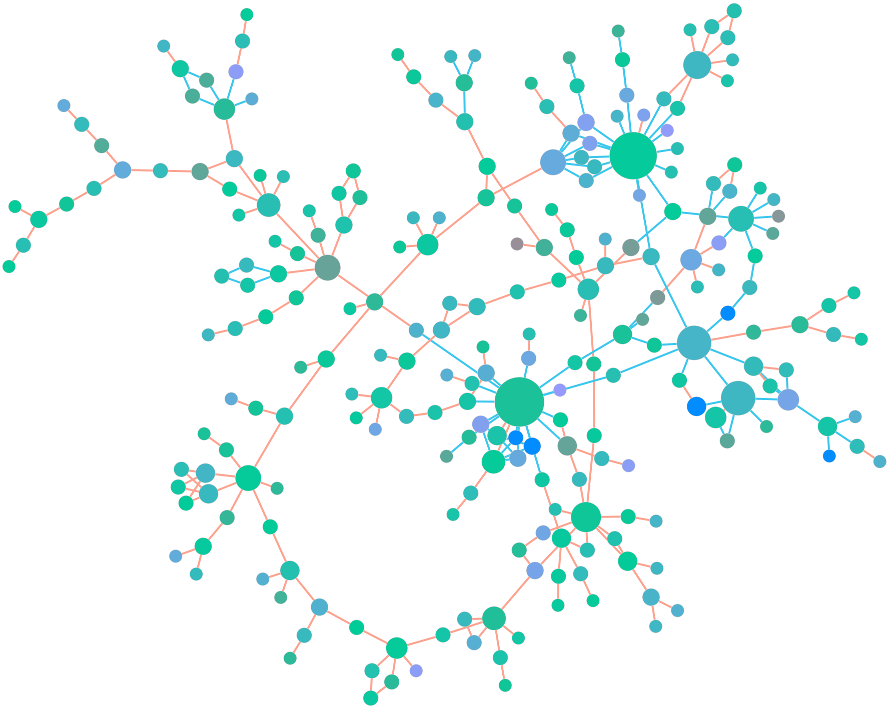
KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Biotechnology

Systems and Network-based Approaches to Complex Metabolic Diseases

MUHAMMAD ARIF



Stockholm, Sweden 2021

Systems and Network-based Approaches to Complex Metabolic Diseases

MUHAMMAD ARIF

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday the 11th June 2021, at 1:00 p.m. in E3, Osquars backe 14, Stockholm.

Doctoral Thesis in Biotechnology
KTH Royal Institute of Technology
Stockholm, Sweden 2021

ISBN: 978-91-7873-880-9
TRITA-CBH-FOU-2021:23

Printed by: Universitetsservice US-AB, Sweden 2021

Abstract

The future of healthcare is personalized medicine, in which disease treatments are tailored based on the individual characteristics of each patient. To reach that objective, we need to obtain a better understanding of diseases. The main facilitator of personalized medicine is systems and data-driven biology, which makes omics data a top commodity in this era. Coupled with computational and biological expertise, omics data can be a useful asset for obtaining mechanistic insights into the biological conundrum, particularly in disease-related contexts. This thesis describes systems biology approaches and their applications in disease-specific contexts. Systems biology assists us in systematically and comprehensively understanding complex biological systems as a whole interconnected system.

The first part of the thesis describes the generation of more than 100 biological networks based on personalized data originated from several different omics, usually referred to as multiomics data, including clinical data and metabolomics, proteomics, and metagenomics data collected from the same individuals. Moreover, we present a web-based multiomics biological network database and visualization platform called iNetModels.

In the second part of the thesis, we describe systems biology frameworks and their applications to the study of various biological questions in disease contexts using single- and multiomics data. First, we present our findings on the integrative view of metabolic activities from multiple tissues after myocardial infarction using transcriptomics data from the heart and other metabolically active tissues. Second, we used transcriptomics data to describe the mechanistic effect of lifelong training on skeletal muscle in both men and women and the role of short-term training in reversing damage from metabolic-related diseases. Third, we deciphered the molecular mechanism of nonalcoholic fatty liver disease (NAFLD) based on clinical data, plasma metabolomics, plasma inflammatory proteomics, and oral and gut metagenomics data. Finally, we elucidated the mechanism of action of CMA supplementation, a potential treatment for NAFLD, based on proteomics and metabolomics data.

In summary, this thesis presents a novel platform for biological network analysis and proven systems biology frameworks to provide mechanistic and systematic understandings of specific diseases using single- and multiomics data.

Sammanfattning

Framtiden för hälsovård är precisionsmedicin; behandling av sjukdom skräddarsys baserat på de individualla egenskaper hos varje enskild patient. För att nå detta mål behöver vi öka vår kunskap om sjukdomar. Det främsta hjälpmedlet för att utveckla precisionsmedicin är system- och datadriven biologi, vilket i sin tur gör omikdata till en viktig resurs i samtiden. Omikdata kan kombineras med expertis inom beräkningsbiologi för att på så vis vara en värdeful tillgång för att få insyn i biologiska mekanismer, särskilt inom sjukdomskontext. Denna avhandling beskriver strategier inom systembiologi, och deras applicering för specifika sjukdomar.

Den första delen av avhandlingen beskriver utvecklandet av mer än 100 biologiska nätverk baserade på personaliserad multiomik-data, inklusive klinisk data samt metabolomik-, proteomik-, och metagenomikdata, insamlat från samma individer. Dessutom presenterar vi en webb-baserad databas innehållande biologiska nätverk byggda från multiomik-data, samt en visualiseringsplatform vid namn iNetModels.

I den andra delen av avhandlingen beskriver vi systembiologiska ramverk och deras applicering för studier av olika sorters biologiska frågor inom sjukdomskontext, genom att använda en eller flera sorters omikdata. Först presenterar vi våra fynd om den integrativa vyn av metaboliska aktiviteter från flertalet vävnader efter hjärtinfarkt, genom att använda transkriptomikdata både från hjärtat och andra metaboliskt aktiva vävnader. Sedan använde vi transkriptomikdata för att beskriva den mekanistiska effekten av livslång träning av skelettmuskel i både män och kvinnor, samt vilken roll kortsiktig träning har i att läka skador från metabolismrelaterade sjukdomar. Efter det dechiffrerade vi den molekylära mekanismen bakom nonalcoholic fatty liver disease (NAFLD), eller fettlever, baserat på kliniska data, plasma-metabolomik, inflammatorisk plasma-proteomik, samt metagenomikdata från månhåla och tarmkanal. Till sist tydliggjorde vi mekanismen av CMA-supplementrering, en potentiell behandling av NAFLD, baserat på proteomik- och metabolomikdata.

Sammanfattningsvis beskriver denna avhandling en ny plattform för biologisk nätverksanalys och bevisade systembiologiska ramverk för att utröna mekanistisk och systematisk förståelse för specifika sjukdomar, genom att använda singel- eller multiomikdata.

Thesis Defense

This thesis will be defended on 11th June 2020 at 13:00, in room E3 at KTH campus, Osquars backe 14, for the degree of Teknologie Doktor (Doctor of Philosophy, PhD) in Biotechnology.

Respondent: Muhammad Arif, MSc in Electrical Engineering, from KTH Royal Institute of Technology, Stockholm, Sweden.

Faculty Opponent: Prof. Dr. Thomas Sauter, Professor in Systems Biology, Department of Life Sciences and Medicine, Faculty of Science, Technology, and Medicine, University of Luxembourg, Luxembourg.

Chairman of the Thesis Defense: Dr. Paul Hudson, Associate Professor, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden.

Evaluation Committee:

Dr. David Moyes, Senior Lecturer, Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral and Craniofacial Sciences, King's College London, United Kingdom.

Dr. Claudia Kutter, Lector, Department of Microbiology, Tumor and Cell Biology, Karolinska Insitute, Stockholm, Sweden.

Dr. Marc Friedländer, Associate Professor, Department of Molecular Biosciences, Wenner-Gren Institute of Stockholm University, Sweden.

Supervisors:

Prof. Dr. Adil Mardinoglu, Professor, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden.

Prof. Dr. Mathias Uhlén, Professor, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden.

List of Publications and Manuscripts

The following articles constitute the basis of the respondent's thesis. All articles are included in the appendices of the thesis.

Paper I

Sunjae Lee#, Cheng Zhang#, **Muhammad Arif**#, Zhengtao Liu, Rui Benfeitas, Gholamreza Bidkhori, Sumit Deshmukh, Mohamed Al Shobky, Alen Lovric, Jan Boren, Jens Nielsen, Mathias Uhlén, Adil Mardinoglu (2018). **TCSBN: a database of tissue- and cancer-specific biological networks.** *Nucleic Acids Research* 46(D1): D595–D600, doi: 10.1093/nar/gkx994

Paper II

Muhammad Arif#, Cheng Zhang#, Xiangyu Li, Cem Güngör, Buğra Çakmak, Metin Arslantürk, Abdellah Tebani, Berkay Özcan, Oğuzhan Subaş, Wenyu Zhou, Brian Piening, Linn Fagerberg, Nathan Price, Leroy Hood, Michael Snyder, Jens Nielsen, Mathias Uhlén, Adil Mardinoglu (2021). **iNetModels 2.0: an interactive visualization and database of multi-omics data.** *Nucleic Acids Research*, doi: 10.1093/nar/gkab254.

Paper III

Muhammad Arif#, Martina Klevstig#, Rui Benfeitas, Stephen Doran, Hasan Turkez, Mathias Uhlén, Maryam Clausen, Johannes Wikström, Damla Etal, Cheng Zhang, Malin Levin, Adil Mardinoglu, Jan Borén. **Integrative transcriptomic analysis of tissue-specific metabolic crosstalk after myocardial infarction**. *eLife*, doi: 10.7554/eLife.66921

Paper IV

Mark A Chapman, **Muhammad Arif**, Eric B Emanuelsson, Stefan M Reitzner, Maléne E Lindholm, Adil Mardinoglu, Carl Johan Sundberg (2020). **Skeletal Muscle Transcriptomic Comparison between Long-Term Trained and Untrained Men and Women**. Cell Reports 31(12): 107808, doi: 10.1016/j.celrep.2020.107808

Paper V

Mujdat Zeybel[#], **Muhammad Arif**[#], Hong Yang[#], Ozlem Altay, Claudia Fredolini, Murat Akyildiz, Burcin Saglam, Mehmet Gokhan Gonenli, Buket Yigit, Burge Ulukan, Dilek Ural, Woonghee Kim, Xiangyu Li, Jochen M. Schwenk, Cheng Zhang, Saeed Shoaie, Hasan Turkez, Jens Nielsen, Mathias Uhlén, Jan Borén, Adil Mardinoglu. **Multi-omics analysis reveals the influence of the oral and gut microbiome on host metabolism in non-alcoholic fatty liver disease.** *Submitted*

Paper VI

Cheng Zhang[#], Elias Bjornson[#], **Muhammad Arif**[#], Abdellah Tebani, Alen Lovric, Rui Benfeitas, Mehmet Ozcan, Kajetan Juszczak, Woonghee Kim, Jung Tae Kim, Gholamreza Bidkhori, Marcus Ståhlman, Per-Olof Bergh, Martin Adiels, Hasan Turkez, Marja-Riitta Taskinen, Jim Bosley, Hanns-Ulrich Marschall, Jens Nielsen, Mathias Uhlén, Jan Borén, Adil Mardinoglu (2020). **The acute effect of metabolic cofactor supplementation: a potential therapeutic strategy against non-alcoholic fatty liver disease**. *Molecular Systems Biology* 16(4): e949, doi: 10.15252/msb.209495

Respondent’s contribution to the included papers

Paper I

The respondent was responsible for the generation of the network and its front- and back-end development. The respondent also contributed to the writing of the manuscript together with the other co-first authors.

Paper II

The respondent was responsible for the pipeline development, network generation, and coordination of the front- and back-end development. The respondent also co-planned the project and contributed to the writing of the manuscript together with the other co-first authors.

Paper III

The respondent was responsible for the computational analysis of the transcriptomics data, the network analysis, and the writing of the manuscript.

Paper IV

The respondent was responsible for the transcriptomics data analysis (quantification, comparison, and functional analysis) and the integration of the data with the genome-scale metabolic model.

Paper V

The respondent was responsible for the clinical and omics data analysis, data integration, and biomarker analysis as well as for the writing of the manuscript together with the other co-first authors.

Paper VI

The respondent was responsible for the plasma proteomics and metabolomics analysis.

Table of Contents

Table of Figures

Chapter I: Introduction

Since its introduction in the early 2000s, systems biology[1,2] has continuously gained increasing traction in the field of biology. Systems biology is a multidisciplinary field that attempts to solve complex biological problems in a holistic manner by interpreting biology at the systems level rather than using the traditional mindset of focusing on a specific biological part or issue. The rise of systems biology has also been propelled by the exponentially increasing number of large-scale biological data, usually called **omics data**, generated in the past decade.

Omics Data

Omics data always involves the simultaneous measurements of a large number of analytes[3], which makes these data superior to other molecular measurements. The emergence of omics data has hugely helped researchers understand the complexity of living cells[4]. The term “omics data” has become a blanket term for data related to studies of molecular biology, ranging from data related to DNA, RNA, protein, and metabolites to data from the gut and oral microbiomes. “Omics” has also been used to refer to a field of study, but in this thesis, the term refers to the data.

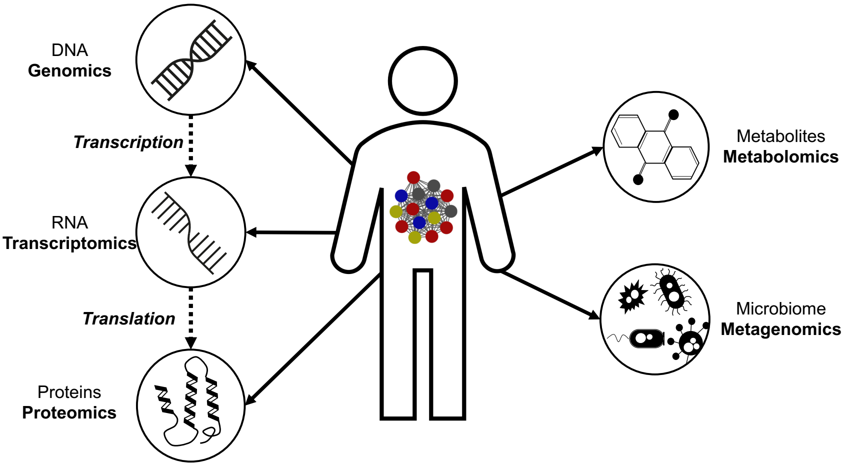


Figure 1 Analytes and their omics.

The use of omics has gained increasing momentum (Figure 2) since the completion of the draft sequence of the human genome was announced by US President Bill Clinton and British Prime Minister Tony Blair in June 2000[5]. This increase is driven mainly by reductions in the costs of next-generation sequencing technology and improvements in the quality of the related tools[6,7]. This development has opened countless new opportunities for studying complex biology in a comprehensive manner at high levels. Here, we discuss the five most popular omics types (Figure 1) and examples of their applications.

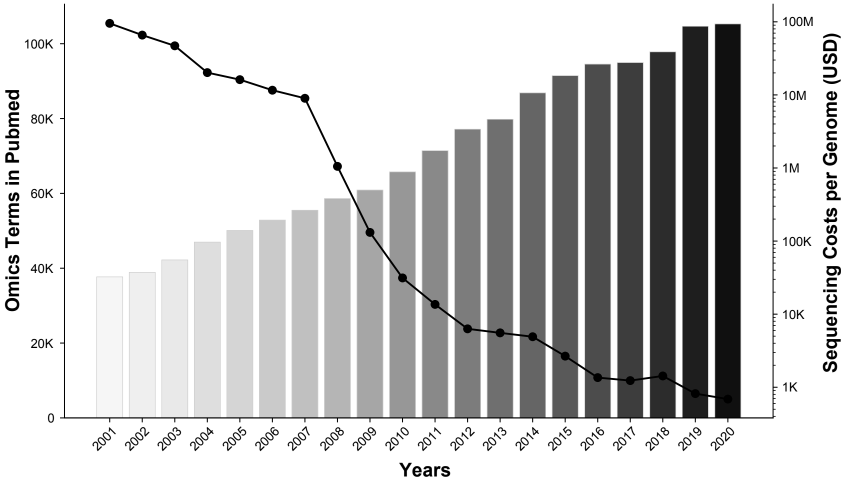


Figure 2 Omics Trend and Decreasing Costs
The decrease in sequencing costs is followed by an increase in the use of omics in biological research (based on a PubMed query, accessed 31-01-2021).

The first omics is **genomics**, which involves the study of the whole genome. These data originate from DNA and consist of DNA profiles and features. Using genomics data, one can retrieve DNA sequences, discover variations in DNA structures or single-nucleotide polymorphisms (SNPs, variations in nucleotides between individuals) between individuals or conditions, identify regulatory factors, and conduct many other analyses. Genomics data have been used to find associations between genetic factors, particularly their variations, and diseases, for example, cardiovascular diseases[8] and diabetes[9]. Many databases for pre-analyzed genomic variations and their association with diseases are currently available, and these include the 1000 Genomes Project database[10] and DisGeNET[11].

Transcriptomics refers to the study of the expression of RNA transcripts, including protein-coding RNA (messenger RNA or mRNA) and noncoding RNA, such as transfer RNA, ribosomal RNA, and microRNA. Instead of retrieving DNA structures, transcriptomics data quantify RNA expression and often focus on protein-coding mRNA transcripts. One can then calculate the significantly differentially expressed transcripts or genes between different conditions (the methods for these analyses are discussed in the next section). This omics approach can successfully be applied to understand the mechanism of action or to the identification of signatures and possible therapeutic targets of diseases, such as myocardial infarction[12] and prostate cancer[13]. The Pathology Atlas[14] is another outstanding example of the utilization of large-scale transcriptomics data and provides a database of prognostic gene markers for 17 different cancer types.

Genomics and transcriptomics data are most commonly generated by two popular high-throughput sequencing (HTS) methods: microarray and next-generation sequencing (NGS). In a microarray, the cDNA of the sample is prepared in a chip with a large array of predefined probes to detect the relative abundance of RNA transcripts[15]. Transcripts can only be detected if they are included in the probe set. In contrast, in NGS, the DNA or RNA samples are fragmented, sequenced, and aligned, and mapped to a reference genome using bioinformatics tools[16]. Even though microarrays are more cost-effective, their transcript detection range is limited by the probes. In contrast, NGS detects all transcripts in a sample without the need to set probes or have any prior knowledge of the system and thus exhibits a larger detection range.

The next omics is **proteomics**, which involves the study of proteins on a large scale. Proteomics data are employed to observe protein behaviors, such as their synthesis, degradation, and modification. Proteomics data can be used to extract protein sequences and quantify protein abundances. Subsequently, one can investigate changes in protein abundance between different conditions through differential expression analysis (the methods are discussed in the next section). There are two popular paradigms in proteomics[17] based on the methods used for data collection: mass spectrometry (MS)-based and affinity-based proteomics. Similar to NGS, MS-based methods detect all available protein signals in a sample, whereas affinity-based methods are limited to the available antibodies in the assay. Based on a search of PubMed, proteomics data are the most popular omics data after genomics data due to their significance and large spectrum of usage. Some examples of the use of proteomics data are the identification of protein biomarkers in Alzheimer's disease[18], the exploration of proteomics changes during aging[19], and the discovery of novel plasma proteins associated with nonalcoholic fatty liver disease[20].

A small set of metabolites, intermediary or end products of metabolic processes, has been used as disease diagnostic tools for a long time [21], but in the omics era, this set has been expanded into a larger set of metabolites to be analyzed, including lipids, amino acids, and fatty acids. The study of metabolites is referred to as **metabolomics**. Based on the methodology used to acquire these data, metabolomics can be divided into two types: untargeted and targeted metabolomics. In untargeted metabolomics, all metabolites, including unknown metabolites, are measured, whereas in targeted metabolites, only characterized and annotated metabolites are measured. In both types of metabolomics, the metabolites are measured by MS. The Human Metabolome Database (HMDB)[22] includes 114,100 metabolites (18,609 metabolites have been detected and quantified), including endogenous, food, microbial, and other metabolites. Metabolomics has been positioned to bridge other omics to the actual phenotype[23]. Naturally, metabolomics has also been popularly used for studies of diseases, particularly metabolic diseases, such as the characterization of fat depots in NAFLD[24] and their association with multiple cardiovascular diseases[25].

Metagenomics refers to the study of the microbiome (also known as our second fingerprint[26]) that lives in the body of the host. The microbiome usually originates from the feces (gut microbiome) or saliva (oral microbiome). Two popular HTS approaches are used to generate metagenomics data: shotgun and 16S sequencing. The shotgun method provides a larger and less focused spectrum of the microbiome, whereas 16S sequencing focuses on the 16S ribosomal RNA gene and is hence restricted to bacteria and archaea[27]. Metagenomics analyses generally start by comparing the taxonomic structure and diversity in response to different perturbations, and functional analyses can then be performed using metabolic modeling or marker genes from the microbial gene catalog. In the past decade, metagenomics has gained popularity in disease research because the microbiome is highly associated with the host conditions[28]. The Human Microbiome Atlas (https://www.microbiomeatlas.org) is a comprehensive database that provides information on human microbiome samples from all over the world, including their association with diseases.

Other types of omics, including **fluxomics**, **interactomics**, and **phenomics**, have been proven to be useful in biological research, particularly research in diseases[29]. However, the amount of data generated can be overwhelming and can blind us to the fact that these types of data are connected. The paradigm of **systems biology** helps us look at and think of a biological system not as singular and detached subsystems but rather as one large interconnected system.

Systems Biology Paradigm

For a long time, biological research, including molecular biology, has been limited by the complexity of the matter itself, which has resulted in the development of simplified approaches to problems. Moreover, exploration of an entire biological system has been impractical due to methodological limitations. Nevertheless, despite these limitations, researchers have been able to unearth amazing insights about life and, more importantly, discover life-saving drugs or therapies for diseases. Due to the -omics data boom, biology has become a data-rich field, which has triggered the emergence of a new field, **systems biology**. There are multiple definitions of systems biology, but the definition I will use in my thesis stems from the Institute for Systems Biology (http://isbscience.org), which defines systems biology using three main keywords: **holistic** (understanding biological systems as a whole interconnected system), **collaborative** (teaming multidisciplinary experts, including biology, physics, computer science, etc.), and **predictive** (predicting changes in the systems due to different conditions).

In contrast to the conventional paradigm that considers a biological system as an aggregation of its subsystems, systems biology considers biological systems as one whole system by building an interconnected model to simulate the complex interaction between different components in molecular biology. These components might include but are not limited to different omics analytes (Figure 1), such as genes, proteins, and metabolites. To build the model accurately, a large amount of data, including multiple subjects, a large number of different analytes (from single- or multiomics), and different perturbations, are naturally required, and as a result, solid computational algorithms and, most definitely, strong computational power are needed for data analysis and model simulation. These requirements tightly couple biology with the field of computer science as well as other supporting disciplines, such as statistics, physics, and chemistry, which makes systems biology a multidisciplinary field. In addition to understanding a system, the objective of building a system-level biological model is to envisage the changes in the system due to variation in conditions. Variation in conditions might translate to the study of different time points to understand the progression of a disease, the comparison of control versus disease samples to elucidate disease mechanisms, the exploration of different therapeutic strategies to find the most suitable treatment for a disease, or other variations. These explorations, comparisons, and predictions are typically approached using three main approaches, namely, statistical inference, machine learning, and network analysis (Figure 3), and these are discussed in more detail in the next chapter.

The application of systems biology and big data has become a norm because this information can help accelerate research processes and ultimately reduce the cost of research[30]. Moreover, it provides broader insights into the disease, its pathophysiological responses, and its molecular mechanisms[31]. Systems biology has been broadly used in this context, including for the discovery of new therapeutic approaches, the repositioning of known drugs to the treatment of multiple diseases[13,32-34], the identification of novel biomarkers, and patient characterization[35-37].

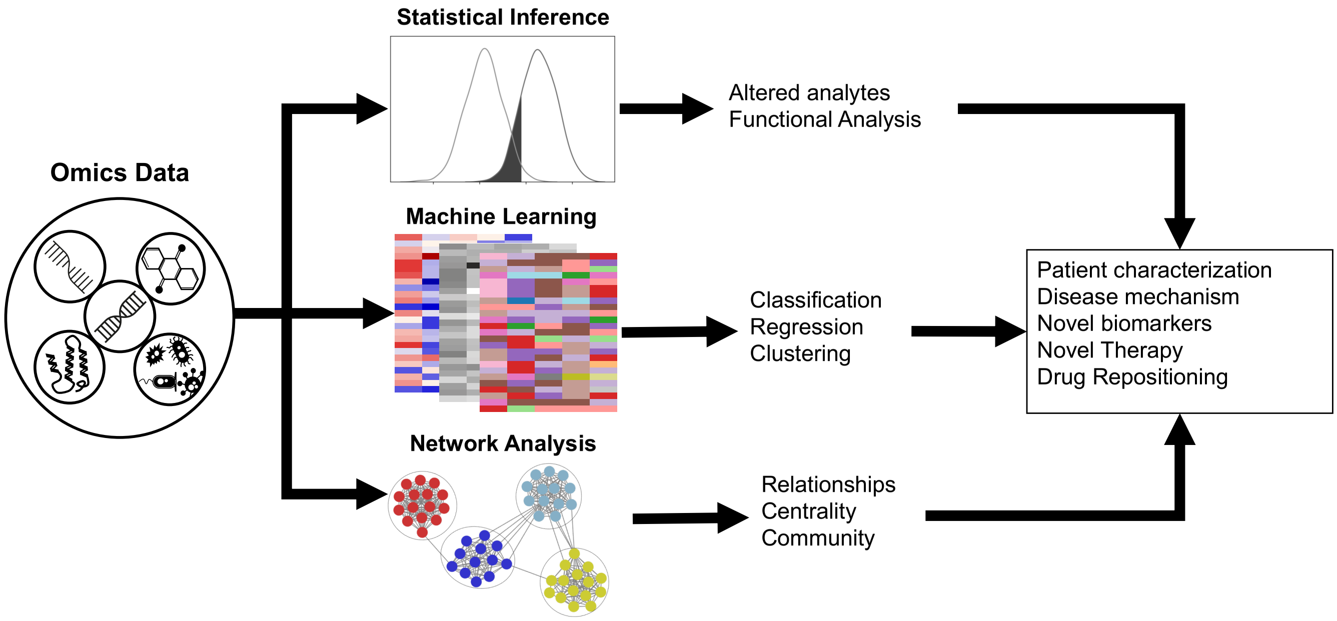


Figure 1 Flow of systems biology

Chapter II: Systems Biology and Biological Networks

In this chapter, three systems biology approaches and their examples are discussed: (1) statistical inference, (2) machine learning, and (3) network analysis, which is our main focus. Moreover, we discuss several biological network types and their components. Finally, at the end of the chapter, we discuss the applications of systems biology using one or combinations of the discussed methods in the context of human diseases.

Statistical Inference and Functional Analysis

The first and most common approach is to perform a side-by-side comparison between different conditions and find significantly different analytes between the conditions. This approach involves the application of inferential statistics methods, ranging from frequentist methods (Student's t-test, Wilcoxon test, ANOVA, and MANOVA) to improved and/or specialized statistical inference methods (LIMMA and DESeq2). Because the comparisons are performed for a large number of analytes, multiple hypothesis testing methods (Bonferroni correction and Benjamini/Hochberg FDR) are often needed to minimize inaccurate inference in the analyses. This approach has been proven to help researchers identify significantly altered analytes, for example, in drug repositioning using transcriptomics data[12] and in metagenomics analysis of asthma patients[38].

Looking at just significantly differentially altered analytes often does not yield a complete systematic view of the changes. Typically, such analyses are followed by a functional analysis to provide more context, such as mapping the data to biochemical pathways, biological functions, disease association, and phenotypes. Many databases, such as KEGG Pathway[39], Gene Ontology[40,41], MetaboAnalyst[42], DisGeNet[43] (for disease association), and Cancer Cell Line Encyclopedia[44], provide contextual annotation for analytes. Further statistical analysis, such as Fisher's exact test, reporter analysis, and overrepresentation analysis, has been employed to show significant changes in the functional context. The combination of statistical inference and functional analysis helps researchers develop hypotheses or find new insights at both the molecular and functional levels, such as finding the therapeutic effect of inhibiting a pathway in cancer[45].

Machine Learning

Machine learning (ML) is described as a data analysis technique where computer algorithms are developed to **learn from the data** and **find hidden insights** from it **without being specifically programmed** to look for them[46]. It is a proven tool for the analysis of large and complex data, including biological data[47]. There are two main objectives in the use of ML in biological data: discovery and prediction. First, ML is used **to discover** patterns in the data and find important features that discriminate two or more conditions. Second, the discovered patterns and important features are used by the ML algorithm to build a model to represent the data that can further be used **to predict** conditions or analyte levels of independent sets of data.

There are two main methods in ML: supervised and unsupervised learning. **Supervised learning** is a method where the discovery and prediction objectives require a known label for the samples or a dependent variable, such as the metadata or phenotypes. There are two subtypes of supervised learning: **classification** (prediction of discrete variables, e.g. stratification of subject conditions) and **regression** (prediction of continuous variables, e.g. prediction of analytes levels). Several examples of supervised learning are support vector machine, decision tree, random forest, and neural network. The opposite is **unsupervised learning**, where the identification of patterns in high-dimension omics data is performed without any prior labeling or known dependent variables. Unsupervised learning includes clustering methods (hierarchical, K-means, and spectral clustering) and dimensionality reduction (PCA, T-SNE, and UMAP). Typically, the latter approach is used in the exploratory stage during data analysis to control the quality of the data and obtain unbiased insights from the data.

There are many examples of successful biological and disease research studies using machine learning. Loomba et al. found a metagenomics signature in NAFLD using random forest[48], and combinations of multiple machine learning algorithms have been used to improve the top-down proteomics data analysis approach[49].

Network Analysis

Networks are often used as a tool to untangle the complexity of biology[50] by understanding the relationships, mainly functional relationships and physical interactions, between the analytes within and/or between different omics. Networks can be inferred using computational calculations and/or experimental data. Before discussing biological networks and their role in systems biology, we first discuss networks and their attributes in general.

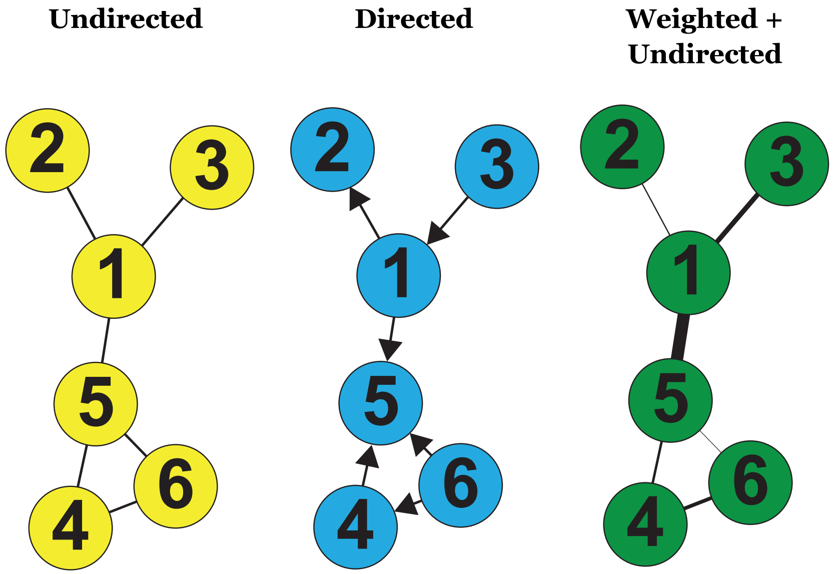
A network is constructed with two main components: **nodes** (or vertices) and **edges** (or links). The individual points in a network are referred to as nodes, which might represent any type of omics component, such as genes, proteins, and metabolites. The relationships between nodes are called edges, and multiple types of edges are used depending on the embedded relationship information (Figure 4).

First, two types of edges are based on directionality: **undirected** and **directed** edges. An undirected edge contains no information on the movement of information in the network; it shows only the presence of a relationship between the connected nodes. In contrast, directed edges show the direction of information flow between the connected nodes. For example, an undirected edge connecting A to B indicates that “nodes A and B have a relationship”, whereas a directed edge between these nodes, i.e., A→B or B→A, indicates that information flows from A to B or B to A, respectively. The other types of edges are **unweighted** and **weighted** edges. Unweighted edges indicate the availability of a relationship, whereas weighted edges provide quantitative values for the relationships. Examples of these quantitative measures are correlation scores[51], evidence scores[52], and experimental or empirical measures[53].

Computationally and mathematically, the topology of networks is represented as a square matrix called an adjacency matrix, in which the rows and columns are the nodes in the network. The matrix contains numerical representations of the edges and can be symmetric or asymmetric (to represent an undirected or directed network, respectively) and binary or nonbinary (to represent an unweighted or weighted network, respectively). Using the adjacency matrix, topological analyses can be performed easily with relatively simple matrix operations. Topological analyses include analyses of the topological distance, centrality, and community and often reveal more insights than those obtained from an analysis of individual network nodes/edges.

$$A_{i,j} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & & A_{2,n} \\ \vdots & & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,n} \end{bmatrix}$$

Equation 1 Adjacency matrix of a network with n nodes.

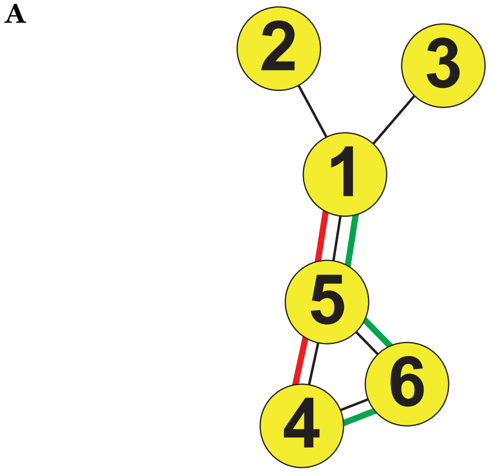


$$\begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 2 & 8 & 0 & 20 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 6 \\ 20 & 0 & 0 & 4 & 0 & 1 \\ 0 & 0 & 0 & 6 & 1 & 0 \end{bmatrix}$$

Figure 4 Network types and their adjacency matrices.

Topological Distance

The interactions between nodes within a specific network are calculated by their topological distance in the network. The topological distance also represents the available path for information flow from one node to another. For example, information can flow between node 1 and node 4 in **Figure 5A** via two paths, which are represented by red and green edges and have lengths of 2 and 3, respectively. The red edge can be referred to as the **shortest path**, which refers to the path with the minimum number of edges between two nodes (denoted as $d_{i,j}$; in this case, between nodes 1 and 4 ($d_{1,4}$)). In an undirected network, the shortest paths between nodes i and j and between j and i ($d_{i,j} = d_{j,i}$) are equal, but this is not always the case in a directed network. The maximum shortest path in a network is called **the network diameter** (D). As shown in **Figure 5B**, the diameter of the network illustrated in **Figure 5A** is D = 3.



B

$$d_{i,j} = \begin{bmatrix} 0 & 1 & 1 & 2 & 1 & 2 \\ 1 & 0 & 2 & 3 & 2 & 3 \\ 1 & 2 & 0 & 3 & 2 & 3 \\ 2 & 3 & 3 & 0 & 1 & 1 \\ 1 & 2 & 2 & 1 & 0 & 1 \\ 2 & 3 & 3 & 1 & 1 & 0 \end{bmatrix}, \sum d_{i,j} = \begin{bmatrix} 7 \\ 11 \\ 11 \\ 10 \\ 7 \\ 10 \end{bmatrix}$$

Figure 5 Shortest Path
(A) Two paths exist between nodes 1 and 4, and these are denoted by red and green paths.
(B) Distance matrix of the shortest path and its sum.

Centrality Analysis

Centrality analysis aims to show the importance of the nodes within a network. Many metrics can be used to measure node importance. The most common parameter used to rank nodes is **degree centrality**, which represents the number of edges that connect the node of interest to other nodes. In undirected networks, the node degree is calculated by the sum of the column or row of the adjacency matrix of that specific node. A directed network has two types of degrees: in-degree and out-degree (both are the same in undirected networks). The in-degree is calculated based on the edges that arrive at the node, and the out-degree is calculated based on the outgoing edges from the node. In other words, the in-degree and out-degree are the sum of the columns and rows of the adjacency matrix, respectively. Using the example from **Figure 4**, node 1 has a degree of 3 (undirected network) as well as an in-degree of 1 and an out-degree of 2 (directed network).

Degree information can also be used to identify whether a network is random or scale-free. In a random network, the degree distribution of the nodes follows a uniform distribution, whereas in a **scale-free** network, this distribution follows a power law. Due to this power-law degree distribution, a scale-free network tends to have several nodes with significantly high degrees (called **hubs**), whereas the rest of the nodes have low degrees. Many real-world networks, including biological networks[54-57], have been shown to exhibit scale-free characteristics. In scale-free networks, hubs are considered important in the network because their removal might cause network dissociation or affect a large number of other nodes. As an example, the deletion of or a pathogenic attack to a hub in a protein-protein interaction network (PPIN) can disrupt the network[58,59] and even cause a lethal phenotype[60,61].

The **clustering coefficient (CC)** or **transitivity** is another centrality parameter. In contrast to degree centrality, which focuses on the edges connecting a node to its neighbors, CC assesses the edges connecting the neighbors of a node. Formally, CC is defined as the fraction of the connected degree of a node's direct neighbors to one another[50]. If k_x is the degree of a specific node x and N_x is the number of edges between the direct neighbors of x, then the CC of node x is defined as

$$CC_x = \frac{2N_x}{k_x(k_x - 1)}$$

For example, for the undirected network shown in **Figure 4,** the CC of node 5 is based on the fraction of edges connecting its direct neighbors (nodes 1, 4, and 6). In this case, k_5 = 3 and N_5 = 1 (only edge between node 4 and 6), hence

$$CC_5 = \frac{2 \times 1}{3(3-1)} = \frac{2}{6} = \frac{1}{3}$$

The pattern of CC has also been found in biological networks, such as networks of disease genes[62], and has been used to define key communities in large integrated biological networks[63] (community analysis is discussed in the next subchapter).

Another important parameter in centrality analysis is **betweenness centrality**. Unlike the previous metrics, which measure the physical connections between nodes, betweenness[64] is calculated by the number of shortest paths passing a specific node. Conceptually, this variable measures the importance of a node to the flow of information in a network. In the undirected network shown in **Figure 4**, node 1 has the highest betweenness score in the network because it is located inside the seven shortest paths in the network (2-3, 2-5, 2-4, 2-6, 3-5, 3-4, and 3-6), whereas node 5 has a centrality score of 6 (1-4, 1-6, 2-4, 2-6, 3-4, and 3-6). Other nodes have a betweenness score of 0. Nodes with high betweenness scores tend to act as bridges between two otherwise separated parts of a network. The deletion of notes with high betweenness scores might cause the network to split into smaller networks. As an example, if we remove node 1 from the network shown in **Figure 4** (undirected), nodes 2 and 3 will be separated from the rest of the network, whereas the removal of nodes 2 or 3 will not affect the network structure. In more modular networks, such as a PPIN, this metric has been found to be essential[65]. In transcriptional regulatory networks (TRNs), the betweenness centrality exhibits a positive correlation with the node degrees[66].

Based on the shortest path concept, another centrality metric, called **closeness centrality**, can be calculated. Closeness is calculated by averaging the shortest path from a node to other nodes in a network and is calculated by dividing the number of neighboring nodes (N-1, where N is the number of nodes in the network) by the sum of the shortest path $\sum d_{ij}$. For the network shown in **Figure 5A**, the closeness scores for nodes 1-6 are 0.714, 0.455, 0.455, 0.5, 0.714, and 0.5, respectively. A node with a higher closeness indicates that the node is closer to the other nodes in the network, and hence, the node has a more central role in the network. The closeness values of genes in a gene coexpression network (GCN), together with their betweenness and degree values, varies more in hepatocellular carcinoma (HCC) samples than in noncancerous samples[67]. Another study showed that central metabolites in a metabolic network can be identified by comparing their closeness centrality scores[55].

Additional centrality metrics that can be used to define the importance of network nodes are not discussed in this thesis. Some examples that have been popularly used in biological contexts are **eigencentrality/eigenvector centrality**[68,69] and **eccentricity**[69], which have been found to be highly positively correlated in several PPINs and TRNs[69].

Centrality analyses have revealed many hidden and novel biological insights. A previous study[70] showed distinct centrality and betweenness patterns in a PPIN of Mendelian and complex disease genes. Another study[71] showed that in three eukaryotic PPINs, proteins with high degree, betweenness, and closeness values are likely to be essential in organisms. A combination of multiple centrality metrics can increase the power of network analysis. Rio et al[72] employed 16 centrality metrics to analyze 18 metabolic networks of *Saccharomyces cerevisiae* with the aim of identifying the essential genes. These researchers concluded that the combination of at least two metrics can accurately predict the essential genes in the network, whereas none of the metrics alone were able to do identify these genes. Similarly, Wang et al.[73] combined seven centrality metrics to successfully predict structurally dominant proteins in a yeast PPIN. These examples, together with the other above-mentioned examples, show the importance of centrality analysis in the assessment of biological networks, including for the discovery of novel biomarkers and, possibly, candidate therapy targets in diseases.

Community Analysis

Because biological networks are built based on the relationships among analytes and molecules, their networks become very complex. Gene coexpression networks (GCNs) can have >10000 gene nodes[51] within a network, PPINs might have >9000 nodes (>64000 edges)[53], and metabolic networks can have >3500 genes, >4000 metabolites, and >13000 reactions[74]. Analyzing the entire network at once can result in dilution of information or accidentally missing important findings due to the overwhelming complexity. To avoid those problems, it is always a good idea to partition a network into smaller subnetworks. This partitioning can also help the elucidation of structural or functional similarities within the subnetworks[75-78]. These subnetworks are called **network communities** or **clusters**.

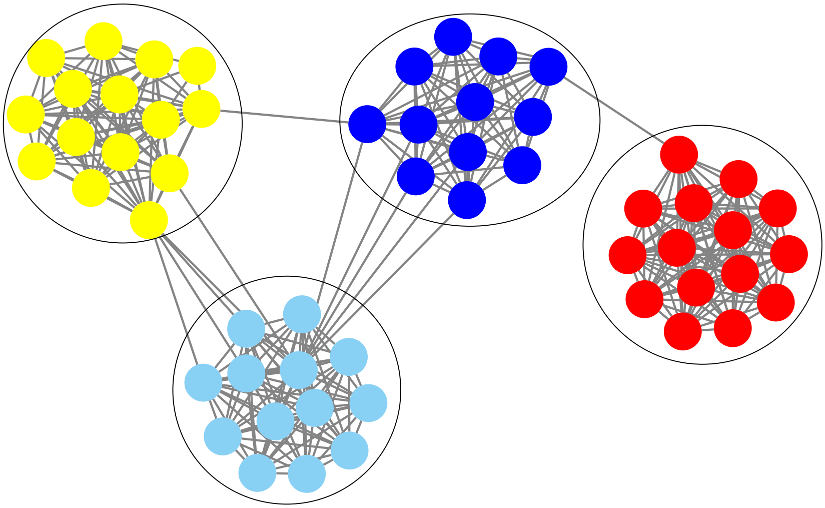


Figure 6 Network communities.
Each node color represents a community/cluster.

A network community is a group of nodes that are highly connected to one another and have fewer connections to nodes from the other group or the rest of the network (Figure 6)[79,80]. In 2002, Ravasz et. al.[81] introduced one of the first proofs about the existence of communities in biological networks. These researchers proved that metabolic networks of 43 organisms are hierarchically constructed of smaller, but dense, subnetworks (communities). Since then, community analysis has evolved to be a powerful tool in biological network analysis, such as for the discovery of disease/tissue-specific genes in gene coexpression networks[63] or for the identification of analytes associated with physical conditions and diseases in multiomics longitudinal networks[82].

Another similar concept that we do not discuss in depth in this thesis is the notion of a **clique**. A clique is a subnetwork in which all the nodes are connected to each other (fully connected subnetwork). The network shown in Figure 4 has seven cliques: six of these cliques have a size of 2 (1-2, 1-3, 1-5, 4-5, 4-6, and 5-6), and the other clique has a size of 3 (4-5-6). The term clique is different from network community, but in some cases, a clique can be synonymous to or part of a community. In a network community, there is no requirement that the subnetworks have to be fully connected. Another main difference between clique and community (in the context of this thesis) is the exclusivity of the membership. A node can be a member of multiple cliques but can only be a member of a single community.

One of the classical approaches for detecting communities within a network is hierarchical clustering[83], which finds the similarities among the nodes based on the network structure (possibly derived from the adjacency matrix). Two main approaches are used for hierarchical clustering: bottom-up (agglomerative or *Ravasz* algorithm)[81] and top-bottom approaches (divisive or *Newman-Girvan* algorithm)[84]. The agglomerative approach starts by assigning each node to a group, and this step is followed by the joining of two groups with the highest similarity. The process is iterated until all nodes are combined into one large group. Multiple approaches have been developed for measuring the similarity between groups: minimum distance (single linkage), minimum distance (complete linkage), and the average distance between the nodes in different groups. In contrast, the divisive approach starts by including all the nodes in one group and then iteratively splits the group into two groups with the lowest similarity. The iteration ends when the preferred number of communities is found. Hierarchical clustering has been popularly used in biological contexts, such as for the identification of tissue similarities based on their transcriptomics profiles[85], and the agglomerative approach has been shown to perform well in detecting communities in gene expression data[86].

During its iterative process, the hierarchical clustering algorithm creates multiple sets of communities. The next question is which set of communities best represents the network. As discussed earlier, we know that a good community is a community with dense connectivity within but sparse connectivity with the rest of the network. In 2006, Newman[87] introduced a concept called **modularity** to define the goodness of a community structure. The modularity of a community c (M_c) is calculated by comparing the actual edge density in the community to the expected edges in a random network with the same degree distribution. The edge density in a community (m_c) is the actual number of edges in the community, L_C, divided by the total possible edges in the same community.

$$m_c = \frac{L_c}{E(L_c)} = \frac{L_c}{n_c(n_c - 1)/2}$$

$$M_c = \left(m_c - E(m_c)\right)$$

where n_c is the number of nodes in community c. Because generating multiple random networks and calculating its edge density to obtain $E(m_c)$ can take considerable computational power and time, the formula can be simplified[50] using information from the network adjacency matrix A_{ij} as follows:

$$M_c = \frac{1}{2L} \sum_{(i,j) \in c} (A_{ij} - \frac{k_i k_j}{2L})$$

where L represents the edges in the entire network and k represents the degree of a specific node. Finally, the network modularity score can be calculated as the sum of all community modularity values M_c,

$$M = \sum_{c=1}^{C} M_c$$

where C is the total number of communities in the network. Naturally, a higher modularity score indicates better network partitioning. For example, to define the best set of communities from the hierarchical cluster, one can calculate the network modularity scores of all sets and obtain the set with the highest modularity score.

Many other community detection algorithms, such as random walk (walktrap)[88], infomap, and spinglass algorithms, have been developed. One of the most popular community detection algorithms is called the **Leiden algorithm**[89], which was built based on network modularity score optimization. The Leiden algorithm takes the basic steps from its predecessor, the Louvain algorithm[90]: (1) local node movement and (2) network aggregation. In the first step, a node is moved around to different communities and stays in the community that yields in the highest increase in the network modularity score. Next, communities are aggregated into larger subnetworks, and the network modularity score is recalculated. These steps are repeated until the network modularity score cannot be further improved. The problem with the original algorithm is that it often results in badly interconnected communities[89]. The Leiden algorithm integrates the smart local move algorithm[91] to give the algorithm the ability to also split, and not only merge, the communities. This improvement was shown to not only increase the modularity score but also eliminate the badly connected communities. Moreover, despite its increased complexity, the new algorithm also increases the calculation speed. The Leiden algorithm (and Louvain) also provides another community quality measure called the constant Potts model (CPM)[92], which is supposed to resolve the resolution limit problem[93]. The resolution limit is the inability of the modularity to detect small communities, which results in the existence of subcommunities.

The Leiden algorithm has been used widely in biological contexts, such as for the identification of microbiota clusters in a multiomics network that respond to vaccination[94]. In single-cell transcriptomics analysis (scRNA-seq), the algorithm plays an integral role in cell cluster identification. This algorithm is embedded as the main clustering algorithm in many famous analysis tools, such as SCANpy[95] and Seurat[96], and has been used in many contexts, such as for understanding the mechanism of action of Alzheimer's disease[97] and COVID-19[98].

Biological Networks

There are many well-known biological networks depending on the type of omics included and how the relationships are inferred. In this section, several networks are discussed, and the discussion includes their omics, how they can be leveraged in biological research (focusing on disease research), and examples of their applications.

Based on Figure 7, the most popular network in biology is the **metabolic network**. Metabolic networks represent the metabolic reactions and processes in specific organisms. In this thesis, we narrowed the definition of metabolic networks to genome-scale metabolic models (GEMs)[74], which are curated networks based on the combination of all metabolic pathways in an organism. GEMs consist of four main components: metabolic reactions, genes, and proteins as the connecting edges and metabolites as the nodes. Many databases, such as Metabolic Atlas (https://metabolicatlas.org) and Virtual Metabolic Human (https://www.vmh.life/), provide well-curated GEMs for different organisms and tissues.

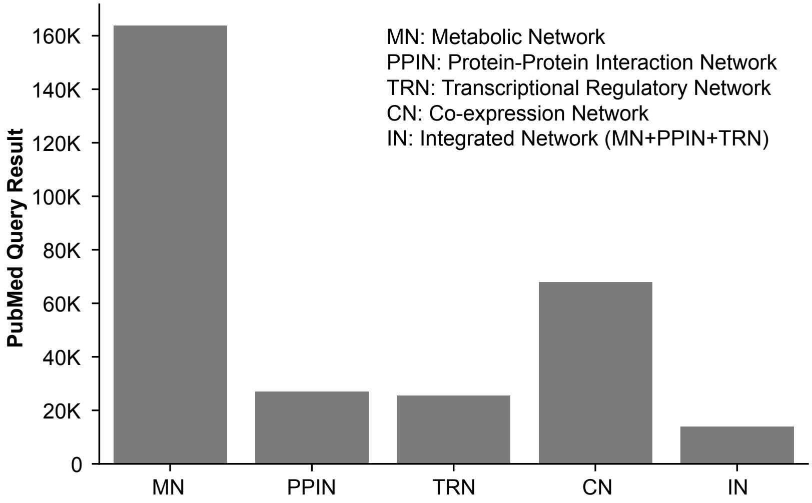


Figure 7 Network Popularity
Network popularity based on a PubMed query (accessed 05-02-2021).

Many analyses can be performed using GEMs, and these include metabolic flux prediction, essentiality analysis, and reporter metabolite analysis. Two well-known approaches have been developed for the prediction of metabolic fluxes: flux balance analysis (FBA)[99] and flux variability analysis (FVA)[100]. FBA predicts metabolic fluxes by using linear programming (LP) optimization to minimize or maximize an objective function, such as biomass growth or ATP production or consumption[101], based on mass-balanced constraints (in-fluxes are equal to out-fluxes). FVA performs both maximization and minimization (double LP problem) to obtain the range of metabolic fluxes that will maintain the mass balance in the model. In addition to flux prediction, we can also perform essentiality analysis to determine the importance of each gene, metabolite, or reaction by blocking them one by one from the model and observing the resulting changes in the objective functional flux. Essentiality analysis has been used for many applications, such as the discovery of antimetabolites that can inhibit the growth of HCC[102] and the prediction of essential genes in renal cell carcinoma[103]. Finally, reporter metabolite (RM)[104] refers to an integration of a network with transcriptomics data. This analysis identifies metabolite alterations based on transcriptional changes among different conditions. Lee et al.[105] successfully employed RM to identify the association between mannose and insulin resistance. Overall, the applications of GEMs have shown that these networks are a proven and powerful tool for simulating metabolic processes[106].

The second network is the **protein-protein interaction network (PPIN)**. A PPIN represents the physical and functional interactions between proteins and consists of proteins as the nodes and their interactions as the edges. Protein interactions are crucial for the proper function of organisms[107]; hence, understanding a PPIN is important. PPIN analysis has been used in many applications, including analyses of the similarity and repositioning of drugs for NAFLD and Alzheimer's disease[108] and, together with differential gene expression analysis, the identification of novel therapeutic targets for lung squamous cell carcinoma[109]. Performing community analysis with a PPIN has also proven to be beneficial for finding biomarkers[110,111]. One of the most comprehensive references of human PPIN is The Human Reference Interactome[53], which has systematically tested up to 17500 proteins, including 9094 proteins and 64006 physical interactions that have been curated in the database. Another important database for PPINs is StringDB[52], which derives interactions from multiple other databases, pathway information, and coexpression networks.

Another important and well-known network is the **transcriptional regulatory network (TRN)**, which represents the relationships between transcription factors (TFs) and their regulated genes. TRNs are generally reconstructed using genomics data[112], such as ATAC-seq, ChIP-seq, and DNase-seq data[105]. TRNs are important for understanding the dysregulation of genes that can lead to diseases[113], such as for the identification of master regulators that can be used as targets for glioblastoma therapy[114]. In most cases, TRNs are used together with other types of networks to amplify their information[115,116]. In 2016, Lee et al.[105] introduced a concept called an **integrated network (IN)** that combines GEMs, PPINs, and TRNs. By combining transcriptomics and RM analyses, these researchers were able to discover that plasma mannose exhibits good insulin resistance, regardless of the subjects' BMI.

Finally, a **gene coexpression network (GCN)** is a network that shows correlations among gene expression based on transcriptomics data. The edges represent the correlation scores between two nodes (genes). WGCNA[117] (weighted gene correlation network analysis) is a popular method for generating and analyzing GCNs, and other researchers[63,118] have used basic correlation analyses to generate GCNs. Similar to the PPIN, the community analysis of a GCN is beneficial for untangling the complexity of the network, such as for the functional annotation of unknown and noncoding genes[119] and the identification of key clusters associated with diseases[63,120]. Because an increasing amount of large multiomics data have been generated in the past few years, the same approaches used with GCNs have been adopted for **multiomics biological networks (MOBNs),** which represent the omics analytes as the nodes and have been used to decipher the complexity of human physiology[121] and diseases[82].

Systems Biology of Complex Diseases

Personalized Medicine

One of the futures of healthcare is P4 (predictive, personalized, preventive, and participatory) medicine[122], which is often called precision or personalized medicine. Personalized medicine is aiming at tailoring treatments of disease to each patient characteristic, as opposed to treating the patients based on the general disease attributes. Rather than using the mindset of "one treatment fits all" for all patients, they are monitored continuously to capture the most accurate disease characteristics and their treatment can also be continuously optimized based on their current state[123]. Not only that this will be more beneficial for the patients, but also be predicted to decrease drug prices and healthcare-related expenses[123-125]. One of the most successful examples of personalized medicine is the discovery of HER2-positive breast cancer type and its targeted therapy that is detected in around 20% of the patients[126].

To facilitate better and suitable individual treatment, more advance and precise patient characterization are required. Coupled with large number of generated omics data, systems biology plays an integral role in accelerating personalized medicine by disentangling the complexity of diseases systematically and holistically. In the next section, we will discuss several examples of systems biology applications that drive the advancement of personalized medicine.

Application of Systems Biology Tools in Personalized Medicine

Before the omics era, patient characterization was mostly performed using patient and family history, imaging (e.g., MRI and ultrasound), electrical signals (e.g., ECG), invasive surgery (e.g., biopsy), and/or molecular data from the blood. Omics and systems biology have contributed significantly in this context by adding further resolutions and layers of information, which might also lead to the **discovery of novel disease mechanisms of action**[29].

Bidkhori et al.[67] discovered new subtypes of HCC tumors based on metabolic networks. These researchers used clinical and transcriptomics data from a publicly available cancer database, The Cancer Genome Atlas (TCGA), and combined these data with an HCC-specific GEM to generate functional gene-gene networks. Based on the network, these researchers identified three HCC subtypes with significantly different molecular and functional signatures that affect patient survival. Benfeitas et al.[127] used a similar approach involving focusing on network analysis and multiomics data integration to stratify HCC into two subtypes based on their redox behavior. Bailey et al.[120] retrieved DNA and RNA data from 382 pancreatic cancer (PC) subjects (and added 74 published samples). Using genomics and gene coexpression network analysis, these researchers discovered four novel PC subtypes with novel gene and pathway signatures for each subtype that could be used as candidate novel biomarkers for PC.

Mardinoglu et al.[128] built personalized GEMs of 86 subjects with hepatic steatosis (HS) to explore the molecular mechanisms of NAFLD. These researchers found alterations in NAD+ and glutathione and negative correlations between glycine and serine in the model. Furthermore, these researchers generated metabolomics data from the same subjects and performed supplementation experiments using mice to validate their findings. They also performed a serine supplementation study with six human subjects and found a decrease in HS. This finding became the basis of a follow-up supplementation study with serine and other metabolic cofactors[129]. Through the integration of GEM, metabolomics, and proteomics data from plasma, these researchers were able to show alterations in lipid, amino acid, and antioxidant metabolism in the subjects, and these findings strengthen the hypothesis that the supplement can be used for NAFLD treatment.

Another successful application of systems biology that is important in the personalized medicine context is **drug repositioning**. Turanli et al.[13] developed a prostate cancer-specific GEM and integrated it with transcriptomics profiles from >1000 drugs. These researchers identified *ifenprodil*, among others, as a candidate drug for prostate cancer and validated their findings by performing *in vitro* experiments. Tian et al.[12] attempted to use valproic acid, a drug prescribed for seizures and bipolar disorder, to treat myocardial infarction. Their experiment in mice showed a successful 50% reduction of infarction. Moreover, by transcriptomics data analysis, these researchers revealed that *Foxm1* is the mediator of the drug responsible for the heart-protective effect. Mannarino et al.[130] explored the mechanism of trabectedin, a drug for sarcoma and ovarian cancer, in leukemia cells by gene expression analysis, and found that *MAFB* is the main transcription factor affected by the drug.

These examples show that omics and systems biology bring us closer to personalized medicine because the studies showed that the approaches can aid in patient characterization. Moreover, the researchers were able to identify novel disease mechanisms and molecular signatures that lead to a better understanding of the analyzed diseases and open doors to new treatment strategies. Furthermore, systems biology provides tools for drug repositioning that can not only decrease the financial requirements but also accelerate drug discovery processes[131].

However, we remain a long way from an era of personalized medicine. Studies have proven that systems biology approaches can characterize patients into subtypes and reveal novel biomarkers. However, translating the results to clinical settings as either diagnostic tools or treatment strategies remains a major challenge, particularly due to unexpected individual variations in the real world that cannot be replicated in experimental settings. To overcome this challenge, we need more *N-of-1* studies, where it considers each individual as an observation unit[132], particularly those with diseases[34].

Chapter III: Present Investigation

As the title suggests, this thesis focuses on the development and application of systems biology to reveal the underlying mechanisms of human diseases and to discover novel biomarkers that can accelerate the discovery of better treatment strategies. The aim of the thesis is to build a general framework for big molecular data analysis that would assist in human disease-related research. In this thesis, we present the results from six research projects, including two papers on biological network platforms (**Papers I-II**, https://inetmodels.com) and four papers on the applications of systems biology in disease- and physiology-related research **(Papers III-VI)**. The papers' aims are summarized below, and the papers can be found in the Appendices.

Paper I – This study aimed to build a database and web-based interactive visualization platform of biological networks for exploring functional relationships between genes and their functions. We generated gene coexpression networks (GCNs) and integrated networks (INs) for >60 tissues and cancers. The platform is further developed in **Paper II**.

Paper II – This study aimed to further develop and expand the platform generated in **Paper I** with multiomics data. We included multiomics biological networks (MOBNs) by including **clinical variables** and **proteomics**, **metabolomics**, and **metagenomics** data from multiple independent studies, including three longitudinal wellness profiling studies and three disease-related studies (including **Paper V**).

Paper III – This study aimed to reveal the metabolic crosstalk between the heart and three metabolically active tissues (liver, skeletal muscle, and adipose tissues) after myocardial infarction (MI). We used **transcriptomics** data and applied systems biology approaches, including GCN analyses, to obtain an integrative view of the tissues.

Paper IV – The study aimed to comprehend the effect of long-term training in both men and women. We analyzed **transcriptomics** data to reveal the shifts between trained and untrained subjects and the differences between both sexes. We further compared our results with publicly available data to predict the effect of short-term exercise in metabolic-related diseases.

Paper V – The study aimed to explain nonalcoholic fatty liver disease (NAFLD) pathogenesis using multiomics data. We integrated **clinical variables**, plasma **metabolomics**, plasma **proteomics**, and gut and oral **metagenomics** data to identify key features of NAFLD. We also developed a multiomics predictive model for the characterization of NAFLD patients.

Paper VI – The study aimed to assess a potential NAFLD therapeutic strategy using metabolic cofactor supplementation. In this study, we performed **metabolomics** and **proteomics** analyses and integrated the data into GEMs to identify the acute effects of the supplementation.

In this thesis, the investigation results are divided into four main areas: **generation of biological networks** followed by the application of systems biology to the **heart, muscle, and liver**. The focus of this chapter is the studies that the coauthors, collaborators, and myself have performed in these areas. The strong point of our work is the combination of strong biological knowledge with computational expertise. This combination results in (1) useful and important biological insights for understanding the diseases and (2) reliable computational methods that can be extended and reproduced by other researchers.

Generation of Biological Networks (Papers I – III and V)

Over the past 20 years, the number of research studies using omics data has risen exponentially (Figure 2). This has opened many new opportunities to explore the molecular mechanism of human diseases. However, we also know that the obtained data are enormous and complex due to, among others, their interconnection within the set. To comprehend these complex and interconnected data, we need the right tools for their appropriate analysis. Biological networks have become a popular tool of choice in the analysis of such data. In these studies, we generated gene coexpression networks (GCNs) and integrated networks (INs). We also generated multiomics biological networks (MOBNs) to show the interplay between clinical, anthropometric, proteomic, metabolomic, and oral and gut metagenomics data at the personalized level.

In **Paper I**, we built the first version of TCSBN, a database of biological networks (https://inetmodels.com) with 63 tissue- and cancer-specific

GCNs and three INs from the liver, muscle, and adipose tissues (Figure 8). The goal of this study was to provide a platform for scientists with any level of bioinformatics background to explore the association and functional relationship between genes in a specific context. The GCNs were generated from publicly available datasets of normal and cancer **transcriptomics** data from The Genotype-Tissue Expression (GTEx)[133] and The Cancer Genome Atlas (TCGA)[134] projects, respectively. The normalized count files were filtered to remove genes with low expression (≤ 1 TPM/FPKM), and correlation analyses were performed. Only the top 100 positive and negative correlations (100 of each) were included in the platform. Moreover, INs were built through the integration of genome-scale metabolic models[74], protein-protein interaction networks (PPINs)[53], and transcriptional regulatory networks (TRNs)[105] from each tissue. The relationships of the genes were derived from PPIN and TRN coregulation.

We further improved the platform in **Paper II** (Figure 8). Specifically, we updated the platform from **Paper I** (https://inetmodels.com) with the latest GTEx and TCGA data, which resulted in 87 GCNs. The greatest improvement in the platform was the inclusion of MOBNs based on three *N-of-1* independent longitudinal wellness studies[82,121,135] and three in-house disease-related studies[*] (COVID-19 and NAFLD supplementation and NAFLD baseline study in **Paper V**), and these networks included gender- and disease-specific networks.

For the MOBNs, we included all available **clinical**, **anthropometric, proteomics, metabolomics**, and oral and gut **metagenomics** data from all studies. The data were corrected by age and gender[†] and matched at the personalized level, which makes this platform the first and only platform that provides MOBNs based on personalized data. We generated cross-sectional networks in all studies and delta networks for the wellness studies[82]. Cross-sectional networks show the analyte correlations throughout all visits and data points, whereas delta networks represent the correlation of the changes in the analytes between visits. For all the networks, we performed Spearman correlation analysis to define the relationships between genes, and only significant correlations (FDR <0.05) were included in the platform. In **Paper I** and **Paper II**, we used

[*] Under review by the time of the thesis writing

[†] Except gender-specific networks

examples from NAFLD-related studies with GCNs to validate the results from a study[63] on fatty acid synthase *(FASN*) and MOBNs to validate the potential of combined metabolic activators (CMAs) as a treatment strategy[129] (**Paper V**).

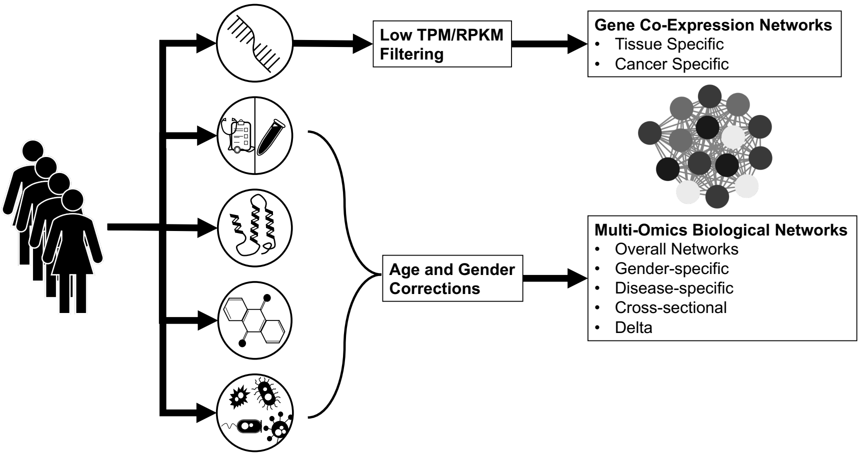


Figure 8 Generation of biological networks in iNetModels.

In **Paper III**, we generated tissue-specific GCNs from the heart and metabolically active tissues (liver, skeletal muscle, and adipose tissues). The same methodology as that described in **Paper I** and **Paper II** was used to build networks from the **transcriptomics** data generated from the mouse models presented in the paper. The aim was to obtain the unique and shared functional relationships between genes in different tissues, particularly to uncover their responses to MI. The networks were further analyzed by performing community analysis using the Leiden algorithm and functional analyses of each community to reveal the community-specific functions. Moreover, we identified the key communities (the highest average clustering coefficient) and tissue-specific communities (based on the tissue-specific genes[136]) of each tissue. Through network analyses and other systems biology approaches, we were able to elucidate the alterations in biological functions and to develop a hypothesis regarding the metabolic crosstalk between the four tissues in this study as a result of MI. We discuss the findings in more detail in the section titled **"Systems Biology of the Heart"**.

Finally, we also generated multiomics data from NAFLD subjects with different degrees of hepatic steatosis (HS) in **Paper V**. A similar methodology as that presented in **Paper II** was used to generate the network, and the network is presented in https://inetmodels.com. The network included analytes from **clinical** data, plasma **metabolomics**, plasma inflammation **proteomics**, and gut and oral **metagenomics** that originated from the same subjects. Through the combination of statistical inference and network analysis, we were able to find key analytes and their relationships with other analytes that were significantly associated with hepatic steatosis. Moreover, through community and functional analyses, we were able to define the role of each community in NAFLD progression. We discuss the findings in more detail in the section titled **"Systems Biology of the Liver"**.

Overall, we have generated >100 biological networks that are collected and presented in iNetModels, an easy-to-use web-based interactive platform (https://inetmodels.com). This platform can be used as an exploration, analysis, and validation tool by anyone, regardless of their bioinformatics background. We also developed a proven framework for biological network generation and a platform for interactive visualization that can be used by anyone. Moreover, we showed the applications of network analysis combined with other systems biology approaches in real disease-related research and show that these can reveal novel disease-related insights. In the next sections, we discuss the results of these applications in more detail.

Systems Biology of the Heart (Paper III)

One of the top causes of death in the world is myocardial infarction (MI)[137], which is generally known by the term "heart attack". Many studies on MI have been performed, and these have provided information on the effect of MI. One of the biggest caveats regarding these studies is that they were limited to a single tissue[138], and as a result, the studies do not provide the best representation of the systemic problems caused by MI. In **Paper III**, we present our work on the integrative analysis of the heart and three metabolically active tissues (liver, skeletal muscle, and adipose tissues) using **transcriptomics** data generated from our MI mouse model (Figure 9). The tissues were obtained 6 and 24 hours after MI or SHAM operation (control).

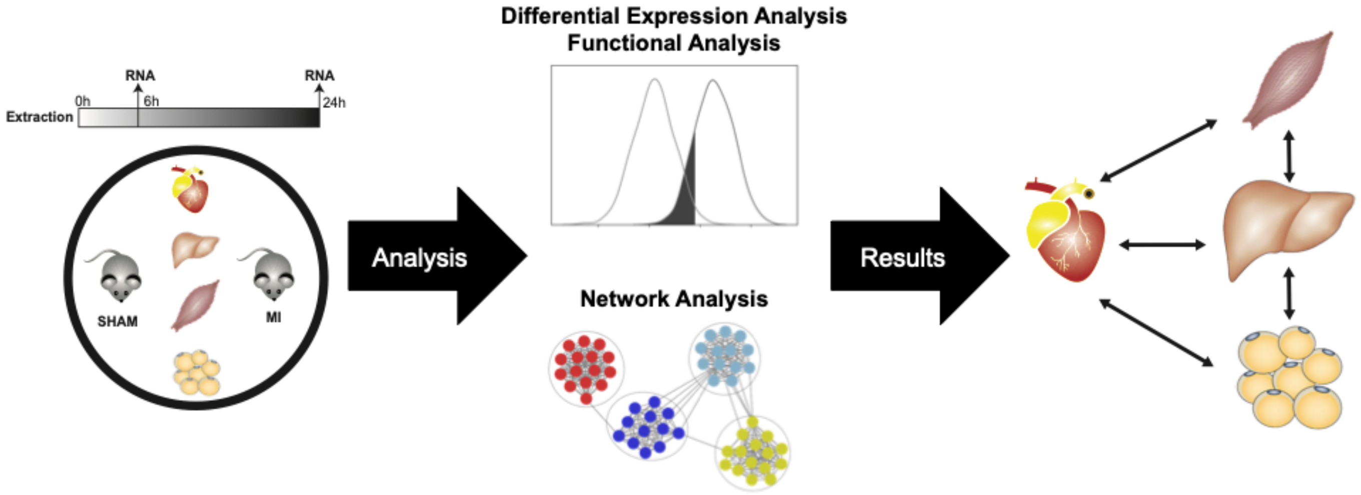


Figure 2 Workflow of Paper III
transcriptomics data were processed using systems biology approaches to discover the metabolic crosstalk between multiple tissues.

We performed the statistical analysis using DESeq2[139] and the functional analysis (KEGG and Gene Ontology) using PIANO[140]. Naturally, the heart showed the highest number of differentially expressed genes (DEGs), and heart-specific functions were downregulated after MI. We further found that retinol metabolism was upregulated in the heart after MI. In addition, processes related to lipid metabolism and inflammatory responses were upregulated in the heart, muscle, and adipose tissue after MI, whereas processes related to mitochondrial functions were downregulated. Furthermore, we detected the inhibition of fatty acid metabolism in the heart and adipose tissue and the activation of oxidative stress in the heart and muscle after MI. Interestingly, the behavior of the liver was unique, as revealed by inhibition of the inflammatory response and enhancement of fatty acid beta-oxidation. Subsequently, we performed a metabolite analysis using GEMs and found that the results supported those obtained from the functional analysis. Interestingly, we found that metabolites related to retinol metabolism were inhibited after MI.

As discussed previously, we generated GCNs for each tissue and performed a community analysis. We embedded the DEGs 24 hours after MI into the communities to obtain their trend. We also performed a functional analysis to determine the biological functions associated with the clusters. In all GCNs, the genes in their key clusters were mainly upregulated and associated with RNA-related functions (transport, processing, and metabolic processes). The heart-specific cluster in the heart GCN consisted of mostly downregulated genes and was associated with heart-specific and mitochondrial functions, whereas the genes in the liver-specific cluster in the liver GCN were connected to retinol and lipid metabolism, and adipose-specific clusters were related to the cell cycle and mRNA processing. Interestingly, we found two muscle-specific clusters, both of which were associated with mitochondrial functions and muscle-specific functions. The genes were also highly linked to multiple metabolic processes and signaling pathways, such as glycolysis, propanoate metabolism, and MAPK signaling pathways.

We also performed centrality analysis and identified several key genes from each tissue-specific cluster that were also identified as DEGs. The key genes in the heart-specific cluster were associated with cardiac muscle regulation and mitochondrial functions. In adipose tissue and skeletal muscle, the key genes were related to fatty acid and lipid metabolism.

Interestingly, the key genes in the liver-specific cluster were significantly associated with retinol metabolism. These findings were consistent with and strengthened the findings from the previous analyses.

Our hypothesis based on the results was that the exchange of fatty acids between adipose tissue, the liver, and skeletal muscle caused the downregulation of fatty acid metabolism in adipose tissue. Moreover, we observed the inhibition of retinol metabolites in the liver. We speculated that this finding was due to the rapid transport of retinol to the heart during MI[141]. These systemic changes caused by MI lead to the deterioration of mitochondrial function and decreased energy production in the heart and skeletal muscle. Our final results from this study revealed several genes with important responses to MI in multiple tissues: *Flnc*, *Prkaca*, *Lgals3*, and *Pprc1*. All of these genes have been previously observed to exhibit strong associations with cardiovascular diseases[142-148], with the exception of *Pprc1*, which is a regulator of mitochondrial biogenesis[149]. We successfully validated all our findings using two independent publicly available mouse datasets.

Overall, we performed a systematic analysis of multi-tissue transcriptomics data from MI mouse models to elucidate the mechanism of MI in the heart and metabolically active tissues. We demonstrated the application of several systems biology approaches (discussed in **Chapter II**), such as DEGs, functional analysis, and GCNs, to explore the systemic effect of MI. We were able to integrate the results and speculate on the metabolic crosstalk between the analyzed tissues.

Systems Biology of Muscle (Paper IV)

Physical activity has been linked to multiple health benefits. This activity not only increases physical wellness and fitness but also reduces the risk of diseases[150], such as heart problems and metabolic-related diseases. Moreover, it has been used as a treatment strategy for many diseases and conditions, including NAFLD[151] and neurological diseases[152]. In **Paper IV**, we performed an in-depth analysis using **transcriptomics** data from skeletal muscle to obtain more in-depth molecular insights into the effect of lifelong training. We also compared trained and untrained women and men to identify and understand the sex differences. Furthermore, we compared our results with publicly available datasets to understand the effect of exercises on the recovery of metabolic-related diseases.

In this study, we collected muscle biopsies from 40 samples, which were obtained from seven to nine subjects in the control (untrained), endurance-trained, and strength-trained groups of both sexes. The subjects were separated based on very robust phenotypic criteria. First, we performed an exploratory analysis of the data using PCA and found that the male control (MC) and female control (FC) groups were clustered together with the male strength (MS) group. Interestingly, the male endurance (ME) group was clustered very tightly with the female endurance (FE) group.

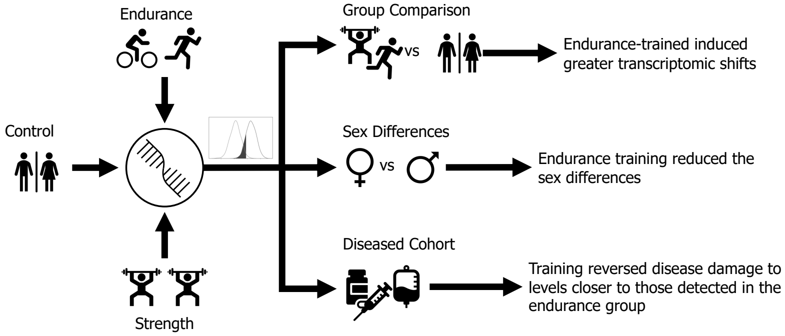

Endurance
Control
Strength
Group Comparison
vs
Endurance-trained induced greater transcriptomic shifts
Sex Differences
vs
Endurance training reduced the sex differences
Diseased Cohort
Training reversed disease damage to levels closer to those detected in the endurance group

Figure 10 Overview of Paper IV.

Subsequently, we performed a differential expression analysis between groups and identified >1000 DEGs in each FE and ME group compared with their respective controls. These genes were associated with many processes and pathways, including cellular respiration and the TCA cycle, in both sexes. Moreover, the FE group showed unique alterations in processes related to protein ubiquitination that were not observed in the ME group. Interestingly, we identified very few DEGs from the comparison of the MS vs MC groups, and we still observed upregulation in cellular respiration in the MS group. Furthermore, we compared male and female subjects to identify the sex differences. The comparison of the control groups (MC vs FC) revealed upregulation of processes related to protein metabolic processes in the MC group, whereas processes related to lipid metabolism and wound healing were found to be upregulated in the FC group. Interestingly, the DEGs obtained from the comparison of the two endurance groups (ME vs FE) were markedly reduced to only 30% of those identified from the MC vs FC comparison. The DEGs identified from the ME vs FE comparison were significantly associated with an increase in mitochondrial functions. We also integrated the transcriptomics data with GEMs by performing RM analysis. The findings emphasized significant upregulation of the TCA cycle as well as BCAA regulation and fatty acid oxidation in both the ME and FE groups.

We then acquired two publicly available transcriptomics datasets of male subjects with type 2 diabetes and women with metabolic syndrome. The data from the subjects were obtained before and after 6-12 months of training. First, we performed hierarchical clustering to obtain the overall view of the data compared with ours. We observed that the pretraining data were highly correlated with our control data. Interestingly, their data after training was more closely correlated to our endurance-trained data than to the other groups. Subsequently, we performed an analysis similar to that performed with our data and retrieved the DEGs in both datasets compared with their respective controls. We found many genes that were altered in opposite directions compared with those identified from the comparison of the ME and FE groups with their respective controls. After training, the number of opposite-direction DEGs decreased significantly, and the change in expression even flipped to the same direction as that found in the endurance-trained group. These flipped genes were related to blood glucose consumption and insulin sensitivity.

In this study, we used systems biology approaches, in collaboration with experts in exercise physiology, to unravel the mechanistic effect of lifelong training in both males and females. We focused on transcriptional changes observed in trained compared with untrained individuals. We then compared our results with the transcriptomic profiles of individuals with metabolic-related diseases and found that short-term training reversed the damage caused by the disease in skeletal muscle.

Systems Biology of the Liver (Papers V-VI)

NAFLD has been labeled "the silent epidemic"[153]. It is one of the most prevalent diseases in the world, affecting approximately 25% of the world's population[153]. NAFLD accounts for hepatic steatosis (HS) and nonalcoholic steatohepatitis (NASH), among other clinical conditions[29], and can progress to irreversible injuries, such as cirrhosis and hepatocellular carcinoma (HCC). Unfortunately, there is currently no approved treatment for this disease. Thus, it is of the utmost importance to understand the characteristics of NAFLD and, ultimately, to develop a potential treatment strategy.

Paper V

In **Paper V**, we generated plasma **metabolomics**, inflammatory **proteomics**, and oral and gut **metagenomics data** as well as **clinical** data to characterize NAFLD. The data were collected from 56 obese subjects with NAFLD who were grouped based on the severity of their HS (none, mild, moderate, and severe) based on MRI results. We also excluded subjects with genetic variants related to NAFLD. We further collected data from a subset of the subjects (22 subjects) 2-3 months after the first visit and used these data for validation. As mentioned in the **"Generation of Biological Networks"** section, we generated a MOBN for this study and deposited it on the iNetModels platform.

The clinical data were analyzed by comparing subjects with mild, moderate, and severe HS with subjects without NAFLD. We observed higher uric acid levels and higher liver enzyme (ALT, AST, and GGT) levels in the severe and moderate groups. These findings were consistent with those obtained in a preceding study[128]. Moreover, higher levels of albumin, creatinine, and creatine kinase were detected in the severe group.

The same comparisons were performed for each omics dataset separately. We analyzed the oral (saliva) and gut (feces) metagenomics data. Particularly in the severe group, we found reductions in the abundance of several gut microbiome species belonging to Actinobacteria (e.g., *Slackia isoflavoniconvertens*), Bacteroidetes, Firmicutes (e.g., *Dorea longicatena* and *Ruminococcus bromii*), and Proteobacteria (*Bilophila wadsworthia*). In the saliva, we observed decreased abundances of several species belonging to Bacteroidetes (e.g., *Porphyromonas endodontalis*) and an increase in the Actinobacteria (*Actinomyces johnsonii*) abundance. In the metabolomics analysis, we found that the majority of the altered metabolites were related to lipid metabolism, which was expected. When focusing on non-lipid-related metabolites, we found that serine- and glycine-related metabolites, which are important in glutathione metabolism, as well as cysteine-glutathione disulfide, were decreased in the severe group. In contrast, the levels of several metabolites related to tryptophan branched-chain amino acids (BCAAs), lysine, uric acid, and the urea cycle were increased. Moreover, a proteomics analysis showed that the majority of the proteins, including LIF-R, CCL20, and CDCP1, were upregulated, particularly in the severe group. The majority of the findings are consistent with those of the previous studies[24,128,129,154-156].

Interestingly, the findings from the single omics analysis were confirmed and can be retrieved from the MOBN. Moreover, with the MOBN, we were able to obtain the functional relationships between different omics data. For example, most glutathione-related metabolites were directly correlated with the liver enzyme GGT but not with other enzymes. Furthermore, we discovered negative correlations between liver fat and the oral microbe *Porphyromonas endodontalis* and several gut microbes, including *Slackia isoflavoniconvertens* and *Bilophila wadsworthia*.

We also observed negative correlations for the NAFLD-associated gut microbe[48] *Dorea longicatena* with AST and ALT and a known protagonist gut microbe[157], *Ruminococcus bromii*, with ALT and uric acid. We also extended our network analysis to centrality and community analyses. The centrality analysis revealed that the hubs were, unsurprisingly, known lipid-associated metabolites and clinical variables, such as ceramide, sphingomyelin, phospholipid, triglyceride, and LDL. The top protein hubs were linked to cytokine-cytokine receptor interactions and several signaling pathways (NF-kappa B, TNF, and IL-17). The community analysis also identified four clusters inside the network. The largest cluster, cluster-0, was associated with amino acid metabolism, whereas cluster-1 was dominated by phospholipid, carbohydrate, and taurine metabolites and the top protein hubs. Cluster-2, the key cluster, was associated with lipid metabolism, whereas the smallest cluster, cluster-3, consisted of analytes related to fatty acid metabolism. All clusters, with the exception of cluster-0 and cluster-1, tended to be positively correlated. These findings show the strength of community and network analyses in general to elucidate the functional relationships of analytes within and between omics types.

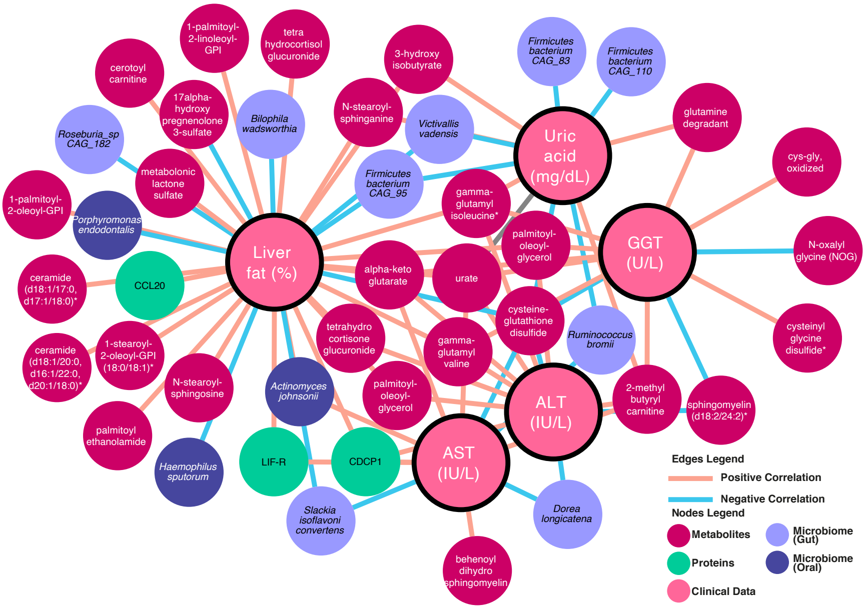

1-palmitoyl-2-linoleoyl-GPI
tetra hydrocortisol glucuronide
cerotoyl carnitine
3-hydroxy isobutyrate
Firmicutes bacterium CAG_83
Firmicutes bacterium CAG_110
17alpha-hydroxy pregnenolone 3-sulfate
Bilophila wadsworthia
N-stearoyl-sphinganine
Victivallis vadensis
Uric acid (mg/dL)
glutamine degradant
Roseburia_sp CAG_182
metabolonic lactone sulfate
Firmicutes bacterium CAG_95
gamma-glutamyl isoleucine*
cys-gly, oxidized
1-palmitoyl-2-oleoyl-GPI
Porphyromonas endodontalis
Liver fat (%)
palmitoyl-oleoyl-glycerol
GGT (U/L)
N-oxalyl glycine (NOG)
ceramide (d18:1/17:0, d17:1/18:0)*
CCL20
alpha-keto glutarate
urate
cysteine-glutathione disulfide
Ruminococcus bromii
cysteinyl glycine disulfide*
ceramide (d18:1/20:0, d16:1/22:0, d20:1/18:0)*
1-stearoyl-2-oleoyl-GPI (18:0/18:1)*
tetrahydro cortisone glucuronide
gamma-glutamyl valine
2-methyl butyryl carnitine
sphingomyelin (d18:2/24:2)*
N-stearoyl-sphingosine
Actinomyces johnsonii
palmitoyl-oleoyl-glycerol
ALT (IU/L)
palmitoyl ethanolamide
Haemophilus sputorum
LIF-R
CDCP1
AST (IU/L)
Slackia isoflavoni convertens
Dorea longicatena
behenoyl dihydro sphingomyelin
Edges Legend
Positive Correlation
Negative Correlation
Nodes Legend
Metabolites
Microbiome (Gut)
Proteins
Microbiome (Oral)
Clinical Data

Figure 11 Network of Liver Fat, Enzymes, and Uric Acid
Top and significantly altered neighbors of liver fat, liver enzymes, and uric acid based on the NAFLD baseline MOBN in iNetModels.

Finally, we developed a multiomics classification model to predict the severity of HS based on multiomics data using random forest. First, we performed a random forest analysis of each omics type and clinical data. We gathered the top features from each of the data types and constructed a combined multiomics predictive model that showed >80% accuracy with both bootstrapped training data and validation data obtained from 22 samples from the validation cohort. Interestingly, removal of the gut and oral metagenomics features from the model decreased the predictive accuracy to ~60%.

In summary, we implemented a wide range of systems biology approaches, which were discussed in the previous chapter, to analyze multiomics data from subjects with NAFLD and varying HS severity. We showed the importance of oral and gut metagenomics data for NAFLD diagnosis and were able to show the biological function alterations due to NAFLD progression and to identify candidate biomarkers.

Paper VI

In **Paper VI**, we performed a calibration study of a candidate therapeutic supplement for NAFLD, which was later named CMA (Combined Metabolic Activator). CMA consists of four natural substances: L-serine, L-carnitine, nicotinamide riboside (NR), and N-acetyl-cysteine (NAC). We generated plasma **metabolomics** and plasma inflammatory **proteomics data** as well as **clinical** variables from 10 and nine male subjects, which comprised the control and supplemented groups, respectively. The study was controlled to minimize confounding factors: diets were controlled throughout the day, the subjects were provided a similar breakfast and then fasted until the end of the study. The data were collected on average every hour for 8 hours. No significant changes in the clinical data were detected before and after supplementation.

We generated targeted metabolomics data for the CMA substances and naturally found increases in their levels in blood. We also generated full-panel untargeted plasma metabolomics data and, as expected, found a high correlation between the targeted and untargeted plasma levels of CMA substances. We found significant downregulation of BCAAs before and after CMA supplementation, and similar findings were obtained with kynurenine, kynurenate, and pyruvate. These alterations contrasted our findings in **Paper V**, where we found upregulation of BCAA metabolites in severe HS. To rule out an effect of fasting, we performed a similar analysis of the control group and found no changes in these metabolites. We also compared the control and supplemented groups at each time point to systematically determine the effect of the supplementation. We found that CMA components and their derivatives, citrulline, and amino acid metabolites showed significant alteration in at least three subsequent time points in the supplemented group. The proteomics analysis revealed that several proteins associated with cytokine receptors and TNF signaling were downregulated.

Finally, we integrated metabolomics data with GEMs to simulate the effect of the supplements in the liver and found increases in fatty acid oxidation, glutathione synthesis and catabolism of BCAAs, and downregulation of glucose consumption. These findings are the exact opposite of our findings in **Paper V** and thus show that CMA supplementation can reduce the severity of HS. Furthermore, we applied pharmacokinetic modeling to calibrate the dosage of each substance.

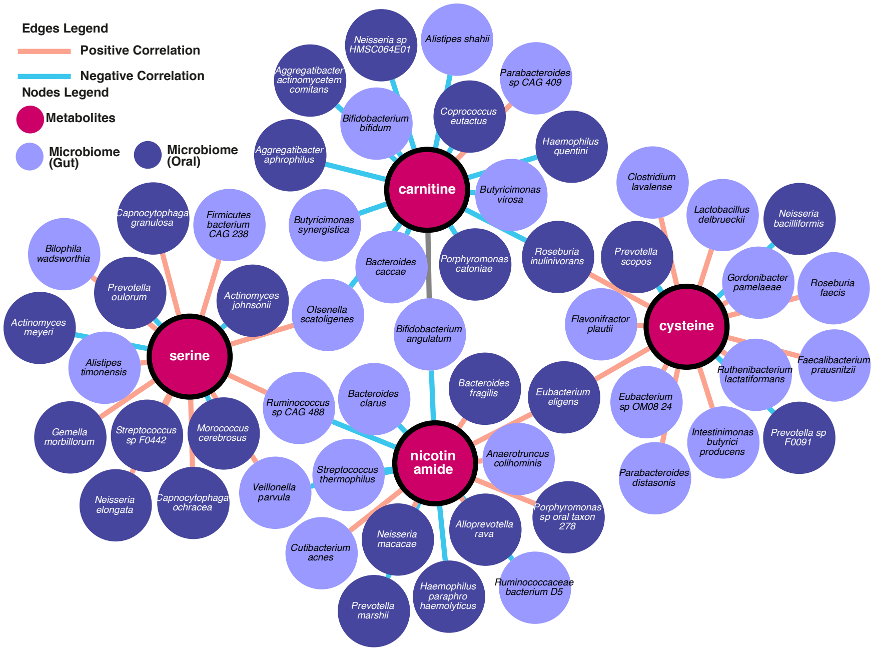


Figure 12 CMA and Microbiome

CMA substances and their top 15 oral and gut microbiome neighbors.

In the use case section presented in **Paper II**, we used a MOBN from an independent wellness study to validate the findings of this study. We found that the available CMA substances were positively associated with BCAA metabolites and negatively correlated with glucose levels. Moreover, we found negative correlations for L-serine with cholesterol-related clinical variables and inflammation markers. We can also use the network generated in **Paper V** to enrich our results and thereby observe the association of the gut and oral metagenomics with CMA components (Figure 12) because their dysbiosis is associated with NAFLD. The results showed that serine is positively correlated with *Bilophila wadsworthia* (gut) and negatively correlated with *Actinomyces johnsonii* (oral), which were found to be correlated with HS in **Paper V**. Carnitine was negatively correlated with *Bacteroides caccae* (gut), whereas cysteine was positively correlated with *Faecalibacterium prausnitzii* (gut). Both of these factors have been previously associated with NAFLD[48,158].

Overall, the study used metabolomics and proteomics data to obtain mechanistic insights into potential treatment supplements for NAFLD. Furthermore, we showed how biological networks can be used to validate and enrich our study. Once again, we also showed the strengths of systems biology approaches for disentangling a complex biological problem.

Chapter IV: Concluding Remarks and Future Perspectives

Personalized medicine is the future, and in this era, every treatment will be tailored to each patient's characteristics. To achieve this goal, we need to obtain a better understanding of diseases, and systems biology can be the main enabler of this information[159]. The rise of systems and data-driven biology has opened and continues to open many new opportunities and approaches in disease-related research, particularly by employing robust algorithms together with strong computational power. With systems biology, we can attempt to decipher the complexity of molecular biological data. Moreover, by coupling these tools with strong biological expertise, we can unravel the mystery of human diseases at the molecular level. One key step to successfully achieve this goal is to understand the interconnection between analytes from one or more omics types, such as genes, proteins, metabolites, and our second fingerprint, the microbiome, to understand their relationships.

This thesis focused on the development and application of systems biology approaches to obtain mechanistic and systematic views of human diseases. In **Papers I** and **II**, we introduced a web-based biological network database, which we envision will be a valuable platform for researchers. Moreover, we developed working frameworks to answer important biological questions related to specific diseases and conditions using single- and multiomics data. In **Paper III**, we integrated transcriptomics data from the heart and other metabolically active tissues to reveal metabolic crosstalk after myocardial infarction, which is associated with one of the highest mortality rates in the world. In **Paper IV,** we analyzed transcriptomics data from skeletal muscle to investigate transcriptional alterations due to lifelong training, including gender differences, and their association with metabolic diseases. In **Paper V**, we used multiomics data to understand the molecular mechanism of NAFLD and build a model that can predict the severity of hepatic steatosis. Furthermore, in **Paper VI**, we used proteomics and metabolomics data to reveal the mechanism of action of potential treatment supplements for NAFLD. During this process, we also generated important datasets and biological networks that can be used for further research. I hope that this thesis can contribute to the realization of personalized medicine.

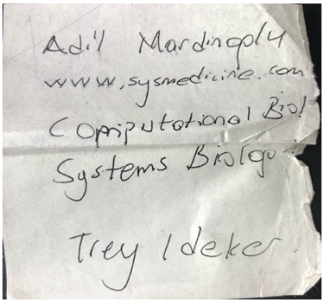
Moving forward, I believe that we need more *N-of-1* studies with multiomics data, particularly studies that focus on the diseased population. These studies will expand our view not only to the differences between the disease vs healthy group but also to the individual variations in the disease group. These individual variations are often not considered in the current models, including those generated in the studies described in this thesis. Moreover, the multiomics data will provide significantly more information and a broader view of the disease, which can lead to more specific patient characterization and a greater opportunity of discovering novel biomarkers. Moreover, having a general framework for both data collections and data analysis is of the utmost importance for maintaining consistency and, more importantly, a high research quality. In my personal opinion, network analysis will play a crucial role in this field. In addition to capturing functional relationships, communities and hubs in a network are likely to be associated with phenotype variations[82]. In contrast, it is impossible to derive a mechanistic understanding from a network alone because "correlation is not causality". Therefore, the incorporation of prior knowledge[160-162] (e.g., pathway and regulatory information) into a network will be beneficial for deriving the causality of the data, which would shorten the analysis cycle. It is exciting to see how this field will advance in the next 10-20 years, and I hope to play a big part in this advancement.

Finally, to reach the ultimate goal of personalized medicine, I believe that we first need to have robust disease models. This goal can only be achieved with a strong data foundation. Thus, I am advocating for more data sharing and collaboration among researchers in disease fields.

Acknowledgments

There are so many people that have helped to make this journey an interesting, very pleasant, and unforgettable one. I'm very grateful to everyone that has made it possible.

First of all, I would like to thank my supervisor, **Adil Mardinoglu**, for giving me the opportunity to join his amazing group. In mid-2016, you gave me a piece of paper with your information (Figure 13), and who would've thought that it would be the beginning of this amazing journey. Your constant support and non-stop encouragement have made this experience very positive. Thank you for giving me freedom during my study and for always be available (as a supervisor and friend) for discussions at any time. I would also like to thank my co-supervisor, **Mathias Uhlén**, for his support, advice, and amazing ideas and vision throughout my PhD. It was a great pleasure to learn from a great and experienced scientist like you.



Adil Mardinoglu
www.sysmedicine.com
Computational Biol
Systems Biolog
Trey Ideker

Figure 13 The "paper"

This journey will not succeed without our great collaborators. I would like to thank all collaborators and co-authors that made this research possible. I would like to specifically thank **Prof. Dr. Jan Boren** for the interesting projects and collaborations. Thanks also to everyone at the **Human Protein Atlas**, **Bash Biotech Inc.,** and **KTH CBH** (including the administrative staff that has been very patient with me) for the help and supports. Special thanks to **Dr. Abdellah Tebani** for all the encouragement, discussions, and great friendship, including the weekly trip to the mosque every Friday.

I would also like to thank all current and ex-members of the Sysmedicine family (Stockholm and London). **Sunjae** and **Cheng** for welcoming me to the group and for the guidance. **Kemal**, **Reza**, **Zhengtao**, **Mohamed**, **Natasa**, **Dorines**, **Feride**, **Beste, Kajetan, Xiangyu, Woonghee, Ozlem,** and others for all the good times at work and outside. I met two of my closest friends in this group: **Rui Benfeitas** and **Alen Lovric**. **Rui**, thanks for being patient with me all the time. I really appreciate all your help and our discussions, about work, life, and random stuff (97.59% of the time). You are definitely one of the most important people that made this possible. **Alen**, thanks for always being really helpful in stressful projects, and good luck with your study! The witty repartee, meme, gif, and video sharing with you both definitely help me to relax, so please keep on sending them. Thanks for the comradery and great times!

Finally, I will never reach this point if it's not because of the hard work of two people that I love the most in the world: my **father** and **mother**. I saw with my own eyes how hard they worked to make sure that we could get the best life and education, and I am eternally grateful for that. Now, it's time for you to relax and let us take care of everything. Thanks also to **my brothers** for all the support, jokes, and helps. I would also like to thank my **extended family** and **in-laws**, especially my **grandparents,** for all the prayers and encouragement. I would like to also thank my Indonesian family in Stockholm, including **PPI Stockholm** and **Futsal Barokah**, for all the fun times, great foods, and for bringing home closer to me. There are many names that I cannot write one-by-one in here, but thanks for everything!

Saving the best for the last, I would like to mention the most special person that has been always on my side and being supportive and understanding throughout this journey: **Fira**. You were always there with your unwavering supports and patients, regardless of how annoying I was especially during stressful times (or all the time?), no questions asked. And thanks for giving me the best thing in the world, the little man **Athif**. This thesis is as much as yours as it is mine.

References

1 Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. Annual review of genomics and human genetics 2, 343-372 (2001).

2 Kitano, H. Systems biology: a brief overview. science 295, 1662-1664 (2002).

3 Lay Jr, J. O., Liyanage, R., Borgmann, S. & Wilkins, C. L. Problems with the "omics". TrAC Trends in Analytical Chemistry 25, 1046-1056 (2006).

4 Palsson, B. In silico biology through "omics". Nature biotechnology 20, 649-650 (2002).

5 Ward, D. C. & White, D. C. (Elsevier Current Trends, 2002).

6 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17, 333 (2016).

7 Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at www.genome.gov/sequencingcostsdata Accessed.

8 Shah, S. et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. Nat Commun 11, 163, doi:10.1038/s41467-019-13690-5 (2020).

9 Xue, A. et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat Commun 9, 2941, doi:10.1038/s41467-018-04951-w (2018).

10 Clarke, L. et al. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res 45, D854-D859, doi:10.1093/nar/gkw829 (2017).

11 Pinero, J. et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford) 2015, bav028, doi:10.1093/database/bav028 (2015).

12 Tian, S. et al. HDAC inhibitor valproic acid protects heart function through Foxm1 pathway after acute myocardial infarction. EBioMedicine 39, 83-94, doi:10.1016/j.ebiom.2018.12.003 (2019).

13 Turanli, B. et al. Discovery of therapeutic agents for prostate cancer using genome-scale metabolic modeling and drug repositioning. EBioMedicine 42, 386-396 (2019).

14 Uhlen, M. et al. A pathology atlas of the human cancer transcriptome. Science 357, doi:10.1126/science.aan2507 (2017).

15 Govindarajan, R., Duraiyan, J., Kaliyappan, K. & Palanisamy, M. Microarray and its applications. J Pharm Bioallied Sci 4, S310-312, doi:10.4103/0975-7406.100283 (2012).

16 Behjati, S. & Tarpey, P. S. What is next generation sequencing? Arch Dis Child Educ Pract Ed 98, 236-238, doi:10.1136/archdischild-2013-304340 (2013).

17 Timp, W. & Timp, G. Beyond mass spectrometry, the next step in proteomics. Sci Adv 6, eaax8978, doi:10.1126/sciadv.aax8978 (2020).

18 Higginbotham, L. et al. Integrated proteomics reveals brain-based cerebrospinal fluid biomarkers in asymptomatic and symptomatic Alzheimer's disease. Sci Adv 6, doi:10.1126/sciadv.aaz9360 (2020).

19 Lehallier, B. et al. Undulating changes in human plasma proteome profiles across the lifespan. Nat Med 25, 1843-1850, doi:10.1038/s41591-019-0673-2 (2019).

20 Niu, L. et al. Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. Mol Syst Biol 15, e8793, doi:10.15252/msb.20188793 (2019).

21 Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. Cold Spring Harb Mol Case Stud 1, a000588, doi:10.1101/mcs.a000588 (2015).

22 Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46, D608-D617, doi:10.1093/nar/gkx1089 (2018).

23 Zampieri, M. & Sauer, U. Metabolomics-driven understanding of genotype-phenotype relations in model organisms. Current Opinion in Systems Biology 6, 28-36 (2017).

24 Lovric, A. et al. Characterization of different fat depots in NAFLD using inflammation-associated proteome, lipidome and metabolome. Sci Rep 8, 14200, doi:10.1038/s41598-018-31865-w (2018).

25 Tzoulaki, I. et al. Serum metabolic signatures of coronary and carotid atherosclerosis and subsequent cardiovascular disease. Eur Heart J 40, 2883-2896, doi:10.1093/eurheartj/ehz235 (2019).

26 Franzosa, E. A. et al. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A 112, E2930-2938, doi:10.1073/pnas.1423854112 (2015).

27 Jovel, J. et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front Microbiol 7, 459, doi:10.3389/fmicb.2016.00459 (2016).

28 Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev Microbiol 14, 508-522, doi:10.1038/nrmicro.2016.83 (2016).

29 Mardinoglu, A., Boren, J., Smith, U., Uhlen, M. & Nielsen, J. Systems biology in hepatology: approaches and applications. Nat Rev Gastroenterol Hepatol 15, 365-377, doi:10.1038/s41575-018-0007-8 (2018).

30 Xia, X. Bioinformatics and drug discovery. Current topics in medicinal chemistry 17, 1709-1726 (2017).

31 Mardinoglu, A., Boren, J., Smith, U., Uhlen, M. & Nielsen, J. Systems biology in hepatology: approaches and applications. Nature Reviews Gastroenterology & Hepatology 15, 365-377 (2018).

32 Mardinoglu, A., Boren, J., Smith, U., Uhlen, M. & Nielsen, J. The employment of systems biology in gastroenterology and hepatology. Nat. Rev. Gastroenterol. Hepatol (2017).

33 Mardinoglu, A. & Nielsen, J. New paradigms for metabolic modeling of human cells. Current Opinion in Biotechnology 34, 91-97 (2015).

34 Nielsen, J. Systems biology of metabolism: a driver for developing personalized and precision medicine. Cell metabolism 25, 572-579 (2017).

35 Benfeitas, R. et al. Characterization of heterogeneous redox responses in hepatocellular carcinoma patients using network analysis. EBioMedicine 40, 471-487 (2019).

36 Bidkhori, G. et al. Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes.

Proceedings of the National Academy of Sciences 115, E11874-E11883 (2018).

37 Lee, S. et al. Integrated network analysis reveals an association between plasma mannose levels and insulin resistance. Cell metabolism 24, 172-184 (2016).

38 Wang, Q. et al. A metagenome-wide association study of gut microbiota in asthma in UK adults. BMC Microbiol 18, 114, doi:10.1186/s12866-018-1257-x (2018).

39 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27-30, doi:10.1093/nar/28.1.27 (2000).

40 Gene Ontology, C. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49, D325-D334, doi:10.1093/nar/gkaa1113 (2021).

41 Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25-29, doi:10.1038/75556 (2000).

42 Chong, J. et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 46, W486-W494, doi:10.1093/nar/gky310 (2018).

43 Pinero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 48, D845-D855, doi:10.1093/nar/gkz1021 (2020).

44 Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature 569, 503-508, doi:10.1038/s41586-019-1186-3 (2019).

45 Camarda, R. et al. Inhibition of fatty acid oxidation as a therapy for MYC-overexpressing triple-negative breast cancer. Nat Med 22, 427-432, doi:10.1038/nm.4055 (2016).

46 Lai, K., Twine, N., O’Brien, A., Guo, Y. & Bauer, D. in Encyclopedia of Bioinformatics and Computational Biology (eds Shoba Ranganathan, Michael Gribskov, Kenta Nakai, & Christian Schönbach) 272-286 (Academic Press, 2019).

47 Xu, C. & Jackson, S. A. Machine learning and complex biological data. Genome Biol 20, 76, doi:10.1186/s13059-019-1689-0 (2019).

48 Loomba, R. et al. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human

Nonalcoholic Fatty Liver Disease. Cell Metab 25, 1054-1062 e1055, doi:10.1016/j.cmet.2017.04.001 (2017).

49 McIlwain, S. J. et al. Enhancing Top-Down Proteomics Data Analysis by Combining Deconvolution Results through a Machine Learning Strategy. J Am Soc Mass Spectrom 31, 1104-1113, doi:10.1021/jasms.0c00035 (2020).

50 Barabási, A.-L. Network science. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371, 20120375 (2013).

51 Lee, S. et al. TCSBN: a database of tissue and cancer specific biological networks. Nucleic Acids Res 46, D595-D600, doi:10.1093/nar/gkx994 (2018).

52 Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47, D607-D613, doi:10.1093/nar/gky1131 (2019).

53 Luck, K. et al. A reference map of the human binary protein interactome. Nature 580, 402-408, doi:10.1038/s41586-020-2188-x (2020).

54 Tong, A. H. et al. Global mapping of the yeast genetic interaction network. Science 303, 808-813, doi:10.1126/science.1091317 (2004).

55 Ma, H. W. & Zeng, A. P. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics 19, 1423-1430, doi:10.1093/bioinformatics/btg177 (2003).

56 Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. The large-scale organization of metabolic networks. Nature 407, 651-654, doi:10.1038/35036627 (2000).

57 Ito, T. et al. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A 97, 1143-1147, doi:10.1073/pnas.97.3.1143 (2000).

58 Ota, M., Gonja, H., Koike, R. & Fukuchi, S. Multiple-Localization and Hub Proteins. PLoS One 11, e0156455, doi:10.1371/journal.pone.0156455 (2016).

59 Wessling, R. et al. Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. Cell

Host Microbe 16, 364-375, doi:10.1016/j.chom.2014.08.004 (2014).

60 Peng, X., Wang, J., Wang, J., Wu, F. X. & Pan, Y. Rechecking the Centrality-Lethality Rule in the Scope of Protein Subcellular Localization Interaction Networks. PLoS One 10, e0130743, doi:10.1371/journal.pone.0130743 (2015).

61 Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. Nature 411, 41-42, doi:10.1038/35075138 (2001).

62 Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A 105, 4323-4328, doi:10.1073/pnas.0701722105 (2008).

63 Lee, S. et al. Network analyses identify liver-specific targets for treating liver diseases. Mol Syst Biol 13, 938, doi:10.15252/msb.20177703 (2017).

64 Golbeck, J. in Introduction to Social Media Investigation (ed Jennifer Golbeck) 221-235 (Syngress, 2015).

65 Joy, M. P., Brock, A., Ingber, D. E. & Huang, S. High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol 2005, 96-103, doi:10.1155/JBB.2005.96 (2005).

66 Potapov, A. P., Voss, N., Sasse, N. & Wingender, E. Topology of mammalian transcription networks. Genome Inform 16, 270-278 (2005).

67 Bidkhori, G. et al. Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes. Proc Natl Acad Sci U S A 115, E11874-E11883, doi:10.1073/pnas.1807305115 (2018).

68 Negre, C. F. A. et al. Eigenvector centrality for characterization of protein allosteric pathways. Proc Natl Acad Sci U S A 115, E12201-E12208, doi:10.1073/pnas.1810452115 (2018).

69 Koschutzki, D. & Schreiber, F. Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene Regul Syst Bio 2, 193-201, doi:10.4137/grsb.s702 (2008).

70 Cai, J. J., Borenstein, E. & Petrov, D. A. Broker Genes in Human Disease. Genome Biology and Evolution 2, 815-825, doi:10.1093/gbe/evq064 (2010).

71 Hahn, M. W. & Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22, 803-806, doi:10.1093/molbev/msi072 (2005).

72 del Rio, G., Koschutzki, D. & Coello, G. How to identify essential genes from molecular networks? BMC Syst Biol 3, 102, doi:10.1186/1752-0509-3-102 (2009).

73 Wang, P., Yu, X. & Lu, J. Identification and evolution of structurally dominant nodes in protein-protein interaction networks. IEEE Trans Biomed Circuits Syst 8, 87-97, doi:10.1109/TBCAS.2014.2303160 (2014).

74 Robinson, J. L. et al. An atlas of human metabolism. Sci Signal 13, doi:10.1126/scisignal.aaz1482 (2020).

75 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. Nat Rev Genet 12, 56-68, doi:10.1038/nrg2918 (2011).

76 Han, J. D. et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430, 88-93, doi:10.1038/nature02555 (2004).

77 Girvan, M. & Newman, M. E. Community structure in social and biological networks. Proc Natl Acad Sci U S A 99, 7821-7826, doi:10.1073/pnas.122653799 (2002).

78 Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. Science 296, 910-913, doi:10.1126/science.1065103 (2002).

79 Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys 70, 066111, doi:10.1103/PhysRevE.70.066111 (2004).

80 Newman, M. E. J. Detecting community structure in networks. The European Physical Journal B 38, 321-330, doi:10.1140/epjb/e2004-00124-y (2004).

81 Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. Hierarchical organization of modularity in metabolic networks. Science 297, 1551-1555, doi:10.1126/science.1073374 (2002).

82 Price, N. D. et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat Biotechnol 35, 747-756, doi:10.1038/nbt.3870 (2017).

83 Newman, M. E. Fast algorithm for detecting community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69, 066133, doi:10.1103/PhysRevE.69.066133 (2004).

84 Newman, M. E. & Girvan, M. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69, 026113, doi:10.1103/PhysRevE.69.026113 (2004).

85 Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. Science 347, 1260419, doi:10.1126/science.1260419 (2015).

86 Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun 9, 1090, doi:10.1038/s41467-018-03424-4 (2018).

87 Newman, M. E. Modularity and community structure in networks. Proc Natl Acad Sci U S A 103, 8577-8582, doi:10.1073/pnas.0601602103 (2006).

88 Pons, P. & Latapy, M. 284-293 (Springer Berlin Heidelberg).

89 Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233, doi:10.1038/s41598-019-41695-z (2019).

90 Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008, doi:10.1088/1742-5468/2008/10/p10008 (2008).

91 Waltman, L. & van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. The European Physical Journal B 86, 471, doi:10.1140/epjb/e2013-40829-0 (2013).

92 Traag, V. A., Van Dooren, P. & Nesterov, Y. Narrow scope for resolution-limit-free community detection. Phys Rev E Stat Nonlin Soft Matter Phys 84, 016114, doi:10.1103/PhysRevE.84.016114 (2011).

93 Fortunato, S. & Barthelemy, M. Resolution limit in community detection. Proc Natl Acad Sci U S A 104, 36-41, doi:10.1073/pnas.0605965104 (2007).

94 Hagan, T. et al. Antibiotics-Driven Gut Microbiome Perturbation Alters Immunity to Vaccines in Humans. Cell 178, 1313-1328 e1313, doi:10.1016/j.cell.2019.08.010 (2019).

95 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15, doi:10.1186/s13059-017-1382-0 (2018).

96 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411-420, doi:10.1038/nbt.4096 (2018).

97 Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature 570, 332-337, doi:10.1038/s41586-019-1195-2 (2019).

98 Ziegler, C. G. K. et al. SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. Cell 181, 1016-1035 e1019, doi:10.1016/j.cell.2020.04.035 (2020).

99 Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? Nat Biotechnol 28, 245-248, doi:10.1038/nbt.1614 (2010).

100 Gudmundsson, S. & Thiele, I. Computationally efficient flux variability analysis. BMC Bioinformatics 11, 489, doi:10.1186/1471-2105-11-489 (2010).

101 Bidkhori, G. et al. Metabolic Network-Based Identification and Prioritization of Anticancer Targets Based on Expression Data in Hepatocellular Carcinoma. Front Physiol 9, 916, doi:10.3389/fphys.2018.00916 (2018).

102 Agren, R. et al. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Mol Syst Biol 10, 721, doi:10.1002/msb.145122 (2014).

103 Gatto, F., Miess, H., Schulze, A. & Nielsen, J. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. Sci Rep 5, 10738, doi:10.1038/srep10738 (2015).

104 Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A 102, 2685-2689, doi:10.1073/pnas.0406811102 (2005).

105 Lee, S. et al. Integrated Network Analysis Reveals an Association between Plasma Mannose Levels and Insulin Resistance. Cell Metab 24, 172-184, doi:10.1016/j.cmet.2016.05.026 (2016).

106 Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. Genome Biol 20, 121, doi:10.1186/s13059-019-1730-3 (2019).

107 Kuzmanov, U. & Emili, A. Protein-protein interaction networks: probing disease mechanisms using model systems. Genome Med 5, 37, doi:10.1186/gm441 (2013).

108 Karbalaei, R., Allahyari, M., Rezaei-Tavirani, M., Asadzadeh-Aghdaei, H. & Zali, M. R. Protein-protein interaction analysis of Alzheimer`s disease and NAFLD based on systems biology methods unhide common ancestor pathways. Gastroenterol Hepatol Bed Bench 11, 27-33 (2018).

109 Li, S., Sun, X., Miao, S., Liu, J. & Jiao, W. Differential protein-coding gene and long noncoding RNA expression in smoking-related lung squamous cell carcinoma. Thorac Cancer 8, 672-681, doi:10.1111/1759-7714.12510 (2017).

110 Li, Z., Qiao, Z., Zheng, W. & Ma, W. Network Cluster Analysis of Protein-Protein Interaction Network-Identified Biomarker for Type 2 Diabetes. Diabetes Technol Ther 17, 475-481, doi:10.1089/dia.2014.0204 (2015).

111 Luo, T., Wu, S., Shen, X. & Li, L. Network cluster analysis of protein-protein interaction network identified biomarker for early onset colorectal cancer. Mol Biol Rep 40, 6561-6568, doi:10.1007/s11033-013-2694-0 (2013).

112 Sun, N. & Zhao, H. Reconstructing transcriptional regulatory networks through genomics data. Stat Methods Med Res 18, 595-617, doi:10.1177/0962280209351890 (2009).

113 Jackson, C. A., Castro, D. M., Saldi, G. A., Bonneau, R. & Gresham, D. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. Elife 9, doi:10.7554/eLife.51254 (2020).

114 Sa, J. K. et al. Transcriptional regulatory networks of tumor-associated macrophages that drive malignancy in mesenchymal glioblastoma. Genome Biol 21, 216, doi:10.1186/s13059-020-02140-x (2020).

115 Padi, M. & Quackenbush, J. Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators. BMC Syst Biol 9, 80, doi:10.1186/s12918-015-0228-1 (2015).

116 Walhout, A. J. Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. Genome Res 16, 1445-1454, doi:10.1101/gr.5321506 (2006).

117 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559, doi:10.1186/1471-2105-9-559 (2008).

118 Anglani, R. et al. Loss of connectivity in cancer co-expression networks. PLoS One 9, e87075, doi:10.1371/journal.pone.0087075 (2014).

119 van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J. P. Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform 19, 575-592, doi:10.1093/bib/bbw139 (2018).

120 Bailey, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 531, 47-52, doi:10.1038/nature16965 (2016).

121 Tebani, A. et al. Integration of molecular profiles in a longitudinal wellness profiling cohort. Nat Commun 11, 4487, doi:10.1038/s41467-020-18148-7 (2020).

122 Hood, L. & Friend, S. H. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 8, 184-187, doi:10.1038/nrclinonc.2010.227 (2011).

123 Ho, D. et al. Enabling Technologies for Personalized and Precision Medicine. Trends Biotechnol 38, 497-518, doi:10.1016/j.tibtech.2019.12.021 (2020).

124 Gavan, S. P., Thompson, A. J. & Payne, K. The economic case for precision medicine. Expert Rev Precis Med Drug Dev 3, 1-9, doi:10.1080/23808993.2018.1421858 (2018).

125 Nassar, S. F., Raddassi, K., Ubhi, B., Doktorski, J. & Abulaban, A. Precision Medicine: Steps along the Road to Combat Human Cancer. Cells 9, doi:10.3390/cells9092056 (2020).

126 Goutsouliak, K. et al. Towards personalized treatment for early stage HER2-positive breast cancer. Nat Rev Clin Oncol 17, 233-250, doi:10.1038/s41571-019-0299-9 (2020).

127 Benfeitas, R. et al. Characterization of heterogeneous redox responses in hepatocellular carcinoma patients using network analysis. EBioMedicine 40, 471-487, doi:10.1016/j.ebiom.2018.12.057 (2019).

128 Mardinoglu, A. et al. Personal model-assisted identification of NAD(+) and glutathione metabolism as intervention target in NAFLD. Mol Syst Biol 13, 916, doi:10.15252/msb.20167422 (2017).

129 Zhang, C. et al. The acute effect of metabolic cofactor supplementation: a potential therapeutic strategy against non-alcoholic fatty liver disease. Mol Syst Biol 16, e9495, doi:10.15252/msb.209495 (2020).

130 Mannarino, L. et al. A systems biology approach to investigate the mechanism of action of trabectedin in a model of myelomonocytic leukemia. Pharmacogenomics J 18, 56-63, doi:10.1038/tpj.2016.76 (2018).

131 Mohammadi, E. et al. Applications of Genome-Wide Screening and Systems Biology Approaches in Drug Repositioning. Cancers (Basel) 12, doi:10.3390/cancers12092694 (2020).

132 Lillie, E. O. et al. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med 8, 161-173, doi:10.2217/pme.11.7 (2011).

133 Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318-1330, doi:10.1126/science.aaz1776 (2020).

134 Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113-1120, doi:10.1038/ng.2764 (2013).

135 Piening, B. D. et al. Integrative Personal Omics Profiles during Periods of Weight Gain and Loss. Cell Syst 6, 157-170 e158, doi:10.1016/j.cels.2017.12.013 (2018).

136 Su, A. I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101, 6062-6067, doi:10.1073/pnas.0400782101 (2004).

137 WHO. Cardiovascular diseases (CVDs) Fact sheets, Available at https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) Accessed.

138 Priest, C. & Tontonoz, P. Inter-organ cross-talk in metabolic syndrome. Nat Metab 1, 1177-1188, doi:10.1038/s42255-019-0145-5 (2019).

139 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550, doi:10.1186/s13059-014-0550-8 (2014).

140 Varemo, L., Nielsen, J. & Nookaew, I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res 41, 4378-4391, doi:10.1093/nar/gkt111 (2013).

141 Palace, V. P., Hill, M. F., Khaper, N. & Singal, P. K. Metabolism of vitamin A in the heart increases after a myocardial infarction. Free Radic Biol Med 26, 1501-1507, doi:10.1016/s0891-5849(99)00013-1 (1999).

142 Zhou, Y. et al. Loss of Filamin C Is Catastrophic for Heart Function. Circulation 141, 869-871, doi:10.1161/CIRCULATIONAHA.119.044061 (2020).

143 Hall, C. L. et al. RNA sequencing-based transcriptome profiling of cardiac tissue implicates novel putative disease mechanisms in FLNC-associated arrhythmogenic cardiomyopathy. Int J Cardiol 302, 124-130, doi:10.1016/j.ijcard.2019.12.002 (2020).

144 Zhong, X., Qian, X., Chen, G. & Song, X. The role of galectin-3 in heart failure and cardiovascular disease. Clin Exp Pharmacol Physiol 46, 197-203, doi:10.1111/1440-1681.13048 (2019).

145 Suthahar, N. et al. Galectin-3 Activation and Inhibition in Heart Failure and Cardiovascular Disease: An Update. Theranostics 8, 593-609, doi:10.7150/thno.22196 (2018).

146 Turnham, R. E. & Scott, J. D. Protein kinase A catalytic subunit isoform PRKACA; History, function and physiology. Gene 577, 101-108, doi:10.1016/j.gene.2015.11.052 (2016).

147 Diviani, D., Dodge-Kafka, K. L., Li, J. & Kapiloff, M. S. A-kinase anchoring proteins: scaffolding proteins in the heart. Am J Physiol Heart Circ Physiol 301, H1742-1753, doi:10.1152/ajpheart.00569.2011 (2011).

148 Bers, D. M. Calcium cycling and signaling in cardiac myocytes. Annu Rev Physiol 70, 23-49, doi:10.1146/annurev.physiol.70.113006.100455 (2008).

149 Ren, J., Pulakat, L., Whaley-Connell, A. & Sowers, J. R. Mitochondrial biogenesis in the metabolic syndrome and cardiovascular disease. J Mol Med (Berl) 88, 993-1001, doi:10.1007/s00109-010-0663-9 (2010).

150 Garber, C. E. et al. American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. Med Sci Sports Exerc 43, 1334-1359, doi:10.1249/MSS.0b013e318213fefb (2011).

151 van der Windt, D. J., Sud, V., Zhang, H., Tsung, A. & Huang, H. The Effects of Physical Exercise on Fatty Liver Disease. Gene Expr 18, 89-101, doi:10.3727/105221617X15124844266408 (2018).

152 Stranahan, A. M. & Mattson, M. P. Recruiting adaptive cellular stress responses for successful brain ageing. Nat Rev Neurosci 13, 209-216, doi:10.1038/nrn3151 (2012).

153 Lazarus, J. V. et al. NAFLD - sounding the alarm on a silent epidemic. Nat Rev Gastroenterol Hepatol 17, 377-379, doi:10.1038/s41575-020-0315-7 (2020).

154 Chu, X. et al. CCL20 is up-regulated in non-alcoholic fatty liver disease fibrosis and is produced by hepatic stellate cells in response to fatty acid loading. J Transl Med 16, 108, doi:10.1186/s12967-018-1490-y (2018).

155 Mardinoglu, A. et al. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat Commun 5, 3083, doi:10.1038/ncomms4083 (2014).

156 Toye, A. A. et al. Subtle metabolic and liver gene transcriptional changes underlie diet-induced fatty liver susceptibility in insulin-resistant mice. Diabetologia 50, 1867-1879, doi:10.1007/s00125-007-0738-5 (2007).

157 Lee, G. et al. Distinct signatures of gut microbiome and metabolites associated with significant fibrosis in non-obese NAFLD. Nat Commun 11, 4982, doi:10.1038/s41467-020-18754-5 (2020).

158 Grabherr, F., Grander, C., Effenberger, M., Adolph, T. E. & Tilg, H. Gut Dysfunction and Non-alcoholic Fatty Liver Disease. Front Endocrinol (Lausanne) 10, 611, doi:10.3389/fendo.2019.00611 (2019).

159 Chen, R. & Snyder, M. Systems biology: personalized medicine for the future? Curr Opin Pharmacol 12, 623-628, doi:10.1016/j.coph.2012.07.011 (2012).

160 Dugourd, A. et al. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol Syst Biol 17, e9730, doi:10.15252/msb.20209730 (2021).

161 Dugourd, A. & Saez-Rodriguez, J. Footprint-based functional analysis of multiomic data. Curr Opin Syst Biol 15, 82-90, doi:10.1016/j.coisb.2019.04.002 (2019).

162 Schubert, M. et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun 9, 20, doi:10.1038/s41467-017-02391-6 (2018).