCAs). These can be used by the host as energy precursors [75] and benefit the immune system [95]. Furthermore, these microbes also supply the host with essential amino acids and vitamins [274]. It has been shown that long-term dietary patterns modulate the composition of the gut microbiota [336, 214, 74, 185, 324] but in a study that tested animal-based and plant-based diets, significant changes in the composition of the microbiome occur in a matter of days [64]. Additionally, the composition of the gut microbiota can be used as a basis for the prediction of blood-sugar level responses and therefore, be used for dietary recommendation [345]. In a later study, a similar predictive method was applied for the blood-glucose level in response to the consumption of different types of bread [173].

Metagenomics techniques are the most common tools to study the composition of the gut microbiome. Analysis of stool samples to derive the microbiome composition are usually achieved by identifying the 16S ribosomal RNA gene content or whole genome shotgun sequencing. Knowing the composition of individual microbiomes allows the association of specific compositions with conditions or dietary patterns. However, the high complexity and inter-individual variability of the gut microbiota composition makes it extremely difficult to understand the underlying mechanisms and specific pathways involved in the observations. Computational methods can support the exploration of these mechanisms and support the generation of models and hypothesis to be tested in subsequent experiments. In the case of nutrition, if these computational tools are able to integrate data from the microbiome and previously discussed omics, then they can provide a unique opportunity for personalized dietary intervention. In this work, we propose constraint-based reconstruction and analysis (COBRA) as the framework to achieve that.

1.4 Constraint Based Reconstruction and Analysis

Constraint-based reconstruction and analysis (COBRA) is an approach that typically uses genome-scale metabolic reconstructions (GENREs) to study metabolic pathways, specific organisms, or metabolic interactions [225, 237]. GENREs represent the full portfolio of metabolic reactions known to be present in a given organism or cell. The starting point of these reconstructions is an annotated list of genes that code for metabolic enzymes [304]. This process can be performed automatically in a time-efficient manner thanks to different available tools and databases [152, 45, 131]. These automatically generated reconstructions are typically referred as "drafts" and usually go through a process of manual curation that addresses issues related with reaction stoichiometric and directionality consistency [85], gene mis-annotations [109], and integration of data derived from experiments [304]. This manual-curation is often very time-consuming and to address this issue several tools and algorithms have been created. Recently in our group, we have also published the largest collection of reconstructions for gut microbes by semi-automatically curating reconstructions [195]. The methods used to generate the AGORA collection are a step forward in the creation of a framework that greatly increases the speed and quality of metabolic reconstruction generation.

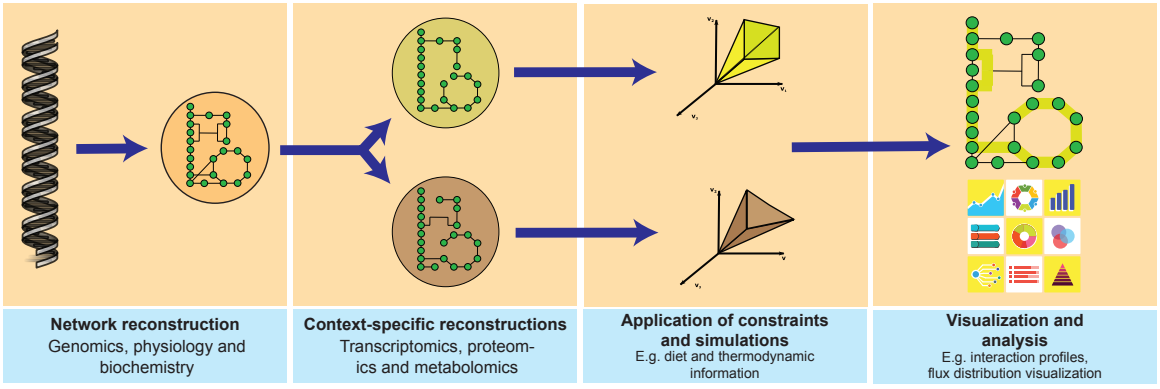

Figure 1.3: The COBRA approach: GENREs are build from the genome annotation and manual curation. Metabolic models are derived by integrating data from different sources. Simulations capture the predicted behavior of the cell/organism under specific conditions.

Metabolic modeling

Metabolic reconstructions can be converted to mathematical representations in the form of a matrix: the stoichiometric matrix (S-matrix). In this matrix, rows represent metabolites and

columns reactions. Each cell, or entry, of the matrix, will contain a value that indicates the stoichiometric coefficient of the metabolite in the reaction. These reconstructions can then be formalized as metabolic models and be used to simulate physiological states. This is achieved by applying condition specific constraints that can specify the medium conditions (e.g. for bacterial growth experiments) or constraining the flux of specific internal reactions according to experimental data. The most commonly used method, Flux Balance Analysis (FBA) [234], specifies an objective function, typically biomass production or ATP maintenance, which produces a "biased" flux vector that represents one of many flux distributions that satisfy the objective and constraints. This set of possible solutions, or solution space, can be investigated using flux variability analysis (FVA) [110] that for each reaction, gives the minimum and maximum flux value in the solution space. Sampling this solution space allows gathering information on the distribution of alternative solutions [116, 269].

These methods are available for the scientific community through software packages such as the CobraToolbox [130] and have been used for modeling human and gut microbiota metabolism. Importantly, the COBRA approach enables the integration of data from previously described *omics* technologies [16, 36, 144], or to use these data to generate context-specific reconstructions [184] (Figure 1.3). Taken together GENREs and COBRA methods have been used to address numerous biomedical questions, including the phenotypic consequences of dietary regimes [262, 280] and enzyme deficiencies [260, 279, 305, 236].

1.5 Scope and aim of the thesis

This thesis describes the development of a knowledge base that aims at integrating different types of information into the COBRA framework and pave the way for its usage as a tool for nutritional recommendations (Figure 1.4). For this purpose, the project was divided into two main objectives: the development of a knowledge-base that integrates human and gut microbiome metabolism and then, the inclusion of information on nutrition and diseases. This thesis describes my work in building the Virtual Metabolic Human (VMH).

It will begin with an overall description of the developed knowledge-base and it's content. After that, using the developed database and its tools, several examples of analysis of the data are shown. A technical description of the database and its application programming

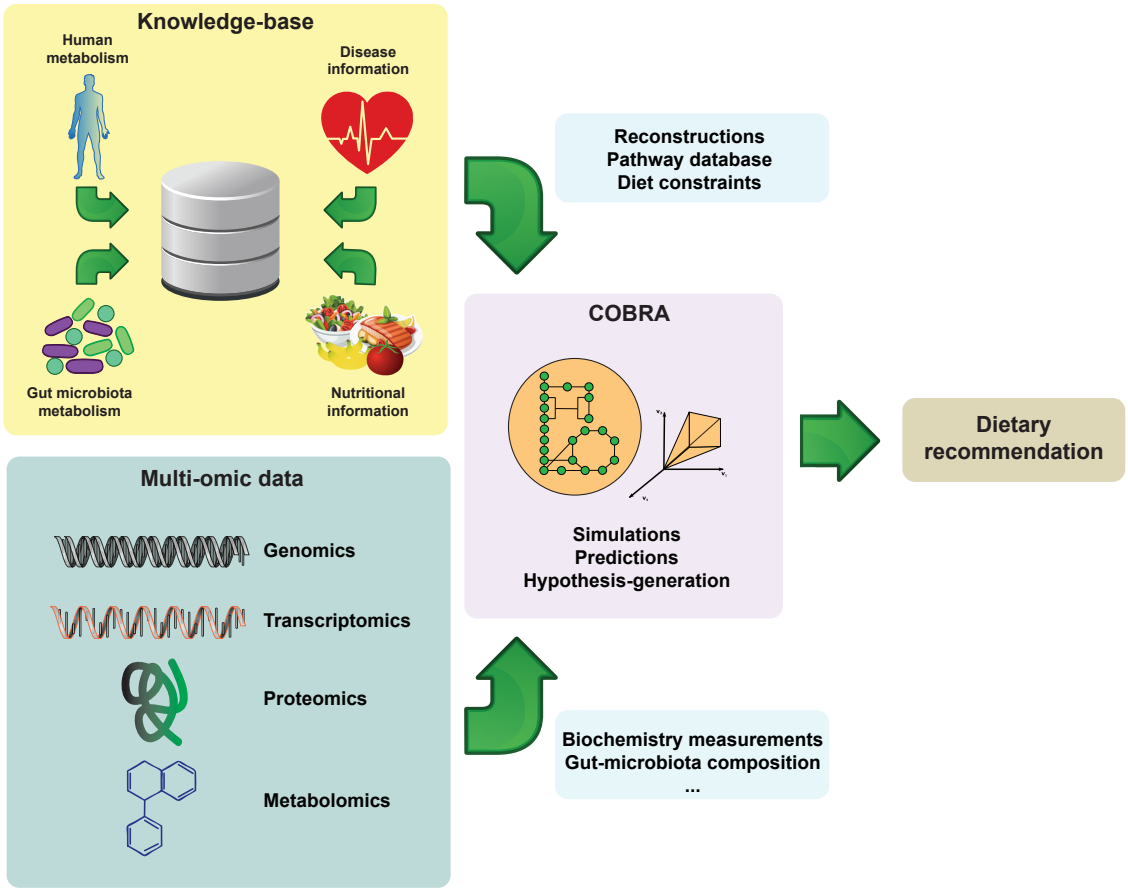

Figure 1.4: The integration of metabolic reconstructions of human and gut microbial metabolism with nutritional information and disease will allow combining such information with multi-omics data. This combination of resources can become a tool for personalized nutrition.

interface (API), as a manual of usage, will also be provided. To conclude, I make a reflection on the challenges and tribulations that the development of a project such as the VMH can bring to researchers in the fields of computational biology and bioinformatics. I discuss my journey in the development and implementation of this project and how adopting agile software development tools and approaches can benefit researchers. I have the hope that these can be more commonly adopted by research teams in the near future.

Below are short descriptions of each chapter and the detailed contributions of the different collaborators involved.

Chapter 2: The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease

Chapter 2 describes the Virtual Metabolic Human (VMH), a database that combines information on genome-scale reconstructions of human and microbial metabolism. The content of the database is described along with several examples of the user interface and it can be used.

Contributions

Alberto Noronha (AN) and Ines Thiele (IT) designed the study. AN, Yohan Jarosz (YJ) and Reinhard Schneider (RS) developed the necessary infrastructure for the project. Jennifer Modamio led the update on ReconMap. AN, IT, Laurent Heirendt, German Preciat, Beatrice Pierson, Hulda S. Harulsdottir, Almut Heinken, Stefania Magnusdottir, Eugen Bauer, and Ronan M. T. Fleming contributed with content to the database. AN developed the database, web-interface, and web API. AN and IT wrote the manuscript. All authors reviewed and approved the text.

Chapter 3: Design and applications of the Virtual Metabolic Human database

Chapter 3 describes the database structure of the VMH database and the development of the web application programming interface that allows third-party access to the database content. Examples of the usage of the API are given. Finally, applications taking advantage of the connectivity of the different resources and tools that compose VMH are shown.

Contributions

IT and AN designed and planned this work. AN, YJ, and RS created the necessary infrastructure. AN developed the database and web API. AN and IT wrote the text.

Chapter 4: Visualization of Metabolic networks and Disease maps

Chapter 4 describes the development of ReconMap, an interactive map of human metabolism and Leigh Map an interactive gene-to-phenotype approach to the diagnosis of Leigh Syn-

drome. This chapter is a combination of the reprints of the ReconMap paper published in Bioinformatics in February 2017 [223] and the Leigh Map paper, published in the journal Annals of Neurology, in January 2017 [250].

Contributions

For the development of ReconMap, IT and Ronan M. T. Fleming (RMTF) were involved in the conception and design of the project. AN, Anna Dröfn Daníelsdóttir, Freyr Jóhannsson, Soffía Jónsdóttir, Sindri Jarlsson, Jón Pétur Gunnarsson, and Sigurður Brynjólfsson manually designed the map. AN, RS and Piotr Gawron supported the integration of ReconMap into the MINERVA framework. All authors read and approved the manuscript.

For the Leigh Map text Shamima Rahman (SR) and IT were involved in the conception and design of the study. Joyeeta Rahman (JR) and AN acquired the data and created the network. SR, IT, JR, and AN drafted the manuscript and the figures.

Chapter 5: Challenges and tribulations in the development of a biological database

Chapter 3 provides an overview of some of the main decisions that need to be made in the development of a biological database. It intends to be a starting guide for researchers involved in similar projects.

Contributions

IT and AN planned, wrote, and reviewed this chapter.

Chapter 6: Concluding remarks

Chapter 6 contains the conclusions of this thesis and the author's personal outlook on the future directions of the use metabolic modeling in the field of nutrition.

Contributions

The text was fully written by AN.

Chapter 2

The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease

Completely or partially as in: Alberto Noronha, Jennifer Modamio, Yohan Jarosz, Laurent Heirendt, German Preciat Gonzàlez, Beatrice Pierson, Hulda S. Harulsdottir, Almut Heinken, Stefania Magnusdottir, Eugen Bauer, Reinhard Schneider, Ronan M. T. Fleming, Ines Thiele. The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. *Manuscript in preparation.*

Abstract

Nutrition plays a key role in metabolic homeostasis and an unbalanced diet is associated with a variety of conditions, such as diabetes and cardiovascular diseases. Metabolism is influenced by genetic and environmental factors and an integrated analysis of data originating from different fields, such as physiology, genetics, and gut microflora is necessary to foster a better understanding of its mechanisms. Genome-scale metabolic models provide a framework for this integration, but a knowledge-base for this purpose is necessary. We have created the Virtual Metabolic Human (VMH), a resource that integrates human and gut-microbe metabolic reconstructions with nutritional and disease information. This integration and the different tools provided by this resource offer a unique environment for the study of the effect of diet on the metabolic system. VMH aims at guiding research in the field of nutrition and support the knowledge gain that could impact the way healthcare and disease prevention is perceived.

2.1 Introduction

Lifestyle parameters, such as diet, are recognized as major modulators of human health and have an important contribution to onset, progression, and severity of various diseases, including cancer, metabolic diseases, and neurodegenerative diseases. To understand these diseases, one needs to account the various factors that influence how the human body processes food. Some of these factors are determined by the genome and patterns of expression of particular genes that translate to the ability - or lack of - to degrade and absorb certain nutrients. Other factors are the composition of the gut microbiota, the diet, and the lifestyle. While multi-omics technologies can support the comprehensive collection of dietary intake data and monitoring of the health status of individuals, the high complexity of these data poses challenges in its integration and interpretation. This integration could indeed lead to insights about the complex processes involved in the digestion of dietary components and how these can contribute or prevent the appearance of the aforementioned conditions.

Databases are a compelling way of storing, connecting, and making available a multitude of information derived from primary literature, experimental data, genome annotations, beyond others. Metabolism-related databases include, but are not limited to the following. For instance, the Kyoto Encyclopedia of Genes and Genomes (KEGG) is an extensive biochemical database covering almost 4000 organisms [152, 153]. BioCyc [45, 161] is a multi-scale knowledge resource containing a collection of 7667 pathway/genome databases. The Human Metabolome Database (HMDB) is the most comprehensive collection of human metabolite data [333, 331, 330], which is also connected to FooDB, a comprehensive resource of nutritional information with 28,000 food components and food additives, and Drugbank, which contains detailed information on FDA approved and experimental drugs [332]. The Human Protein Atlas contains protein expression and RNA-seq data for numerous human tissues and cell lines [315]. The BiGG knowledge-base [164] is a resource for centralized storing of genome-scale metabolic reconstructions, providing search functionalities, pathway visualization via Escher [163], and a comprehensive application programming interface.

However, despite the wealth of biochemical databases, there is no database that explicitly connects human metabolism with genetics, (gut) microbial metabolism, nutrition, and diseases. One reason for this may be the use of non-standardized nomenclature that complicate

their integration. Moreover, manual curation of database content is time-consuming and requires expert domain knowledge. Genome-scale metabolic reconstructions represent the full repertoire of known metabolism occurring in a given organism and describe the underlying network of genes, proteins, and biochemical reactions. High-quality reconstructions go through an intensive curation process that follows established protocols to ensure a high quality and coverage of available information about the organism [304]. Thus, metabolic reconstructions represent valuable knowledge bases summarizing current metabolic knowledge about organisms.These reconstructions enable the integration of data originating from different "-omics" technologies [16, 36, 144]. Moreover, several algorithms exist that use these "-omics" data to generate context-specific reconstructions [232]. This so-called constraint-based modeling approach (COBRA) is completed by a plethora of methods that use condition-specific models derived from these reconstructions to simulate the phenotypic behavior of the cell or organism under different conditions [234, 237].

Here, we describe the Virtual Metabolic Human database (VMH, http://vmh.life), which has at its core the manually curated human metabolic reconstruction, Recon 3D, which has been developed by the systems biology community over the past decade [39, 73, 300, 305]. Recon 3D describes the underlying network of genes, proteins, and biochemical reactions present in at least one human cell, as encoded by 17% of the protein-coding part of the human genome. Using Recon 3D as a docking station, we could connect manually-curated genome-scale metabolic reconstructions for more than 770 human gut microbes thanks to an overlapping metabolite and reaction nomenclature [195]. We then linked over 200 Mendelian metabolic diseases [260] to the genes present in Recon 3D as well as the molecular composition of more than 8000 food items from the USDA National Nutrient Database for Standard Reference [317]. Moreover, all VMH entries are connected to external databases, making VMH a unique reference database for human metabolism. A comprehensive, Google-like map of the human metabolism, ReconMap [223] and a Leigh-disease specific map [250] are hosted on VMH permitting the visualization of simulation results. VMH is composed of three layers, a MySQL relational database (for information storage), a representational state transfer application-programming interface (API), and a user-friendly web interface for browsing, querying, and downloading the VMH database content. Users can provide feedback through the different platforms of the website, which will be curated and integrated into the

knowledge base. Taken together, VMH represents a novel, comprehensive, multi-faceted overview of human metabolism.

2.2 The Virtual Metabolic Human

The VMH consists of four resources: “Human Metabolism”, “Gut Microbiome”, “Disease”, and “Nutrition”. These are interlinked based on shared nomenclature and database entries for metabolites, reactions, or genes (Figure 2.1).

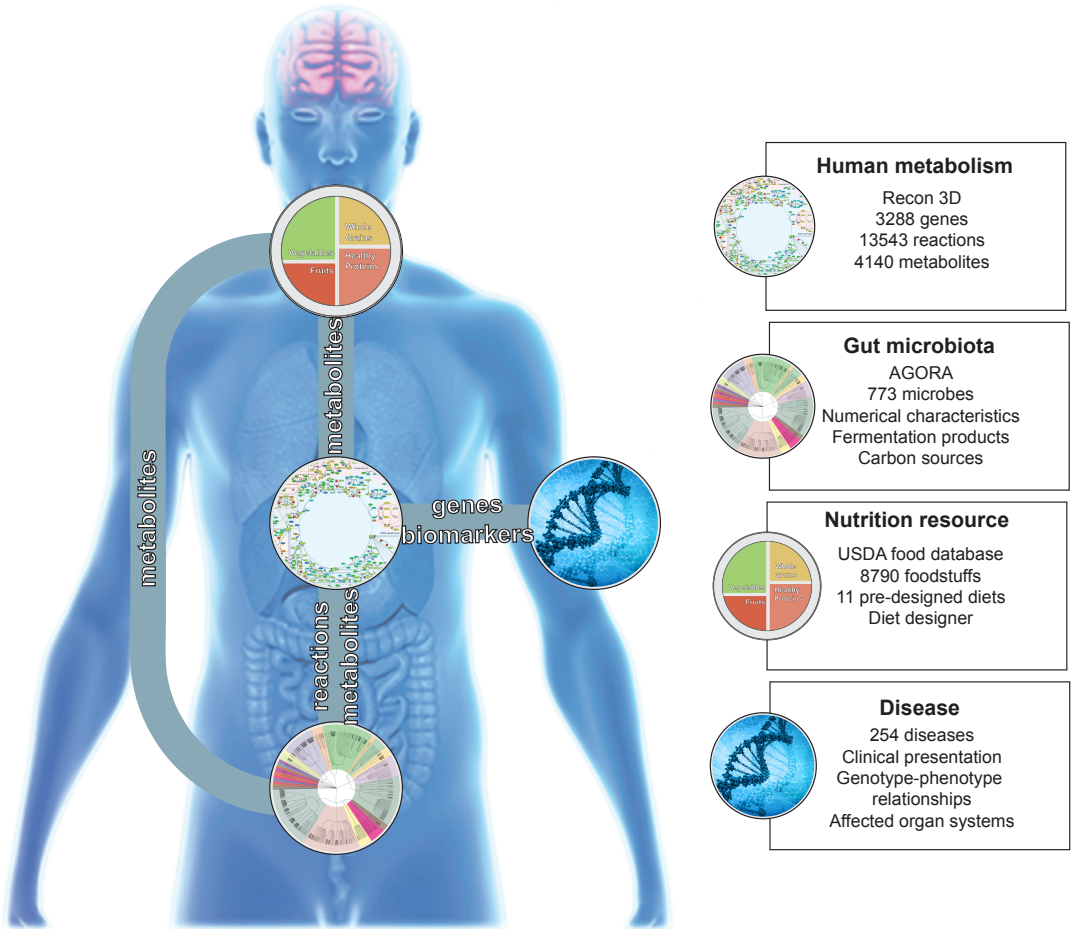

Figure 2.1: Overview of the Virtual Metabolic Human. The database is composed of 4 resources: "Human Metabolism", "Gut Microbiome", "Disease", and "Nutrition". The 4 resources are connected with each other through entities sharing nomenclature.

Overall, the VMH contains 18,107 unique reactions, 5,222 unique metabolites, 3,288 human genes, and 486,471 microbial genes as well as 255 diseases, 773 microbes, and 8,790

food items. The underlying database architecture allows for easy navigation between the four resources. For instance, one can connect the reaction and metabolite content between the “Human” and the “Gut microbiome” resources to identify common or specific metabolic modules across organisms as well as their complex interactions. The "Disease" resource is connected with the "Human" resource by disease-affected genes as well as biomarker information in the form of metabolites [260]. Finally, the “Nutrition” resource is connected with the "Human" and "Gut microbiome" resources by mapping food nutrients to 100 metabolites (Figure 2.1). Each resource is “one-click-away” and all search results and database content are downloadable. Each entity of the database (e.g., metabolite, reaction, and gene) has a detail page where additional information is provided, connections with other entities of the database, and links to external resources. In the following, we briefly describe the content of each resource and detail pages.

2.3 Human metabolism

The VMH hosts the most recent version of human metabolic network reconstruction, named Recon 3D [39], which accounts for 13,543 metabolic reactions distributed across 126 subsystems, 4,140 unique metabolites, and 3,695 genes. The content of Recon3D has been assembled through extensive literature review over the past 10 years, and is continuously updated by us and others. Each reaction, metabolite, and gene contains its own detailed page, with additional information of supporting evidence in the literature, as well as their relations with other entities of the database. Great emphasis has been put into collecting a comprehensive set of database dependent and independent identifiers, allowing the identification of each entry and its cross-reference to other, external resources, such as KEGG and HMDB.

The visualization of metabolic pathways is an essential tool to understand the biological processes. We have generated a substantially updated metabolic map of ReconMap, which visualizes the extended and refined content captured Recon3D. , as well as a generic, constrained model of Recon3D, can be downloaded in different formats, e.g., in the systems biology markup language (.smbl) or in the proprietary Matlab (.mat) format, from the download page and the API.

2.4 Gut microbime

This resource contains the AGORA collection of 773 semi-automatically curated strain-specific metabolic reconstructions, belonging to 205 genera and 605 species [195]. All microbial reconstructions were based on literature-derived experimental data and comparative genomics. A typical reconstruction contains an average of 771 (±262) genes, 1198 (±241) reactions, and 933 (±139) metabolites. We provide detailed information for each strain and reconstruction along with known fermentation products and carbon sources.

2.5 Nutrition

This resource contains the molecular composition information for 8,790 food items distributed in 25 food groups, which was obtained from the USDA National Nutrient Database for Standard Reference [317]. Of the 150 nutritional constituents, 100 could be mapped onto the metabolites present in the VMH (Supplementary Table A.1). Within this resource, we provide 11 diets, which were defined based on real-life examples and literature. For instance, an "EU diet" was designed based on information from an Austrian Survey, on which about 100 people from different ages [77]. The composition of each meal (e.g., eggs and bread for breakfast) is given in the appropriate portion sizes. The molecular composition can be downloaded in g per person (70kg) per day or as flux rate (in millimole per person per day), which can be directly integrated with, e.g., the human metabolic model in the COBRA toolbox.

The 11 pre-designed diets available in VMH were designed with the support of a nutrition professional to follow the caloric content based on the average recommended daily intake (around 2500 calories for a male person). The diets consist of a one-day meal plan and include information about energy content, fatty acids, amino acids, carbohydrates, dietary fibers, vitamins, minerals, and trace elements. The information for the nutritional composition or the foods and dishes has been provided by the “Österreichische Nährwerttabelle” (Gatternig, Maierhofer et al). The calculation of the fluxes is made by converting the nutrient amount present in the foodstuff portions from grams to millimole per human per day. For each metabolite, its molecular masses were calculated. After a conversion of units, we determine

the amount of that metabolite in the portion of food, using the database nutritional information:

$$metaboliteamount = \frac{databasevalue \times portion}{100} \tag{2.1}$$

After this we convert this value to a flux using the following formula:

$$flux = \frac{metaboliteamount}{metabolitemass} \times 1000 \tag{2.2}$$

Diet designer

The available diets are a good starting point but they limit the freedom with which researchers can test changes to a diet. Manually calculating the fluxes is a laborious task and for that reason, we have created the "Diet Designer" tool. This tool allows users to design their own diets. The interface is divided into two lists: "Available foods" and "Selected foods". Users can search and select any food from the available 8,790 foods and add them to the list of selected by specifying a portion size. While the user designs the diet overall information is updated on a panel on top of the selected list of foods with information on total calories, lipids, proteins, and carbohydrates, and weight. When finished, the user can see and download the corresponding molecular composition as well as flux values (Figure 2.2).

2.6 Disease

Our resource includes 254 inherited metabolic diseases (IMDs), which are rare genetic disorders leading to a defective or abnormal enzyme function [260]. A total of 288 unique genes and 1872 unique reactions are associated with these IMDs. We compiled clinical presentation, genotype-phenotype relationships, and the affected organ systems associated with these IMDs from multiple literature and database resource.

The VMH also hosts the LeighMap [250], which represents a computational gene-to-phenotype diagnosis support tool for mitochondrial disorders. The Leigh Map comprises 87 genes and 234 phenotypes, expressed in Human Phenotypic Ontology (HPO) terms [170], providing sufficient phenotypic and genetic variation to test the network's diagnostic capability. The Leigh Map can be queried to generate a list of candidate genes and aims to

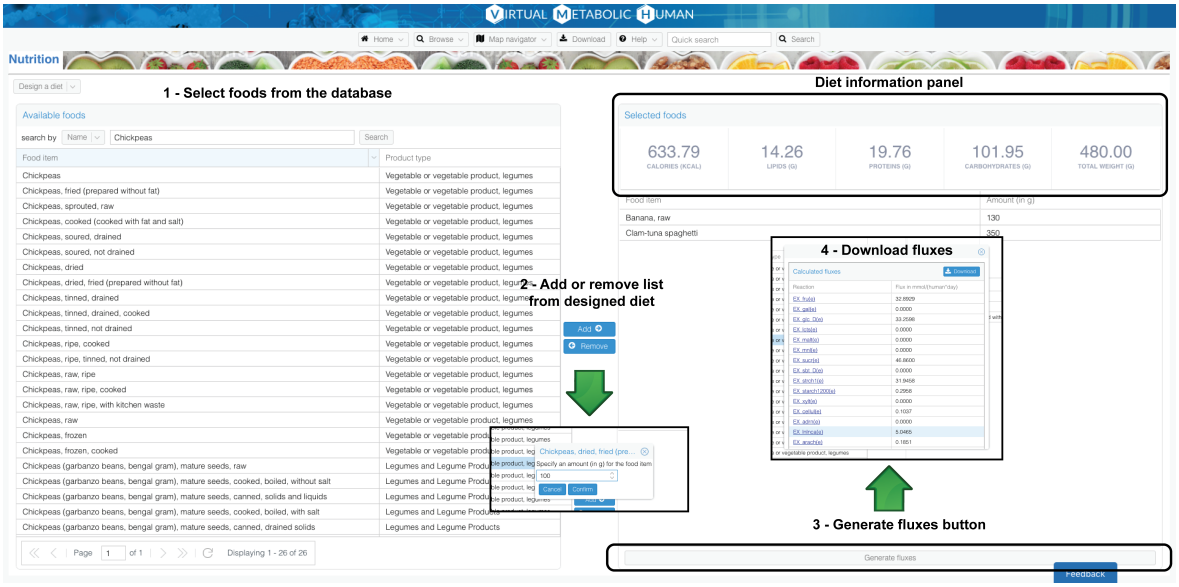

Figure 2.2: Overview of the Diet Designer. The interface is split into two panels. The list of available foods (1) and the list of selected foods. Users can select a food from this list and specify a portion in grams. When the list is finished users can download the flux values to be integrated into their simulations.

support clinicians by providing faster and more accurate diagnoses for patients. This will facilitate taking appropriate measures for further treatments and demonstrates the efficacy of computational support tools for mitochondrial disease.

2.7 Detail Pages

VMH contains detailed information for each entity in the database, including internal connections and internal resources. Through the user interface, a user can easily search the different resources and navigate the various levels of detail, e.g. from disease information to low-level metabolite biomarker information and chemical structure. In this section, we will provide details on each of these detail pages.

2.7.1 Metabolite detail page

Each metabolite in VMH is represented by an abbreviation that uniquely identifies a specific molecule involved in, at least, one metabolic reaction present in the database. Each metabolite also contains a name that better identifies that specific molecule, and description and synonyms extracted from HMDB when available. The formula displayed in VMH is

often different from other databases, and this is due to the fact that metabolites in VMH can represent the acid/base form of the neutral molecule. Therefore, there is always a charge value associated with it. Inchi string and Smiles are also available for most of the metabolites in VMH thanks to the work of Preciat et al. [106] in which mol files were generated for all Recon3D metabolites, and users can visualize (and interact with) the structure of these metabolites. The *mol* files are also available for download on the detail page of a metabolite.

There is an extensive list of external links displayed for each metabolite when that information is available. There are cases where some of these were not identified but we hope that the feedback functionality of VMH will support a community effort in the completion of these missing values. Available external databases are KEGG [152, 153], PubChem [162, 325], Chebi [118], HMDB [333, 331, 330], Foodb, ChemSpider [243], BioCyc [45, 161], DrugBank [332], and Wikipedia. A biochemical and disease maps section is also available where we map these molecules to visualization tools. This feature currently displays identifiers for ReconMap and PD-Map [93] when available, but we envision an expansion as new maps become available.

Thermodynamic information is displayed, when available [222]. For the metabolites, Standard Gibbs energy information is presented for different compartments. The information about the presence of the metabolites in human biofluids was extracted from HMDB, literature sources [50, 135, 145, 278] and the Netherlands Metabolomics Centre (NMC - http://www.metabolomicscentre.nl/). Information can be qualitative (presence) or quantitative if a range of values is specified. The sources of the information are specified in each row of values. Biomarker information when available connects the metabolite with diseases. In addition, each metabolite has the information about the number of human and microbial reactions where it is involved, as well as if it used as a carbon source or is a fermentation product to any of the microbes available in the "Gut Microbiota" resource.

A detailed view of the metabolite detailed page can be seen in Figure 2.3.

2.7.2 Reaction detail page

A reaction in VMH is represented by its abbreviation and a more detailed description, which is usually the name of the associated enzyme and on occasions, the cellular location where the

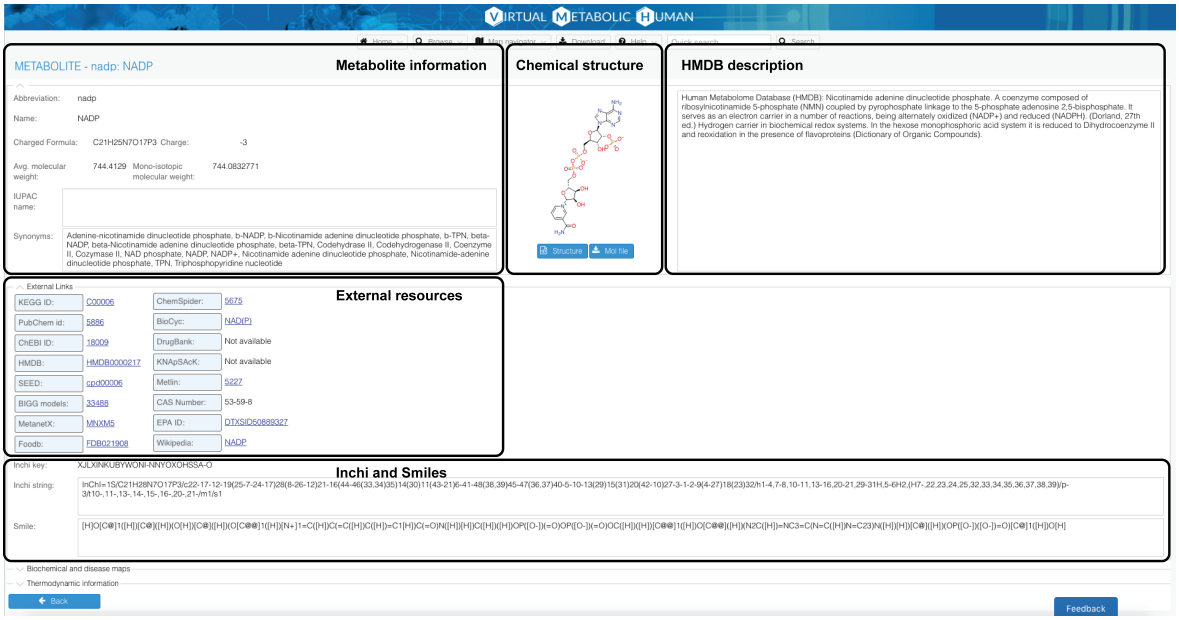

Figure 2.3: Overview of the metabolite detail page. The interface contains additional detailed information of the metabolite, including a visualization of the chemical structure. In addition, this page includes connections to external resources, and to related internal entities (e.g. reactions in which the metabolite is present).

reaction occurs. This is a particularity of GENREs, where reactions occurring in different cellular compartments will be represented by different reaction entities. In consequence, metabolites in the reaction formulas are represented by the metabolite abbreviation and a letter between squared brackets, identifying the compartment. This is necessary, for instance, to represent the transport reaction with the identifier *D_3AIBt* (D-3-Amino-Isobutyrate Transport transport from the cytosol to the extracellular environment):

$$D\_3AIBt : 3aib\_D[c] \rightarrow 3aib\_D[e] \tag{2.3}$$

Associated with each reaction subsystem information, notes added by curators, a confidence score [304], and literature sources is displayed on the reaction detail page. The reaction is also graphically displayed in an atom mapped fashion, and its structure available. KEGG [152, 153], ReconMap [223], and COG [302] identifiers along with the Enzyme Commission number are displayed under “External Links”. Standard reaction Gibbs energy is displayed when available. Finally, from the detail page, a user can also navigate to associated genes, microbes, and diseases.

An example detail page for the reaction Hexokinase 1 is shown in Figure 2.4.

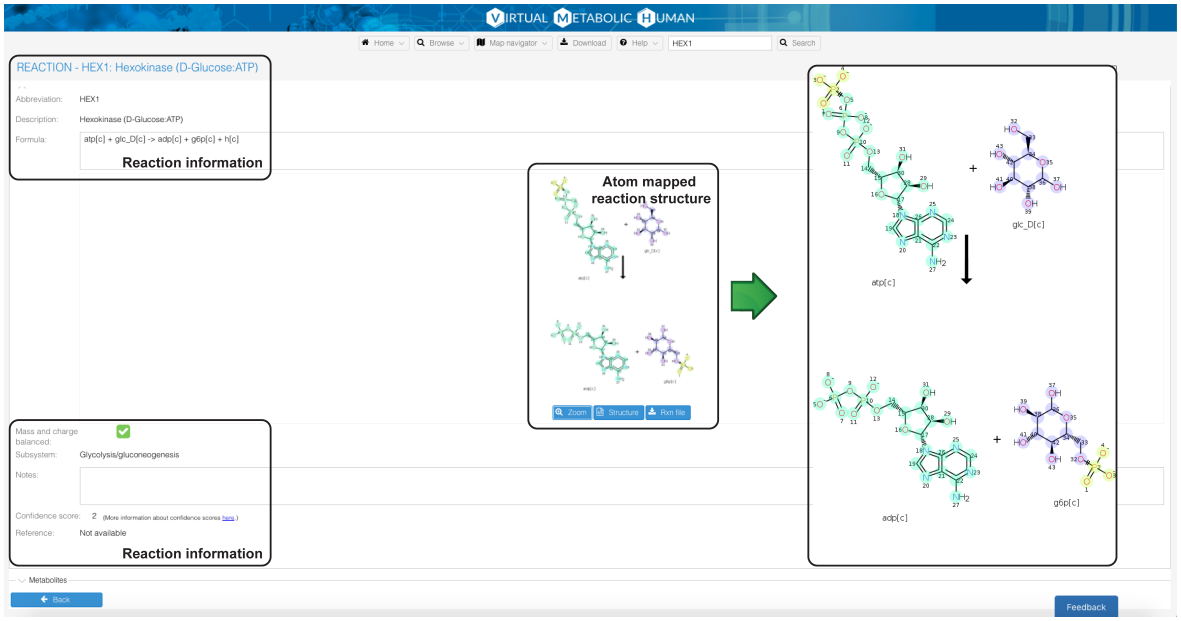

Figure 2.4: Overview of the reaction detail page. The interface contains additional detailed information of the reaction, including a visualization of the atom mapped chemical structure. In addition, this page includes connections to external resources, and to related internal entities (e.g. metabolites present in the reaction).

2.7.3 Gene detail page

Human gene detail page

The gene number used in VMH to identify human genes is a combination of Entrez Gene identifier and the transcript number. This explains why the number of genes that are displayed in the web interface is higher than the number of "unique" genes indicated in this manuscript. Each detail page contains additional information for each gene, external links to several resources, such as Ensembl [62], HGNC [108], ChEMBL [30], Uniprot [59], Entrez Gene [194], OMIM [9, 115], Human Protein Atlas [315], UCSC [314], WikiGene [134], and Gene Ontology [57]. Furthermore, connections with diseases and associated reactions are included in a similar fashion as with the other database entities.

2.7.4 Microbe gene detail page

The microbe gene detail page is considerably simpler than the human gene. Each gene is uniquely associated with one microbe and the detail page displays the sequence and associated reactions.

2.7.5 Microbe detail page

The microbe detail contains information about phylogeny (e.g. kingdom, order, phylum).External resources connect to SEED [72], IMG [199], NCBI [329], and KBASE [14]. In addition, each microbe has associated a set of numerical characteristics extracted from their reconstructions with a visualization of the S-matrix (Figure 2.5). Internal connections display the reaction, metabolite, and gene content. Regarding the curation process, a list of fermentation products and carbon sources and detailed pathway curation status information is also available. Finally, in each microbe page, it is also possible to download the correspondent reconstruction in different formats, as well as the genome in FASTA format.

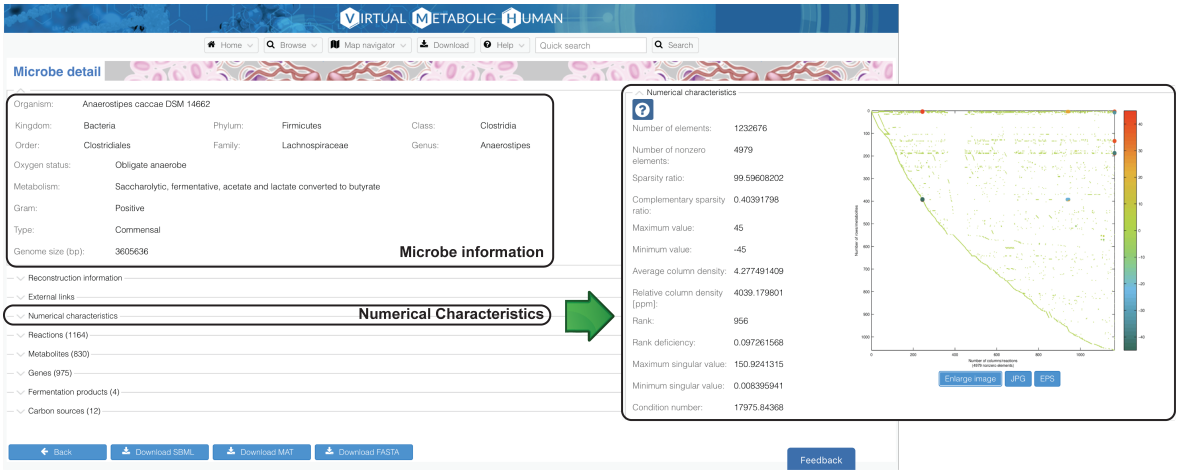

Figure 2.5: Overview of the microbe detail page. The interface contains additional detailed information about the microbe, including numerical characteristics. In addition, this page includes connections to external resources, and the content of microbe metabolic reconstruction.

2.8 Discussion

The VMH captures in a unique manner information for human and gut microbial metabolism and links it to hundreds of diseases and to nutritional data. As such, the VMH addresses an increasing need to enable the fast analysis and interpretation of complex data arising from large-scale biomedical studies. For instance, an increasing number of studies link the microbial composition to diet and disease [53, 345]. However, the generation of novel hypothesis about functional implications of observed correlations, e.g., between microbial

abundances in certain disease states, is slowed by the lack of facilitating, online databases. In particular, the "Diet Designer" tool permits, in conjunction with computational modeling, to test *in silico* causative hypotheses that could then be experimentally tested. The use of synthetic microbial communities is of great value for hypotheses testing and the VMH can facilitate the design of defined microbial communities with specified metabolic capabilities. The VMH provides an environment by making diverse data along the diet-gut-health access available to the biomedical community.

The visualization of complex "omics" data is crucial for their interpretation. Such data can be overlaid with ReconMap [223] as well as of the Parkinson's disease map [93]. As the metabolic elements in these maps are connected to the VMH, the omics data can be put into the larger context of human metabolism. Importantly, the disease map concept is now extended to other important diseases, which can be directly linked to the content of the VMH, rendering it a unique hub for human metabolism in health and disease. Linking the "Disease resource" as well as the maps to clinical phenotypes, expressed in Human Phenotypic Ontology (HPO) terms [170], would allow for the investigation of genotype-phenotype relationships from molecular-level omics data within one knowledge base. An integral part of systems biology is computational modeling, with COBRA modeling gaining increasing attention by a broad scientific community. At the foundation of the VMH lie genome-scale metabolic reconstructions. Thus, VMH serves directly the growing COBRA community and their needs by providing a user-friendly interface to the reconstructions' content, providing the reconstructions in multiple standard formats (e.g., SBML [141]) for download, allowing the access of the entire knowledge base via the API, and enabling the formulation of various *in silico* personalized diets via the "Diet designer", which can then directly be integrated with the human or microbial metabolic reconstructions using the COBRA toolbox [130]. Simulation results based on these diets or based on the integration of "omics" data, e.g., metabolomics [16] or transcriptomics [232], can then be visualized and interpreted in the context of the human metabolic map, ReconMap, or a disease-specific map. We are closely working with the COBRA community to further expand the value of the VMH for biomedical applications based on computational modeling.

VMH integrates a considerable part of the components influencing metabolic homeostasis but there is still a long road ahead. As it is, VMH has little coverage of regulation and epi-

genetics, which are of high importance to completely understand how, for instance, the same diet can differently affect individuals of the same genetic background. There are approaches that combine gene expression and metabolism in the COBRA framework [226, 303], but these models are still scarce and the computational power required to study them is very demanding. The modeling of xenobiotics can be combined with the COBRA methodology by integrating PBPK modeling and adding a time dimension to these simulations. These efforts are still at an initial stage, but more studies combining these techniques are becoming available [111]. For this purpose, physiological data could also be stored in VMH, such as blood flow rate, glomerular filtration rate, cardiac output, hematocrit values, and oxygen uptake for the reference man and woman [231] as well for specific populations, such as infants [22, 294], pregnant women [2, 341], and elderly people [307]. Such "Physiological resource" would expand the value of the VMH for the quantitative pharmacology community, which could link predicted pharmacodynamics properties of drugs to the metabolism of specific populations The effect of drug treatment varies significantly among individuals, and genetic differences alone are insufficient to explain the observed inter-individual differences in drug response [217]. Human gut microbes metabolize many drugs [114, 202]; however, their contribution to an individual's drug response and safety is poorly understood. Diet does not only modulate the microbiota composition and biochemical functions but also alters drug bioavailability [270]. Hence, a valuable expansion of the VMH would be to add a "Drug resource", which would allow investigating FDA approved drugs in the context of the human metabolic reconstruction, as well as of the microbial reconstructions. The corresponding data have been collected for Recon 2 [261] as well as off-target and side effects have been investigated using the human metabolic reconstruction [47]. It would be of great value to connect such resource with the numerous online resources that capture i) drug information, such as DrugBank DB [332], ii) gene-drug interaction data [256], iii) adverse reactions: SIDER database [176], VigiAccess [275], and EudraVigilance [5]; and iv) drug-disease information: DIDB [of Washington]. Moreover, the drug entries could be linked to clinical trials [229, 4]. The inclusion of these data and connections to external knowledge bases would permit the users to exploit the increasing knowledge on the human gut microbiota as well as microbiota- and diet-related interpersonal variability for drug development and clinical trial design.

The integrative nature of VMH, and in particular the addition of nutritional information

in the context of metabolic modeling offers a new perspective in the field and is a first step towards establishing a methodology that will potentiate the understanding of the mechanisms of metabolic homeostasis, and how its disruption can lead to the occurrence of diseases. We hope that the inclusion of these missing factors, such as lifestyle, into the metabolic modeling framework will be facilitated by VMH.

Chapter 3

Design and applications of the Virtual Metabolic Human database

Abstract

Biological databases are important tools in the life sciences and biomedical fields. These are typically accessible through a web-interface and, in some cases, through Application Programming Interfaces (APIs). These APIs, if accessible through the web, allow access of third-party applications to the database content without the constraints of a web browser. In this chapter, we describe the structure of the Virtual Metabolic Human database and its web API. Additionally, we showcase how this tool can be used to perform analysis combining the different resources available. A detailed description of the functionalities of the Virtual Metabolic Human's API is available at `vmh.uni.lu/api/docs`.

3.1 Introduction

As data in the life sciences and biomedical fields become increasingly complex, biological databases gain relevance as they promote knowledge organization and dissemination. Biological databases are typically accessible through web-interfaces. These interfaces have, over the years, become increasingly sophisticated but they are bound to the perspective of the development team that has to predict what kind of usage visitors desire. Additionally, data analysis and visualization is bound to a web-browser and integration with other tools is often limited. Programmatic access to databases, on the other hand, enables third-party applications or user-made scripts to access the content in a more unrestricted fashion. This can be achieved with the use of web services commonly known as Application Programming Interfaces (APIs).

An API is the means by which third-party applications can write code that interfaces with other code. A web service is an API that works across the internet (or a network) using HTTP or other protocols. By implementing these web services biological databases give toolboxes (e.g. CobraToolbox) or generic statistics software (e.g. R, Matlab) direct access to the database content without the bounds of a previously defined interface, granting a higher degree of freedom for data analysis and visualization. Several biological databases implement these type of interfaces, such as ChEBI [118], Enrichr [177], or BIGG [164].

In this chapter, we introduce the architecture and web API of VMH. In an API information is accessed through a series of URL endpoints with parameters that support filtering, search, and pagination of results. All URL patterns and additional parameters are available at `vmh.uni.lu/_api/docs` where end-users can live test the different functionalities available. To finalize, we highlight how these resources can be combined by showcasing several applications of VMH that include studying the complex interaction profiles of different gut microbes, mechanisms of drug detoxification, and potential disease treatment.

3.2 Methods

The VMH database, presented in Chapter 2 can be described as a system of 3 layers. On its foundation, we find a MySQL relational database for information storage. Management

and access to this database are accomplished using a Python framework called Django [89]. Django is a web-framework that allows the development of fast and secure web applications by providing an abstraction layer for structuring and manipulating the data of its applications in the form of "Models". These "Models" are sources of information and typically map to a single database table, each attribute mapping a database field. Building a set of these "Models" in Django will automatically create the corresponding database tables.

To access the database content, VMH presents a Representational state transfer (REST), or RESTful, web service. A RESTful web service allows submitting requests to access and manipulate data using a predefined set of operations and retrieve data in several formats such as XML, HTML, or JSON [83]. To develop this webs ervice, the Django Rest Framework [51] provides powerful and flexible tools combined with the advantages of Django previously described. This REST API allows other software to interact with VMH.

The top layer is the front-end reachable via a web-browser at `http://vmh.life`. The interface was developed in Sencha ExtJS 5.1 [46], a JavaScript application framework that is used to build interactive cross-platform web applications. This framework includes pre-tested and integrated components sparing developers the effort of building a web-interface from scratch. This 3-layer architecture and a simplified database schema are displayed in Figure 3.1.

3.2.1 Database structure

In Figure 3.1 the Database layer is represented by a simplified conceptual schema (detailed database schema in Supplementary Figure A.1) highlighting the main data structures stored in VMH. At the core of the database are genome-scale metabolic reconstructions (GENREs). In a GENRE, the main structure that represents the metabolic network is the Stoichiometric Matrix (S-matrix). In this matrix, each row corresponds to a metabolite and a column to a metabolic or transport reaction.

Metabolites

The metabolite "Model" was defined as shown in Figure 3.2. This "Model" has two unique identifiers: an internal one, automatically generated for internal reference in the database,

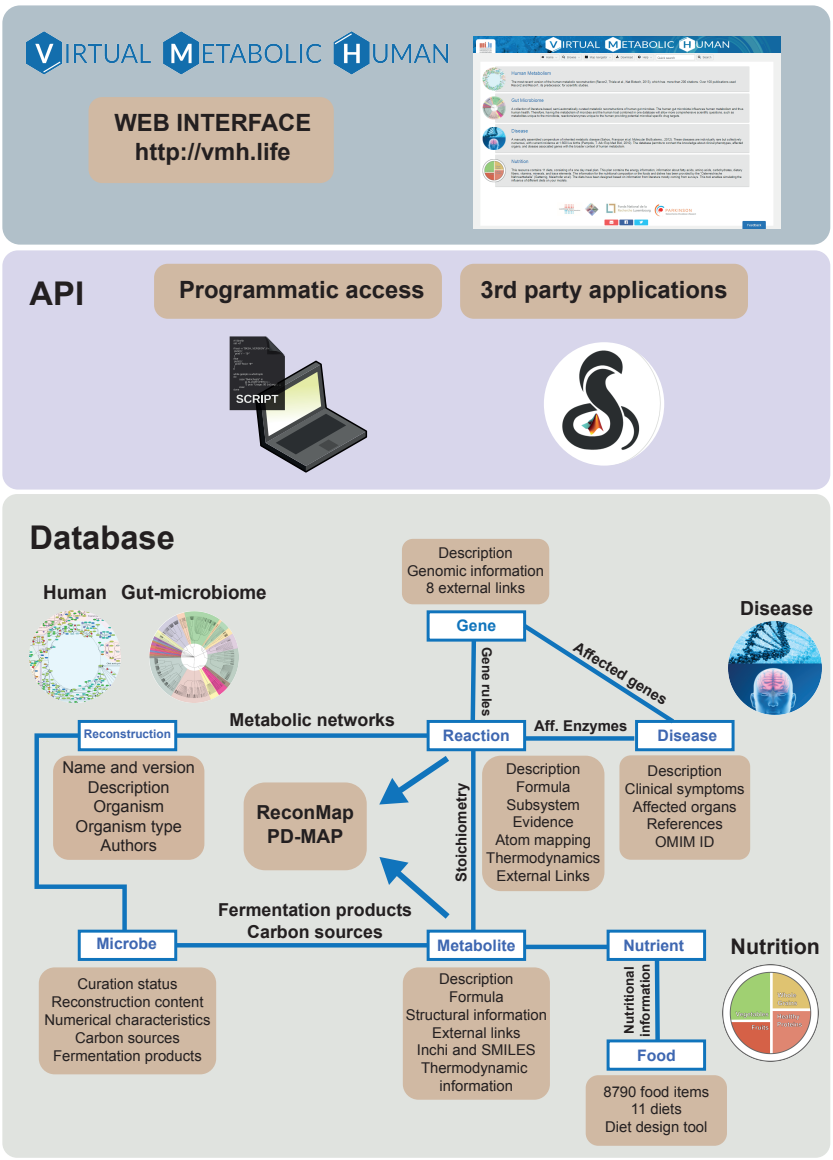

Figure 3.1: Overview of the Virtual Metabolic Human. The resource is divided in two interfaces and its database containing 4 resources. Users can interact with the database using the two available interfaces: (i) a user-friendly web-interface and (ii) an application programming interface that allows the programmatic access to the information contained in the database. At the core of the database is the representation of reconstructions as sets of reactions. The database connects 4 resources through shared nomenclature: (i) the Human metabolism and Gut microbiota resources share metabolites and reactions, (ii) the nutrients in the Nutrition resource are mapped to metabolites that can be shared by the human and gut microbes, and (iii) the diseases in the Disease resource include mutated genes and metabolite biomarkers present in the Human resource.

and an abbreviation which is used in the GENRE. The fields *fullName*, *description*, and *synonyms* describe each metabolite in more detail. Additionally, this "Model" contains a charged formula and an associated charge. In VMH the formulas can represent the acid/base form of the neutral molecule, hence the charge can yield a different formula when compared to other databases. Several links to external resources are also stored but for simplicity, they were omitted from Figure 3.2.

Reactions

An enzyme can catalyze the same reaction in different locations and with different co-factors. In VMH these are considered different entities. For instance, hexokinase can have 3 different cofactors (ATP, UDP, GDP), each form represented by separately. The Reaction "Model" was created as seen in Figure 3.2. This "Model", similarly to the metabolite, possesses two unique identifiers and additional information in the form of a description, formula, notes, and reversibility. In addition, it can contain references, a confidence score, and a mass and charge balance status associated with the reconstruction and curation process [304].

S-Matrix

Following the Entity-relationship model principles [49], the relationship between the previously defined Metabolite and Reaction "Models" is of cardinality many-to-many (A reaction can have *many* metabolites and a metabolite can be present in *many* reactions). In a relational database system such as the VMH database, these relationships are typically implemented using associative tables.

In a GENRE, this relationship is defined by the S-matrix. Each cell of the S-matrix contains the stoichiometric value of a given metabolite in a reaction. In this representation, metabolites with negative values are the reactants and positive values the products of biochemical transformations present in the GENRE. A metabolite can occur in different reactions and each instance of a metabolite in different cellular locations will have a corresponding row. Each of these cellular locations, or compartments, is represented by a letter code (e.g. 'x' for peroxisome) and associated with the metabolite identifier (e.g h2o[x]).

```
class Metabolite(models.Model):
    met_id = models.AutoField(primary_key=True,
    ↪  db_column="met_id")
    abbreviation = models.CharField(max_length=250,
    ↪  db_index=True)
    fullName = models.TextField(null=True)
    description = models.TextField(null=True)
    synonyms = models.TextField(null=True)
    chargedFormula = models.CharField(max_length=250)
    charge = models.IntegerField()
    avgmolweight = models.FloatField(null=True)
    monoisotopicweight = models.FloatField(null=True)
```

```
class Reaction(models.Model):
    rxn_id = models.AutoField(primary_key=True,
    ↪  db_column="rxn_id", max_length=50)
    abbreviation = models.CharField(max_length=250,
    ↪  db_index=True)
    description = models.TextField(null=True)
    formula = models.TextField(null=True)
    reversible = models.IntegerField()
    mcs = models.IntegerField(null=True)
    notes = models.TextField(null=True)
    ref = models.TextField(null=True)
    massbalance = models.IntegerField()
```

```
class Smatrix(models.Model):
    id = models.AutoField(primary_key=True, db_column="id")
    rxn = models.ForeignKey(Reaction, related_name="smatrix",
    ↪  db_column='rxn_id')
    met = models.ForeignKey(Metabolite, related_name="smatrix",
    ↪  db_column='met_id')
    value = models.FloatField()
    comp = models.CharField(max_length=5)
```

Figure 3.2: Metabolite, Reaction, and Smatrix models in Django.

To define this relationship a "Model" called "Smatrix" was created (Figure 3.2). Each row of the "Smatrix" table represents one cell of the S-matrix. To retrieve all metabolites involved in a reaction, one needs to retrieve all rows with a specific reaction identifier and all rows with the same metabolite value will result in all reactions that metabolite takes part.

Genes

There are two types of genes in the VMH database: human and microbe. Both types are associated with reconstructions through associative models (*ReconToGene* and *ReconToMicrobeGene*). In VMH, genes can be specific to a reconstruction and therefore, the gene-protein-rules (GPRs) as well. For this reason, genes are not associated with the Reaction "Model" directly but rather with the entity that connects the Reaction and Reconstruction "Models".

Reconstructions

In VMH, a GENRE is represented by a list of reactions. The reconstruction "Model" is summarized in Figure 3.3. The connection of a reaction with a reconstruction is not merely associative and needs to contain specific information such as, a gene-protein-rule, dependent on the organism. The reaction nomenclature is shared across reconstructions but the genes and reaction bounds are treated individually. This association between a reaction and a reconstruction is represented by a model called Recon as shown in Figure 3.3 (not to confuse with the global reconstruction of human metabolism Recon).

In order to accommodate different types reconstructions, the fields organism and organism type were added. However, there is additional information about each organism that might be of interest. At the same time, this information can vary significantly depending on the organism type, becoming desirable to store it in a resource-specific table. While for human metabolism it is convenient to store information for specific reconstructions that are subsets of Recon, such as the organ reconstructions, for the Gut Microbiota resource, species-specific information might be useful. For that purpose, we have created two "Models": Organ and Microbe each mapping to their own tables (Figure 3.3). Each of these contains a reconstruction in a one-to-one association to guarantee that each reconstruction is associated with a single organism.

```
class Recon(models.Model):
    id = models.AutoField(primary_key=True)
    reconstruction = models.ForeignKey(Reconstruction)
    rxn = models.ForeignKey(Reaction)
    lb = models.FloatField()
    ub = models.FloatField()
    cs = models.IntegerField(null=True)
    gpr = models.TextField(null=True)
    subsystem = models.TextField(null=True)
    ref = models.TextField(null=True)
```

```
class Reconstruction(models.Model):
    model_id = models.AutoField(primary_key=True)
    name = models.CharField(max_length=100, db_index=True)
    organism = models.CharField(max_length=250)
    organismtype = models.CharField(max_length=100, null=True)
    author = models.CharField(max_length=150, null=True)
    version = models.CharField(max_length=25, null=True)
```

```
class Microbe(models.Model):
    id = models.AutoField(primary_key=True)
    reconstruction = models.OneToOneField(Reconstruction)
    organism = models.CharField(max_length=250, null=True)
    kingdom = models.CharField(max_length=250, null=True)
    phylum = models.CharField(max_length=250, null=True)
    mclass = models.CharField(max_length=250, null=True)
    order = models.CharField(max_length=250, null=True)
    family = models.CharField(max_length=250, null=True)
    genus = models.CharField(max_length=250, null=True)
    mtype = models.CharField(max_length=250, null=True)
    species = models.CharField(max_length=250, null=True)
```

```
class Organ(models.Model):
    id = models.AutoField(primary_key=True)
    reconstruction = models.OneToOneField(Reconstruction)
    organ_abbreviation = models.CharField(max_length=150)
    name = models.CharField(max_length=150)
    descname = models.CharField(max_length=250)
    description = models.TextField(null=True)
```

Figure 3.3: Recon, Reconstruction, Microbe, and Organ "Models" in Django.

Nutrition

The Nutrition resource in the database is composed of a dataset of food items extracted from the USDA Nutrient Database for Standard Reference [317]. Food information is represented by the Food model which translates into the database Food table. Each food item contains information fields that include the data source from where the nutritional data was collected to accommodate future inclusions of additional food composition databases. The nutritional data is stored using the Nutrient "Model" (Figure 3.4), that stores nutrient definitions and mapping to Metabolite entities. This connection is what enables the calculation of fluxes from diets. The amounts of nutrient per Food are, however, stored in a different "Model" called NutritionData which solves the many-to-many relationship cardinality between the Nutrient and Food "Models".

Disease

The Disease resource of the VMH stores data about more than 200 diseases. This data is stored via the Disease "Model" and corresponding table. The "Model" contains fields that store information such as the mode of inheritance, prevalence and typical phenotypes. In addition, there are 3 connections to other models. The mutated genes connect the Gene "Model". This information is curated [260] and only genes existing in VMH are stored (this means that it is possible that the disease affects more genes than indicated). The reactions associated with these genes through the GPRs are then stored in the *rxns* field. Finally, known biomarkers existing in Recon were also mapped and stored in the metabolites field. through the Biomarker "Model". A detailed description of the Disease and Biomarker "Models" is available in Figure 3.4

```python
class Nutrient(models.Model):
    id = models.AutoField(primary_key=True)
    nut_no = models.CharField(max_length=10, unique=True)
    unit = models.CharField(max_length=10)
    tag_name = models.CharField(max_length=50, null=True)
    description = models.CharField(max_length=100)
    common_name = models.CharField(max_length=100)
    category = models.CharField(max_length=100, null=True)
    subcategory = models.CharField(max_length=100, null=True)
    mets = models.ForgeignKey(Metabolite)
```

```python
class Disease(models.Model):
    id = models.AutoField(primary_key=True)
    abbreviation = models.CharField(null=True, max_length=45)
    name = models.CharField(max_length=250)
    dtype = models.CharField(max_length=150, null=True)
    subtype = models.CharField(max_length=150, null=True)
    inheritance = models.CharField(max_length=150, null=True)
    omim = models.CharField(max_length=50, null=True)
    phenotype = models.TextField(null=True)
    prevalence = models.TextField(null=True)
    organ = models.CharField(max_length=250, null=True)
    references = models.TextField(null=True)
    ghr = models.CharField(max_length=250, null=True)
    orphanet = models.CharField(max_length=250, null=True)
    genes = models.ManyToManyField(Gene)
    rxns = models.ManyToManyField(Reaction)
    metabolites = models.ManyToManyField(Metabolite,
    ↪  through="Biomarker")
```

```python
class Biomarker(models.Model):
    metabolite = models.ForeignKey(Metabolite)
    disease = models.ForeignKey(Disease)
    name = models.CharField(null=True, max_length=250)
    value = models.CharField(max_length=25)
    normalConcentration = models.CharField(null=True)
    rangeConcentration = models.CharField(null=True)
    reference = models.TextField(null=True)
    ramedis = models.TextField(null=True)
```

Figure 3.4: Disease and Biomarker models in Django.

3.2.2 RESTful API

The VMH API can be reached at `http://vmh.uni.lu/_api`. This page displays some of the available resources that can be used to retrieve data. Each of these is reachable through an Uniform Resource Identifier (URI) which provides data in different formats, such as HTML, JSON or text format (CSV). As an example, the URI *'metabolites'* returns the list of metabolites in the database. For each of these identifiers, additional filters can be applied which allow to further refine the search. In the first snippet of code snippet of code in Figure 3.5 a filter to the metabolite abbreviation field is used, so the response will only retrieve the metabolite with the abbreviation value of *h2o*. The additional parameter, *format*, specifies that the response should be in JavaScript Object Notation (JSON) format. Alternatively, it is possible to use an interface from a programming language to interact with the web API. In this Chapter, we provide examples using Python and the Core API (`http://www.coreapi.org/`) Python implementation, a format-independent Document Object Model that allows interaction with web APIs in a robust and meaningful way. It allows the integration into applications and avoiding the need of constructing specific HTML requests and decoding the server responses.

```
curl -X GET http://vmh-internal.uni.lux/_api/metabolites/
↪  ?abbreviation=h2o&format=json
```

```
import coreapi

# Initialize a client & load the schema document
client = coreapi.Client()
schema = client.get("http://vmh.uni.lu/_api/docs")

# Interact with the API endpoint
action = ["metabolites", "list"]
params = {"abbreviation": "h2o"}
result = client.action(schema, action, params=params)
```

Figure 3.5: Two examples of how to fetch a specific metabolite from the VMH Web API. The first using the CURL command from a shell environment, while the second uses the Python package Core API.

The detailed description of all available calls and parameters are available at `http:`

//vmh.uni.lu/_api/docs. In this interface, users can directly interact with the API. In the next sections, we will demonstrate how the API can be used to perform complex queries to the various resources available in VMH.

Reconstruction information

Exchange reactions represent the ability of an organism or cell to interact with the external environment. In VMH, the abbreviations of the exchange reactions follow the same naming pattern. The code to retrieve the list of exchange reactions present in any microbe reconstruction is given in Figure 3.6.

In the first snippet of Figure 3.6, two parameters are used: the *organismtype* which, if specified, filters the resulting content to specific types of organisms from 3 available: "Human" (for the general reconstruction), "Human organ", or "Microbe"; the second filter was applied to the abbreviation using the keyword *icontains* to search for abbreviations that contain the specific pattern 'EX_' in a case-insensitive way.

```
# list of reactions
action = ["reactions", "list"]
# Two parameters: is in a microbe and abbreviation contains 'EX_'
params = {"organismtype":"microbe",
 ↪  "abbreviation__icontains":"EX_"}
results = client.action(schema, action, params=params)
```

```
>>> results.get("count")
>>> 393
```

```
# list of reactions
action = ["microbes", "list"]
params = {"phylum":"Bacteroidetes"}
results = client.action(schema, action, params=params)
```

```
>>> results["count"]
>>> 112
```

Figure 3.6: VMH API interactions. The first snippet retrieves all exchange reactions in microbes. The second snippet of code retrieves all microbes of the Bacteroidetes phylum

Analyzing exchange reactions allows understanding if an organism or cell has the potential to consume or produce given compounds. This is interesting, for instance, for microbial organisms. In this context, VMH also contains a curated list of carbon sources and fermentation products from different experiments. This can be a useful screening tool for the design of synthetic microbiota communities as we will show later.

For microbial species, it is also possible to filter according to phylogenetic information using the *microbe* end-point of the API, which allows filtering microbes by many different parameters, such as phylum, class, or kingdom as shown in the last snippet of code of Figure 3.6.

For pathway-associated information, users can take advantage of the subsystem information in the *Recon* model described before. For this, one needs to then be able to fetch all reactions present in a specific reconstruction. Using the code from the second snippet of Figure 3.6 will retrieve 112 Bacteroidetes and in each of them, the reconstruction name is available. With this value, it is possible to retrieve all associations between reactions and that reconstruction through the 'rxntomodel' endpoint as shown in Figure 3.7. Each of these associations contains the subsystem information which is a group of reactions involved in a specific pathway. This allows comparing reconstruction content in regards to pathway content.

Nutrition information

The nutrition resource in the VMH database is composed of three main resources: the food composition database, a set of pre-designed diets, and the diet designer tool. With the diet designer, users can create specific diets and download flux values to be integrated into metabolic simulations in tools such as the CobraToolbox [130]. This feat can also be achieved through the API making it possible to integrate this feature into other software tools. To calculate fluxes from nutritional information, several API calls are necessary:

- *foods*: retrieves the list of foods
- *nutrients*: retrieves the list of nutrients
- *nutritiondata*: retrieves the amount per 100 grams of a nutrient in a specific foods

```
# previous result
print "Organism: " + results.get("results")[0].get("organism))
one_reconstruction =
 ↪  results.get("results")[0].get("reconstruction"]])

# with reconstruction value get all reactions
action = ["rxntomodel", "list"]
params =
 ↪  {"model":results.get("results")[0].get("reconstruction")}
recon_results = client.action(schema, action, params=params)
print "Total reactions in reconstruction "+
 ↪  results.get("results")[0].get("reconstruction") + ": " +
 ↪  str(recon_results.get("count"))
```

```
>>> 'Bacteroides caccae ATCC 43185'
>>> 'Total reactions in reconstruction
 ↪  Bacteroides_caccae_ATCC_43185: 1225'
```

Figure 3.7: API interaction that retrieves all reactions in the reconstruction of Bacteroides caccae ATCC 43185.

- *mmass*: contains the molecular mass value of metabolites and associated exchange reaction

To generate a flux value for a selected food item and a portion in grams, it is necessary to get the nutritional data for each nutrient present in that food. Nutrients in VMH can have associated metabolites and with the respective molecular mass, it is possible to convert the portion weight in grams to milimol. This value is then associated with an exchange reaction. The corresponding example code for such a task is shown in Figure 3.8.

```
foodName = "Apples, raw, with skin"
portion = 200
fluxes = dict()
action = ["foods", "list"]
params = {"name": "Apples, raw, with skin"}
result = client.action(schema, action, params=params)
food_code = result.get('results')[0].get('food_id')

# With the food_code get all nutritional data on food
action = ["nutritiondata", "list"]
params = {"food": food_code}
foodNutData = client.action(schema, action, params=params)

# For each nutrient calculate the fluxes if they have metabolites
↪  associated
for nutrientData in foodNutData.get('results'):
        nutrient = nutrientData.get('nutrient')
        amountInFood = nutrientData.get('nutr_value')
        for met in nutrient.get('mets'):
                # get molecular mass for that nutrient
                action = ["mmass", "list"]
                params = {"metabolite": met}
                mass = client.action(schema, action,
                ↪  params=params).get('results')[0]
                exchangeReaction = mass.get('reaction')
                massValue = mass.get('molecularmass')
                unitfactor = 1
                if nutrient.get('unit') == 'mg':
                        unitfactor = 1000
                elif nutrient.get('unit') == 'microg':
                        unitfactor = 1000000

                amountInGrams = ((amountInFood / unitfactor) /
                ↪  100) * portion
        flux = (amountInGrams / massValue) * 1000;
        fluxes[exchangeReaction] = flux
```

Figure 3.8: API interaction that converts the nutritional information of a food item into flux values.

3.2.3 Pagination

All query results from the VMH API are paginated. This ensures that operations run smoothly and avoid high connection load. Each response from the server contains 50 results per page

but the total amount of results is displayed with URIs for the next and previous page. It is possible to modify the page size of the response, but we advise users to moderate the use of high page number definitions for a smooth experience and adapt the scripts to accommodate for multiple page requests. Iterating over response pages can be done by checking if a next page exists or by calculating the total number of pages by dividing the *count* parameter by the page size. The code previously presented in Figure 3.7 retrieves all 112 Bacteroidetes. Results of this call are, therefore, divided into 3 different pages. To access each of them, users need to add the parameter *page* to the API call as shown in Figure 3.9.

```
# list of reactions
action = ["microbes", "list"]
params = {"phylum":"Bacteroidetes"}

# page counter
page = 1
results = client.action(schema, action, params=params)
print "Page " + str(page) + " has " +
 ↪  str(len(results.get("results"))) + " entries."
while results.get("next"):
    page = page + 1
    params = {"phylum":"Bacteroidetes", "page":page}
    results = client.action(schema, action, params = params)
    print "Page " + str(page) + " has " +
     ↪  str(len(results.get("results"))) + " entries."
```

```
>>> Page 1 has 50 entries.
>>> Page 2 has 50 entries.
>>> Page 3 has 12 entries.
```

Figure 3.9: API interaction that retrieves all microbes of the Bacteroidetes phylum and iterates through all result pages.

3.3 Results

VMH grants users a unique view of the metabolism of the gut microbiota and the host. In this section we will showcase some of the applications enabled by the tools described in the last two chapters.

3.3.1 Exploring the complex interactions between microbes, nutrition, and host metabolism

Microbial metabolic interactions represent driving forces for the microbial community composition as well as emergent metabolic properties, e.g., short chain fatty acids production, which serve as energy source for the human body and play an important immunomodulating role [259]. Using VMH, we can systematically query for shared metabolite exchanges between human, microbes, and the nutrition resource. The portfolio of exchange reactions, which define metabolites that an organism can exchange with its environment, gives us an organism-specific "interaction profile" and comparing these profiles provides a better understanding of the roles that specific organisms can play in complex systems, such as the human gut. All 773 gut microbes share a common set of 16 exchange reactions, whereas each microbe has an average of 129 +/- 26 exchange reactions. When comparing the presence of exchange reactions across these 773 gut microbes, using tSNE for visualization of the high dimensionality data [193], we find a clear separation of the 112 representatives belonging to the phyla Bacteroidetes (Figure 3.10-A), indicating that these microbes share a unique set of exchange reactions, which is distinct from the other phyla.

In contrast, the other phyla overlap in their exchange reactions and thus in their interaction potential. For instance, the 356 *Firmicutes* representatives are broadly distributed, overlap with *Actinobacteria*, but are well separated from the *Bacteroidetes*. We then compared the complete metabolic repertoire, i.e., all metabolites and reactions, of the *Bacteroidetes* and *Firmicutes* representatives. As expected, most of the repertoire is shared between the phyla (Figure 3.10-B). Of the 194 metabolites *Firmicutes* does not share with *Bacteroidetes*, 38 have corresponding exchange reactions, while 51 of the 64 unique metabolites of *Bacteroidetes* can be exchanged. In VMH, it is possible to retrieve information about the pathways these specific metabolites are involved in, through the "Subsystem" attribute of the reactions involving those metabolites. *Firmicutes* specific metabolites are involved in 15 subsystems, which are mainly associated with the metabolism of amino acids, while *Bacteroidetes* display a unique ability to degrade plant polysaccharides and proteoglycans, components of cell walls. Additionally, *Bacteroidetes* are also the phyla that, in general, overlap the least with the 100 metabolites defined in the nutrition resource, indicating that it might be more dependent on

the availability of a smaller number of dietary compounds than the rest of the compared phyla (Figure 3.10-C). At the same time, it also highlights that plant polysaccharide and proteoglycan content in foodstuff is currently not reported in nutritional databases, such as the USDA, on which we are basing our nutrition database. It is noteworthy that more recent efforts focusing on metabolomically and comprehensively characterizing foodstuff, such as FooDB. Overall, we find that the gut microbes overlap in average with 42 +/- 5 of the 100 defined food components, and these overlapping metabolites belong mostly to the subsystems (excluding transports) Fatty acid oxidation, Cholesterol metabolism, and Fatty acid synthesis (Figure 3.10-C). For the overlap with the Human resource, we find an average of 90 +/- 16 interactions predominately in the Nucleotide interconversion, Glycerophospholipid metabolism, and Methionine and cysteine metabolism (Figure 3.10-D).The gut microbiota is characterized by functional redundancy, i.e., the same functions can be performed by multiple bacteria that may be either closely or distantly related [212]. This redundancy also extends to diet-metabolizing genes in multiple species across phyla. Hence, a microbiota-wide systematic approach to exploiting and characterizing the capabilities of the gut microbiota to modulate dietary and host metabolism is enabled with the VMH.

3.3.2 Designing synthetic microbial communities with VMH

Synthetic microbial communities are commonly used to test biomedically-relevant hypotheses and to gain novel insight into microbial ecology. Various gut microbial communities have been developed, capturing key properties of more complex communities [29, 70]. VMH can be used to design synthetic gut microbial communities with, e.g., a particular glycan-degradation profile. Recently, Desai et al. [70] have demonstrated that four of 13 species that they included into a synthetic gut microbiota composed for their capability could grow on mucus-O-glycans. The exchange profiles of the corresponding four microbial reconstructions were in agreement with the experimental data for almost all glycans (Figure 3.11-A). Only the metabolic reconstruction of *Barnesiella intestinihominis* is missing exchange reactions, and metabolic degradation reactions, for mucus O-glycans but contains an exchange reaction for cellobiose while no growth in vivo was found. This comparison highlights the importance of the manual, literature-based curation effort that had been undertaken for our microbial

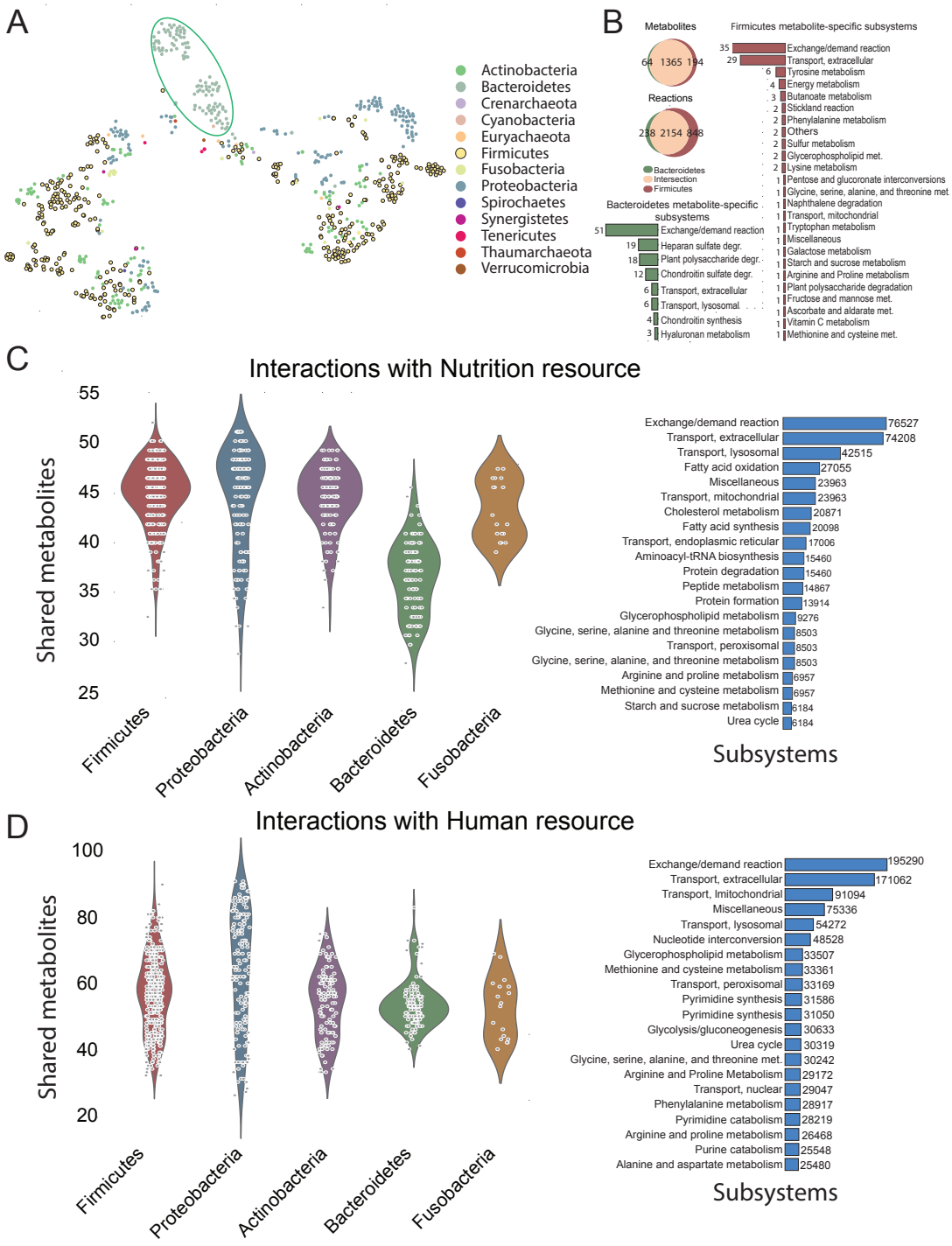

Figure 3.10: Using the gut microbiome resource of VMH to compare and analyze the capabilities of different gut bacteria. A - tSNE of the interaction profile of the microbes in VMH. Only exchange reactions were considered, as these represent potential interactions; B - Reaction and metabolite content comparison of the two most abundant phyla in VMH: *Bacteroidetes* and *Firmicutes*; C - Comparison of interactions between phyla and the Human resource (Recon 3D) and the Nutrition resource.

reconstructions but also how new experimental data can further refine them. We identified 14 further gut microbes, mostly belonging to *Bacteroides* and *Bifidobacterium*, with an overlapping mucus-O-glycans utilization profile as the four microbes (Figure 3.11-B). From the VMH, we can retrieve an extended glycan and polysaccharide utilization profile, which could be used to broaden the carbon source utilization capabilities of the synthetic microbiota. This example illustrates that VMH enables researchers to analyze the in silico potential of different microbes and supports experiment design, taking advantage of the collection of literature curated "Fermentation Products" and "Carbon Sources" available.

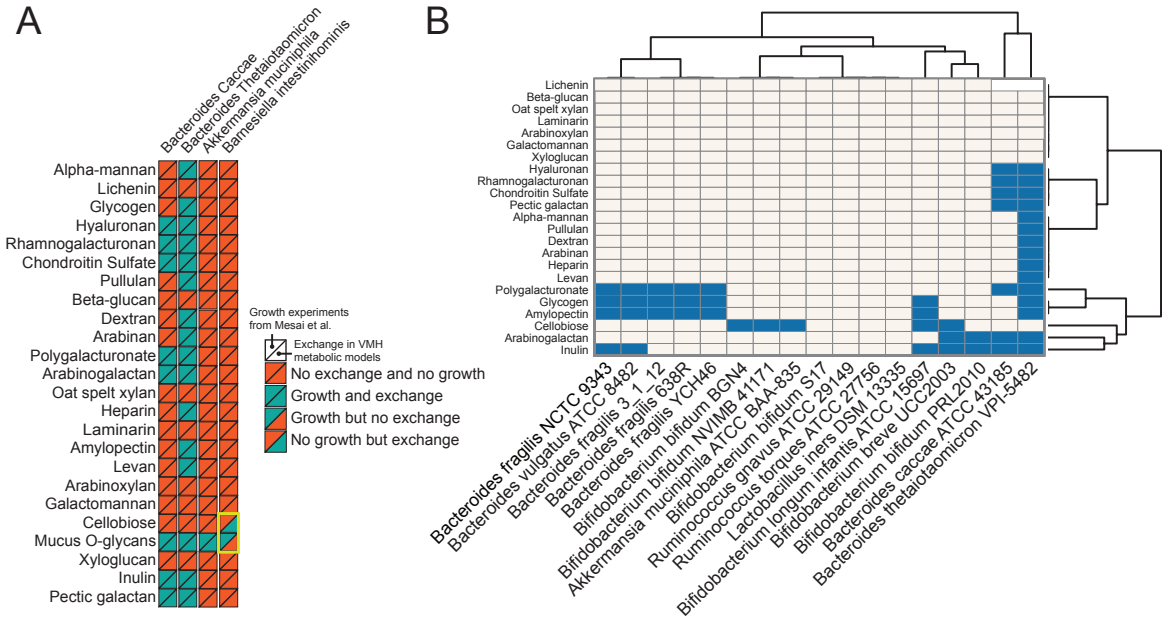

Figure 3.11: A - Comparison between AGORA models and experimental results; the existence of exchange reaction (ability to uptake a given compound) was compared against single carbon source growth experiments (Desai et al., 2016). Full concordance was found, except with *Barnesiella intestinihominis*. B – Other microbes in VMH displaying the ability to uptake Mucus O-Glycans, showcasing how the resource can be used for designing experiments of synthetic microbiotas.

3.3.3 Drug detoxification and retoxification

Xenobiotic metabolism often involves the process of glucuronidation of drugs, which is an important mechanism for drug detoxification and subsequent elimination through bile or urine [264]. UDP-glucuronic acid (VMH ID: udpglcur), formed in the liver, is an essential intermediate in this process (Figure 3.12-A). It has been shown, in rats, that its availability

can be rate limiting for the elimination of exogenous and endogenous toxins [138]. Within VMH, 18 genes encode for enzymes carrying out the liver glucuronidation of 37 endo- and exogenous metabolites, including 18 drugs [261]. To identify potential dietary intervention strategies to alleviate UDP-glucuronic acid limitation, we use VMH to investigate the different metabolic routes by which UDP-glucuronic acid is synthesized and identify how UDP-glucuronic acid availability may be increased, e.g., through targeted dietary supplementation. UDP-glucuronic acid is synthesized from UDP-glucose (VMH: udpg) by the reaction of the UDPglucose 6-dehydrogenase (VMH ID: UDPGD), which in turn is synthesized by the UTP-glucose-1-phosphate uridylyltransferase (VMH reaction: GALU) from glucose-1-phosphate (VMH ID: g1p) and UDP-glucose (Figure 3.12-A). The next step is to investigate all sources of glucose-1-phosphate in VMH, which leads us, with the assistance of ReconMap (Figure 3.12-B), to the pathways "Gluconeogenesis" and "Glycogenolysis". At least in rats, it has been shown that UDP-glucuronic acid for glucuronidation is predominantly derived from glycogen [21]. Accordingly, a high dosage of acetaminophen can deplete the liver glycogen storage [138]. The importance of the Gluconeogenesis for glycogen storage is highlighted by the fact that 2 out of 14 Mendelian glycogen storage diseases, listed in VMH, are due to defects in enzymes along this pathway. Additionally, liver glycogen storage can be effectively replenished by carbohydrates, such as glucose and fructose, after exercise [56]. Interestingly, it has been found that maltodextrin (MD) drinks containing galactose or fructose were double as effective then MD drinks rich in glucose to restore on postexercise liver glycogen synthesis [68]. However, maltodextrin has a higher glycemic index than sugar and it can impair intestinal anti-bacterial responses and defense mechanisms [221], e.g., by increasing the survivability of Salmonella [220]. Since the absorption of fructose is facilitated when ingested in combination with glucose [311], we searched VMH for foodstuffs that are high in fructose and glucose (Table 1). Naturally occurring foodstuffs include honey, medjool dates, and raisins (Table 1-A). The content of galactose is considerably lower in most food items but honey and Greek yogurt are among the best choices (Table 1-B). Thus, it it is possible to use naturally occurring foodstuffs to replenish glycogen stores by providing the necessary glycogen precursors for the gluconeogenesis. Once glucuronidated, drug derivates are excreted either via urine or the enterohepatic route. In the latter case, the glucuronidated drug, such as the cancer drug irinotecan, can be retoxified through the action of microbial beta- glucuronidase [293, 301].

We can investigate how many gut microbes could use the product of the beta-glucuronidases catalyzed reaction glucoronate (VMH ID: glcur) as a carbon source. For 35 out of the 733 gut microbes, glucoronate has been reported to be a carbon source (Vos et al., 2010), most of which belong to *Proteobacteria* (16 species), *Bacteroidetes* (14 species), and *Actinobacteria* (3 species). Additionally, 114 microbes encode for genes to transport glucoronate in and out of the cell via proton symport (VMH ID: GLCURt2r) in their metabolic reconstructions. A total of 256 microbes encode for the glucuronate isomerase converting glucoronate into Fructuronate (VMH ID: fruur). Thus, there are potentially 256 of the 773 gut microbial strains that could use glucoronate as a carbon source. However, a preliminary analysis of the 773 gut microbial genomes suggests that only 13 of those genomes encode for the beta-glucuronidase. These examples demonstrate how VMH can provide a novel, multi-faceted view to human drug metabolism, and its nutritional and microbial aspects.

A - Fructose-rich foods		**Values in g per 100g**		
Food	**Manufactor**	**Fructose**	**Glucose**	**Total Sugar**
Sweetener, syrup, agave		55.6	12.43	68.03
Agave, dried (Southwest)		42.83	3.48	68.03
Honey		40.94	35.75	82.12
Dates, medjool		31.95	33.68	66.47
Raisins, seedless		29.68	27.75	59.19
Cranberries, dried, sweetened		26.96	29.69	72.56
Figs, dried, uncooked		22.93	24.79	47.92
Figs, dried, uncooked		22.93	24.79	47.92
Lemonade-flavor drink, powder		22.73	2.26	97.15
Jujube, Chinese, fresh, dried		20.62	18.28	0
Lemonade, frozen concentrate, pink		20.06	18.57	46.46
Dates, deglet noor		19.56	19.87	63.35

Lemonade, frozen concentrate, white		17.99	16.3	44.46
Agave, cooked (Southwest)		17.57	1.58	20.87
Sauce, barbecue, SWEET BABY RAY'S, original	Sweet baby Ray's, Inc.	17.52	20.85	38.37
Beverages, Lemonade, powder		17.5	2.75	94.7
Formulated bar, POWER BAR, chocolate		15.96	11.94	30.07
McDONALD'S, Sweet 'N Sour Sauce	McDonald's Corporation	15.63	18.76	35.79
McDONALD'S, Barbeque Sauce	McDonald's Corporation	15.44	18.27	34.31
Sauce, barbecue, KRAFT, original	Kraft Foods, Inc.	14.58	16.65	32.26
Sauce, barbecue		14.17	16.39	33.24
B - Galactose-rich foods		**Values in g per 100g**		
Food	**Manufactor**	**Galactose**	**Glucose**	**Total Sugar**
Formulated bar, SLIM-FAST OPTIMA meal bar, milk chocolate peanut	Slim-Fast Foods Company	5.62	1.24	25
Honey		3.1	35.75	82.12
Dulce de Leche		1.03	1.7	49.74
Celery, cooked, boiled, drained, without salt		0.85	0.71	2.37
Celery, cooked, boiled, drained, with salt		0.85	0.71	2.37
Beets, canned, regular pack, solids and liquids		0.8	0.28	6.53

Yogurt, Greek, nonfat, vanilla, CHOBANI	Chobani	0.68	0.3	7.61
Yogurt, Greek, vanilla, nonfat		0.6	0.32	9.54
Yogurt, Greek, vanilla, lowfat		0.6	0.32	9.54
Cherries, sweet, raw		0.59	6.59	12.82
Yogurt, Greek, nonfat, strawberry, DANNON OIKOS	Danone	0.56	0.25	11.63
Yogurt, Greek, nonfat, strawberry, CHOBANI	Chobani	0.55	0.77	10.86
Yogurt, Greek, strawberry, nonfat		0.55	0.65	11.27
Yogurt, Greek, strawberry, DANNON OIKOS	Danone	0.54	0.3	11
Yogurt, Greek, nonfat, vanilla, DANNON OIKOS	Danone	0.54	0.27	11.4
Yogurt, Greek, strawberry, lowfat		0.53	0.54	11.23
Celery, raw		0.48	0.4	1.34
T.G.I. FRIDAY'S, fried mozzarella	T.G.I Friday's	0.4	0.5	1.45
Spices, onion powder		0.36	0.73	6.63
Corn, sweet, white, canned, whole kernel, drained solids		0.36	0.83	2.42

Table 3.1: Foodstuff in VMH with the highest concentration of fructose and galactose. Source of food nutritional information: US Department of Agriculture, Agricultural Research Service, Nutrient Data Laboratory. USDA National Nutrient Database for Standard Reference, Release 28. Version Current: September 2015.

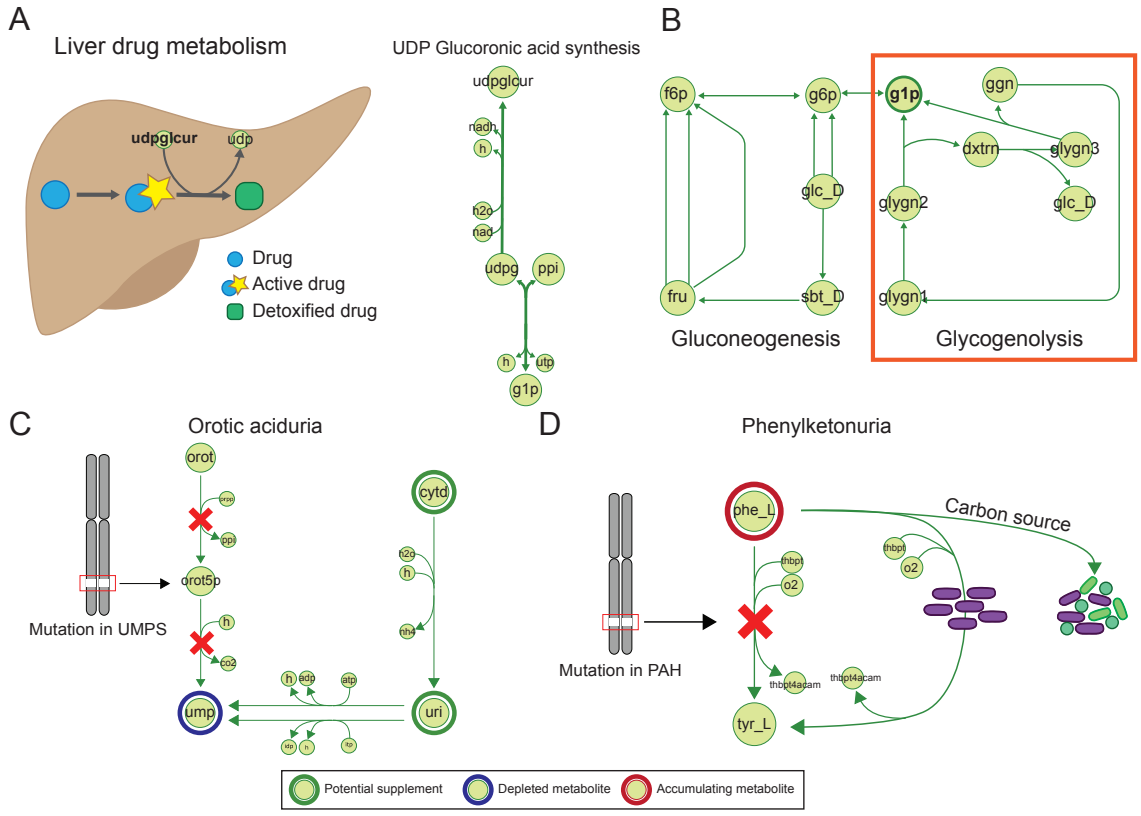

Figure 3.12: Using VMH to investigate mechanisms of disease and drug metabolism. A - UDP-glucuronic acid (VMH metabolite: udpglcur), formed in the liver, is an essential intermediate in the glucuronidation of drugs. UDPglucose 6-dehydrogenase (VMH reaction: UDPGD) converts UDP-glucose (VMH metabolite: udpg) to UDP-glucoronic acid, and UTP-glucose-1-phosphate uridylyltransferase (VMH reaction: GALU) converts glucose-1-phosphate (g1p) to UPD-glucose. B - Sources of g1p found in ReconMap; it was shown in rats that glycogenolysis is the source of UDP-glucoronic acid in the process of glucoronidation (Bánhegyi et al., 1988) C - Mechanism of Orotic Aciduria: mutation in UMPS affects reactions that transform orotic acid into uridine monophosphate. Mechanisms found in VMH point to known treatment with the use of uridine and cytidine; D - Phenylketonuria mechanism: mutation of PAH causes incapability of degrading phenylalanine. Some gut microbes show the capability to degrade phenylalanine or use it as carbon source. A treatment strategy might involve gut microbiome community engineering.

3.3.4 Probiotic approaches to rare disease treatment

Orotic Aciduria (OMIM 258900) is an autosomal recessive disorder caused by a mutation in the uridine monophosphate synthetase gene (EntrezGene ID: 7372). In VMH, this gene associated with two reactions (VMH ID: ORPT and OMPDC) that transform orotic acid (VMH ID: orot) into uridine monophosphate (VMH ID: ump; Figure 3.12-C), consistent with the two enzymatic activities encoded by this gene [224]. The gene deficiency leads to pyrimidine

starvation that can be efficiently treated with uridine or cytidine (Figure 3.12-C). However, the supplemented uridine competes for intestinal absorption with dietary pyrimidines, or purines [282, 337]. VMH accounts for the corresponding facilitated transport reactions associated with the SLC29A1 (EntrezGene ID: 2030) and SLC29A2 (EntrezGene ID: 3177) as well as the sodium-dependent transport reaction enabled by SLC28A3 (EntrezGene ID: 64078). We have previously predicted that the human commensal gut microbe B. thetaiotamicron could also supplement the host with uridine [128]. Using VMH, we can readily identify further 415 gut microbes that could potentially supplement the human host with uridine as they encode for the 5'-Nucleotidase (VMH ID: NTD2, E.C. 3.1.3.5) as well as a uridine transporter (VMH ID: URIt2r). Of those microbes, 18 have been classified as probiotics in VMH and include 15 *Bifidobacterium* strains, two *Clostridium butyricum* strains, and a *Lactobacillus reuteri* strain. These probiotics are commonly found in yogurts, fermented food products, and probiotic formulations. While we could not find evidence for probiotic use in orotic aciduria, recent guidelines for management of methylmalonic and propionic acidemia included the use of probiotics [27]. Furthermore, researchers have demonstrated that the benefit of engineered *L. reuteri* strains in a murine phenylketonuria (PKU) model [76].

PKU is caused by a mutation in the gene PAH (EntrezGene ID: 5053) leading an inability to degrade phenylalanine (Figure 3.12-D). The life-long treatment consists of a diet low in phenylalanine. An alternative strategy could be to engineer the gut microbiota such that it consumes the excess of dietary phenylalanine. One option is the aforementioned engineering of probiotics, where the researchers introduced the phenylalanine ammonia-lyase to L. reuteri. This enzyme is ubiquitous in higher plants but rare in microbes, and VMH does not account for the corresponding reaction (although that does not necessarily mean that none of the 773 microbial genomes encode for this gene). However, an alternative pathway (VMH IDs: PHETA1, PLACOR, PLACD) converting phenylalanine to trans-cinnamic acid exists in six *Clostridium* strains, including four *Clostridium difficile* strains. Another option is to "replace" the mutated PAH gene with a microbial counterpart. In VMH, there are 26 microbes encoding for the microbial version of the genes, including two commensal *Bacillus cereus* strains and one probiotic strain (Lactobacillus reuteri SD2112). While B. cereus is known to be a causative agent in a minority of foodborne illnesses [174], the *L. reuteri* strain has been added to yogurt formulation, with the aim to improve oral hygiene

[210]. Additionally, the literature-derived carbon source table in VMH lists additional three commensal microbes that use phenylalanine as a carbon source: *Clostridium barletti*, *Anaerobaculum hydrogeniformans*, and *Gordonibacter pamelaeae*. The latter two have been recently patented to be used as probiotics for the inhibition of clostridial caused inflammation [31]. Taken together, VMH can be used to identify candidate microbes that could be used in addition or as a replacement for current dietary intervention strategies used in the treatment of certain inborn errors of metabolism.

3.4 Discussion

In this Chapter, we have described some aspects of the technical implementation of the VMH database. The architecture can be represented as a 3-layer infrastructure, with the database at its base. Building a database using Django allowed the conception of data "Models" that are inter-connected easing the accessing of data across resources. This connectivity is reflected in the interface where different levels of knowledge are accessible via each of the entity's detail page.

In addition, VMH provides access to an API which opens a unique window to the database content, allowing integration with other software. To demonstrate the potential of this tool, we have shown how it can be used to perform different analysis on the various resources available. In this context, we have shown how the generated tools can be used to perform complex analysis combining the available resources. VMH enables exploration of the different levels of interactions between microbes and host, providing an additional connection with the nutrition resource. We have additionally shown, that the gut microbiota resource can be used as a support tool in the design of synthetic microbial communities, an important research tool used to mimic the behavior of complex communities [29, 70]. Complex biochemical mechanisms, such as drug detoxification and retoxification can also be investigated with VMH with the advantage that potential microbial interactions with drugs can also be screened, an area of research that will increasingly attract more attention in the future. Finally, an example of how to use VMH to investigate potential treatments for diseases, including beneficial microbial community composition design strategies (with the inclusion of probiotics) was shown.

Taken together, the assembly of knowledge and tools in VMH gives researchers a uniquely integrated environment that allows performing complex analysis of metabolism. As VMH expands we believe that it will become increasingly important for multiple research communities.

Chapter 4

Visualization of Metabolic networks and Disease maps

Abstract

Visualization tools in research, provide support in knowledge search and interpretation of research data. Network visualization, in particular, is a typical approach used by Systems Biology researchers to try to understand how different biological processes are connected and the mechanisms behind those interactions. In the case of the human metabolic network, no intuitive map exists that is aesthetically pleasing and that enables integration of omics data and simulation results. In this Chapter, we introduce ReconMap 2.0, a visualization of the human metabolic network consistent with the content of Recon 2, the generic human metabolic reconstruction of metabolism. In addition, we explore a different visualization mechanism that is becoming increasingly popular: disease maps. We have created a prototype for a gene-to-phenotype map of mitochondrial disorders, using Leigh Syndrome, the most common phenotype of mitochondrial disease. These two resources are integrated into the Virtual Metabolic Human database available at `http://vmh.life`.

4.1 Introduction

Analyzing and deriving knowledge from biological networks is a challenging task for researchers. Visualization tools can provide much needed support especially if they allow the visualization of simulation and experimental data in a given context. Different tools exist for the visualization of biological networks such as CellDesigner [94], Cytoscape [276], Escher [163], MetDraw [149], MINERVA [100]. In general, these tools support the automatic generation of layouts, but for large-scale networks the results of these algorithms are often not aesthetically pleasing, hindering the analysis. The human metabolic network is such an example for which there was no intuitive visualization. For this purpose, we have created ReconMap, a comprehensive, manually curated map of human metabolism integrated into MINERVA, which uses the Google Maps Application Programming Interface (API) for highly responsive interactive navigation within a platform that facilitates queries and custom data visualization. ReconMap can be accessed via http://vmh.uni.lu, with network export in a Systems Biology Graphical Notation compliant format released under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. A Constraint-Based Reconstruction and Analysis (COBRA) Toolbox extension to interact with ReconMap is available via `https://github.com/opencobra/cobratoolbox`.

Another type of visualization approach that has gained relevance in recent years is disease maps (e.g. PD Map [93]). The main idea behind these maps is to provide a visual representation of all molecular components and pathways involved in a disease pathogenesis and progression. Mapping Mitochondrial disorders would be of interest as these are severe diseases with high clinical, biochemical, and phenotypic diversity, and no curative therapies. Additionally, diagnosis is extremely challenging due to the involvement of two genomes, varying ages of onset, and genetic diversity. Among the various mitochondrial disorders, Leigh Syndrome, a progressive neurodegenerative disorder, is the most common phenotype. Patients typically suffer from spongiform lesions in the basal ganglia and/or brainstem. Leigh syndrome is genetically heterogeneous and its diagnosis is very complex. For these reasons we have initiated the effort of mapping Mitochondrial disorders with Leigh Syndrome by addressing the diagnosis challenge with the creation of a gene-to-phenotype Leigh Syndrome map.

4.2 ReconMap: An interactive visualisation of human metabolism

Completely or partially as in: Alberto Noronha, Anna Dröfn Daníelsdóttir, Piotr Gawron, Freyr Jóhannsson, Soffía Jónsdóttir, Sindri Jarlsson, Jón Pétur Gunnarsson, Sigurður Brynjólfsson, Reinhard Schneider, Ines Thiele, and Ronan M. T. Fleming. **ReconMap: An interactive visualisation of human metabolism**.*Bioinformatics*, 2017.

A genome-scale metabolic reconstruction represents the full portfolio of metabolic and transport reactions that can occur in a given organism. A mathematical model can be derived from such a reconstruction, allowing one to simulate an organism's phenotypic behavior under a particular condition [237]. Recon 2 [305] is a very comprehensive knowledge-base of human metabolism and has been applied for numerous biomedical studies, including the mapping and analysis of omics datasets [305]. However, despite numerous visualization efforts using automated layouts [149], there is no genome-scale and biochemically intuitive human metabolic map available for visualization of omic data in its network context. Here, we release ReconMap, a comprehensive, manually curated map of human metabolism presented utilizing the Google Maps Application Programming Interface (API) for highly responsive interactive navigation within a platform that facilitates queries and custom data visualization.

4.2.1 Features

ReconMap content was derived from Recon 2.04, obtained from the Virtual Metabolic Human database (VMH, `http://vmh.uni.lu`). Reactions (hyperedges) were manually laid out using the biochemical network editor CellDesigner [94]. Each metabolite (node) was designated by its abbreviation and a letter corresponding to the compartment, in which the reaction occurs (e.g., '[c]' for cytosol). Metabolites present in a high number of reactions, e.g., common cofactors, were replicated across the map to minimize hyperedge crossover. ReconMap is presented using Molecular Interaction NEtwoRk visualization (MINERVA [100]), a standalone web service built on the Google Maps API, that enables low latency web display and interactive navigation of large-scale molecular interaction networks. Each metabolite and reaction in ReconMap links to the corresponding curated content provided by the VMH database. Moreover, MINERVA functionality connects ReconMap to external

databases, such as the CHEMBL database [30].

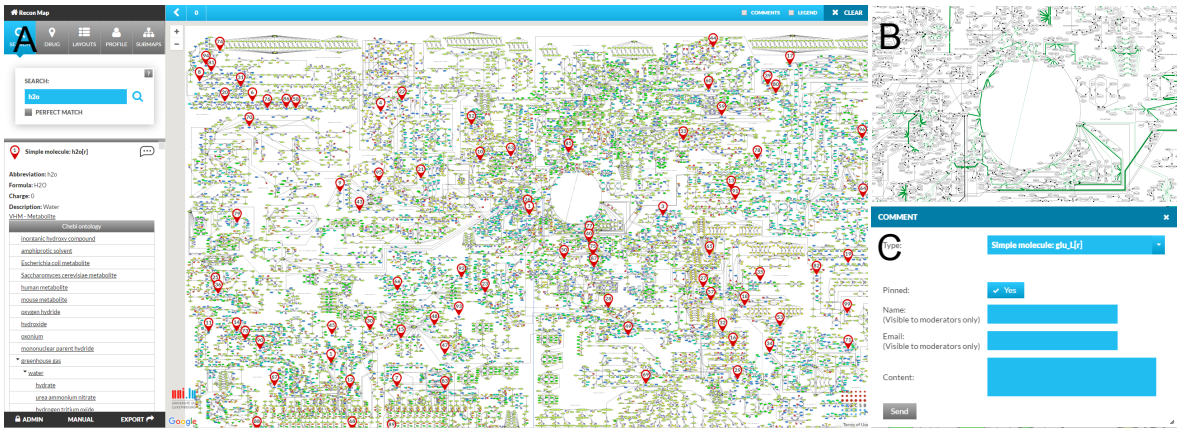

Figure 4.1: Web interface of ReconMap with search functionality. Information retrieved for on a specific molecule are shown, along with external links; B - overlay of a flux distribution, using differential thickness and color of the edges; C - Feedback interface that allows users to provide suggestions and corrections to entities of the ReconMap and Recon2.

Overlay of simulation results and multi-omics datasets

Recon-derived simulation results can be visualized on ReconMap using a new extension to the COBRA Toolbox [130]. By submitting an account request through the "ADMIN" area of ReconMap, the user can perform a simulation, e.g., Flux Balance Analysis, using the COBRA toolbox function *optimizeCBmodel*, then call the function *buildFluxDistLayout* to write the input file for a context-specific ReconMap Overlay. This permits the user to translate each flux value into a custom thickness and color within a simple tab-delimited file to highlight certain reactions. Similarly, registered users can display omic data on ReconMap via the "Overlay" menu, by uploading a tab-delimited file assigning a different color and thickness to each node and reaction.

Community-driven refinement of ReconMap & Recon

All users may post suggestions for refinement and expansion that are linked to a specific metabolite or reaction in specific locations of the map (right click then select "Add comment"). Each suggestion is forwarded to VMH curators for consideration when planning further curation effort. As such, ReconMap enables the community-driven refinement of human metabolic reconstruction and visualization.

Connecting ReconMap and PDMap

The Parkinson's disease map (PDMap [93], http://pdmap.uni.lu) displays molecular interactions known to be involved in the pathogenesis of Parkinson's disease. A total of 168 metabolites connect ReconMap and PDMap via standard identifiers. These connections are available in the metabolites description as well as in their detail pages in the VMH website. This feature is particularly interesting when mapping omics datasets on both maps, thereby allowing the simultaneous investigation of metabolic and non-metabolic pathways relevant for Parkinson's and other neurodegenerative diseases.

Implementation and usage example

ReconMap was drawn using CellDesigner and is displayed using the MINERVA platform, built on the Google Map API, using human reconstruction content from the VMH database `http://vmh.uni.lu`. Matlab scripts for analysis of COBRA Toolbox simulation results using ReconMap are freely available in the COBRA Toolbox `https://opencobra.github.io/cobratoolbox`. This combination of tools is aimed at allowing the user to visualize what cannot be appreciated at first with model simulation outputs.

In order to access remotely to ReconMap, the user has to be registered by requesting access at the VMH map page (`http://vmh.uni.lu/#reconmap`). Using these credentials, the user can then configure the MATLAB 'minerva' structure to access ReconMap as shown in Figure 4.2. After this step, the user needs to initialize the CobraToolbox and load a metabolic model (in this case Recon 2.04).

```
load('minerva.mat')
minerva.model = 'ReconMap-2.01';
minerva.login = 'user_name';
minerva.password = 'user_password';
```

Figure 4.2: Setup of ReconMap credentials in the CobraToolbox

1. Overlay a flux distribution

As an example of a layout, we would like to visualize the fluxes when maximizing ATP production through complex V (ATP synthase) in the Electron Transport Chain. To do so, we

use Flux Balance Analysis (FBA) and set as the objective function the reaction responsible for this process ('ATPS4m'). To do this two CT functions are necessary:

- *ChangeObjective*: function, changes the objective function of a constraint-based model
- *optimizeCbModel*: function solves a flux balance analysis problem.

```
initCobraToolbox;
changeCobraSolver('glpk', 'LP');
model = readCbModel('Recon2.v04.mat')
% Rename the model.
model_atp_production = model
model_atp_production = changeObjective(model_atp_production,
↪  'ATPS4m');
solution_atp_prod_max_regularised =
↪  optimizeCbModel(model_atp_production, 'max',1e-6);
solution_atp_prod_max_sparse =
↪  optimizeCbModel(model_atp_production, 'max','zero');
```

Figure 4.3: Setting up FBA simulations for ATP production through complex V (ATP Synthase) with Recon 2.04.

The *buildFluxDistLayout* function creates a layout that is automatically sent to the ReconMap website. After this, results can be visualized at `http://vmh.uni.lu/#reconmap` by selecting the "Overlays" section as displayed in Figure 4.4.

2. Overlay a subsystem

There is also the possibility of highlighting specific subsystems by using the function *generateSubsytemsLayout*. A subsystem is a group of metabolic reactions involved in the same metabolic pathway, such as glycolysis, oxidative phosphorylation, or citric acid cycle. For instance, to highlight the TCA cycle in ReconMap and obtain the resulting overlay as in Figure 4.5, users can execute commands as illustrated in Figure 4.5.

Alternatively, the user can generate a layout of all common subsystems between the model and ReconMap using the function *generateSubsystemLayouts*.

```
serverResponse = buildFluxDistLayout(minerva, model,
↪  solution_atp_production_max_regularised,
↪  'atp_prod_max_regularised3');
serverResponse = buildFluxDistLayout(minerva, model,
↪  solution_atp_production_max_sparse, 'atp_prod_max_sparse4');
```

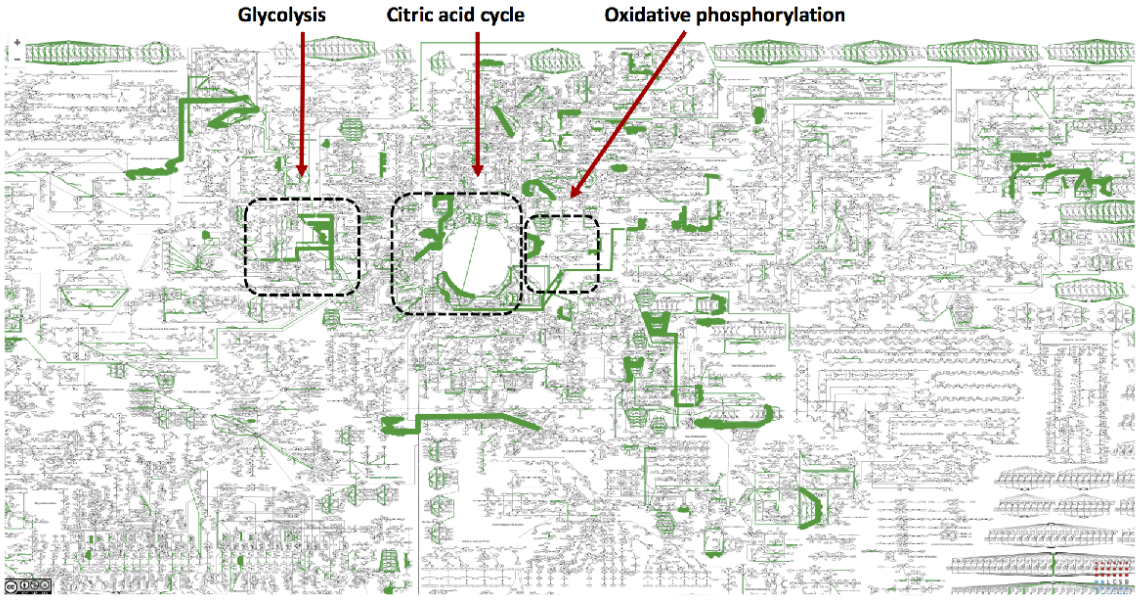

Figure 4.4: Remote overlay submission to ReconMap. Web interface of ReconMap shows resulting overvays.

4.3 Leigh map: A novel computational diagnostic resource for mitochondrial disease

Completely or partially as in: Rahman J., Noronha A,, Thiele I., Rahman S. **Leigh map: A novel computational diagnostic resource for mitochondrial disease**.*Annals of Neurology*, 2017.

Mitochondrial disorders are among the most severe metabolic disorders wherein patients suffer from multisystemic phenotypes, often resulting in early death [187]. Clinical, biochemical, and genetic heterogeneity among individuals, together with poor understanding of gene-to-phenotype relationships, pose significant diagnostic and therapeutic challenges for clinicians. In light of recent advances in next generation sequencing technologies, whole ex-

```
generateSubsytemsLayout(minerva, model, 'Citric acid cycle',
↪  '#6617B5');
```

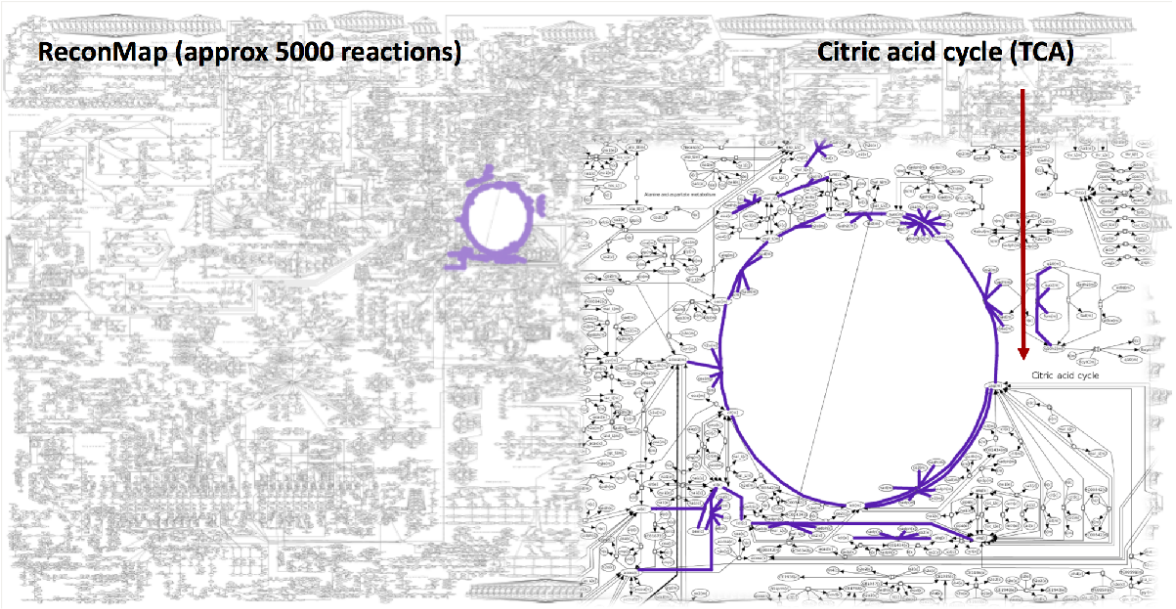

Figure 4.5: Code for generating subsystems overlays and web interface of ReconMap displaying a subsystem overlay.

ome sequencing (WES) is emerging as the new global standard for the diagnosis of monogenic disorders, including mitochondrial diseases [335]. However, owing to genetic heterogeneity of mitochondrial disorders and ongoing discovery of novel disease genes, WES data may not provide clinicians with enough certainty for a definitive diagnosis.

With these challenges in mind, we present the Leigh Map, a novel computational gene-to-phenotype network to be used as a diagnostic resource for mitochondrial disease, using Leigh syndrome (Mendelian Inheritance in Man 256000), the most genetically heterogeneous and most frequent phenotype of pediatric mitochondrial disease [182, 251], as a prototype. Leigh syndrome is a progressive neurodegenerative disorder defined neuropathologically by spongiform basal ganglia and brainstem lesions [251, 183]. Clinical manifestations include psychomotor retardation, with regression, and progressive neurological abnormalities related to basal ganglia and/or brainstem dysfunction, often resulting in death within 2 years of initial presentation [251, 288]. However, many patients may also present with multisystemic (eg, cardiac, hepatic, renal, or hematological) phenotypes. To date, there are 89 genes known to

cause Leigh syndrome, the majority of which are difficult to definitively differentiate from each other, either biochemically or clinically. We hypothesized that these multisystemic features may help to distinguish different genetic subtypes of Leigh syndrome.

The Leigh Map (freely available at `vmh.uni.lu/#leighmap`), was built on the Molecular Interaction NEtwoRks VisuAlization (MINERVA) platform [100] previously used to construct networks of Parkinson disease and human metabolism [93, 223, 305]. The network comprises 89 genes and 236 phenotypes, expressed in Human Phenotypic Ontology (HPO) terms [171, 170], providing sufficient phenotypic and genetic variation to test the network's diagnostic capability. The Leigh Map aims to enhance the interpretation of WES data to aid clinicians in providing faster and more accurate diagnoses for patients so that appropriate measures can be taken for optimal management. The phenotypic components of the Leigh Map can be queried to generate a list of candidate genes. In addition, the genetic components of the Leigh Map may also be queried to browse a list of all reported phenotypes associated with a particular gene defect. We propose that this functionality can be used to enhance clinical surveillance of patients with an established genetic diagnosis. Blinded validation of test cases containing clinical and biochemical, but not genetic, data demonstrated that 2 independent testers were able to predict the correct causative gene using this method in 80% of cases. The success of the Leigh Map demonstrates the efficacy of computational networks as diagnostic aids for mitochondrial disease (Figure 4.6).

4.3.1 Creation of the Leigh Map

Systematic Literature Review

The genetic and phenotypic information gathered in this study came from an initial knowledgebase of >900 publications, collected from PubMed (latest search November 2016) and the senior author's personal archive. To facilitate data collection from this large breadth of literature associated with Leigh syndrome, we performed systematic literature mining with QDA Miner Lite (v1.4.2; Provalis Research, Montreal, Quebec, Canada) to generate a list of genes reported to cause Leigh syndrome or Leigh-like syndromes, and their corresponding phenotypes. Phenotypic information was standardized by manually entering each reported phenotype into Phenomizer (compbio.charite.de/phenomizer), [171, 170] a free on-

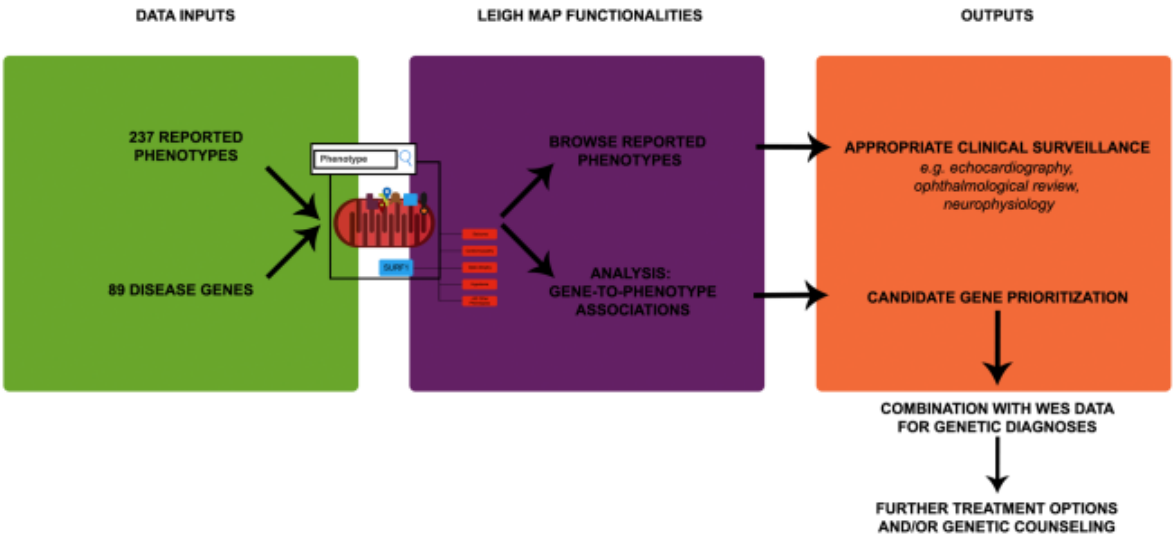

Figure 4.6: Conceptualization of the Leigh Map. The Leigh Map is a novel computational resource that effectively integrates a large amount of phenotypic and genetic data from the literature and synthesizes it into a comprehensive resource that has the potential to improve diagnostic outcomes and more vigilant clinical surveillance for patients with Leigh syndrome. WES = whole exome sequencing.

line resource, which catalogues thousands of standardized human phenotypes, to obtain the appropriate HPO term and number. In addition to obtaining individual Leigh syndrome genes and phenotypes, we collected information on additional parameters that will give users further insight for an informed diagnosis. Such parameters include modes of inheritance, magnetic resonance imaging findings, and patient demographic information. These data were then organized into an Excel file. Although we aimed to rely solely on text mining to obtain these data, some publications required manual clarification, owing to formatting errors on QDA Miner, which were especially prevalent in publications with large tables. In total, we consulted >500 publications to create the Leigh Map. A simplified version of the gene-to-phenotype knowledgebase is provided in Tables 1 and 2.

Mitochondrial Dysfunction	**Genes (mode of inheritance)**	**Example Phenotypes**
OXPHOS subunits		
Complex I	NDUFA1 (XL); NDUFA2, NDUFA9, NDUFA10, NDUFA12, NDUFS1, NDUFS2, NDUFS3, NDUFS4, NDUFS7, NDUFS8, NDUFV1, NDUFV2 (AR); MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND5, MT-MD6 (maternal)	DDwR, FTT, hypertrichosis, HCM, LA, LD, liver failure, myopathy, OA, PN, renal tubulopathy, SNHL, SZ
Complex II	SDHA (AR)	DDwR, FTT, HCM, LA, OA, paraganglioma, pheochromocytoma, SZ
Complex III	UQCRQ (AR)	Ataxia, dementia, DD, dystonia, myopathy
Complex IV	COX8A, NDUFA4 (AR); MT-CO3 (maternal)	Ataxia, DD, DR, diabetes mellitus, LA, LD, microcephaly, PN, SNHL, SZ
Complex V	MT-ATP6 (maternal)	DDwR, FTT, HCM, LA, LD, myopathy, OA, SZ
OXPHOS assembly		
Complex I assembly	NDUFAF2, NDUFAF4, NDUFAF5, NDUFAF6, C17ORF89, FOXRED1, NUBPL(AR)	Anemia, DDwR, FTT, HCM, LA, liver failure, myopathy, OA, SNHL, SZ
Complex II assembly	SDHAF1 (AR)	DDwR, LA, LD, liver failure, myopathy
Complex III assembly	BCS1L, TTC19 (AR)	DDwR, FTT, LD, LA, liver failure, renal tubulopathy, SNHL, SZ

Complex IV assembly	SURF1, SCO2, COX10, COX15, PET100 (AR)	DDwR, FTT, hypertrichosis, HCM, LA, LD, myopathy, OA, renal tubulopathy, SNHL, SZ
Cofactor biosynthesis and metabolism		
CoQ10biosynthesis	COQ9, PDSS2 (AR)	DDwR, FTT, HCM, hypotonia, myopathy, nephrotic syndrome, renal tubulopathy, SZ
Lipoic acid biosynthesis	LIAS, LIPT1 (AR)	DDwR, dystonia, FTT, hypertension, LA, LD, OA, SZ
Thiamine metabolism	SLC19A3, TPK1 (AR)	DDwR, dystonia, microcephaly, hypoglycemia, LD, OA, SZ
Biotinidase	BTD (AR)	Ataxia, DR, hypotonia, LA, spastic tetraplegia
Other metabolic dysfunction		
Pyruvate dehydrogenase complex	PDHA1 (XL); PDHX, PDHB, DLAT, DLD (AR)	DD, FTT, LA, LD, microcephaly, myopathy, OA, PN, SZ
Amino acid metabolism	HIBCH, ECHS1 (AR)	Abnormal plasma acylcarnitines, DDwR, FTT, LA, LD, microcephaly, myopathy, OA, SZ

AR = autosomal recessive; DD = developmental delay; DDwR = developmental delay with regression; DR = developmental regression; FTT = failure to thrive; HCM = hypertrophic cardiomyopathy; LA = lactic acidosis; LD = leukodystrophy; OA = optic atrophy; OXPHOS = oxidative phosphorylation; PN = peripheral neuropathy; SNHL = sensorineural hearing loss; SZ = seizures; XL = X-linked.

Table 4.1: Leigh Syndrome Disease Genes and Phenotypes Associated with Metabolism

Mitochondrial Dysfunction	**Genes (mode of inheritance)**	**Example Phenotypes**
Mitochondrial DNA maintenance	POLG, SUCLA2, SUCLG1, FBXL4 (AR)	DDwR, FTT, HCM, LA, LD, methylmalonic aciduria, myopathy, OA, renal tubulopathy, SZ
Mitochondrial translation	GFM1, GFM2, TSFM, TRMU, MTFMT, GTPBP3, TACO1, C12ORF65, LRPPRC, EARS2, FARS2, IARS2, NARS2 (AR); MT-TI, MT-TK, MT-TL1, MT-TL2, MT-TV, MT-TW (maternal)	Anemia, DDwR, FTT, hypoglycemia, HCM, LA, LD, OA, renal tubulopathy, SZ
Mitochondrial dynamics	SLC25A46 (AR), DNM1L (AD)	Ataxia, DDwR, FTT, hypotonia, microcephaly, LA, SZ
Mitochondrial import	SLC25A19 (AR)	DD, FTT, hypotonia, microcephaly, PN, SZ
Membrane phospholipids	SERAC1 (AR)	3-Methylglutaconic aciduria, DDwR, FTT, LA, liver failure, OA, SNHL, SZ
Mitochondrial sulfur dioxygenase	ETHE1 (AR)	DDwR, ethylmalonic aciduria, LA, renal tubulopathy, SZ

Oligomeric AAA + ATPase	CLPB (AR)	DDwR, FTT, HCM, LD, OA, SZ
Apoptosis	AIFM1 (AR)	DDwR, HCM, hypoglycemia, SNHL, SZ
RNA import	PNPT1 (AR)	DR, dystonia, muscle weakness, SNHL, SZ
RNA-specific adenosine deaminase	ADAR (AR)	DDwR, microcephaly, SZ, skin hyperpigmentation
Nuclear translocation pathway	RANBP2 (AR)	Ataxia, cognitive impairment, myopathy, SZ
Nuclear pore complex protein	NUP62 (AR)	FTT, DR, OA, SZ
Manganese transporter	SLC39A8 (AR)	DD, FTT, LA, SNHL, SZ
AD = autosomal dominant; AR = autosomal recessive; ATPase = adenosine triphosphatase; DD = developmental delay; DDwR = developmental delay with regression; DR = developmental regression; FTT = failure to thrive; HCM = hypertrophic cardiomyopathy; LA = lactic acidosis; LD = leukodystrophy; OA = optic atrophy; PN = peripheral neuropathy; SNHL = sensorineural hearing loss; SZ = seizures.		

Table 4.2: Leigh Syndrome Disease Genes and Phenotypes Associated with Other Mitochondrial Functions

4.3.2 Structure and Functionality of Leigh Map

The Leigh Map was manually assembled using CellDesigner (v4.4) [94] by incorporating phenotypic, genetic, and demographic data collected through literature mining. The map layout loosely follows mitochondrial structure. The outermost compartment represents the cytosol, where it is possible to find the nucleus and the mitochondrion. Three nuclear genes, nuclear envelope protein NUP62, nuclear export protein RANBP2, and adenosine deaminase

ADAR, have been included in our network as genes causing a clinical and radiological phenotype closely resembling Leigh syndrome [23, 283, 190]. The mitochondrion is visualized in its double membrane structure, and mitochondrial genes are grouped according to function and can be found in their submitochondrial location (eg, outer membrane, matrix). To represent gene-to-phenotype associations, a submap was created for each gene, displaying all phenotypes associated with any given gene defect. Also incorporated at this stage are links to external databases (eg, Uniprot [59] and HGNC [108]) and modes of inheritance. This approach enables a modular overview of the map, avoiding overwhelming the user with the "hairball" effect caused by the high connectivity of the network. All submaps were integrated in the MINERVA framework [100], which makes use of the Google Maps application programming interface, enables content query, and allows a low-latency interactive navigation of the network and its submodules simply by clicking a specific gene and opening the embedded submap window available on the interface.

Navigation through the network is similar to that of Google Maps, wherein the user can reveal increasingly specific components of information by zooming in on the different compartments (Fig 2, Supplementary Figs 1–4). Additional data (patient demographics, modes of inheritance, external annotations, etc) can be accessed by clicking an element of the map. The corresponding data will be displayed in the left panel. The search functionality enables the query of multiple genes and phenotypes. The query results are displayed in the information panel and are also highlighted on the map. When searching for multiple phenotypes, all genes associated with each phenotype will be listed. Opening the submap for any given gene will display 1 or more of the highlighted phenotype elements, providing an immediate visual interpretation of the search results.

The Leigh Map provides data about 89 genes reported to cause Leigh syndrome and Leigh-like syndromes, the highest number of Leigh syndrome genes that has been collated to date, as well as 236 associated phenotypes. The network consists of >1,700 interactions, all of which can be manually queried by the user. To facilitate access, causative Leigh syndrome genes are segregated according to gene function and arranged on a simplified schematic of the mitochondrion. Genes with similar functions are grouped together in subcategories. Examples of gene categories that can be found on the Leigh Map include genes involved in oxidative phosphorylation (eg, NDUFA1, SDHA) and genes that maintain mitochondrial

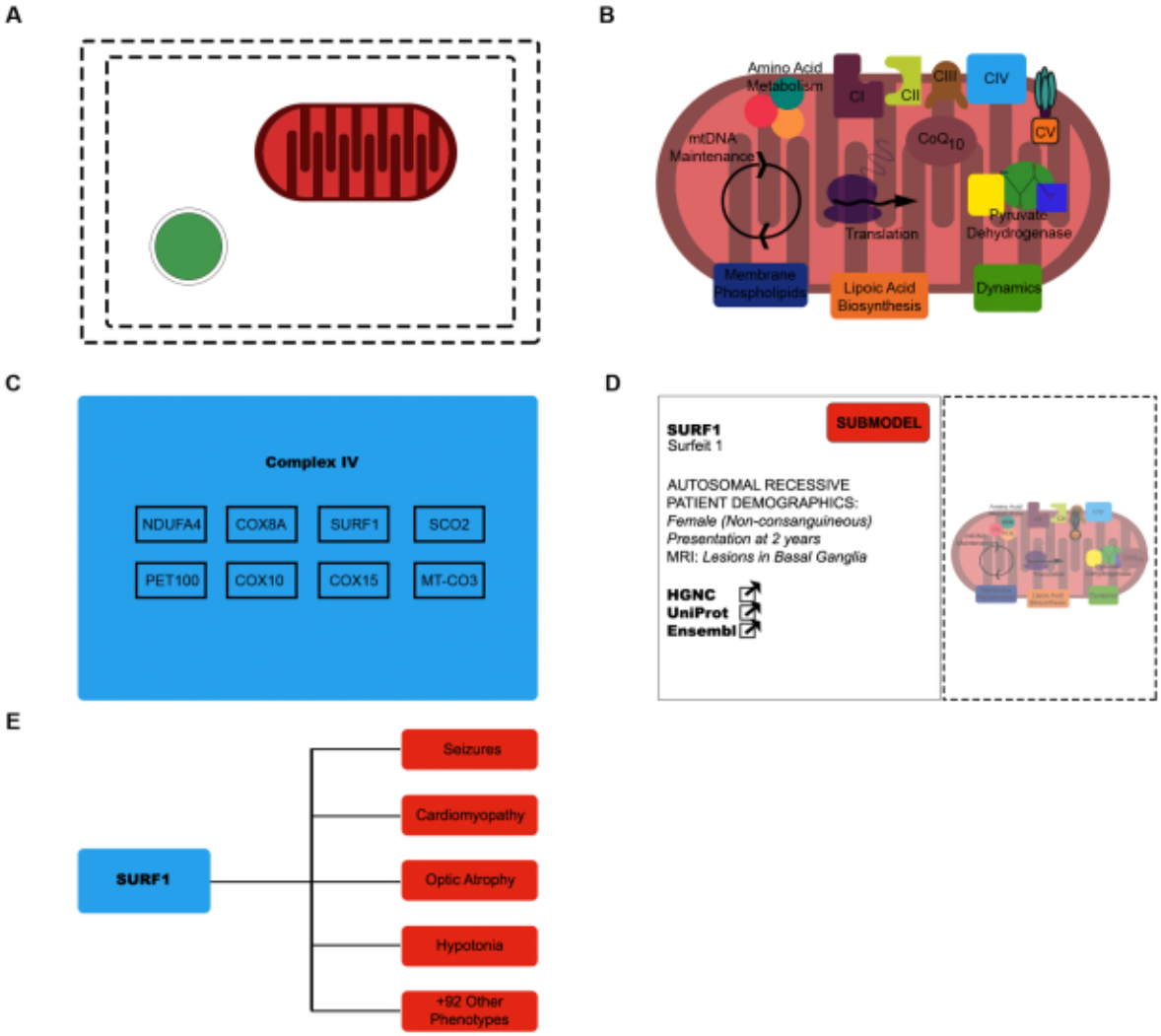

Figure 4.7: Schematic layout of the Leigh Map. The Leigh Map is a novel gene-to-phenotype network that can be used as a diagnostic resource for Leigh syndrome. The layout and navigation of the Leigh Map are similar to those of Google Maps, wherein the user zooms in on components to reveal further layers of information. (A) The outermost part of the Leigh Map is a simplified diagram of the cell. (B, C) Clicking on a compartment (eg, the mitochondrion) reveals categories of genes associated with Leigh syndrome (B), and zooming in on subcompartments within the mitochondrion reveals individual genes (C). (D) Detailed information about a specific gene defect can be accessed by clicking on a gene (SURF1 in this example), which will display a left-hand panel that provides additional information and external annotations. (E) Each gene contains a "submodel" that can be accessed by clicking. Gene submodels display all phenotypes associated with the gene of interest (a total of 96 phenotypes in the case of SURF1 deficiency). Live screenshots of the Leigh Map are provided in Supplementary Figure 4.6.

DNA (eg, POLG, SUCLA2; see Fig 2). Expression of Leigh syndrome phenotypes in HPO terms serves to normalize the network, thereby eliminating discrepancies in clinical jargon for phenotypes for which >1 synonym exists. “Leukodystrophy,” for example, can

be described alternatively as “leukoencephalopathy” or “white matter changes.” The use of different nomenclature varies among clinicians and in different geographical regions; therefore, the use of a single HPO term (leukodystrophy; HP: 0002415) simplifies the Leigh Map and encourages its widespread utilization (Figure 4.8).

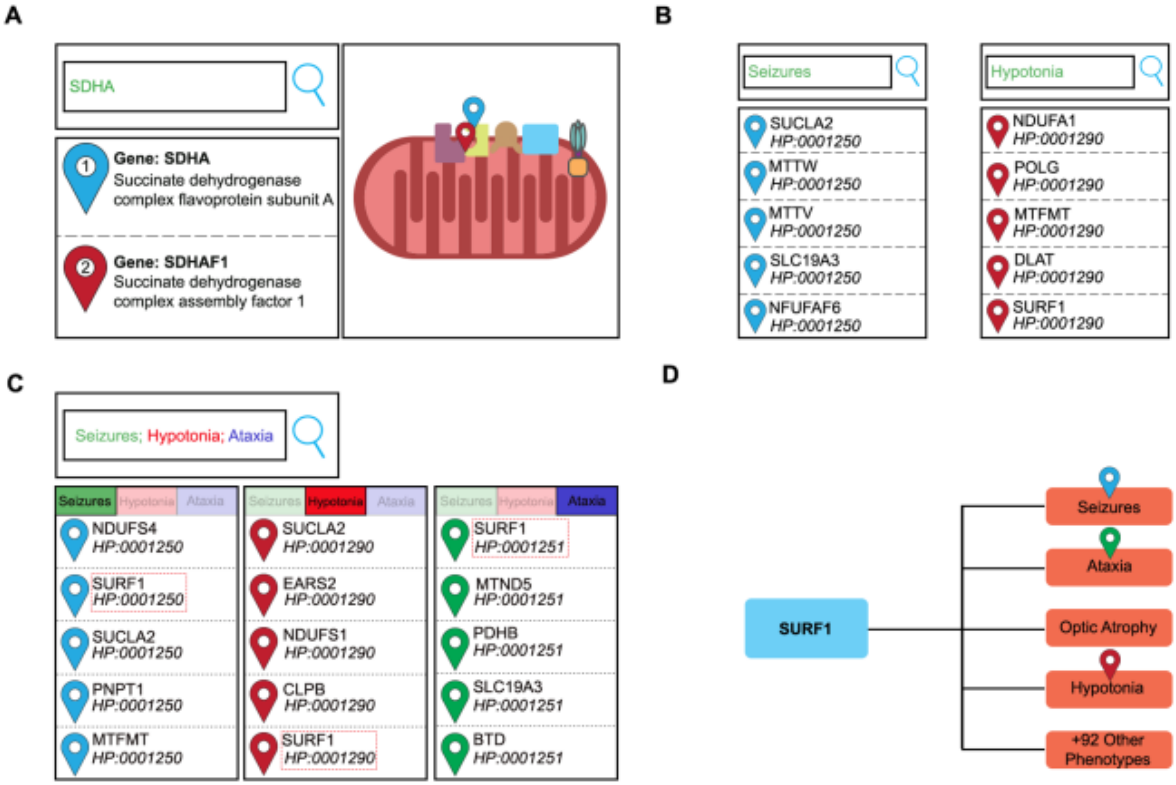

Figure 4.8: Querying the Leigh Map. (A–C) All phenotypic and genetic components of the Leigh Map can be queried using the search function in the left-hand panel. The user can query a particular gene by typing the name of the gene or any known alias into the search box. The results of the search will be displayed in the left-hand panel, and the matching gene(s) will become marked on the network (A). Phenotypes can be queried in the same way. The results of a phenotype search will display all genes associated with the queried phenotype (B). Multiple phenotypes can be queried simultaneously by separating phenotypes with a semicolon. The results of a multiple phenotype search will be displayed in different tabbed panels through which the user can navigate (C). (D) Clicking on the gene’s submodel in any multiple phenotype search will display all highlighted phenotypes from the query.

4.3.3 The Efficacy of the Leigh Map as a Diagnostic Resource

Blinded validation by 2 nonclinical investigators using a series of anonymized test cases revealed that the Leigh Map was able to identify the correct gene for 16 of 20 cases. The first and second authors, who both lack formal clinical expertise, acted as independent blinded

testers of the network. The anonymized test cases were obtained from the senior author's clinical practice, a national mitochondrial disease clinic where patients with Leigh syndrome who have diverse clinical presentations and genetic causes are diagnosed and managed. The criteria for these test cases were patients who had a definitive genetic diagnosis of Leigh syndrome, confirmed by Sanger sequencing or WES. Testers were provided with clinical vignettes and biochemical data, without genetic information. All corresponding phenotypes identified from each test case were entered into the query box of the Leigh Map, each separated by a semicolon. The search tool then generated a list of candidate genes for each phenotype in individual panels, which were then manually browsed to establish a list of candidate genes (see Fig 3). We define "candidate genes" as those that include >50% of the queried phenotypes. Due to the immense number of phenotypes on the network, every test case generated a list of potentially causative genes. For 10 cases, the Leigh Map was able to identify the correct gene as the "top hit," that is, the gene corresponding to the highest number of matched phenotypes. The network also predicted the correct gene for an additional 6 cases, in which they were not the top hit. In the remaining 4 test cases, the Leigh Map failed to produce the correct gene as one of the generated candidate genes. In all cases, the Leigh Map produced a shortlist of no more than 8 candidate genes, effectively eliminating 90% of the genes in the network. Multiple advanced search is not yet possible on this platform, so some manual deduction is required for the use of the Leigh Map at this time.

4.3.4 Future Prospects

Due to its high success rate in predicting causative genes by nonclinical testers, we conclude that the Leigh Map is an efficacious diagnostic resource that, in combination with WES data and metabolic testing, can be used by clinicians to provide patients with accurate diagnoses or to direct further biochemical investigation. Increased certainty of the genetic causes of mitochondrial disease has significant implications, because it could potentially attenuate the need for invasive diagnostic procedures, namely muscle biopsy with an attendant general anesthetic, which could pose risk to pediatric patients. It is important to iterate that we do not propose that the Leigh Map act as a substitute for WES data or other relevant functional studies, but rather as a supplement to these techniques.

The computational nature of the Leigh Map allows for the addition of novel disease genes or phenotypes with relative ease; thereby, clinicians have access to a database of all current causative genes, which can enhance the interpretation of WES data. Ideally, we will update both the phenotypic and genetic components of the Leigh Map concurrently with the literature and also develop a facility wherein experts can submit additional genetic or phenotypic information. This is especially beneficial within the context of mitochondrial diseases, because novel genes are constantly being identified. For Leigh syndrome specifically, one-third of the causative genes were identified within the past 5 years.[3]

Currently, the most significant limitation of the Leigh Map is the lack of a multiple advanced search facility. Although the absence of this feature does not detract from the network's accuracy, it does reduce its ease of use. Future work aims to implement this feature into the network. Furthermore, the efficacy of the Leigh Map is affected by the breadth of literature available for individual genes. SURF1, one of the earliest mitochondrial disease genes to be identified and the most common nuclear genetic cause of Leigh syndrome, is the subject of numerous publications [326]. Thus, SURF1 is associated with > 90 phenotypes in the Leigh Map, the largest number for any single gene. In contrast, the recently characterized complex I assembly gene C17ORF89 [86] only features in a small section of a larger publication and accordingly is associated with only 2 phenotypes on the Leigh Map, although patients who harbor this mutation may display other phenotypes.

Expanding the current gene-to-phenotype binary of the Leigh Map is a future prospect that can further improve its usefulness as a diagnostic resource. Although there are no current curative therapies for mitochondrial disease, there are numerous compounds that are aimed at symptomatic management, including anticonvulsant drugs used to manage epilepsy and cofactor and vitamin supplements, such as coenzyme Q10, thiamine, and biotin, used to treat corresponding deficiencies. The addition of drug targets (a current feature of the MINERVA platform) to the Leigh Map could potentially provide insight into the effectiveness of various agents in treating mitochondrial disease in specific genetic contexts. For example, patients with SLC19A3 mutations respond dramatically to biotin and thiamine therapy [81], whereas those with HIBCH mutations may benefit from N-acetyl cysteine [82]. cDNA and protein mutations and annotations regarding animal models are also useful potential supplements to the Leigh Map. Leigh syndrome is a defined disorder [183] wherein certain phenotypes

appear almost ubiquitously, including hypotonia (91% of patients), developmental delay (82%), lactic acidosis (78%), and failure to thrive (61%). The failure to deduce the correct candidate genes for a minority of our test cases was due to the predominant presence of these common Leigh syndrome phenotypes and a lack of discriminating phenotypes. We found more success in "diagnosing" cases that presented with less frequently observed phenotypes such as cardiomyopathy (59%), optic atrophy (47%), or renal tubulopathy (15%). Therefore, the addition of these extra elements can be helpful in narrowing down a large list of candidate genes, thereby increasing the predictive power of the Leigh Map. An alternative approach to increase diagnostic power for common phenotypes is to incorporate a scoring system, which is a common element in other bioinformatics resources such as BLAST [8]. In the context of our network, we propose "common" phenotypes be scored lower than less frequently observed phenotypes. The addition of a scoring system would complement the more sophisticated advanced search feature that we aim to implement in the future.

4.4 Conclusions

In this Chapter, we have shown two applications of network visualization for different contexts that can be used by researchers through the VMH.

ReconMap allows for efficient visualization of manually curated human metabolic reactions and metabolites from the VMH database, with numerous connections to complimentary online resources. ReconMap is a generic visualization of human metabolism and serves as a template for the generation of cell-, tissue-, and organ-specific maps. Moreover, omics data and flux distributions resulting from simulations can be visualized in ReconMap in a network context via an extension to The COBRA Toolbox. ReconMap can be readily connected to disease-specific maps, such as the Parkinson's disease map, thereby enabling investigations beyond metabolic pathways. Future directions include multiscale visualization, conserved moiety tracing [117], drug target search, and increased synergy with simulation tools.

On another front, the progressive improvements in sequencing technologies and increased global cooperation have allowed for the generation of copious amounts of genetic and clinical information pertaining to mitochondrial disease. The Leigh Map effectively integrates these clinical and scientific data into an efficacious diagnostic resource for a genetically het-

erogeneous disorder, the success of which provides the basis for the construction of larger computational networks for a wider scope of mitochondrial and metabolic diseases.

In the future, we expect that multi-layer maps will become a reality. Information represented in maps and networks following different approaches as those shown in this chapter, will start overlapping. Integrating detailed information on metabolic pathways, combined with gene-to-phenotype relationships, will enable researchers to interactively visualize, for instance, pathways affected by specific mutations and how clinical phenotypes translate into metabolic states.

Chapter 5

Challenges and tribulations in the development of a biological database

Abstract

Biological databases are important tools that allow organizing and sharing the increasing amounts of data generated by new technologies and research projects. As the need for additional biological databases arises, researchers will face various design and technical challenges. Small teams and budget limitations are often a factor contributing to the difficulties of execution of such projects. For this reason, we believe that the research community will benefit from a starting guide aimed at researchers planning to develop a biological database. This work highlights some of the decisions that need to be taken and issues that need addressing when creating a biological database accessible through a web page. These instructions are not a complete guide for database development but they are a result of our experience in the development of the Virtual Metabolic Human database.

5.1 Introduction

The progress in technologies used in life sciences and biomedical fields led to an increase in the amounts and complexity of data generated. In response to this, biological databases became important tools to organize and share data collected from scientific experiments, omics technologies, literature, and different analyses. Over the years biological databases have increased in numbers and popularity. The NAR online Molecular Biology Database Collection keeps a list of active databases and publishes a yearly database issue [97]. It has been recognized that a biological database does not live only of its data and that an intuitive web interface is an essential component [24]. Web application programming interfaces (web APIs) have become ubiquitous and are also gaining relevance for biological databases. These APIs allow access to database content in a more efficient way and enable programmatic access of third-party applications allowing analysis that go beyond the ones provided by pre-defined web interfaces.

The development of a biological database is, therefore, an effort that involves analyzing, combining, and structuring biological data but carries several technical challenges due to the need of combining different technologies and coding in different programming languages. To make matters worse, it is fairly common that research groups do not have dedicated teams for software/database development and maintenance, which further accentuates this problem. While for software libraries there are a considerable number of articles aimed at computational biologists and bioinformaticians, that cover topics such as best practices and workflows [342, 189, 63], such resource for the development of biological databases is, to the best of our knowledge, still lacking.

In this Chapter, we will discuss strategies that can be taken in the development of a biological database. We will focus on examples from the development of the Virtual Metabolic Human (VMH) and possible future improvements. We will cover some definitions about databases, web interface programming, and Web APIs. Software and database development are ever changing, and therefore the advice presented and choice of technologies is not set in stone. We do hope that they can still provide a clear picture of the typical problems and strategies to address them in projects of this scope.

5.2 Choosing the database system

The selection of the database system should be the first task on a developers' head. There are several types of databases available and as expected, they fit different roles. For instance, there are databases that use memory instead of disk to store information. They are extremely efficient but also extremely expensive.

Typically, a biological database has well-defined content and write/delete operations occur at specific points in time (minor or major updates). In addition, user interaction is often restricted to reading information. For this reason, a general purpose database will be adequate in most scenarios (e.g. MySQL, PostgreSQL, Oracle). For the development of VMH, we have selected MySQL for its simplicity and efficiency.

5.2.1 Database management systems (DBMS)

The main tasks that the developer(s) of a biological database is assigned are typically updating/creating content, and database maintenance. These tasks can be performed using specific commands, normally in a variation of the SQL standard, or through user interfaces provided by most DBMS.

For VMH, the database management is made using Django, a Python-based server-side web framework. One of the most attractive features of Django is that it provides the tools to create and manage database content for different database systems (MySQL, PostgreSQL, and Oracle). Django greatly facilitates database maintenance due to its migrations system. Migrations keep track of changes in the database without the need to implement any SQL-like code. These migrations allow version-control of the database structure in a streamlined and simplified way.

5.3 Database content and access

One important aspect to include in a biological database is connections with other resources. Aggregating information from other sources is a good idea if due credit is given and no licensing terms are breached. Support for standards is encouraged as this will enable other users and databases to access and use your data more easily. The MIRIAM registry [151]

provides location-independent identifiers for data used in the biomedical domain and most known biological databases have been registered there.

Biological databases are normally accessible through a web-interface. In addition, we recommend that the content is made available for download as flat files and that programmatic access is enabled by a web service.

5.3.1 Web interface

Choosing a language and a framework to develop a web page can be a daunting task. There are literally dozens of choices to pick from. In the context of this work, the main concern should be choosing a framework without a steep learning curve, with good documentation, and importantly, a large community of developers. Web resources such as StackOverflow (`https://stackoverflow.com/`), the largest online community for developers, can be a good reference point for the size of the community using a specific framework. Highly active communities mean that most of the problems that developers will encounter were probably solved by another person at some point in time. These resources allow saving great amounts of time by avoiding replication of effort. Finally, when choosing a framework it might be necessary to consider the associated licensing costs. JavaScript frameworks, in particular, are increasingly popular, such as Bootstrap, Angular, ReactJS, or ExtJS. These frameworks, one way or another, simplify the development of web pages by providing pre-defined modules and components that work across browsers and systems.

5.3.2 Programmatic access

Programmatic access to database content can be enabled through a web application programming interface (web API). These type of interfaces are extremely useful as they allow other applications or user-made scripts to access the database content. There are, as similar to web interfaces, several frameworks to choose from. In our perspective, the same considerations discussed before should be taken. In the case of VMH, we have decided to use the *Django Rest Framework* package as this enabled combining the database management with the web API development.

In a web API information is accessed through a series of URL endpoints. These URL

endpoints should also support searching, filtering, and pagination. All these URL patterns and additional parameters need to be well documented (e.g. `vmh.uni.lu/_api/docs`). When a web API becomes available for broad use it will possibly be adopted by other applications and databases. This needs to be taken into consideration when updating/changing functionality. For this reason, web APIs development frameworks also support versioning.

5.3.3 Domain name, DNS, and hosting

Let us assume that a database and website are ready to go live. At this point three things should be taken into consideration: the domain, the Domain Name Servers (DNS)/System to use, and where to host the website.

The domain is the name of your website (for instance `vmh.uni.lu`) with which users will visit the website. For this name to work, it needs to be registered with an accredited registrar so it becomes an alias to the IP of your server, uniquely identifying the server. For this redirection to work (translation of the website address into an IP) a nameserver or DNS needs to be set (Figure 5.1). DNS are like internet phone books, where IPs replace telephone numbers.

Setting up the domain name and DNS is a very complex task, and that is one of the reasons why universities and research institutes usually have dedicated teams that handle these. In the case that this is not possible, there are plenty of registrars (domain name vendors) that combine both services for reasonable prices. It is worth mentioning that it is possible to register several domain names for one web page. For VMH, we have acquired additional domain names (e.g. `http://vmh.life`) that redirect to the default one. When all is done, a server needs to be set up with the actual content of the web page and the database. Nowadays, it is becoming more common to use virtual machines as servers. This is especially handy for a development/production environment, where a website can be tested in a more realistic scenario.

5.4 Agile Implementation

In 2001, a group of influential software developers published the *Agile for Software Manifesto* [28]. This manifesto advocated for a shift in software development to emphasize rapid

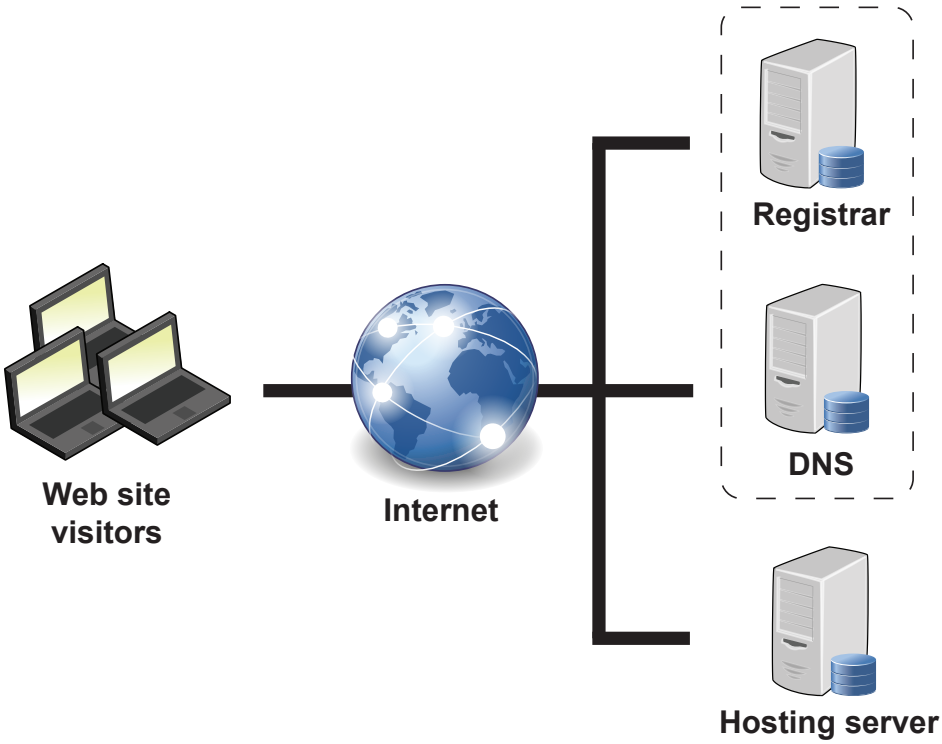

Figure 5.1: Web site access with the domain name depends on 3 services. Domain name registar, DNS name resolving to a physical IP, and the host server.

delivery and response. These methods became particularly attractive for ICT start-ups. These companies are usually composed of small teams and require continuous development on their products based on user feedback. This has also led to the popularization of concepts such as the Lean Startup and Lean Software Development [255, 203] and Scrum [299].

On that same note, a research group can greatly benefit from adopting similar strategies. The development team of a biological database project is often small (or individual) and the interdisciplinary nature of such projects makes it close to impossible to accurately predict the exact needs of the end users. For these reasons, we advocate for an agile development approach focusing on fast release iterations with feedback mechanisms put in place that will allow collecting information on bugs, the incorrectness of information, and suggestions for additional features and their rapid implementation.

Experimentalists need to follow strict protocols for their research to be reproducible and for this reason, we find most researchers to be perfectionists. With agile software development, ideally, one tests and implements changes in a fast manner without much concern on delivering

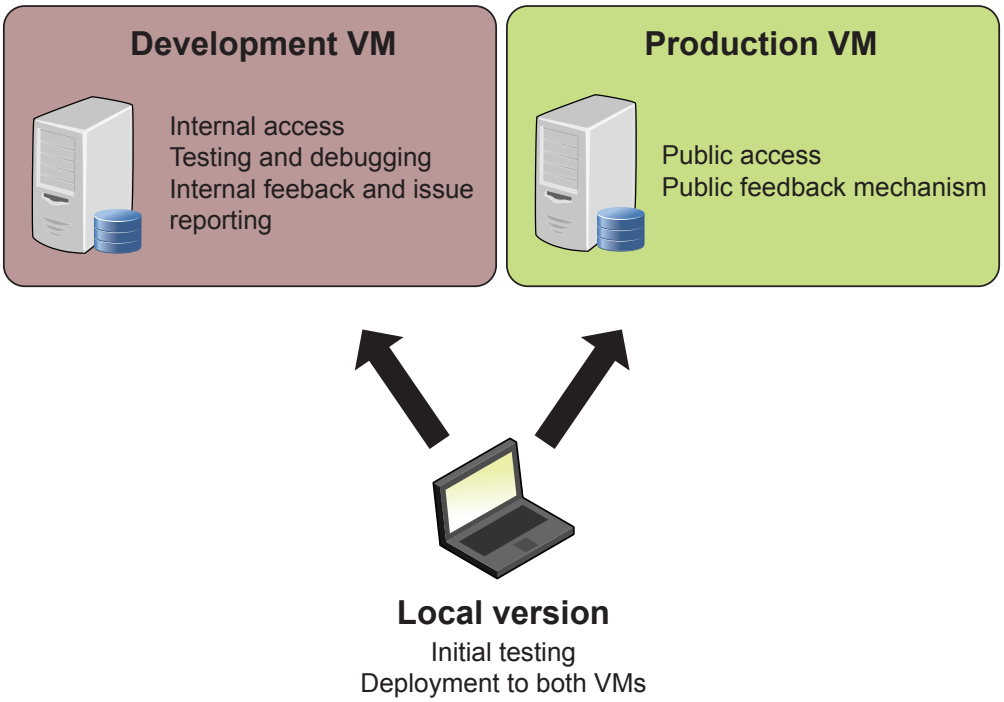

Figure 5.2: Proposed development and productions environments. A development virtual machine hosting an internal version of the website to be tested by the research group or institute. A production virtual machine hosting the public version with a general feedback mechanism.

a finalized product. This means that a compromise between these two somehow opposing views needs to be found. In the development of VMH, we have not released changes to the public environment often. We did, however, started testing the database and website with potential users at a very early stage. To achieve this we set up a development/production environment as shown in Figure 5.2. The development environment is a server running on a VM that hosts an internal version of the database and website available to our research institute through an internal domain name. To collect internal user feedback, we have used our institute's GitLab instance. GitHub and Gitlab are collaborative software development platforms that are based on Git, a version control software.

Each project on GitHub or GitLab has an *Issues* section, where users can report bugs or suggest new features. Another interesting feature is that it is possible to organize the issues in a similar fashion to a Scrum Task Board (Figure 5.3). In this board tasks/issues are organized in three categories: To-Do, Doing, and Closed. This allows the team to better organize and plan their development while keeping users informed on the progress of the development

cycle. For the public version of VMH, we have added a Feedback button to the main page where users can send their suggestions and feedback. Additionally, for ReconMap, we use the MINERVA framework feedback mechanisms that allow users to specify locations in the map and leave comments or report errors.

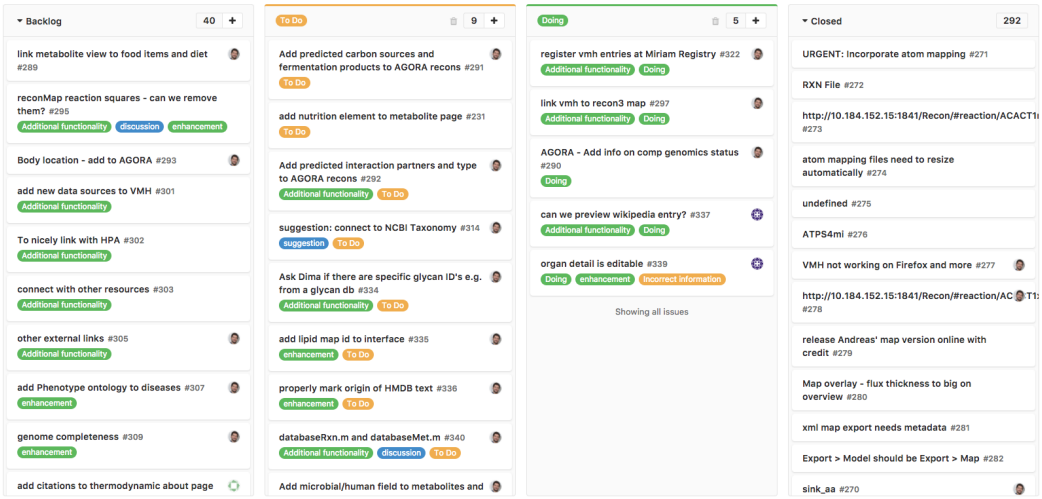

Figure 5.3: Gitlab issue board.

5.5 Discussion

Databases are important tools in the life sciences and biomedicine fields. As new technologies arise and additional data is generated new data resources will become necessary. Developing and maintaining a database is often challenging for research groups due to small teams, lack of proper infrastructure, or the wide range of different skills necessary.

In this Chapter, we have highlighted some of the main technologies and tasks that need to be covered when developing such a project. We have, in some cases, adopted said strategies in the development of the Virtual Metabolic Human database. As such, this Chapter is not intended to be viewed as a rule book but rather a guide that can be a starting point for researchers involved in the development a biological database. We believe that the development strategy of such projects is very dependent on the context. For this reason, we advocate the adoption of agile strategies in the development of software for research purposes. We believe these techniques can bring great benefits and better results in the future.

Chapter 6

Concluding remarks

The increase in the incidence of non-communicable disease (NCDs) is one of the main challenges society and research face nowadays. These diseases are very closely associated with lifestyle, and in particular, with diet. Due to the influence of many factors on the interaction between the human body and ingested nutrients, understanding the mechanisms behind the effect of specific dietary patterns in health is an extremely complex task. Dietary assessment tools and studies of nutrition have inherent limitations that are being addressed with Systems Biology approaches and *omics* technologies.

The usage of *omics* technologies can rapidly and comprehensively measure health-related markers. As discussed in Chapter 1 metabolomics technologies can be used to measure small metabolites and nutrients available in biological fluids (e.g., blood and urine). These technologies have been used for dietary assessment and nutritional recommendation [228]. On another hand, the gut microbiome is also closely associated with dietary patterns [336] and general well being [55]. The composition of these communities can be determined using sequencing technologies (e.g., 16S-RNA, shotgun sequencing) and changes in composition influence how the host processes certain food components. This was shown, for instance, for blood sugar level responses to different foods [345]. Together, these technologies can support the collection of dietary intake data and monitoring of the health status of individuals. More needs to be done, however, to promote the understanding of the mechanisms behind individual responses. Being able to do so, and predict the impact of dietary patterns based on biological fluid measurements, will pave the way for a truly personalized dietary recommendation approach.

Constraint-based reconstruction and analysis (COBRA) uses genome-scale metabolic reconstructions (GENREs), the collection of all known metabolic reactions to occur in an organism, as a basis for the creation of metabolic models that can be used predict the metabolic responses to specific conditions. These models have been used for various applications and can serve as a docking station for data from different sources (e.g. metabolomics). Applying this approach to nutrition presents a unique window to the mechanistic effects of specific dietary components. Together with maps of metabolism and disease, they represent an innovative approach to studying the effect of nutrition on health. The work of this thesis describes the development of a resource that takes the first step in that direction.

Chapters 2 and 3 describe a knowledge base that connects genome-scale metabolic reconstructions of human [39] and a collection of typical gut microbes [195] with nutrition and disease. In addition, we exemplify how connecting the different resources allowed a unique view of metabolism using the different tools made available with VMH. One such example is the *Diet designer*, which allows the integration of nutritional data from food into *in silico* simulations to predict the impact of different dietary compositions. In Chapter 4 we have described the creation of ReconMap and LeighMap, visualization tools that can support researchers. Finally, Chapter 5 discusses some of the challenges and decisions necessary to be taken to perform a project such as described in this thesis, while trying to provide general guidelines that can be of support for researchers involved in similar efforts. Taken together, the work of this thesis demonstrates how COBRA and VMH can be relevant tools and resources in the study of human nutrition and health.

6.1 A knowledge base integrating metabolism, nutrition, and disease information

Metabolism is influenced by genetic and environmental factors. For its study, an integrated analysis of data originating from different fields is necessary. Genome-scale metabolic models provide a framework for this integration and for this reason, we have created the Virtual Metabolic Human (VMH). VMH is a resource that integrates human and gut-microbe metabolic reconstructions with nutritional and disease information. VMH hosts the most

recent version of the human metabolic network reconstruction, Recon 3D, and an high-quality collection of typical human-gut microbial reconstructions of metabolism, the AGORA collection. Each entity of the database has a detailed page with external links that connect with other sources. A collection of diseases is also available and their connections with the human metabolic network. Finally, the Nutrition resource is composed by nutritional information for more than 8000 food items extracted from the USDA food composition database, a set of *in silico* diets, and a *Diet designer* tool that enables the creation of user-defined *in silico* diets.

6.1.1 Biological database development

With the increasing amount of biological databases and analysis tools, interchangeability of data and interaction between applications becomes a central concern. For this purpose, the developers of a biological database should ensure the connection of their database content with other resources and provide tools that allow other researchers to use the knowledge they have compiled. In Chapter 3 we describe the 3-layer architecture of VMH. The core of VMH is its database, which structure is based on the underlying metabolic network as represented in genome-scale reconstructions of metabolism. We have also shown how the resources in VMH are connected based on this structure and how this structure connects with external resources.

The 2 remaining layers are the access points to VMH: it's web interface and the API. The API allows programmatic access to the database, which means that other applications and databases can access these different resources in a customizable way and without the need to download the full database. Based on this combination of tools, we exemplify how VMH can be used to perform complex analysis such as how to explore the complex interactions between microbes, nutrition, and host metabolism. Synthetic microbial communities are developed to mimic the behavior of more complex communities [29, 70] and VMH can be used to screen potential compositions to be used in experiments. We have furthermore, used VMH to study the mechanism of drug detoxification and retoxification, and finally, showed how the disease resource, when combined with the other elements of VMH, can be used as a tool to hypothesize treatment strategies.

Taken together, the tools available with VMH aim at accelerating research and promote interchangeability of knowledge in the field. In that perspective, in Chapter 5 we complement this work with a general guide to biological database development based on the experience acquired during this project.

6.2 Metabolic and disease maps

In this day and age analysis of biological data requires managing large data-sets and advanced statistical analysis. For this reason, visualization of data becomes attractive as it can simplify this analysis by giving a visual context to the data. For this reason, biochemical pathway visualization is of great interest, but due to the complexity of the human metabolic network, no intuitive map with the capabilities of overlaying simulation and experimental data was available. In Chapter 4, we introduce ReconMap, a comprehensive, manually curated map of human metabolism [223]. ReconMap was integrated into MINERVA [100], a tool built on the Google Maps API, that enables interactive overlaying of experimental and simulation data. An extension to the CobraToolbox that allows remote interaction with ReconMap.

Disease maps are gaining relevance in the biomedicine field as they provide visualization of disease mechanisms. Mitochondrial disorders are severe and diverse metabolic diseases for which diagnosis is challenging. We have initiated the effort of mapping Mitochondrial disorders with Leigh Syndrome [250] by developing a gene-to-phenotype map.

The further development of such maps and tools holds potential for combining visualization approaches. It would be of interest to integrate the network visualization of simulation and experimental data with clinical and mechanistic disease information. For instance, associating specific phenotypes from a disease map with flux visualizations from the metabolic network, to correlate clinical features with metabolic states through the integration of experimental data.

6.3 Challenges and the way forward

Studying the effect of specific dietary pattern in health, especially long-term, is a very difficult task. As described in Chapter 1, there are inherent limitations to nutrition assessment tools

and studies. Current efforts in the identification of dietary intake biomarkers are using omics technology, and gut microbiome research is growing rapidly. In my point of view, an approach that manages to integrate these two approaches can promote the understanding of the underlying mechanisms of the effects of diet in health.

While VMH captures, in a unique manner, information for human and gut microbial metabolism and links it to hundreds of diseases and nutritional data, the COBRA approach offers methods and tools to perform analysis and simulations to further study these metabolic reconstructions. The combination and further improvement of these two can offer the means to address some of the limitations of nutrition research.

There are studies that characterized dietary patterns using metabomolics [136, 235], and several changes in the composition of the gut microbiota are associated with dietary patterns [65, 53, 313]. It would be interesting to use VMH and the COBRA approach to simultaneously integrate these complex data. In doing so, one could investigate what changes in the metabolome are caused by diet itself or how they correlate with the specific gut microbiota composition through the creation of community models [196]. An additional layer of complexity can be added by using VMH's resources to design *in silico* diets and predict how the system will respond to different dietary compositions. Being able to characterize these responses, the next step is to predict the effect of specific diets based on biofluid data measurements of an individual, paving the way for a truly personalized dietary recommendation mechanism.

Another promising application would be to understand if this approach could be applied to disease treatment. The metabolism of several drugs is included in VMH and Recon3D [39]. A combination of physiologically based pharmacokinetic (PBPK) and COBRA modeling predicted the positive impact on the efficacy of a drug for Parkinson's Disease treatment if administered with a serine-rich diet [111] and more recently, the usage of the gut microbe models of VMH was used to predict potential treatment strategies for Crohn's Disease [25]. VMH needs to accompany this progression and include additional information that is relevant for these purposes, such as the "Physiological resource" and "Drug resource" discussed in Chapter 2.

For metabolic modeling applications to be further translated to practice, additional validation of this approach must be pursued. Data obtained from nutrition studies using

metabolomics technologies and/or gut microbiome sequencing, such as diet efficacy tests, or nutritional biomarker studies can be used for this purpose. Replicating observations computationally can give mechanistic insights into the studies' results and will foster an improvement of the available models and tools. In addition, *in vitro* modeling technologies that mimic the gut environment are becoming more advanced [204]. These could be a means of validating these approaches by testing the effect of different diets or nutrients and support the creation of strategies for gut microbiota modulation through diet. These validations could then lead to further *in vivo* experiments or clinical trials and an eventual translation of metabolic modeling to healthcare applications.

Bibliography

[foo] Foodball: The food biomarker alliance - http://foodmetabolome.org/.

[2] Abduljalil, K., Furness, P., Johnson, T. N., Rostami-Hodjegan, A., and Soltani, H. (2012). Anatomical, physiological and metabolic changes with gestational age during normal pregnancy. *Clinical pharmacokinetics*, 51(6):365–396.

[3] Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J., and Hebert, J. R. (2005). The effect of social desirability and social approval on self-reports of physical activity. *American journal of epidemiology*, 161(4):389–398.

[4] Agency, E. M. Eu clinical trials register - https://www.clinicaltrialsregister.eu/.

[5] Agency, E. M. European database of suspected adverse drug reaction reports - http://www.adrreports.eu/.

[6] Alkerwi, A., Sauvageot, N., Donneau, A.-F., Lair, M.-L., Couffignal, S., Beissel, J., Delagardelle, C., Wagener, Y., Albert, A., and Guillaume, M. (2010). First nationwide survey on cardiovascular risk factors in grand-duchy of luxembourg (oriscav-lux). *BMC Public Health*, 10(1):468.

[7] Allen, N. E., Grace, P. B., Ginn, A., Travis, R. C., Roddam, A. W., Appleby, P. N., and Key, T. (2008). Phytanic acid: measurement of plasma concentrations by gas–liquid chromatography–mass spectrometry analysis and associations with diet and other plasma fatty acids. *British journal of nutrition*, 99(3):653–659.

[8] Altschul, S. F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *Journal of molecular evolution*, 36(3):290–300.

[9] Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2008). Mckusick's online mendelian inheritance in man (omim®). *Nucleic acids research*, 37(suppl_1):D793–D796.

[10] Andersson, A., Marklund, M., Diana, M., and Landberg, R. (2011). Plasma alkylresorcinol concentrations correlate with whole grain wheat and rye intake and show moderate reproducibility over a 2-to 3-month period in free-living swedish adults. *The Journal of nutrition*, 141(9):1712–1718.

[11] Arab, L., Tseng, C.-H., Ang, A., and Jardack, P. (2011). Validity of a multipass, web-based, 24-hour self-administered recall for assessment of total energy intake in blacks and whites. *American journal of epidemiology*, 174(11):1256–1265.

[12] Arab, L., Wesseling-Perry, K., Jardack, P., Henry, J., and Winter, A. (2010). Eight self-administered 24-hour dietary recalls using the internet are feasible in african americans and whites: the energetics study. *Journal of the American Dietetic Association*, 110(6):857–864.

[13] Argyri, K., Miller, D. D., Glahn, R. P., Zhu, L., and Kapsokefalou, M. (2007). Peptides isolated from in vitro digests of milk enhance iron uptake by caco-2 cells. *Journal of agricultural and food chemistry*, 55(25):10221–10225.

[14] Arkin, A. P., Stevens, R. L., Cottingham, R. W., Maslov, S., Henry, C. S., Dehal, P., Ware, D., Perez, F., Harris, N. L., Canon, S., et al. (2016). The doe systems biology knowledgebase (kbase). *bioRxiv*, page 096354.

[15] Arsenault, L. N., Matthan, N., Scott, T. M., Dallal, G., Lichtenstein, A. H., Folstein, M. F., Rosenberg, I., and Tucker, K. L. (2009). Validity of estimated dietary eicosapentaenoic acid and docosahexaenoic acid intakes determined by interviewer-administered food frequency questionnaire among older adults with mild-to-moderate cognitive impairment or dementia. *American journal of epidemiology*, 170(1):95–103.

[16] Aurich, M. K., Fleming, R. M., and Thiele, I. (2016). Metabotools: A comprehensive toolbox for analysis of genome-scale metabolic models. *Frontiers in physiology*, 7.

[17] Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *science*, 307(5717):1915–1920.

[18] Bais, P., Moon, S. M., He, K., Leitao, R., Dreher, K., Walk, T., Sucaet, Y., Barkan, L., Wohlgemuth, G., Roth, M. R., et al. (2010). Plantmetabolomics. org: a web portal for plant metabolomics experiments. *Plant physiology*, 152(4):1807–1816.

[19] Baldrick, F. R., Woodside, J. V., Elborn, J. S., Young, I. S., and McKinley, M. C. (2011). Biomarkers of fruit and vegetable intake in human intervention studies: a systematic review. *Critical reviews in food science and nutrition*, 51(9):795–815.

[20] Bandini, L. G., Schoeller, D. A., Cyr, H. N., and Dietz, W. H. (1990). Validity of reported energy intake in obese and nonobese adolescents. *The American journal of clinical nutrition*, 52(3):421–425.

[21] Bánhegyi, G., Garzó, T., Antoni, F., and Mandl, J. (1988). Glycogenolysis-and not gluconeogenesis-is the source of udp-glucuronic acid for glucuronidation. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 967(3):429–435.

[22] Barrett, J., Della Casa Alberighi, O., Läer, S., and Meibohm, B. (2012). Physiologically based pharmacokinetic (pbpk) modeling in children. *CliniCAl PhArMACology & TherAPeuTiCS*, 92(1):40–49.

[23] Basel-Vanagaite, L., Muncher, L., Straussberg, R., Pasmanik-Chor, M., Yahav, M., Rainshtein, L., Walsh, C. A., Magal, N., Taub, E., Drasinover, V., et al. (2006). Mutated nup62 causes autosomal recessive infantile bilateral striatal necrosis. *Annals of neurology*, 60(2):214–222.

[24] Bateman, A. (2007). Bioinformatics editorial. section "what makes a good database?". *Nucleic acids research*, 35.

[25] Bauer, E. and Thiele, I. (2017). From metagenomic data to personalized computational microbiotas: Predicting dietary supplements for crohn's disease. *arXiv preprint arXiv:1709.06007*.

[26] Baum, F., Fedorova, M., Ebner, J., Hoffmann, R., and Pischetsrieder, M. (2013). Analysis of the endogenous peptide profile of milk: identification of 248 mainly casein-derived peptides. *Journal of proteome research*, 12(12):5447–5462.

[27] Baumgartner, M. R., Hörster, F., Dionisi-Vici, C., Haliloglu, G., Karall, D., Chapman, K. A., Huemer, M., Hochuli, M., Assoun, M., Ballhausen, D., et al. (2014). Proposed guidelines for the diagnosis and management of methylmalonic and propionic acidemia. *Orphanet journal of rare diseases*, 9(1):130.

[28] Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., et al. (2001). Manifesto for agile software development.

[29] Becker, N., Kunath, J., Loh, G., and Blaut, M. (2011). Human intestinal microbiota: characterization of a simplified and stable gnotobiotic rat model. *Gut Microbes*, 2(1):25–33.

[30] Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., et al. (2014). The chembl bioactivity database: an update. *Nucleic acids research*, 42(D1):D1083–D1090.

[31] Berry, D., Kaplan, J., and Rahman, S. (2017). Probiotic compositions containing clostridiales for inhibiting inflammation. US Patent 9610307B2.

[32] Bingham, S. (1997). Dietary assessments in the european prospective study of diet and cancer (epic). *European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP)*, 6(2):118–124.

[33] Bingham, S., Cassidy, A., Cole, T., Welch, A., Runswick, S., Black, A., Thurnham, D., Bates, C., Khaw, K.-T., Key, T., et al. (1995). Validation of weighed records and other methods of dietary assessment using the 24 h urine nitrogen technique and other biological markers. *British Journal of Nutrition*, 73(4):531–550.

[34] Block, G., Thompson, F., Hartman, A., Larkin, F., and Guire, K. (1992). Comparison of two dietary questionnaires validated against multiple dietary records collected during a 1-year period. *Journal of the American Dietetic Association*, 92(6):686–693.

[35] Blumberg, J., Heaney, R. P., Huncharek, M., Scholl, T., Stampfer, M., Vieth, R., Weaver, C. M., and Zeisel, S. H. (2010). Evidence-based criteria in the nutritional context. *Nutrition reviews*, 68(8):478–484.

[36] Bordbar, A. and Palsson, B. O. (2012). Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *Journal of internal medicine*, 271(2):131–141.

[37] Bourgeois, M., Jacquin, F., Savois, V., Sommerer, N., Labas, V., Henry, C., and Burstin, J. (2009). Dissecting the proteome of pea mature seeds reveals the phenotypic plasticity of seed protein composition. *Proteomics*, 9(2):254–271.

[38] Boushey, C. J., Coulston, A. M., Rock, C. L., and Monsen, E. (2001). *Nutrition in the Prevention and Treatment of Disease*. Academic Press.

[39] Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Aurich, M., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G., Prlić, A., Sastry, A., Danielsdottir, A. D., Heinken, A., Noronha, A., Rose, P. W., Burley, S. K., Fleming, R. M., Nielsen, J., Thiele, I., and Palsson, B. O. (2017). Recon3d: A resource enabling a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology (Accepted)*.

[40] Burdge, G. C. and Lillycrop, K. A. (2010). Nutrition, epigenetics, and developmental plasticity: implications for understanding human disease. *Annual review of nutrition*, 30:315–339.

[41] Buzzard, I. M., Faucett, C. L., Jeffery, R. W., McBANE, L., McGOVERN, P., Baxter, J. S., Shapiro, A. C., Blackburn, G. L., T CHLEBOWSKI, R., Elashoff, R. M., et al. (1996). Monitoring dietary change in a low-fat diet intervention study: advantages of using 24-hour dietary recalls vs food records. *Journal of the American Dietetic Association*, 96(6):574–579.

[42] Cani, P. D., Lecourt, E., Dewulf, E. M., Sohet, F. M., Pachikian, B. D., Naslain, D., De Backer, F., Neyrinck, A. M., and Delzenne, N. M. (2009). Gut microbiota fermentation of prebiotics increases satietogenic and incretin gut peptide production with consequences for appetite sensation and glucose response after a meal. *The American journal of clinical nutrition*, 90(5):1236–1243.

[43] Carroll, R. J., Midthune, D., Subar, A. F., Shumakovich, M., Freedman, L. S., Thompson, F. E., and Kipnis, V. (2012). Taking advantage of the strengths of 2 different dietary assessment instruments to improve intake estimates for nutritional epidemiology. *American journal of epidemiology*, 175(4):340–347.

[44] Casey, P. H., Goolsby, S. L., Lensing, S. Y., Perloff, B. P., and Bogle, M. L. (1999). The use of telephone interview methodology to obtain 24-hour dietary recalls. *Journal of the American Dietetic Association*, 99(11):1406–1411.

[45] Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., et al. (2007). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 36(suppl_1):D623–D631.

[46] Chandan, R. (2011). Sencha/extjs: Object oriented javascript - `https://www.sencha.com`.

[47] Chang, R. L., Xie, L., Xie, L., Bourne, P. E., and Palsson, B. Ø. (2010). Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS computational biology*, 6(9):e1000938.

[48] Cheatham, C. L., Goldman, B. D., Fischer, L. M., da Costa, K.-A., Reznick, J. S., and Zeisel, S. H. (2012). Phosphatidylcholine supplementation in pregnant women consuming moderate-choline diets does not enhance infant cognitive function: a randomized, double-blind, placebo-controlled trial. *The American journal of clinical nutrition*, 96(6):1465–1472.

[49] Chen, P. P.-S. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, 1(1):9–36.

[50] Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y., Chen, R., Miriami, E., Karczewski, K. J., Hariharan, M., Dewey, F. E., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307.

[51] Christie, T. (2011). Django rest framework - `http://www.django-rest-framework.org/`.

[52] Cianferoni, A. and Spergel, J. M. (2009). Food allergy: review, classification and diagnosis. *Allergology International*, 58(4):457–466.

[53] Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., Harris, H. M., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., et al. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410):178–184.

[54] Clarke, R., Halsey, J., Lewington, S., Lonn, E., Armitage, J., Manson, J. E., Bønaa, K. H., Spence, J. D., Nygård, O., Jamison, R., et al. (2010). Effects of lowering homocysteine levels with b vitamins on cardiovascular disease, cancer, and cause-specific mortality: meta-analysis of 8 randomized trials involving 37 485 individuals. *Archives of internal medicine*, 170(18):1622–1631.

[55] Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):1258–1270.

[56] Conlee, R. K., Lawler, R. M., and Ross, P. E. (1987). Effects of glucose or fructose feeding on glycogen repletion in muscle and liver after exercise or fasting. *Annals of nutrition and metabolism*, 31(2):126–132.

[57] Consortium, G. O. et al. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261.

[58] Consortium, H. M. P. et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

[59] Consortium, U. et al. (2014). Uniprot: a hub for protein information. *Nucleic acids research*, page gku989.

[60] Corella, D., Carrasco, P., Sorlí, J. V., Estruch, R., Rico-Sanz, J., Martínez-González, M. Á., Salas-Salvadó, J., Covas, M. I., Coltell, O., Arós, F., et al. (2013). Mediterranean diet reduces the adverse effect of the tcf7l2-rs7903146 polymorphism on cardiovascular risk factors and stroke incidence. *Diabetes Care*, 36(11):3803–3811.

[61] CUMMINGS, S. R., BLOCK, G., McHENRY, K., and BARON, R. B. (1987). Evaluation of two food frequency methods of measuring dietary calcium intake. *American Journal of Epidemiology*, 126(5):796–802.

[62] Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2015. *Nucleic acids research*, 43(D1):D662–D669.

[63] da Veiga Leprevost, F., Barbosa, V. C., Francisco, E. L., Perez-Riverol, Y., and Carvalho, P. C. (2014). On best practices in the development of bioinformatics software. *Frontiers in genetics*, 5.

[64] David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563.

[65] De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from europe and rural africa. *Proceedings of the National Academy of Sciences*, 107(33):14691–14696.

[66] De Lorgeril, M., Salen, P., Martin, J.-L., Monjaud, I., Delaye, J., and Mamelle, N. (1999). Mediterranean diet, traditional risk factors, and the rate of cardiovascular complications after myocardial infarction. *Circulation*, 99(6):779–785.

[67] de Oliveira, F. P., Mendes, R. H., Dobbler, P. T., Mai, V., Pylro, V. S., Waugh, S. G., Vairo, F., Refosco, L. F., Roesch, L. F. W., and Schwartz, I. V. D. (2016). Phenylketonuria and gut microbiota: A controlled study based on next-generation sequencing. *PloS one*, 11(6):e0157513.

[68] Décombaz, J., Jentjens, R., Ith, M., Scheurer, E., Buehler, T., Jeukendrup, A., and Boesch, C. (2011). Fructose and galactose enhance postexercise human liver glycogen synthesis. *Medicine and science in sports and exercise*, 43(10):1964–1971.

[69] Delage, B. and Dashwood, R. H. (2008). Dietary manipulation of histone structure and function. *Annu. Rev. Nutr.*, 28:347–366.

[70] Desai, M. S., Seekatz, A. M., Koropatkin, N. M., Kamada, N., Hickey, C. A., Wolter, M., Pudlo, N. A., Kitamoto, S., Terrapon, N., Muller, A., et al. (2016). A dietary fiber-deprived gut microbiota degrades the colonic mucus barrier and enhances pathogen susceptibility. *Cell*, 167(5):1339–1353.

[71] Development Initiatives (2017). *Global Nutrition Report 2017: Nourishing the SDGs.* Development Initiatives.

[72] Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A. A., and Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the seed and model seed. *Systems Metabolic Engineering: Methods and Protocols*, pages 17–45.

[73] Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782.

[74] Duncan, S. H., Belenguer, A., Holtrop, G., Johnstone, A. M., Flint, H. J., and Lobley, G. E. (2007). Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Applied and environmental microbiology*, 73(4):1073–1078.

[75] Duncan, S. H., Scott, K. P., Ramsay, A. G., Harmsen, H. J., Welling, G. W., Stewart, C. S., and Flint, H. J. (2003). Effects of alternative dietary substrates on competition between human colonic bacteria in an anaerobic fermentor system. *Applied and environmental microbiology*, 69(2):1136–1142.

[76] Durrer, K. E., Allen, M. S., and von Herbing, I. H. (2017). Genetically engineered probiotic for the treatment of phenylketonuria (pku); assessment of a novel treatment in vitro and in the pahenu2 mouse model of pku. *PloS one*, 12(5):e0176286.

[77] Elmadfa, I. (2012). Österreichischer ernährungsbericht 2012. 1. auflage, wien. *Zugriff am*, 27:2013.

[78] Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.-I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J., et al. (2013). Primary prevention of cardiovascular disease with a mediterranean diet. *New England Journal of Medicine*, 368(14):1279–1290.

[79] Fang, M., Chen, D., and Yang, C. S. (2007). Dietary polyphenols may affect dna methylation. *The Journal of nutrition*, 137(1):223S–228S.

[80] Farnaud, S. and Evans, R. W. (2003). Lactoferrin—a multifunctional protein with antimicrobial properties. *Molecular immunology*, 40(7):395–405.

[81] Fassone, E., Wedatilake, Y., DeVile, C. J., Chong, W. K., Carr, L. J., and Rahman, S. (2013). Treatable leigh-like encephalopathy presenting in adolescence. *BMJ case reports*, 2013:bcr2013200838.

[82] Ferdinandusse, S., Waterham, H. R., Heales, S. J., Brown, G. K., Hargreaves, I. P., Taanman, J.-W., Gunny, R., Abulhoul, L., Wanders, R. J., Clayton, P. T., et al. (2013). Hibch mutations can cause leigh-like disease with combined deficiency of multiple mitochondrial respiratory chain enzymes and pyruvate dehydrogenase. *Orphanet journal of rare diseases*, 8(1):188.

[83] Fielding, R. T. and Taylor, R. N. (2000). *Architectural styles and the design of network-based software architectures*. University of California, Irvine Doctoral dissertation.

[84] FLEGAL, K. M. and LARKIN, F. A. (1990). Partitioning macronutrient intake estimates from a food frequency questionnaire. *American journal of epidemiology*, 131(6):1046–1058.

[85] Fleming, R. M., Vlassis, N., Thiele, I., and Saunders, M. A. (2016). Conditions for duality between fluxes and concentrations in biochemical networks. *Journal of theoretical biology*, 409:1–10.

[86] Floyd, B. J., Wilkerson, E. M., Veling, M. T., Minogue, C. E., Xia, C., Beebe, E. T., Wrobel, R. L., Cho, H., Kremer, L. S., Alston, C. L., et al. (2016). Mitochondrial protein interaction mapping identifies regulators of respiratory chain function. *Molecular cell*, 63(4):621–632.

[87] Forouzanfar, M. H., Afshin, A., Alexander, L. T., Aasvang, G. M., Bjertness, E., Htet, A. S., Savic, M., Vollset, S. E., Norheim, O. F., and Weiderpass, E. (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the global burden of disease study 2015. *The Lancet*.

[88] Forster, H., Fallaize, R., Gallagher, C., O'Donovan, C. B., Woolhead, C., Walsh, M. C., Macready, A. L., Lovegrove, J. A., Mathers, J. C., Gibney, M. J., et al. (2014). Online dietary intake estimation: the food4me food frequency questionnaire. *Journal of medical Internet research*, 16(6).

[89] Foundation, D. S. (205). Django: The web framework for perfectionists with deadlines - `https://www.djangoproject.com/`.

[90] Frank, D. N., Amand, A. L. S., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences*, 104(34):13780–13785.

[91] Fraser, G. E. (2003). A search for truth in dietary epidemiology. *The American journal of clinical nutrition*, 78(3):521S–525S.

[92] Fuchs, D., Erhard, P., Rimbach, G., Daniel, H., and Wenzel, U. (2005). Genistein blocks homocysteine-induced alterations in the proteome of human endothelial cells. *Proteomics*, 5(11):2808–2818.

[93] Fujita, K. A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., Crespo, I., Perumal, T. M., Jurkowski, W., Antony, P. M., et al. (2014). Integrating pathways of parkinson's disease in a molecular interaction map. *Molecular neurobiology*, 49(1):88–102.

[94] Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). Celldesigner 3.5: a versatile modeling tool for biochemical networks. *Proceedings of the IEEE*, 96(8):1254–1265.

[95] Furusawa, Y., Obata, Y., Fukuda, S., Endo, T. A., Nakato, G., Takahashi, D., Nakanishi, Y., Uetake, C., Kato, K., Kato, T., et al. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory t cells. *Nature*, 504(7480):446–450.

[96] Galas, D. J. and McCormack, S. J. (2003). An historical perspective on genomic technologies. *Curr Issues Mol Biol*, 5(4):123–127.

[97] Galperin, M. Y., Fernández-Suárez, X. M., and Rigden, D. J. (2017). The 24th annual nucleic acids research database issue: a look back and upcoming changes. *Nucleic acids research*, 45(D1):D1–D11.

[98] Gao, L., Wang, A., Li, X., Dong, K., Wang, K., Appels, R., Ma, W., and Yan, Y. (2009). Wheat quality related differential expressions of albumins and globulins revealed by two-dimensional difference gel electrophoresis (2-d dige). *Journal of proteomics*, 73(2):279–296.

[99] Gareau, M. G., Sherman, P. M., and Walker, W. A. (2010). Probiotics and the gut microbiota in intestinal health and disease. *Nature Reviews Gastroenterology and Hepatology*, 7(9):503–514.

[100] Gawron, P., Ostaszewski, M., Satagopam, V., Gebel, S., Mazein, A., Kuzma, M., Zorzan, S., McGee, F., Otjacques, B., Balling, R., et al. (2016). Minerva—a platform for visualization and curation of molecular interaction networks. *npj Systems Biology and Applications*, 2:16020.

[101] Gersovitz, M., Madden, J. P., and Smiciklas-Wright, H. (1978). Validity of the 24-hr. dietary recall and seven-day record for group comparisons. *Journal of the American Dietetic Association*, 73(1):48–55.

[102] Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., Duperier, C., Tremblay-Franco, M., Martin, J.-F., Jacob, D., et al. (2014). Workflow4metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics*, 31(9):1493–1495.

[103] Gibbons, H. and Brennan, L. (2017). Metabolomics as a tool in the identification of dietary biomarkers. *Proceedings of the Nutrition Society*, 76(1):42–53.

[104] Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *science*, 312(5778):1355–1359.

[105] Godfrey, K. M., Gluckman, P. D., and Hanson, M. A. (2010). Developmental origins of metabolic disease: life course and intergenerational perspectives. *Trends in Endocrinology & Metabolism*, 21(4):199–205.

[106] Gonzalez, G. A. P., El Assal, L. R., Noronha, A., Thiele, I., Haraldsdóttir, H. S., and Fleming, R. M. (2017). Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to recon 3d. *Journal of Cheminformatics*, 9(1):39.

[107] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

[108] Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., and Bruford, E. A. (2014). Genenames. org: the hgnc resources in 2015. *Nucleic acids research*, 43(D1):D1079–D1085.

[109] Green, M. and Karp, P. (2005). Genome annotation errors in pathway databases due to semantic ambiguity in partial ec numbers. *Nucleic acids research*, 33(13):4035–4039.

[110] Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC bioinformatics*, 11(1):489.

[111] Guebila, M. B. and Thiele, I. (2016). Model-based dietary optimization for late-stage, levodopa-treated, parkinson's disease patients. *NPJ Systems Biology and Applications*, 2:16013.

[112] Guerrero, A., Dallas, D. C., Contreras, S., Chee, S., Parker, E. A., Sun, X., Dimapasoc, L., Barile, D., German, J. B., and Lebrilla, C. B. (2014). Mechanistic peptidomics: factors that dictate specificity in the formation of endogenous peptides in human milk. *Molecular & Cellular Proteomics*, 13(12):3343–3351.

[113] Guo, A. C., Jewison, T., Wilson, M., Liu, Y., Knox, C., Djoumbou, Y., Lo, P., Mandal, R., Krishnamurthy, R., and Wishart, D. S. (2012). Ecmdb: the e. coli metabolome database. *Nucleic acids research*, 41(D1):D625–D630.

[114] Haiser, H. J. and Turnbaugh, P. J. (2013). Developing a metagenomic view of xenobiotic metabolism. *Pharmacological research*, 69(1):21–31.

[115] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517.

[116] Haraldsdóttir, H. S., Cousins, B., Thiele, I., Fleming, R. M., and Vempala, S. (2017). Chrr: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743.

[117] Haraldsdóttir, H. S. and Fleming, R. M. (2016). Identification of conserved moieties in metabolic networks by graph theoretical analysis of atom transition networks. *PLoS computational biology*, 12(11):e1004999.

[118] Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., et al. (2012). The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463.

[119] Haubrock, J., Nöthlings, U., Volatier, J.-L., Dekkers, A., Ocké, M., Harttig, U., Illner, A.-K., Knüppel, S., Andersen, L. F., Boeing, H., et al. (2011). Estimating usual food intake distributions by using the multiple source method in the epic-potsdam calibration study. *The Journal of nutrition*, 141(5):914–920.

[120] Hauser, A.-T. and Jung, M. (2008). Targeting epigenetic mechanisms: potential of natural products in cancer chemoprevention. *Planta medica*, 74(13):1593–1601.

[121] Heady, J. A. (1961). Development of a method of classifying the diets of individuals for use in epidemiological studies. *J. R. Stat. Soc. Ser.*, 124:336–371.

[122] Hebert, J. R., Clemow, L., Pbert, L., Ockene, I. S., and Ockene, J. K. (1995). Social desirability bias in dietary self-report may compromise the validity of dietary intake measures. *International journal of epidemiology*, 24(2):389–398.

[123] Hebert, J. R., Ebbeling, C. B., Matthews, C. E., Hurley, T. G., Yunsheng, M., Druker, S., and Clemow, L. (2002). Systematic errors in middle-aged women's estimates of energy intake: comparing three self-report measures to total energy expenditure from doubly labeled water. *Annals of epidemiology*, 12(8):577–586.

[124] Hébert, J. R., Frongillo, E. A., Adams, S. A., Turner-McGrievy, G. M., Hurley, T. G., Miller, D. R., and Ockene, I. S. (2016). Perspective: Randomized controlled trials are not a panacea for diet-related research. *Advances in Nutrition: An International Review Journal*, 7(3):423–432.

[125] Hebert, J. R., Hurley, T. G., Peterson, K. E., Resnicow, K., Thompson, F. E., Yaroch, A. L., Ehlers, M., Midthune, D., Williams, G. C., Greene, G. W., et al. (2008). Social desirability trait influences on self-reported dietary measures among diverse participants in a multicenter multiple risk factor trial. *The Journal of nutrition*, 138(1):226S–234S.

[126] Hebert, J. R., Ma, Y., Clemow, L., Ockene, I. S., Saperia, G., Stanek III, E. J., Merriam, P. A., and Ockene, J. K. (1997). Gender differences in social desirability and social approval bias in dietary self-report. *American journal of epidemiology*, 146(12):1046–1055.

[127] Hébert, J. R., Peterson, K. E., Hurley, T. G., Stoddard, A. M., Cohen, N., Field, A. E., and Sorensen, G. (2001). The effect of social desirability trait on self-reported dietary measures among multi-ethnic female health center employees. *Annals of epidemiology*, 11(6):417–427.

[128] Heinken, A., Sahoo, S., Fleming, R. M., and Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut microbes*, 4(1):28–40.

[129] Heinzmann, S. S., Merrifield, C. A., Rezzi, S., Kochhar, S., Lindon, J. C., Holmes, E., and Nicholson, J. K. (2011). Stability and robustness of human metabolic phenotypes in response to sequential food challenges. *Journal of proteome research*, 11(2):643–655.

[130] Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdottir, H. S., Keating, S. M., Vlasov, V., Wachowiak, J., et al. (2017). Creation and analysis of biochemical constraint-based models: the cobra toolbox v3. 0. *arXiv preprint arXiv:1710.04038*.

[131] Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982.

[132] Hodgson, J. M., Chan, S. Y., Puddey, I. B., Devine, A., Wattanapenpaiboon, N., Wahlqvist, M. L., Lukito, W., Burke, V., Ward, N. C., Prince, R. L., et al. (2004). Phenolic acid metabolites as biomarkers for tea-and coffee-derived polyphenol exposure in human subjects. *British journal of nutrition*, 91(2):301–305.

[133] Hoffman, D., Lowenstein, H., Marsh, D. G., Platts-Mills, T. A., Thomas, W., et al. (1994). Allergen nomenclature. *International archives of allergy and immunology*, 105(3):224–233.

[134] Hoffmann, R. (2008). A wiki for the life sciences where authorship matters. *Nature genetics*, 40(9):1047–1051.

[135] Holle, R., Happich, M., Löwel, H., Wichmann, H., study group, M., et al. (2005). Kora-a research platform for population based health research. *Das Gesundheitswesen*, 67(S 01):19–25.

[136] Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K., Chan, Q., Ebbels, T., De Iorio, M., Brown, I. J., Veselkov, K. A., et al. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*, 453(7193):396–400.

[137] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714.

[138] Howell, S., Hazelton, G. A., and Klaassen, C. D. (1986). Depletion of hepatic udp-glucuronic acid by drugs that are glucuronidated. *Journal of Pharmacology and Experimental Therapeutics*, 236(3):610–614.

[139] Huan, T., Forsberg, E. M., Rinehart, D., Johnson, C. H., Ivanisevic, J., Benton, H. P., Fang, M., Aisporna, A., Hilmers, B., Poole, F. L., et al. (2017). Systems biology guided by xcms online metabolomics. *Nature Methods*, 14(5):461–462.

[140] Huan, T., Tang, C., Li, R., Shi, Y., Lin, G., and Li, L. (2015). Mycompoundid ms/ms search: Metabolite identification using a library of predicted fragment-ion-spectra of 383,830 possible human metabolites. *Analytical chemistry*, 87(20):10619–10626.

[141] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., et al. (2003). The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

[142] Hummel, J., Selbig, J., Walther, D., and Kopka, J. (2007). The golm metabolome database: a database for gc-ms based metabolite profiling. *Metabolomics*, pages 75–95.

[143] Humphrey, L. L., Fu, R., Rogers, K., Freeman, M., and Helfand, M. (2008). Homocysteine level and coronary heart disease incidence: a systematic review and meta-analysis. In *Mayo Clinic Proceedings*, volume 83, pages 1203–1212. Elsevier.

[144] Hyduke, D. R., Lewis, N. E., and Palsson, B. Ø. (2013). Analysis of omics data with genome-scale models of metabolism. *Molecular BioSystems*, 9(2):167–174.

[145] Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B. S., Mewes, H.-W., et al. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nature genetics*, 42(2):137–141.

[146] Ioannidis, J. P. (2016). We need more randomized trials in nutrition—preferably large, long-term, and with negative results. *The American journal of clinical nutrition*, 103(6):1385–1386.

[147] Jackson, A. A., Burdge, G. C., and Lillycrop, K. A. (2010). Diet, nutrition and modulation of genomic expression in fetal origins of adult disease. In *Personalized Nutrition*, volume 101, pages 56–72. Karger Publishers.

[148] Jackson, R. D., LaCroix, A. Z., Gass, M., Wallace, R. B., Robbins, J., Lewis, C. E., Bassford, T., Beresford, S. A., Black, H. R., Blanchette, P., et al. (2006). Calcium plus vitamin d supplementation and the risk of fractures. *New England Journal of Medicine*, 354(7):669–683.

[149] Jensen, P. A. and Papin, J. A. (2014). Metdraw: automated visualization of genome-scale metabolic network reconstructions and high-throughput data. *Bioinformatics*, 30(9):1327–1328.

[150] Johnson, C. H., Ivanisevic, J., and Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7):451–459.

[151] Juty, N., Le Novère, N., and Laibe, C. (2011). Identifiers. org and miriam registry: community resources to provide persistent identification. *Nucleic acids research*, 40(D1):D580–D586.

[152] Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

[153] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2013). Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205.

[154] Kaput, J. (2004). Diet–disease gene interactions. *Nutrition*, 20(1):26–31.

[155] Kaput, J. (2008). Nutrigenomics research for personalized nutrition and medicine. *Current opinion in biotechnology*, 19(2):110–120.

[156] Kaput, J., Klein, K. G., Reyes, E. J., Kibbe, W. A., Cooney, C. A., Jovanovic, B., Visek, W. J., and Wolff, G. L. (2004). Identification of genes contributing to the obese yellow avy phenotype: caloric restriction, genotype, diet× genotype interactions. *Physiological genomics*, 18(3):316–324.

[157] Kaput, J., Noble, J., Hatipoglu, B., Kohrs, K., Dawson, K., and Bartholomew, A. (2007a). Application of nutrigenomic concepts to type 2 diabetes mellitus. *Nutrition, metabolism and cardiovascular diseases*, 17(2):89–103.

[158] Kaput, J., Perlina, A., Hatipoglu, B., Bartholomew, A., and Nikolsky, Y. (2007b). Nutrigenomics: concepts and applications to pharmacogenomics and clinical medicine. *Pharmacogenomics*.

[159] Kaput, J., Swartz, D., Paisley, E., Mangian, H., Daniel, W. L., and Visek, W. J. (1994). Diet-disease interactions at the molecular level: an experimental paradigm. *The Journal of nutrition*, 124(8 Suppl):1296S–1305S.

[160] Karlic, H., Thaler, R., Gerner, C., Grunt, T., Proestling, K., Haider, F., and Varga, F. (2015). Inhibition of the mevalonate pathway affects epigenetic regulation in cancer cells. *Cancer genetics*, 208(5):241–252.

[161] Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., and López-Bigas, N. (2005). Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic acids research*, 33(19):6083–6089.

[162] Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. (2015). Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

[163] King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015a). Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS computational biology*, 11(8):e1004321.

[164] King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. (2015b). Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522.

[165] Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J., and Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65(4):1003–1010.

[166] Kirk, H., Cefalu, W. T., Ribnicky, D., Liu, Z., and Eilertsen, K. J. (2008). Botanicals as epigenetic modulators for mechanisms contributing to development of metabolic syndrome. *Metabolism*, 57:S16–S23.

[167] Kirkpatrick, S. I., Subar, A. F., Douglass, D., Zimmerman, T. P., Thompson, F. E., Kahle, L. L., George, S. M., Dodd, K. W., and Potischman, N. (2014). Performance of the automated self-administered 24-hour recall relative to a measure of true intakes and to an interviewer-administered 24-h recall. *The American journal of clinical nutrition*, 100(1):233–240.

[168] Kitts, D. D. and Weiler, K. (2003). Bioactive proteins and peptides from food sources. applications of bioprocesses used in isolation and recovery. *Current pharmaceutical design*, 9(16):1309–1323.

[169] Koeberl, M., Clarke, D., and Lopata, A. L. (2014). Next generation of food allergen quantification using mass spectrometric systems. *Journal of proteome research*, 13(8):3499–3509.

[170] Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C., Brown, D. L., Brudno, M., Campbell, J., et al. (2013). The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(D1):D966–D974.

[171] Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464.

[172] Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., Dörmann, P., Weckwerth, W., Gibon, Y., Stitt, M., et al. (2004). Gmd@ csb. db: the golm metabolome database. *Bioinformatics*, 21(8):1635–1638.

[173] Korem, T., Zeevi, D., Zmora, N., Weissbrod, O., Bar, N., Lotan-Pompan, M., Avnit-Sagi, T., Kosower, N., Malka, G., Rein, M., et al. (2017). Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. *Cell Metabolism*, 25(6):1243–1253.

[174] Kotiranta, A., Lounatmaa, K., and Haapasalo, M. (2000). Epidemiology and pathogenesis of bacillus cereus infections. *Microbes and infection*, 2(2):189–198.

[175] Krug, S., Kastenmüller, G., Stückler, F., Rist, M. J., Skurk, T., Sailer, M., Raffler, J., Römisch-Margl, W., Adamski, J., Prehn, C., et al. (2012). The dynamic range of the human metabolome revealed by challenges. *The FASEB Journal*, 26(6):2607–2619.

[176] Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2015). The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079.

[177] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97.

[178] Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

[179] Kundu, P., Blacher, E., Elinav, E., and Pettersson, S. (2017). Our gut microbiome: The evolving inner self. *Cell*, 171(7):1481–14903.

[180] Kussmann, M., Panchaud, A., and Affolter, M. (2010). Proteomics in nutrition: status quo and outlook for biomarkers and bioactives. *Journal of proteome research*, 9(10):4876–4887.

[181] Lafay, L., Mennen, L., Basdevant, A., Charles, M., Borys, J., Eschwege, E., and Romon, M. (2000). Does energy intake underreporting involve all kinds of food or only specific food items? results from the fleurbaix laventie ville sante (flvs) study. *International Journal of Obesity & Related Metabolic Disorders*, 24(11).

[182] Lake, N. J., Compton, A. G., Rahman, S., and Thorburn, D. R. (2016). Leigh syndrome: one disorder, more than 75 monogenic causes. *Annals of neurology*, 79(2):190–203.

[183] Leigh, D. (1951). Subacute necrotizing encephalomyelopathy in an infant. *Journal of neurology, neurosurgery, and psychiatry*, 14(3):216.

[184] Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291–305.

[185] Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature*, 444(7122):1022–1023.

[186] Li, X., Gianoulis, T. A., Yip, K. Y., Gerstein, M., and Snyder, M. (2010). Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. *Cell*, 143(4):639–650.

[187] Lightowlers, R. N., Taylor, R. W., and Turnbull, D. M. (2015). Mutations causing mitochondrial disease: What is new and what challenges remain? *Science*, 349(6255):1494–1499.

[188] Link, A., Balaguer, F., and Goel, A. (2010). Cancer chemoprevention by dietary polyphenols: promising role for epigenetics. *Biochemical pharmacology*, 80(12):1771–1792.

[189] List, M., Ebert, P., and Albrecht, F. (2017). Ten simple rules for developing usable software in computational biology. *PLoS computational biology*, 13(1):e1005265.

[190] Livingston, J. H., Lin, J.-P., Dale, R. C., Gill, D., Brogan, P., Munnich, A., Kurian, M. A., Gonzalez-Martinez, V., De Goede, C. G., Falconer, A., et al. (2013). A type i interferon signature identifies bilateral striatal necrosis due to mutations in adar1. *Journal of medical genetics*, pages jmedgenet–2013.

[191] Lloyd, A. J., Beckmann, M., Favé, G., Mathers, J. C., and Draper, J. (2011). Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. *British Journal of Nutrition*, 106(6):812–824.

[192] Losonczy, K. G., Harris, T. B., and Havlik, R. J. (1996). Vitamin e and vitamin c supplement use and risk of all-cause and coronary heart disease mortality in older persons: the established populations for epidemiologic studies of the elderly. *The American journal of clinical nutrition*, 64(2):190–196.

[193] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

[194] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58.

[195] Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., et al. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1):81.

[196] Magnúsdóttir, S. and Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Current opinion in biotechnology*, 51:90–96.

[197] Maki, K. C., Slavin, J. L., Rains, T. M., and Kris-Etherton, P. M. (2014). Limitations of observational evidence: implications for evidence-based dietary recommendations. *Advances in Nutrition: An International Review Journal*, 5(1):7–15.

[198] Mares-Perlman, J. A., Brady, W. E., Klein, B. E., Klein, R., Haus, G. J., Palta, M., Ritter, L. L., and Shoff, S. M. (1995). Diet and nuclear lens opacities. *American journal of epidemiology*, 141(4):322–334.

[199] Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., et al. (2011). Img: the integrated microbial genomes database and comparative analysis system. *Nucleic acids research*, 40(D1):D115–D122.

[200] Marshall, T. A., Gilmore, J. M. E., Broffitt, B., Levy, S. M., and Stumbo, P. J. (2003). Relative validation of a beverage frequency questionnaire in children ages 6 months through 5 years using 3-day food and beverage diaries. *Journal of the American Dietetic Association*, 103(6):714–720.

[201] Matar, C., Amiot, J., Savoie, L., and Goulet, J. (1996). The effect of milk fermentation by lactobacillus helveticus on the release of peptides during in vitro digestion. *Journal of Dairy Science*, 79(6):971–979.

[202] Maurice, C. F., Haiser, H. J., and Turnbaugh, P. J. (2013). Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell*, 152(1):39–50.

[203] Maurya, A. (2012). *Running lean: iterate from plan A to a plan that works*. " O'Reilly Media, Inc.".

[204] May, S., Evans, S., and Parry, L. (2017). Organoids, organs-on-chips and other systems, and microbiota. *Emerging Topics in Life Sciences*, 1(4):385–400.

[205] McCarthy, M. I. (2010). Genomics, type 2 diabetes, and obesity. *New England Journal of Medicine*, 363(24):2339–2350.

[206] Mennen, L. I., Sapinho, D., Ito, H., Bertrais, S., Galan, P., Hercberg, S., and Scalbert, A. (2006). Urinary flavonoids and phenolic acids as biomarkers of intake for polyphenol-rich foods. *British Journal of Nutrition*, 96(1):191–198.

[207] Mezgec, S. and Koroušić Seljak, B. (2017). Nutrinet: A deep learning food and drink image recognition system for dietary assessment. *Nutrients*, 9(7):657.

[208] Miller, R. L. and Ho, S.-m. (2008). Environmental epigenetics and asthma: current concepts and call for studies. *American journal of respiratory and critical care medicine*, 177(6):567–573.

[209] Molag, M. L., de Vries, J. H., Ocké, M. C., Dagnelie, P. C., van den Brandt, P. A., Jansen, M. C., van Staveren, W. A., and van't Veer, P. (2007). Design characteristics of food frequency questionnaires in relation to their validity. *American journal of epidemiology*, 166(12):1468–1478.

[210] Mollstam, B. and Connolly, E. (2005). Product containing lactobacillus reuteri strain attc pta-4965 or pta-4964 for inhibiting bacteria causing dental caries. US Patent 6,872,565.

[211] Monsen, E. R. and Van Horn, L. (2007). *Successful approaches*. American Dietetic Associati.

[212] Moya, A. and Ferrer, M. (2016). Functional redundancy-induced stability of gut microbiota subjected to disturbance. *Trends in microbiology*, 24(5):402–413.

[213] Moyer, M. W. (2014). Vitamins on trial. *Nature*, 510(7506):462.

[214] Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., Fontana, L., Henrissat, B., Knight, R., and Gordon, J. I. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, 332(6032):970–974.

[215] Myint, T., Fraser, G. E., Lindsted, K. D., Knutsen, S. F., Hubbard, R. W., and Bennett, H. W. (2000). Urinary 1-methylhistidine is a marker of meat consumption in black and in white california seventh-day adventists. *American journal of epidemiology*, 152(8):752–755.

[216] Nakahata, Y., Kaluzova, M., Grimaldi, B., Sahar, S., Hirayama, J., Chen, D., Guarente, L. P., and Sassone-Corsi, P. (2008). The nad+-dependent deacetylase sirt1 modulates clock-mediated chromatin remodeling and circadian control. *Cell*, 134(2):329–340.

[217] Nebert, D. W., Zhang, G., and Vesell, E. S. (2008). From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. *Drug metabolism reviews*, 40(2):187–224.

[218] Neuhouser, M. L., Tinker, L., Shaw, P. A., Schoeller, D., Bingham, S. A., Horn, L. V., Beresford, S. A., Caan, B., Thomson, C., Satterfield, S., et al. (2008). Use of recovery biomarkers to calibrate nutrient consumption self-reports in the women's health initiative. *American journal of epidemiology*, 167(10):1247–1259.

[219] Nicholson, J. K., Lindon, J. C., and Holmes, E. (1999). 'metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data. *Xenobiotica*, 29(11):1181–1189.

[220] Nickerson, K. P., Chanin, R., and McDonald, C. (2015). Deregulation of intestinal anti-microbial defense by the dietary additive, maltodextrin. *Gut microbes*, 6(1):78–83.

[221] Nickerson, K. P. and McDonald, C. (2012). Crohn's disease-associated adherent-invasive escherichia coli adhesion is enhanced by exposure to the ubiquitous dietary polysaccharide maltodextrin. *PLoS One*, 7(12):e52132.

[222] Noor, E., Haraldsdóttir, H. S., Milo, R., and Fleming, R. M. (2013). Consistent estimation of gibbs energy using component contributions. *PLoS computational biology*, 9(7):e1003098.

[223] Noronha, A., Daníelsdóttir, A. D., Gawron, P., Jóhannsson, F., Jónsdóttir, S., Jarlsson, S., Gunnarsson, J. P., Brynjólfsson, S., Schneider, R., Thiele, I., et al. (2016). Reconmap: an interactive visualization of human metabolism. *Bioinformatics*, 33(4):605–607.

[224] Nyhan, W. L. (2005). Disorders of purine and pyrimidine metabolism. *Molecular genetics and metabolism*, 86(1):25–33.

[225] Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular systems biology*, 5(1):320.

[226] O'brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., and Palsson, B. Ø. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology*, 9(1):693.

[227] Ocke, M. C., Bueno-de Mesquita, H. B., Pols, M. A., Smit, H. A., van Staveren, W. A., and Kromhout, D. (1997). The dutch epic food frequency questionnaire. ii. relative validity and reproducibility for nutrients. *International journal of epidemiology*, 26(suppl_1):S49.

[228] O'donovan, C. B., Walsh, M. C., Nugent, A. P., McNulty, B., Walton, J., Flynn, A., Gibney, M. J., Gibney, E. R., and Brennan, L. (2015). Use of metabotyping for the delivery of personalised nutrition. *Molecular nutrition & food research*, 59(3):377–385.

[229] of Health, U. N. I. et al. (2012). Clinicaltrials.gov - `https://clinicaltrials.gov/`.

[of Washington] of Washington, U. Drug interaction database program - `https://www.druginteractioninfo.org`.

[231] on Radiological Protection. Task Group, I. C. and Snyder, W. S. (1975). *Report of the task group on reference man*. Pergamon.

[232] Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D. C., and Lewis, N. E. (2017). A systematic evaluation of methods for tailoring genome-scale metabolic models. *Cell Systems*, 4(3):318–329.

[233] Ortega-Azorín, C., Sorlí, J. V., Asensio, E. M., Coltell, O., Martínez-González, M. Á., Salas-Salvadó, J., Covas, M.-I., Arós, F., Lapetra, J., Serra-Majem, L., et al. (2012). Associations of the fto rs9939609 and the mc4r rs17782313 polymorphisms with type 2 diabetes are modulated by diet, being higher when adherence to the mediterranean diet pattern is low. *Cardiovascular diabetology*, 11(1):137.

[234] Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, 28(3):245–248.

[235] O'Sullivan, A., Gibney, M. J., and Brennan, L. (2010). Dietary intake patterns are reflected in metabolomic profiles: potential role in dietary assessment studies–. *The American journal of clinical nutrition*, 93(2):314–321.

[236] Pagliarini, R., Castello, R., Napolitano, F., Borzone, R., Annunziata, P., Mandrile, G., De Marchi, M., Brunetti-Pierri, N., and di Bernardo, D. (2016). In silico modeling of liver metabolism in a human disease reveals a key enzyme for histidine and histamine homeostasis. *Cell reports*, 15(10):2292–2300.

[237] Palsson, B. and Palsson, B. Ø. (2015). *Systems biology*. Cambridge university press.

[238] Panchaud, A., Kussmann, M., and Affolter, M. (2005). Rapid enrichment of bioactive milk proteins and iterative, consolidated protein identification by multidimensional protein identification technology. *Proteomics*, 5(15):3836–3846.

[239] Patti, G. J., Yanes, O., and Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews Molecular cell biology*, 13(4):263–269.

[240] Paul, D. S. and Beck, S. (2014). Advances in epigenome-wide association studies for common diseases. *Trends in molecular medicine*, 20(10):541–543.

[241] Pavlidis, C., Lanara, Z., Balasopoulou, A., Nebel, J.-C., Katsila, T., and Patrinos, G. P. (2015a). Meta-analysis of genes in commercially available nutrigenomic tests denotes lack of association with dietary intake and nutrient-related pathologies. *Omics: a journal of integrative biology*, 19(9):512–520.

[242] Pavlidis, C., Patrinos, G. P., and Katsila, T. (2015b). Nutrigenomics: A controversy. *Applied & translational genomics*, 4:50–53.

[243] Pence, H. E. and Williams, A. (2010). Chemspider: an online chemical information resource.

[244] Pérez-Jiménez, J., Neveu, V., Vos, F., and Scalbert, A. (2010). Identification of the 100 richest dietary sources of polyphenols: an application of the phenol-explorer database. *European journal of clinical nutrition*, 64:S112–S120.

[245] Pijls, L. T., Vries, H. d., Donker, A. J., and Eijk, J. T. M. v. (1999). Reproducibility and biomarker-based validity and responsiveness of a food frequency questionnaire to estimate protein intake. *American journal of epidemiology*, 150(9):987–995.

[246] Pisani, P., Faggiano, F., Krogh, V., Palli, D., Vineis, P., and Berrino, F. (1997). Relative validity and reproducibility of a food frequency dietary questionnaire for use in the italian epic centres. *International journal of epidemiology*, 26(suppl_1):S152.

[247] Potischman, N. (2003). Biologic and methodologic issues for nutritional biomarkers. *The Journal of nutrition*, 133(3):875S–880S.

[248] Preis, S. R., Spiegelman, D., Zhao, B. B., Moshfegh, A., Baer, D. J., and Willett, W. C. (2011). Application of a repeat-measure biomarker measurement error model to 2 validation studies: examination of the effect of within-person variation in biomarker measurements. *American journal of epidemiology*, 173(6):683–694.

[249] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

[250] Rahman, J., Noronha, A., Thiele, I., and Rahman, S. (2017). Leigh map: A novel computational diagnostic resource for mitochondrial disease. *Annals of neurology*, 81(1):9–16.

[251] Rahman, S., Blok, R., Dahl, H.-H., Danks, D., Kirby, D., Chow, C., Christodoulou, J., and Thorburn, D. (1996). Leigh syndrome: clinical features and biochemical and dna abnormalities. *Annals of neurology*, 39(3):343–351.

[252] Rapola, J. M., Virtamo, J., Ripatti, S., Huttunen, J. K., Albanes, D., Taylor, P. R., and Heinonen, O. P. (1997). Randomised trial of α-tocopherol and β-carotene supplements on incidence of major coronary events in men with previous myocardial infarction. *The Lancet*, 349(9067):1715–1720.

[253] Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., et al. (2015). Accurate, fully-automated nmr spectral profiling for metabolomics. *PLoS One*, 10(5):e0124219.

[254] Rayman, M. P., Infante, H. G., and Sargent, M. (2008). Food-chain selenium and human health: spotlight on speciation. *British Journal of Nutrition*, 100(2):238–253.

[255] Reis, E. (2011). The lean startup. *New York: Crown Business*.

[256] Relling, M. and Klein, T. (2011). Cpic: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clinical Pharmacology & Therapeutics*, 89(3):464–467.

[257] Ross, A. B., Bourgeois, A., Macharia, H. N., Kochhar, S., Jebb, S. A., Brownlee, I. A., and Seal, C. J. (2012). Plasma alkylresorcinols as a biomarker of whole-grain food consumption in a large population: results from the wholeheart intervention study. *The American journal of clinical nutrition*, 95(1):204–211.

[258] Rossen, N. G., MacDonald, J. K., de Vries, E. M., D'Haens, G. R., de Vos, W. M., Zoetendal, E. G., and Ponsioen, C. Y. (2015). Fecal microbiota transplantation as novel therapy in gastroenterology: a systematic review. *World journal of gastroenterology: WJG*, 21(17):5359.

[259] Round, J. L. and Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9(5):313–323.

[260] Sahoo, S., Franzson, L., Jonsson, J. J., and Thiele, I. (2012). A compendium of inborn errors of metabolism mapped onto the human metabolic network. *Molecular BioSystems*, 8(10):2545–2558.

[261] Sahoo, S., Haraldsdóttir, H. S., Fleming, R. M., and Thiele, I. (2015). Modeling the effects of commonly used drugs on human metabolism. *The FEBS journal*, 282(2):297–317.

[262] Sahoo, S. and Thiele, I. (2013). Predicting the impact of diet and enzymopathies on human small intestinal epithelial cells. *Human molecular genetics*, 22(13):2705–2722.

[263] Sales, N., Pelegrini, P., and Goersch, M. (2014). Nutrigenomics: definitions and advances of this new science. *Journal of nutrition and metabolism*, 2014.

[264] Sanchez, R. and Kauffman, F. (2010). *Comprehensive Toxicology: Regulation of Xenobiotic Metabolism in the Liver*. Elsevier.

[265] Sauer, S. and Luge, T. (2015). Nutriproteomics: Facts, concepts, and perspectives. *Proteomics*, 15(5-6):997–1013.

[266] Sawaya, A. L., Tucker, K., Tsay, R., Willett, W., Saltzman, E., Dallal, G. E., and Roberts, S. B. (1996). Evaluation of four methods for determining energy intake in young and older women: comparison with doubly labeled water measurements of total energy expenditure. *The American journal of clinical nutrition*, 63(4):491–499.

[267] Scagliusi, F. B., Ferriolli, E., Pfrimer, K., Laureano, C., Cunha, C. S., Gualano, B., Lourenço, B. H., and Lancha, A. H. (2008). Underreporting of energy intake in brazilian women varies according to dietary assessment: a cross-sectional study using doubly labeled water. *Journal of the American Dietetic Association*, 108(12):2031–2040.

[268] Schap, T., Zhu, F., Delp, E. J., and Boushey, C. J. (2014). Merging dietary assessment with the adolescent lifestyle. *Journal of human nutrition and dietetics*, 27(s1):82–88.

[269] Schellenberger, J. and Palsson, B. Ø. (2009). Use of randomized sampling for analysis of metabolic networks. *Journal of Biological Chemistry*, 284(9):5457–5461.

[270] Schmidt, L. E. and Dalhoff, K. (2002). Food-drug interactions. *Drugs*, 62(10):1481–1502.

[271] Schoeller, D. A. (1995). Limitations in the assessment of dietary energy intake by self-report. *Metabolism*, 44:18–22.

[272] Seppo, L., Jauhiainen, T., Poussa, T., and Korpela, R. (2003). A fermented milk high in bioactive peptides has a blood pressure–lowering effect in hypertensive subjects. *The American journal of clinical nutrition*, 77(2):326–330.

[273] Serra-Majem, L., Roman, B., and Estruch, R. (2006). Scientific evidence of interventions using the mediterranean diet: a systematic review. *Nutrition reviews*, 64(s1).

[274] Shafquat, A., Joice, R., Simmons, S. L., and Huttenhower, C. (2014). Functional and phylogenetic assembly of microbial communities in the human microbiome. *Trends in microbiology*, 22(5):261–266.

[275] Shankar, P. R. (2016). Vigiaccess: Promoting public access to vigibase. *Indian journal of pharmacology*, 48(5):606–607.

[276] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

[277] Shephard, R. J. (2003). Limits to the measurement of habitual physical activity by questionnaires. *British journal of sports medicine*, 37(3):197–206.

[278] Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nature genetics*, 46(6):543–550.

[279] Shlomi, T., Cabili, M. N., and Ruppin, E. (2009). Predicting metabolic biomarkers of human inborn errors of metabolism. *Molecular systems biology*, 5(1):263.

[280] Shoaie, S., Ghaffari, P., Kovatcheva-Datchary, P., Mardinoglu, A., Sen, P., Pujos-Guillot, E., de Wouters, T., Juste, C., Rizkalla, S., Chilloux, J., et al. (2015). Quantifying diet-induced metabolic changes of the human gut microbiome. *Cell metabolism*, 22(2):320–331.

[281] Shriver, B. J., Roman-Shriver, C. R., and Long, J. D. (2010). Technology-based methods of dietary assessment: recent developments and considerations for clinical practice. *Current Opinion in Clinical Nutrition & Metabolic Care*, 13(5):548–551.

[282] Simmonds, H., Webster, D., Becroft, D., and Potter, C. (1980). Purine and pyrimidine metabolism in hereditary orotic aciduria: some unexpected effects of allopurinol. *European journal of clinical investigation*, 10(4):333–339.

[283] Singh, R. R., Sedani, S., Lim, M., Wassmer, E., and Absoud, M. (2015). Ranbp2 mutation and acute necrotizing encephalopathy: 2 cases and a literature review of the expanding clinico-radiological phenotype. *European Journal of Paediatric Neurology*, 19(2):106–113.

[284] Slupsky, C. M., Rankin, K. N., Wagner, J., Fu, H., Chang, D., Weljie, A. M., Saude, E. J., Lix, B., Adamko, D. J., Shah, S., et al. (2007). Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Analytical chemistry*, 79(18):6995–7004.

[285] Smacchi, E. and Gobbetti, M. (2000). Bioactive peptides in dairy products: synthesis and interaction with proteolytic enzymes. *Food Microbiology*, 17(2):129–141.

[286] Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., and Siuzdak, G. (2005). Metlin: a metabolite mass spectral database. *Therapeutic drug monitoring*, 27(6):747–751.

[287] Sofi, F., Abbate, R., Gensini, G. F., and Casini, A. (2010). Accruing evidence on benefits of adherence to the mediterranean diet on health: an updated systematic review and meta-analysis. *The American journal of clinical nutrition*, 92(5):1189–1196.

[288] Sofou, K., De Coo, I. F., Isohanni, P., Ostergaard, E., Naess, K., De Meirleir, L., Tzoulis, C., Uusimaa, J., De Angst, I. B., Lönnqvist, T., et al. (2014). A multicenter study on leigh syndrome: disease course and predictors of survival. *Orphanet journal of rare diseases*, 9(1):52.

[289] Solanky, K. S., Bailey, N. J., Beckwith-Hall, B. M., Bingham, S., Davis, A., Holmes, E., Nicholson, J. K., and Cassidy, A. (2005). Biofluid 1 h nmr-based metabonomic techniques in nutrition research—metabolic effects of dietary isoflavones in humans. *The Journal of nutritional biochemistry*, 16(4):236–244.

[290] Sperber, H., Mathieu, J., Wang, Y., Ferreccio, A., Hesson, J., Xu, Z., Fischer, K. A., Devi, A., Detraux, D., Gu, H., et al. (2015). The metabolome regulates the epigenetic landscape during naive-to-primed human embryonic stem cell transition. *Nature cell biology*, 17(12):1523–1535.

[291] Stampfer, M. J., Hennekens, C. H., Manson, J. E., Colditz, G. A., Rosner, B., and Willett, W. C. (1993). Vitamin e consumption and the risk of coronary disease in women. *New England Journal of Medicine*, 328(20):1444–1449.

[292] Stevens, J., Taber, D. R., Murray, D. M., and Ward, D. S. (2007). Advances and controversies in the design of obesity prevention trials. *Obesity*, 15(9):2163–2170.

[293] Stringer, A. M. (2009). *Chemotherapy-induced mucositis: the role of gastrointestinal microflora and mucins in the luminal environment.* PhD thesis.

[294] Strolin Benedetti, M., Whomsley, R., and Baltes, E. L. (2005). Differences in absorption, distribution, metabolism and excretion of xenobiotics between the paediatric and adult populations. *Expert opinion on drug metabolism & toxicology*, 1(3):447–471.

[295] Subar, A. F., Crafts, J., Zimmerman, T. P., Wilson, M., Mittl, B., Islam, N. G., McNutt, S., Potischman, N., Buday, R., Hull, S. G., et al. (2010). Assessment of the accuracy of portion size reports using computer-based food photographs aids in the development of an automated self-administered 24-hour recall. *Journal of the American Dietetic Association*, 110(1):55–64.

[296] Subar, A. F., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D. A., Bingham, S., Sharbaugh, C. O., Trabulsi, J., Runswick, S., Ballard-Barbash, R., et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the open study. *American journal of epidemiology*, 158(1):1–13.

[297] Subar, A. F., Kirkpatrick, S. I., Mittl, B., Zimmerman, T. P., Thompson, F. E., Bingley, C., Willis, G., Islam, N. G., Baranowski, T., McNutt, S., et al. (2012). The automated self-administered 24-hour dietary recall (asa24): a resource for researchers, clinicians, and educators from the national cancer institute. *Journal of the Academy of Nutrition and Dietetics*, 112(8):1134–1137.

[298] Subar, A. F., Thompson, F. E., Potischman, N., Forsyth, B. H., Buday, R., Richards, D., McNutt, S., Hull, S. G., Guenther, P. M., Schatzkin, A., et al. (2007). Formative research of a quick list for an automated self-administered 24-hour dietary recall. *Journal of the American Dietetic Association*, 107(6):1002–1007.

[299] Sutherland, J. and Sutherland, J. (2014). *Scrum: the art of doing twice the work in half the time*. Crown Business.

[300] Swainston, N., Smallbone, K., Hefzi, H., Dobson, P. D., Brewer, J., Hanscho, M., Zielinski, D. C., Ang, K. S., Gardiner, N. J., Gutierrez, J. M., et al. (2016). Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12(7):1–7.

[301] Takakura, A., Kurita, A., Asahara, T., Yokoba, M., Yamamoto, M., Ryuge, S., Igawa, S., Yasuzawa, Y., Sasaki, J., Kobayashi, H., et al. (2012). Rapid deconjugation of sn-38 glucuronide and adsorption of released free sn-38 by intestinal microorganisms in rat. *Oncology letters*, 3(3):520–524.

[302] Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28(1):33–36.

[303] Thiele, I., Fleming, R. M., Que, R., Bordbar, A., Diep, D., and Palsson, B. O. (2012). Multiscale modeling of metabolism and macromolecular synthesis in e. coli and its application to the evolution of codon usage. *PloS one*, 7(9):e45635.

[304] Thiele, I. and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121.

[305] Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., et al. (2013). A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419–425.

[306] Thomas, G. H. (2001). Metabolomics breaks the silence. *Trends in microbiology*, 9(4):158.

[307] Thompson, C. M., Johns, D. O., Sonawane, B., Barton, H. A., Hattis, D., Tardif, R., and Krishnan, K. (2009). Database for physiologically based pharmacokinetic (pbpk) modeling: physiological data for healthy and health-impaired elderly. *Journal of Toxicology and Environmental Health, Part B*, 12(1):1–24.

[308] Thompson, F. E., Subar, A. F., et al. (2008). Dietary assessment methodology. *Nutrition in the Prevention and Treatment of Disease*, 2:3–39.

[309] Thompson, F. E., Subar, A. F., Loria, C. M., Reedy, J. L., and Baranowski, T. (2010). Need for technological innovation in dietary assessment. *Journal of the American Dietetic Association*, 110(1):48.

[310] Thomson, C. A., Giuliano, A., Rock, C. L., Ritenbaugh, C. K., Flatt, S. W., Faerber, S., Newman, V., Caan, B., Graver, E., Hartz, V., et al. (2003). Measuring dietary change in a diet intervention trial: comparing food frequency questionnaire and dietary recalls. *American journal of epidemiology*, 157(8):754–762.

[311] Truswell, A. S., Seach, J. M., and Thorburn, A. (1988). Incomplete absorption of pure fructose in healthy subjects and the facilitating effect of glucose. *The American journal of clinical nutrition*, 48(6):1424–1430.

[312] Tulipani, S., Llorach, R., Jáuregui, O., López-Uriarte, P., Garcia-Aloy, M., Bullo, M., Salas-Salvadó, J., and Andrés-Lacueva, C. (2011). Metabolomics unveils urinary changes in subjects with metabolic syndrome following 12-week nut consumption. *Journal of proteome research*, 10(11):5047–5058.

[313] Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science translational medicine*, 1(6):6ra14–6ra14.

[314] Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C. M., Gibson, D., Gonzalez, J. N., Guruvadoo, L., et al. (2016). The ucsc genome browser database: 2017 update. *Nucleic acids research*, 45(D1):D626–D634.

[315] Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250.

[316] Ulanovskaya, O. A., Zuhl, A. M., and Cravatt, B. F. (2013). Nnmt promotes epigenetic remodeling in cancer by creating a metabolic methylation sink. *Nature chemical biology*, 9(5):300–306.

[317] US Department of Agriculture, Agricultural Research Service, N. D. L. (2016). Usda national nutrient database for standard reference, release 28.

[318] van der Werf, M. J., Overkamp, K. M., Muilwijk, B., Coulier, L., and Hankemeier, T. (2007). Microbial metabolomics: toward a platform with full metabolome coverage. *Analytical biochemistry*, 370(1):17–25.

[319] Vaquero, A. and Reinberg, D. (2009). Calorie restriction and the exercise of chromatin. *Genes & development*, 23(16):1849–1869.

[320] vel Szic, K. S., Declerck, K., Vidaković, M., and Berghe, W. V. (2015). From inflammaging to healthy aging by dietary lifestyle choices: is epigenetics the key to personalized nutrition? *Clinical epigenetics*, 7(1):33.

[321] Vereecken, C., Covents, M., Sichert-Hellert, W., Alvira, J. F., Le Donne, C., De Henauw, S., De Vriendt, T., Phillipp, M., Beghin, L., Manios, Y., et al. (2008). Development and evaluation of a self-administered computerized 24-h dietary recall method for adolescents in europe. *International Journal of Obesity*, 32:S26–S34.

[322] Verkasalo, P. K., Appleby, P. N., Allen, N. E., Davey, G., Adlercreutz, H., and Key, T. J. (2001). Soya intake and plasma concentrations of daidzein and genistein: validity of dietary assessment among eighty british women (oxford arm of the european prospective investigation into cancer and nutrition). *British Journal of Nutrition*, 86(3):415–421.

[323] Villeneuve, L. M. and Natarajan, R. (2010). The role of epigenetics in the pathology of diabetic complications. *American Journal of Physiology-Renal Physiology*, 299(1):F14–F25.

[324] Walker, A. W., Ince, J., Duncan, S. H., Webster, L. M., Holtrop, G., Ze, X., Brown, D., Stares, M. D., Scott, P., Bergerat, A., et al. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *The ISME journal*, 5(2):220–230.

[325] Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., Thiessen, P. A., He, S., and Zhang, J. (2016). Pubchem bioassay: 2017 update. *Nucleic acids research*, 45(D1):D955–D963.

[326] Wedatilake, Y., Brown, R. M., McFarland, R., Yaplito-Lee, J., Morris, A. A., Champion, M., Jardine, P. E., Clarke, A., Thorburn, D. R., Taylor, R. W., et al. (2013). Surf1 deficiency: a multi-centre natural history study. *Orphanet journal of rare diseases*, 8(1):96.

[327] Weinberg, E. G. (2011). The wao white book on allergy 2011-2012. *Current Allergy & Clinical Immunology*, 24(3):156–157.

[328] Wellen, K. E., Hatzivassiliou, G., Sachdeva, U. M., Bui, T. V., Cross, J. R., and Thompson, C. B. (2009). Atp-citrate lyase links cellular metabolism to histone acetylation. *Science*, 324(5930):1076–1080.

[329] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., et al. (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1):D13–D21.

[330] Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., et al. (2012). Hmdb 3.0—the human metabolome database in 2013. *Nucleic acids research*, 41(D1):D801–D807.

[331] Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., et al. (2008). Hmdb: a knowledgebase for the human metabolome. *Nucleic acids research*, 37(suppl_1):D603–D610.

[332] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672.

[333] Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., et al. (2007). Hmdb: the human metabolome database. *Nucleic acids research*, 35(suppl_1):D521–D526.

[334] Wolf, S., Schmidt, S., Müller-Hannemann, M., and Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, 11(1):148.

[335] Wortmann, S. B., Koolen, D. A., Smeitink, J. A., van den Heuvel, L., and Rodenburg, R. J. (2015). Whole exome sequencing of suspected mitochondrial patients in clinical practice. *Journal of inherited metabolic disease*, 38(3):437–443.

[336] Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108.

[337] Wurtman, R. J., Regan, M., Ulus, I., and Yu, L. (2000). Effect of oral cdp-choline on plasma choline and uridine levels in humans. *Biochemical pharmacology*, 60(7):989–992.

[338] Xia, J., Sinelnikov, I. V., Han, B., and Wishart, D. S. (2015). Metaboanalyst 3.0—making metabolomics more meaningful. *Nucleic acids research*, 43(W1):W251–W257.

[339] Xia, J. and Wishart, D. S. (2016). Using metaboanalyst 3.0 for comprehensive metabolomics data analysis. *Current Protocols in Bioinformatics*, pages 14–10.

[340] Yoon, K.-H., Lee, J.-H., Kim, J.-W., Cho, J. H., Choi, Y.-H., Ko, S.-H., Zimmet, P., and Son, H.-Y. (2006). Epidemic obesity and type 2 diabetes in asia. *The Lancet*, 368(9548):1681–1688.

[341] Young, J. F., Branham, W. S., Sheehan, D. M., Baker, M. E., Wosilait, W. D., and Luecke, R. H. (1997). Physiological "constants" for pbpk models for pregnancy. *Journal of toxicology and environmental health*, 52(5):385–401.

[342] Yurkovich, J. T., Yurkovich, B. J., Dräger, A., Palsson, B. O., and King, Z. A. (2017). A padawan programmer's guide to developing software libraries. *Cell systems*, 5(5):431–437.

[343] Yusuf, S., Dagenais, G., Pogue, J., Bosch, J., and Sleight, P. (2000). Vitamin e supplementation and cardiovascular events in high-risk patients. *The New England journal of medicine*, 342(3):154–160.

[344] Zamboni, N., Saghatelian, A., and Patti, G. J. (2015). Defining the metabolome: size, flux, and regulation. *Molecular cell*, 58(4):699–706.

[345] Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094.

[346] Zeisel, S. H. (2012). Diet-gene interactions underlie metabolic individuality and influence brain development: implications for clinical practice derived from studies on choline metabolism. *Annals of Nutrition and Metabolism*, 60(Suppl. 3):19–25.

[347] Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., Parameswaran, P., Crowell, M. D., Wing, R., Rittmann, B. E., et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences*, 106(7):2365–2370.

[348] Zoetendal, E., Rajilić-Stojanović, M., and De Vos, W. (2008). High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut*, 57(11):1605–1615.

Appendix A

Supplementary Material

A.1 Mapping of nutritional data with VMH metabolites

Nutrient information from the USDA National Nutrient Database for Standard Reference [317] was mapped to VMH metabolites. Table A.1 shows all nutrient definitions present in the nutritional composition database and, when that was possible, the corresponding metabolite abbreviation. Additionally, we have categorized each nutrient for display purposes in the detail pages of VMH.

Tag name	Nutrient description	Metabolites VMH	Category	Subcategory
PROCNT	Protein		Proteins	Total protein
FAT	Total lipid (fat)		Lipids	Total lipids
CHOCDF	Carbohydrate, by difference		Carbohydrates	Total carbohydrate
ASH	Ash		Minerals and trace elements	Ash
ENERC_KCAL	Energy		Energy content	Energy in kcal
STARCH	Starch	strch1, strch2, starch1200	Carbohydrates	Carbohydrate
SUCS	Sucrose	sucr	Carbohydrates	Disaccharide
GLUS	Glucose (dextrose)	glc_D	Carbohydrates	Monosaccharide
FRUS	Fructose	fru	Carbohydrates	Monosaccharide
LACS	Lactose	lcts	Carbohydrates	Disaccharide

MALS	Maltose	malt	Carbohydrates	Disaccharide
ALC	Alcohol, ethyl	etoh	Other	Alcohol
WATER	Water	h2o	Other	Water
	Adjusted Protein		Proteins	
CAFFN	Caffeine		Other	Total caffeine
THEBRN	Theobromine		Other	
ENERC_KJ	Energy		Energy content	Energy in kj
SUGAR	Sugars, total		Carbohydrates	Total sugar
GALS	Galactose	gal	Carbohydrates	Monosaccharide
FIBTG	Fiber, total dietary		Dietary Fibers	Total dietary fibers
CA	Calcium, Ca	ca2	Minerals and trace elements	Mineral
FE	Iron, Fe	fe2, fe3	Minerals and trace elements	Trace element
MG	Magnesium, Mg	mg2	Minerals and trace elements	Mineral
P	Phosphorus, P	pi	Minerals and trace elements	Mineral
K	Potassium, K	k	Minerals and trace elements	Mineral
NA	Sodium, Na	na1	Minerals and trace elements	Mineral
ZN	Zinc, Zn	zn2	Minerals and trace elements	Mineral
CU	Copper, Cu	cu2	Minerals and trace elements	Trace element
FLD	Fluoride, F		Minerals and trace elements	Trace element

MN	Manganese, Mn	mn2	Minerals and trace elements	Mineral
SE	Selenium, Se	sel	Minerals and trace elements	Mineral
VITA_IU	Vitamin A, IU		Vitamins	
RETOL	Retinol	retinol	Vitamins	Fat soluble vitamin
VITA_RAE	Vitamin A, RAE		Vitamins	
CARTB	Carotene, beta	caro	Vitamins	Fat soluble vitamin
CARTA	Carotene, alpha		Vitamins	Fat soluble vitamin
TOCPHA	Vitamin E (alpha-tocopherol)	avite1	Vitamins	
VITD	Vitamin D		Vitamins	
ERGCAL	Vitamin D2 (ergocalciferol)	vitd2	Vitamins	Fat soluble vitamin
CHOCAL	Vitamin D3 (cholecalciferol)	vitd3	Vitamins	
VITD	Vitamin D (D2 + D3)		Vitamins	
CRYPX	Cryptoxanthin, beta		Vitamins	
LYCPN	Lycopene		Vitamins	
LUT+ZEA	Lutein + zeaxanthin		Vitamins	
TOCPHB	Tocopherol, beta	bvite	Vitamins	Fat soluble vitamin
TOCPHG	Tocopherol, gamma	yvite	Vitamins	Fat soluble vitamin
TOCPHD	Tocopherol, delta		Vitamins	Fat soluble vitamin
TOCTRA	Tocotrienol, alpha	avite2	Vitamins	Fat soluble vitamin
TOCTRB	Tocotrienol, beta		Vitamins	Fat soluble vitamin

TOCTRG	Tocotrienol, gamma		Vitamins	Fat soluble vitamin
TOCTRD	Tocotrienol, delta		Vitamins	Fat soluble vitamin
VITC	Vitamin C, total ascorbic acid	ascb_L	Vitamins	Water soluble vitamin
THIA	Thiamin	thm	Vitamins	Water soluble vitamin
RIBF	Riboflavin	ribflv	Vitamins	Water soluble vitamin
NIA	Niacin	nac, ncam	Vitamins	Water soluble vitamin
PANTAC	Pantothenic acid	pnto_R	Vitamins	Water soluble vitamin
VITB6A	Vitamin B-6	pydam, pydx, pydxn	Vitamins	Water soluble vitamin
FOL	Folate, total	fol, 10fthf, 5mthf, thf	Vitamins	Water soluble vitamin
VITB12	Vitamin B-12	adocbl	Vitamins	Water soluble vitamin
CHOLN	Choline, total	chol	Vitamins	Water soluble vitamin
MK4	Menaquinone-4		Vitamins	Fat soluble vitamin
VITK1D	Dihydrophyl-loquinone		Vitamins	Fat soluble vitamin
VITK1	Vitamin K (phylloqui-none)	phyQ	Vitamins	Fat soluble vitamin
FOLAC	Folic acid	fol	Vitamins	Water soluble vitamin
FOLFD	Folate, food		Vitamins	Water soluble vitamin
FOLDFE	Folate, DFE		Vitamins	Water soluble vitamin
BETN	Betaine	glyb	Proteins	Amino acid
TRP_G	Tryptophan	trp_L	Proteins	Amino acid
THR_G	Threonine	thr_L	Proteins	Amino acid
ILE_G	Isoleucine	ile_L	Proteins	Amino acid
LEU_G	Leucine	leu_L	Proteins	Amino acid
LYS_G	Lysine	lys_L	Proteins	Amino acid
MET_G	Methionine	met_L	Proteins	Amino acid

CYS_G	Cystine	cys_L	Proteins	Amino acid
PHE_G	Phenylalanine	phe_L	Proteins	Amino acid
TYR_G	Tyrosine	tyr_L	Proteins	Amino acid
VAL_G	Valine	val_L	Proteins	Amino acid
ARG_G	Arginine	arg_L	Proteins	Amino acid
HISTN_G	Histidine	his_L	Proteins	Amino acid
ALA_G	Alanine	ala_L	Proteins	Amino acid
ASP_G	Aspartic acid	asp_L	Proteins	Amino acid
GLU_G	Glutamic acid	glu_L	Proteins	Amino acid
GLY_G	Glycine	gly	Proteins	Amino acid
PRO_G	Proline	pro_D, pro_L	Proteins	Amino acid
SER_G	Serine	ser_L	Proteins	Amino acid
HYP	Hydroxyproline	4hpro_LT	Proteins	Amino acid
	Vitamin E, added		Vitamins	Fat soluble vitamin
	Vitamin B-12, added		Vitamins	Water soluble vitamin
CHOLE	Cholesterol	chsterol	Lipids	Cholesterol
FATRN	Fatty acids, total trans		Lipids	Fatty acids, total trans
FASAT	Fatty acids, total saturated		Lipids	Total saturated fatty acids
F4D0	4:0	but	Lipids	Fatty acid
F6D0	6:0	caproic	Lipids	Fatty acid
F8D0	8:0	octa	Lipids	Fatty acid
F10D0	10:0	dca	Lipids	Fatty acid
F12D0	12:0	ddca	Lipids	Fatty acid
F14D0	14:0	ttdca	Lipids	Fatty acid
F16D0	16:0	hdca	Lipids	Fatty acid
F18D0	18:0	ocdca	Lipids	Fatty acid
F20D0	20:0	arach	Lipids	Fatty acid
F18D1	18:1 undifferentiated	ocdcea	Lipids	Fatty acid
F18D2	18:2 undifferentiated	lnlc	Lipids	Fatty acid
F18D3	18:3 undifferentiated	lnlc	Lipids	Fatty acid
F20D4	20:4 undifferentiated	arachd	Lipids	Fatty acid

F22D6	22:6 n-3 (DHA)	crvnc, CE0328	Lipids	Fatty acid
F22D0	22:0	docosac	Lipids	Fatty acid
F14D1	14:1	ttdcea	Lipids	Fatty acid
F16D1	16:1 undifferentiated	hdcea	Lipids	Fatty acid
F18D4	18:4	strdnc	Lipids	Fatty acid
F20D1	20:1	CE2510	Lipids	Fatty acid
F20D5	20:5 n-3 (EPA)	tmndnc	Lipids	Fatty acid
F22D1	22:1 undifferentiated	doco13ac	Lipids	Fatty acid
F22D5	22:5 n-3 (DPA)	clpnd	Lipids	Fatty acid
PHYSTR	Phytosterols		Lipids	Total phytosterols
STID7	Stigmasterol		Lipids	Phytosterol
CAMD5	Campesterol		Lipids	Phytosterol
SITSTR	Beta-sitosterol		Lipids	Phytosterol
FAMS	Fatty acids, total monounsaturated		Lipids	Total monounsaturated fatty acids
FAPU	Fatty acids, total polyunsaturated		Lipids	Total polyunsaturated fatty acids
F15D0	15:0	ptdca	Lipids	Fatty acid
F17D0	17:0	hpdca	Lipids	Fatty acid
F24D0	24:0	lgnc	Lipids	Fatty acid
F16D1T	16:1 t		Lipids	Fatty acid
F18D1T	18:1 t		Lipids	Fatty acid
F22D1T	22:1 t		Lipids	Fatty acid
	18:2 t not further defined		Lipids	Fatty acid
	18:2 i		Lipids	Fatty acid
F18D2TT	18:2 t,t		Lipids	Fatty acid
F18D2CLA	18:2 CLAs		Lipids	Fatty acid
F24D1C	24:1 c	nrvnc	Lipids	Fatty acid
F20D2CN6	20:2 n-6 c,c	eidi1114ac	Lipids	Fatty acid
F16D1C	16:1 c		Lipids	Fatty acid
F18D1C	18:1 c		Lipids	Fatty acid
F18D2CN6	18:2 n-6 c,c		Lipids	Fatty acid

F22D1C	22:1 c		Lipids	Fatty acid
F18D3CN6	18:3 n-6 c,c,c	lnlncg	Lipids	Fatty acid
F17D1	17:1	M00003, M01238	Lipids	Fatty acid
F20D3	20:3 undifferentiated	dlnlcg	Lipids	Fatty acid
FATRNM	Fatty acids, total trans-monoenoic		Lipids	Total trans-monoenoic fatty acids
FATRNP	Fatty acids, total trans-polyenoic		Lipids	Total trans-polyenoic fatty acids
F13D0	13:0	M03051	Lipids	Fatty acid
F15D1	15:1		Lipids	Fatty acid
F18D3CN3	18:3 n-3 c,c,c (ALA)	lnlnca	Lipids	Fatty acid
F20D3N3	20:3 n-3		Lipids	Fatty acid
F20D3N6	20:3 n-6		Lipids	Fatty acid
F20D4N6	20:4 n-6		Lipids	Fatty acid
	18:3i		Lipids	Fatty acid
F21D5	21:5		Lipids	Fatty acid
F22D4	22:4		Lipids	Fatty acid
F18D1TN7	18:1-11 t (18:1t n-7)		Lipids	Fatty acid

Table A.1: Mapping of nutritional information from the USDA National Nutrient Database for Standard Reference, Release 28 with metabolites from VMH.

A.2 VMH detailed schema

At the core of VMH is a MySQL relational database. This database contains 59 tables and takes around 1 GB of disk space. Figure A.1 shows a detailed schema of the database containing 44 of the 59 tables. The excluded tables are related with user administration, website definitions, and database migrations, a form of version control provided by the Django framework which allows tracking changes to the database structure without the need to write SQL code. This schema was generated using the software MySQL Workbench 6.0 Community edition, available at `https://www.mysql.com/products/workbench/`.

A.3 Leigh Map interface

The Leigh Map is integrated in the MINERVA framework and accessible at `http://vmh.uni.lu/#leighmap`. Figure A.2 shows the interface of the Leigh Map. In Figure A.2-A the conceptual overview of the mitochondria is visible. By zooming in additional detail will be revealed. Users can search for genes and pehnotypes in the left panel, as shown in Figure

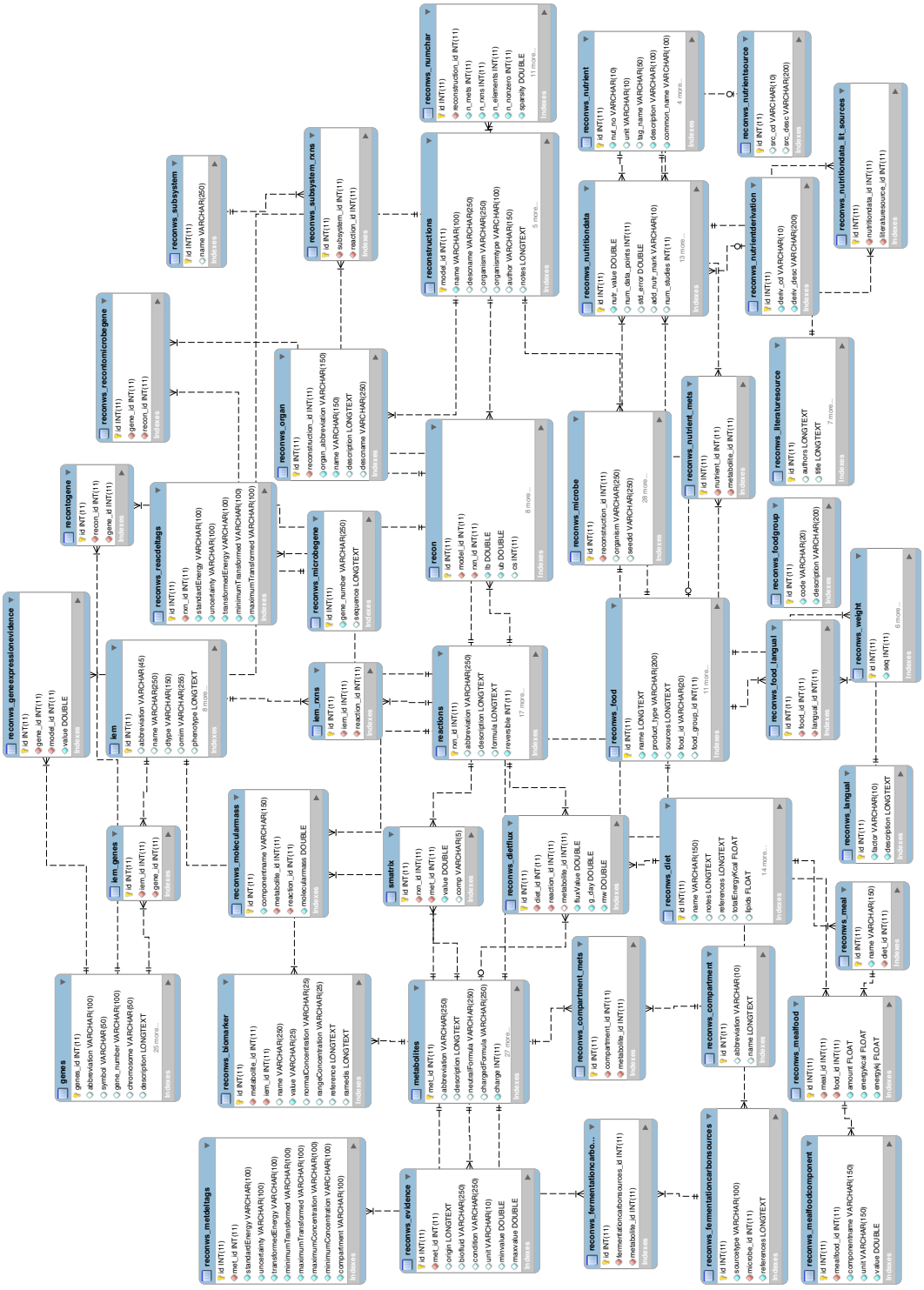

Figure A.1: Detailed schema of the VMH database

A.2-B. Each gene has an associated submap that shows every phenotype association with a marker on the searched ones Figure A.2-C.

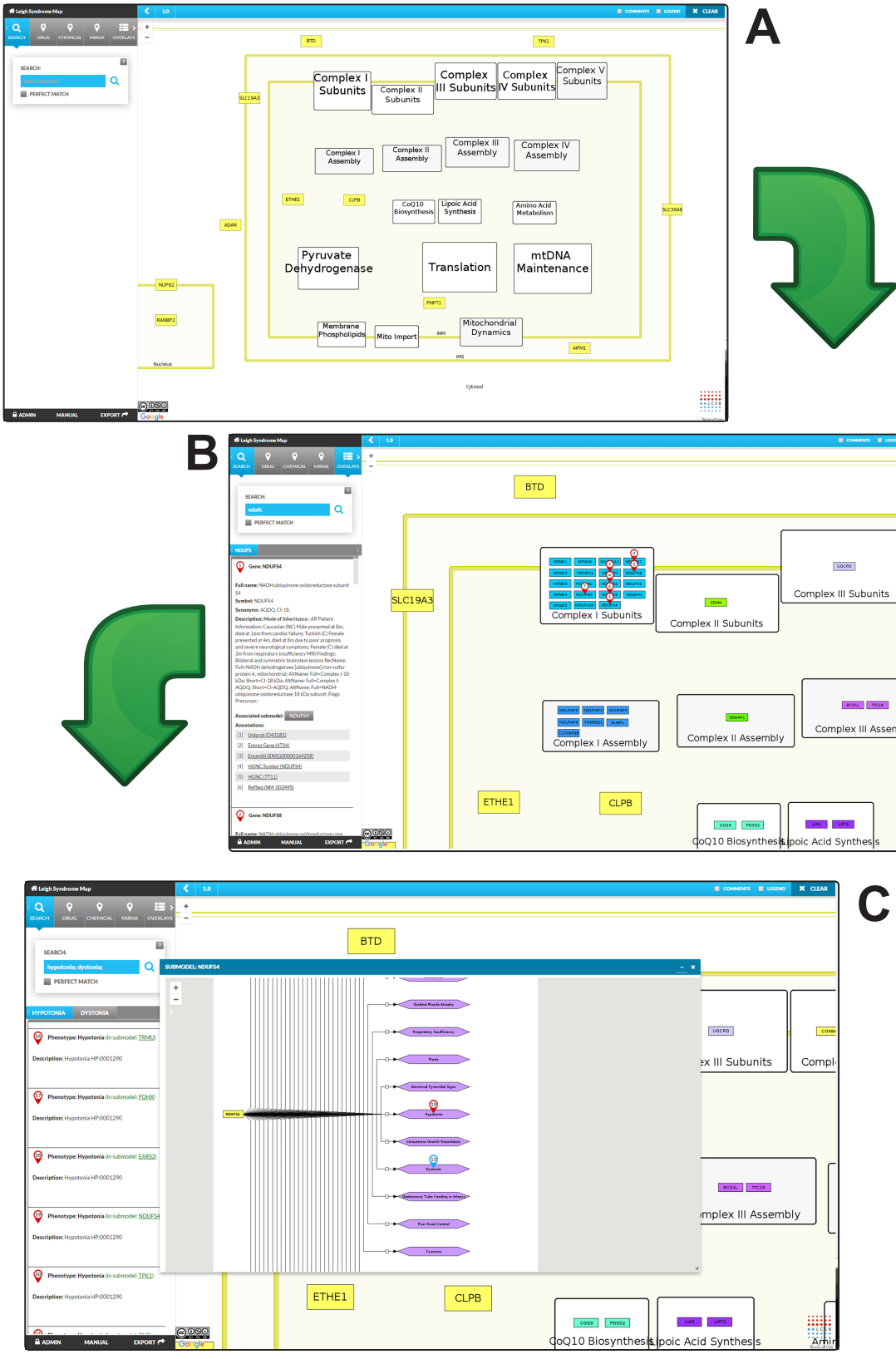

Figure A.2: Interface of the Leigh Map. A - conceptual overview of the mitochondria. B - search functionality. C - gene-submap displaying associated phenotypes.