Turun yliopisto
University of Turku

ANALYSIS OF HIGH-DIMENSIONAL AND LEFT-CENSORED DATA WITH APPLICATIONS IN LIPIDOMICS AND GENOMICS

Maiju Pesonen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. A I osa - tom. 548 | Astronomica - Chemica - Physica - Mathematica | Turku 2016

University of Turku

Faculty of Mathematics and Natural Sciences
Department of Mathematics and Statistics
Doctoral Programme in Mathematics and Computer Sciences

Supervised by

Professor Jaakko Nevalainen, PhD
School of Health Sciences
University of Tampere, Tampere, Finland

Professor Susmita Datta, PhD
Department of Biostatistics
University of Florida, Gainesville, FL, USA

Reviewed by

Professor Jukka Corander, PhD
Department of Biostatistics
Faculty of Medicine
University of Oslo, Oslo, Norway

Assistant Professor Samiran Ghosh, PhD
Center for Molecular Medicine and Genetics
School of Medicine
Wayne State University, Detroit, MI, USA

Opponent

Professor Ernst Wit, PhD
Johann Bernoulli Institute for Mathematics and
Computer Science
University of Groningen, Groningen, Netherlands

ISBN 978-951-29-6642-4 (PRINT)
ISBN 978-951-29-6643-1 (PDF)
ISSN 0082-7002 (Print)
ISSN 2343-3175 (Online)
Painosalama Oy - Turku, Finland 2016

Abstract

Recently, there has been an occurrence of new kinds of high-throughput measurement techniques enabling biological research to focus on fundamental building blocks of living organisms such as genes, proteins, and lipids. In sync with the new type of data that is referred to as the omics data, modern data analysis techniques have emerged. Much of such research is focusing on finding biomarkers for detection of abnormalities in the health status of a person as well as on learning unobservable network structures representing functional associations of biological regulatory systems. The omics data have certain specific qualities such as left-censored observations due to the limitations of the measurement instruments, missing data, non-normal observations and very large dimensionality, and the interest often lies in the connections between the large number of variables.

There are two major aims in this thesis. First is to provide efficient methodology for dealing with various types of missing or censored omics data that can be used for visualisation and biomarker discovery based on, for example, regularised regression techniques. Maximum likelihood based covariance estimation method for data with censored values is developed and the algorithms are described in detail. Second major aim is to develop novel approaches for detecting interactions displaying functional associations from large-scale observations. For more complicated data connections, a technique based on partial least squares regression is investigated. The technique is applied for network construction as well as for differential network analyses both on multiple imputed censored data and next-generation sequencing count data.

Tiivistelmä

Uudet mittausteknologiat ovat mahdollistaneet kokonaisvaltaisen ymmärryksen lisäämisen elollisten organismien molekyylitason prosesseista. Niin kutsutut omiikka-teknologiat, kuten genomiikka, proteomiikka ja lipidomiikka, kykenevät tuottamaan valtavia määriä mittausdataa yksittäisten geenien, proteiinien ja lipidien ekspressio- tai konsentraatiotasoista ennennäkemättömällä tarkkuudella. Samanaikaisesti tarve uusien analyysimenetelmien kehittämiselle on kasvanut. Kiinnostuksen kohteena ovat olleet erityisesti tiettyjen sairauksien riskiä tai prognoosia ennustavien merkkiaineiden tunnistaminen sekä biologisten verkkojen rekonstruointi.

Omiikka-aineistoilla on useita erityisominaisuuksia, jotka rajoittavat tavanomaisten menetelmien suoraa ja tehokasta soveltamista. Näistä tärkeimpiä ovat vasemmalta sensuroidut ja puuttuvat havainnot, sekä havaittujen muuttujien suuri lukumäärä. Tämän väitöskirjan ensimmäisenä tavoitteena on tarjota räätälöityjä analyysimenetelmiä epätäydellisten omiikka-aineistojen visualisointiin ja mallin valintaan käyttäen esimerkiksi regularisoituja regressiomalleja. Kuvailemme myös sensuroidulle aineistolle sopivan suurimman uskottavuuden estimaattorin kovarianssimatriisille. Toisena tavoitteena on kehittää uusia menetelmiä omiikka-aineistojen assosiaatiorakenteiden tarkasteluun. Monimutkaisempien rakenteiden tarkasteluun, visualisoimiseen ja vertailuun esitetään erilaisia variaatioita osittaisen pienimmän neliösumman menetelmään pohjautuvasta algoritmista, jonka avulla voidaan rekonstruoida assosiaatioverkkoja sekä multi-imputoidulle sensuroidulle että lukumääräaineistoille.

Acknowledgements

According to an American study, 67% of the PhD students feel hopeless at least once a year (reference not needed). Here, I would like to express my sincerest gratitude for those, who have supported me throughout this journey and kept reminding that great things never come from comfort zones, and those, who have shared both the despair and the success, and made this act more *Singing in the rain* than *Les Misérables*.

The completion of this thesis would not have been possible without the expertise and infinite (best estimate) patience of my supervisor Professor Jaakko Nevalainen. I would also like to thank my second supervisor Professor Susmita Datta not only for guiding me to expand my research towards the fascinating areas of genomics and association networks but also for her wonderful hospitality during my research visit in Louisville. I also acknowledge the two pre-examiners Professor Jukka Corander and Assistant Professor Samiran Ghosh for taking the time to read the thesis and providing valuable proposals for improvement.

This research was funded by the Finnish Doctoral Program in Stochastics and Statistics, UTU Doctoral Programme in Mathematics and Computer Sciences, and Zora Biosciences Ltd. Special thanks to Zora for sharing their expertise on lipidomic research and for providing interesting data to work with. Financial support was also received from the following foundations: Turun yliopistosäätiö, Emil Aaltosen säätiö, Oskar Öflunds Stiftelse, and Suomalainen Konkordia-liitto.

I thank everyone at UTU Stats staff for collegial support, and especially MSc (soon PhD) Ilmari Ahonen for friendship, lunches filled

with general rant, and moments of shared thesis related anxiety. I would like to thank tilastotytöt for giving me updates on off-campus rumours and Tynnyri people for filling up my wine glass regularly. Thank you Mum and Dad for always encouraging me to read and study but still letting me to choose my own way. (Thank you also for knowing when not to ask about the progress of the thesis.) Thanks to my two sisters Sanna and Kati for being the best company, I always have splendid time with you two!

Lastly, I thank Mr. Kashyap Gupta for canceling his talk in JSM2011's contributed session "Bayesian Modeling in Physics and Engineering" so that certain Henri Pesonen, PhD, had some unexpected spare time to stroll around the poster session at exhibition hall D. That act marked the beginning of the greatest adventure of my life.

Turku, 13.10.2016

Maiju Pesonen

Contents

Abstract i

Tiivistelmä iii

Acknowledgements v

List of Original Publications ix

1 Introduction 1
- 1 The omics revolution . . . 1
- 2 Genomics data . . . 3
  - 2.1 The blueprint for life . . . 3
  - 2.2 Gene regulation and expression . . . 4
  - 2.3 Detecting gene expression . . . 5
  - 2.4 Next-generation sequencing . . . 5
  - 2.5 Characteristics of RNA-seq data . . . 7
- 3 Lipidomics data . . . 8
  - 3.1 Lipid species beyond total cholesterol . . . 8
  - 3.2 Emerging field of lipidomics . . . 9
  - 3.3 Quantification of lipid concentrations . . . 10
  - 3.4 Characteristics of lipidomic data . . . 12
- 4 Types of omics studies . . . 14
  - 4.1 Differential expression . . . 14
  - 4.2 Biomarker discovery . . . 15
  - 4.3 Interaction studies and pathway discovery . . . 16
- 5 Setup and aims . . . 17

2 Regression analyses on incomplete omics data 21

1 On the missing data terminology 21
2 Censored data analysis approaches 22
2.1 Maximum likelihood based approach 23
2.2 Multiple imputation approach 24
2.3 ML or MI? 26
3 Regularised regression 26
4 Dimension reduction techniques 29
5 Models for non-continuous responses 31
5.1 Models for count data 31
5.2 Regularised generalised linear models 32

3 Association networks and differential network analysis 35
1 From differential expression to differential networking 35
2 Extracting meaning from network structures 36
3 Reconstruction of the association network 38
3.1 Correlation networks 38
3.2 Bayesian networks 39
3.3 Gaussian graphical models 40
3.4 Model based approaches 41
3.5 Edge selection 42
4 Differential network analysis 43

4 Conclusions 47

Summaries of the Original Publications 49

References 51

Publications 71

List of Original Publications

This thesis consists of an introduction and the following publications:

P1. M Kujala, J Nevalainen (2015): A case study of normalization, missing data and variable selection methods in lipidomics. *Statistics in Medicine*, 34(1): 59-73.

P2. M Pesonen, H Pesonen, J Nevalainen (2015): Covariance matrix estimation for left-censored data. *Computational Statistics and Data-analysis*, 92: 13-25.

P3. M Kujala, J Nevalainen, W März, R Laaksonen, S Datta (2015): Differential network analysis with multiply imputed lipidomic data. *PLoS ONE*, 10(3): e0121449.

P4. M Pesonen, J Nevalainen, S Potter, S Datta, S Datta (2016): A combined PLS and negative binomial regression model for inferring association networks from next generation sequencing count data. *Manuscript*.

CHAPTER 1

Introduction

1 The omics revolution

As long as there has been data, there has been a challenge to transform the data into a meaningful information. Data, in its different forms such as figures, signals, RNA sequences, time series, videos, or functions, is examined for patterns which are seen as information and can be further refined to knowledge. During the most of the twentieth century, science has been based on a reductionistic approach, aiming at understanding complex phenomena by reducing them to smaller, simpler or more fundamental fragments, such as individual genes or proteins. However, during the first decade of the twenty-first century, several omics disciplines have emerged, aiming at analysing a living organism in its entirety: genomics to sequence, assemble, and understand the functions and structure of whole genomes and genes, [93], proteomics to study the structures and functions of proteins produced in a cell [164], and metabolomics to study the chemical processed involving metabolites [35], to name a few. This development has been enabled by the advancements in the modern measuring technology, such as DNA sequencing and quantitative mass spectrometry.

The name “omics” has become a common term referring to a collection of studies of entities [155]. As seen in Figure 1, the explosion

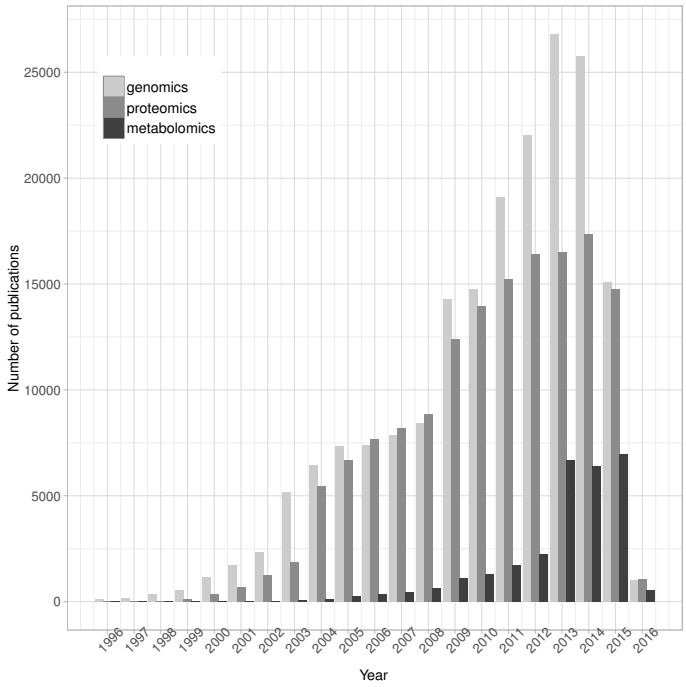

Figure 1: Number of omics-related publications found by Web of Knowledge topic search for the past twenty years. For 2016, the number of publications indicates papers published before March 19, 2016.

of the omics research has led to the rapid increase of the collected data. At the same time, the data is becoming more complex. Data complexity appears as non-linearities, high dimensionality, missing or censored values, non-normality of the measured variables, or dependencies between observational units. Even though the vast amount of data holds a potential to approach research questions systematically and on a grand scale, it poses a huge challenge to statisticians as the tools to understand and make efficient use of these data are not developing at the same pace. The research presented in this thesis aims at developing efficient methodology for multivariate data sets with dependent and/or incomplete observations. The theoretical development is motivated by biomedical applications in genomics and lipidomics.

2 Genomics data

2.1 The blueprint for life

Three fundamental macromolecules being essential to all living organisms currently known, are 1) deoxyribonucleic acid (DNA) along ribonucleic acid (RNA) for storing and decoding genetic instructions, 2) proteins formed by chains of amino acid residues for performing and catalysing various vital functions such as metabolic reactions and DNA replication, and 3) carbohydrates for energy storage and as structural components. From these three, DNA, carried almost in every cell of human body, is the blueprint for performing actions that make things *living*: how to maintain a constant state, how to transform energy, how to grow, how to adapt, how to respond to stimuli, and how to reproduce [84].

All the information stored in DNA is coded by four different nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The order of the bases determines the information needed for building and maintaining an organism. When the bases are attached together with sugar and phosphate molecules, they form a nucleotide. Further, the bases pair up with each other, A with T and C with G, resulting in a ladder like structure that forms a spiral called double helix. DNA is divided into different regions, genes, that are the basic physical and functional units of heredity. Each gene encodes a functional product, such as RNA or protein. Even though large parts of the genome are shared between individuals of the same organism, there may be multiple variants (alleles) of any given gene, leading to polymorphism. These small differences in genome contribute to each individual's unique physical features, the phenotype.

Approximately only two percent of the human genome includes protein coding genes [1]. However, when studying hereditary diseases they are usually targets of the greatest interest, as mutations in them are easily detectable. Even a change as small as one nucleic acid being replaced by another can change one amino acid to another one and that changes the whole end product protein. However, many disease-related mutations also happen in the regions that do not directly encode proteins but rather regulate how the genes behave. Identifying portions of the genome that do and do not code pro-

teins and attaching biological information to these regions is called genome annotation. This process sums up the fundamental pursuit of genetics: determining the genotypes giving rise to different phenotypes.

2.2 Gene regulation and expression

All the information needed to determine the properties and functions of each single cell is encoded in DNA. Regardless of nearly every cell in an organism containing the same set of genes, in a given cell, only a small portion of these genes are active at a given time. During an early development of an organism, cells start to take on specific functions, for which they find the instructions from the blueprint stored in their nucleus, DNA. This carefully controlled pattern is guided by gene regulation, which gives a cell the control over the structure and function, by turning on appropriate genes on and off at proper times. Gene regulation is what makes a liver cell different from a skin cell, and a healthy cell different from a cancer cell. By gene regulation, an organism can also respond and adapt to its environment.

When a gene is turned on, the information it contains is delivered through a two-step course, the transcription into RNA or messenger RNA (mRNA), and the translation of the resulting RNA into proteins [5]. This process is called gene expression and it reflects the activity of a given gene and the rate it passes information to carry its function. Thus, it is the most fundamental level at which genotype revises the observable trait, in other words, the phenotype. By observing changes in gene expression and activity, researchers can potentially identify previously unknown, molecularly characterised diseases and discover biomarkers that predict the risk of a specific condition or response to a given treatment. Eventually, the results of the genome-based research could be implemented as highly effective diagnostic tools, personalised medicine or targeted lifestyle interventions. An interesting question remains to be determined by the researchers aiming at understanding the complex mechanisms behind some specific conditions of interest: which genes are turned on and when?

2.3 Detecting gene expression

The central dogma of molecular biology, first described by Francis Crick [29, 30], is often formulated as follows: "DNA is transcribed into RNA and RNA is translated into proteins in the ribosomes". In this form, the dogma hypothesises that transferring the sequential information stored in DNA is a one way process. Even though exceptions to this central flow are numerous, this basic principle provides a way to get a snapshot of the state of a cell at a specific time, developmental stage or under different environmental conditions. The array of RNA reflects the expression levels of the related genes and provides a measure of gene activity [44].

The most precise estimate of the gene expression would be achieved by detecting the final gene product, but it is often easier to quantify some of the precursors, typically mRNA, and estimate the true expression from these measurements. Measuring the abundance of RNA in a cell or tissue utilises an important exception to the direction of the genetic information flow in the Crick's central dogma: an RNA template can be transcripted to complementary DNA (cDNA) by reverse transcription polymerase chain reaction (RT-PCR) [46] . When combined with real-time polymerase chain reaction (qPCR), which is used to measure the amplification of DNA using fluorescent probes, RT-PCR can be also used to quantify the relative abundance of RNA being present in a cell and thus, describe gene expression by a single number, expression level [107]. Compared to other RNA quantification methods, such DNA microarrays [61] or northern blots [19], the RT-PCR is considered to be the most efficient and sensitive.

2.4 Next-generation sequencing

Through all years available in Web of Knowledge databases (1900 onwards), the keyword "genomics" gives a first hit for a paper published in 1988. In the early years, genetics research focused on individual or a small subset of genes at a time, formulating hypotheses from existing descriptive theories and testing them through wet lab experiments. While producing valuable information, such methods are generally time-consuming. The project aiming at sequencing the whole human genome, launched in 1990, both accelerated and relied on the advances in new, affordable technologies, from fast sequencing tech-

niques to computing methods handling enormous amounts of data. The Human Genome Project was announced completed in April 2003 [76] and only a year later, in 2004, the so-called *next-generation sequencing* (NGS) instruments being capable of producing millions of DNA sequence reads in a single run became commercially available [98].

RNA profiling methodologies based on NGS technologies, usually referred as RNA-seq, typically involve isolating and randomly fragmenting mRNA, translating mRNA to cDNA by reverse transcription and preparing cDNA for sequencing [152]. Sequencing means identifying the nucleotides of a given DNA molecule and converting the result into a *read* consisting of a sequence of letters A, T, C, and G. The fragments of cDNA are simultaneously sequenced to produce hundreds of millions of short reads. The most commonly used NGS sequencing technology solution is the Illumina sequencer (*www.illumina.com*) [15]. To infer gene expression, the RNA-reads are aligned to a known reference genome sequence. Quantitative measures of gene expression levels are then achieved by counting the number of the reads aligned between the beginning and the end of each region in the genome annotation (genes) [44].

The NGS techniques quickly changed the field of genomics by expanding the genomic studies from previously focused readouts to genome-wide scale, by accumulating unprecedented amounts of data and enabling the study of gene-protein interactions, mutation mapping, polymorphism and noncoding RNA discoveries [96]. These techniques have had a major impact on exploring and answering genome-wide biological questions: Figure 2 shows the increasing number of NGS related publications since 2004, as given by the Web of Knowledge topic search. The development of the technology is reflected in the number of publications, as is the decrease in costs. According to the data provided by National Human Genome Research Institute (*www.genome.gov/sequencingcosts*), in 2004 the cost of sequencing on human genome was on the scale of tens of millions of US dollars, as today, in 2016, we are approaching the limit of one thousand US dollars. The fast sequencing technologies have not only led to a huge increase in genetic information but also placed bioinformatics and biostatistics at the leading edge of the novel pipeline devel-

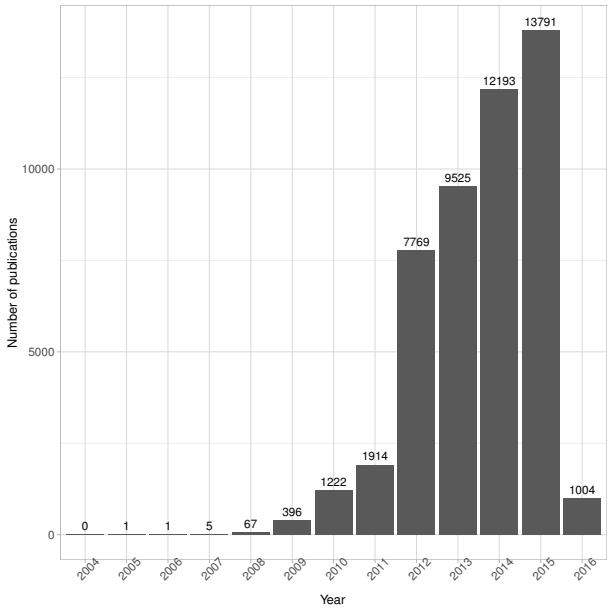

Figure 2: Number of NGS related publications found by Web of Knowledge topic search, since 2004. For 2016, the number of publications indicates papers published before March 19, 2016.

opment for storing, analysing, interpreting and visualising petabyte-scale datasets [96].

2.5 Characteristics of RNA-seq data

One main feature uniting all omics data, including RNA-seq data, is the high dimensionality. Since the discovery of the protein-coding genes, the estimates of their number have been shrinking. However, an RNA-seq data originating from human genome can still contain approximately 19,000-20,500 genes [41, 28], usually measured from a relatively low number of biological replicates. Valid statistical inference requires enough replicates to estimate error [45]. Thus, RNA-seq experiments should include at least three biological replicates per comparison group [10].

The previous golden standard method to detect gene expression, DNA microarrays, generated continuous measurements that represented the concentration of the mRNA molecules detected in the

mRNA assay. Whereas the log ratio expression values from microarray data are typically assumed to be normally distributed, same assumptions can not be made for RNA-seq data that measures the expression levels with positive integers, counts. Count data is often analysed using methods based on Poisson distribution. However, many studies have shown that the variance grows faster than the mean in RNA-seq data, a phenomenon known as overdispersion.

In the extensive literature of statistical methods for RNA-seq data, some sort of data pre-processing, often referred to as normalisation, is required [37, 23]. Normalising data aims at ensuring that the gene expression levels are comparable within and across samples. Different sequencing depths, represented by varying total read counts per sample (library sizes), are one of the most apparent factor causing the read counts of the samples being measured on different scales, and thus incomparable. In other words, two genes having similar expression can have very different read count values depending on the gene length, as a longer transcript will have more reads mapping to it [111]. The most common solution is to transform the read count to reads per kilobase per million mapped reads (RPKB) scale [104].

3 Lipidomics data

3.1 Lipid species beyond total cholesterol

Human plasma is composed of nucleic acids, amino acids (mainly in the form of proteins), carbohydrates (sugars), and lipids (fats, waxes, sterols, fat-soluble vitamins) [119]. The first three components are widely studied, whereas lipids stand out due to their structural diversity, function, and a vast number of individual molecular species. Estimates on the number of different lipid species vary from a few thousand up to hundred thousand lipid species [129]. On April 2, 2015, over 40,000 individual lipid structures were indexed in the most comprehensive lipid database, LIPID MAPS [142]. Individual lipid species are divided into lipid classes sharing similar structures and biological functions. They hold vital roles in biological physiology not only as energy storagers but as well as signaling molecules and structural components of cell membranes [140]. In terms of mass,

lipids are the most important constituent of the human brain, and the second most important of all other soft tissues [105].

Overall, lipids are considered as metabolites, and hence, lipidomics belongs to the general field of metabolomics. Nevertheless, lipidomics is regarded as a distinct discipline due to the uniqueness and functional specificity of lipids relative to other metabolites. In contrast to proteins, there are no genes coding for lipids as such. Lipids are obtained for example from our diet and they are further modified by gene coded enzymes. Similarly to the end products of molecular pathways initiated at genomic, transcriptomic and proteomic levels, lipids serve as valuable indicators of both genetic and environmental factors. Therefore, lipids have been proposed to be as important for life as proteins and genes and their importance for life and health has been recognised [119, 129].

As the rich spectrum of individual lipid species all have defined roles in the support and sustenance of cellular functions of human body, it is only natural that lipid metabolism has been related to several human diseases, such as diabetes [101, 87], cancer [173, 103, 162], cardiovascular disease [135, 134], brain injuries [132], and Alzheimer's disease [161, 118]. As a consequence, the exploration of lipid profiles holds the potential to provide a readout of biomarkers for an early detection of a disease [163, 66].

Drugs targeted against lipid-metabolising enzymes are not new to the pharmaceutical industry: statins, the cholesterol-lowering agents, are a multibillion business solely in the United States [144]. However, detailed analyses of lipid profiles are also expected to reveal information that will stretch beyond the knowledge obtained with the current routine clinical lipidology tools, such as total triglyceride levels and high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol [86]. This has inspired the emergence of a new field of omics research, lipidomics.

3.2 Emerging field of lipidomics

The rapidly expanding field of lipidomics complements the breakthroughs made in genomics and proteomics [34, 156]. Despite the fact that lipidomics is considered yet another omics technology, it

delivers novel type of data and information. A distinguishing characteristic of lipidomic data is that lipids can be considered as intermediate phenotypes providing much more detailed information about the state of an organism determined by a combination of genetic regulation, functionality of protein machinery, and environmental factors, compared to, for example, genetic information alone [86].

After completing the sequencing of the human genome, research has expanded to postgenomic technologies, including metabolomics [66]. The Web of Knowledge topic search with keyword "metabolomics" gave the first three hits for year 2000, as seen in Figure 3. During the same year, approximately three genomics and one proteomics related papers were published every day. The early metabolomics studies predominantly focused on metabolites that were easier to detect and quantify. Lipids gained less attention as their comprehensive analysis was hindered by the sheer complexity of the lipidome with tens of thousands of different lipid species, requiring different instruments to examine the lipidome, and leading to labor-intensive workflows. Thus, in comparison to other omics technologies, the emergence of lipidomics has been slower, as reflected in the limited number of publications in Figure 3. The first hit for keyword "lipidomics" appeared in 2002, but the first article providing an in-depth description of the human lipidome was published in 2010 [119]. Still in 2015, only 13.4% of all metabolomics-related publications concerned lipidomics, as seen in Figure 3.

Novel analytical technologies, especially liquid chromatography and mass spectrometry, and the more widespread availability of regents and tools, such as synthetic lipid standards, analogues of natural lipids, and lipid affinity probes, have spurred the study of lipid metabolism and enabled conducting analyses in a high-throughput format [54, 58, 156]. Lipidomics can be expected to contribute in various areas of biomedical research, with various applications in drug and biomarker development, and support in inferring structure and function of biological systems.

3.3 Quantification of lipid concentrations

Quantifying lipid concentrations from an aliquot of serum typically starts by adding constant amounts of chemical substances called

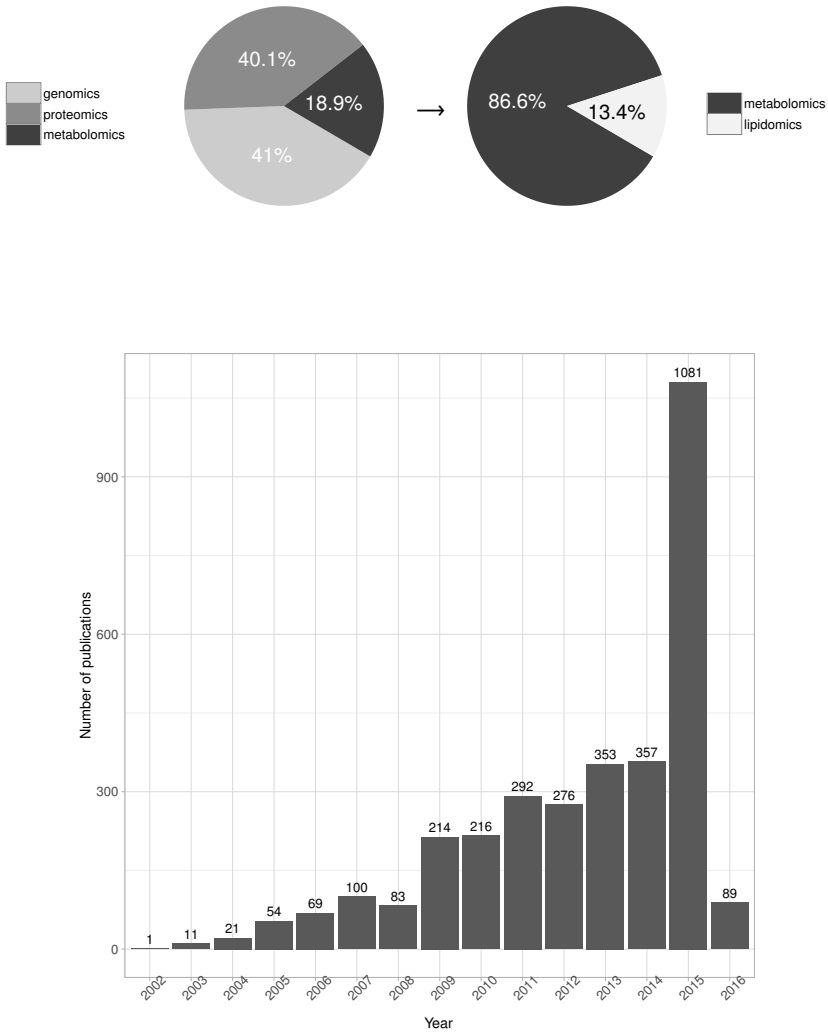

Figure 3: The number of lipidomics-related publications found by Web of Knowledge topic search since the first hit in 2002 (bottom). For 2016, the number of publications indicates papers published before March 19, 2016. The pie charts (top) show the relative proportions of genomics, proteomics and metabolomics publications from all omics publications in 2015. During that year, still only 13.4% of all metabolomics-related publications concerned lipidomics.

internal standards to serum samples. The internal standard is selected to match the lipids analysed so that it produces a signal similar to the signal of lipids, but still, different enough for the measuring instrument to be able to separate the two. The serum lipid concentrations can then be derived from a calibration curve by comparing the signals originating from the lipids to the signal originating from the internal standard. After adding the internal standard, the small molecule metabolome, including lipids, are extracted from the insoluble material (proteins).

Rapid advances in technologies such as mass spectrometry (MS) contribute significantly to the research of lipidomics as the structures and functions of the lipids on a molecule level can nowadays be efficiently identified and quantified [157, 18, 81, 91]. MS is an analytical technology that determines the mass-to-charge ratio of individual analytes. The samples introduced to mass spectrometer are first ionised in the ion source and then detected after being separated according to their mass-to-charge ratios by the mass analyser system. The detected signals are finally displayed in a mass spectrum, a plot of ion intensity versus the mass-to-charge ratio. The intensities of individual ions are achieved from the mass spectrum [57].

Comprehensive lipidomic studies demanded a repertoire of many different analytical platforms [86]. For each platform containing a set of samples, a stringent cut-off is applied for separating background noise from actual lipid peaks in the mass spectrum. This cut-off value is called a lower limit of detection (LLOD). Acquired mass spectrometry data is processed using bioinformatic tools that convert masses and counts of detected peaks into corresponding lipid names and abbreviations, usually assigned according to Lipid MAPS nomenlecture [42]. Quality control samples are also included to monitor the overall quality of the lipid extraction and MS analyses by removing technical outliers and lipid species that were detected below the lipid class based LLOD.

3.4 Characteristics of lipidomic data

New technologies are producing vast amounts of lipidomic data and thus, have created a great demand for sophisticated tools for statistical analysis and inference. Similarly to other omics data, lipidomics

data is characterised by a substantial number of individual lipid species analysed from relatively few samples or biological replicates. A feature specific for lipidomics data is that the measured lipids comprise small groups called lipid classes. Since lipids belonging to the same class and even different classes of lipids share similar biological functions and structure, typically, there exists high co-regulation across different lipids [106]. In lipidomics data, this is reflected by groups of mutually correlated lipids. Therefore, the correlation structure of the lipidomics data should be considered when analysing and visualising the data.

During the start of the research projects summarised in this thesis, at the beginning of 2011, comprehensive lipidomic studies demanded a repertoire of many different analytical platforms. Combined with sample handling and analytical accuracy, there might be variability between plates of samples analysed on different platforms that does not reflect true biological differences. To eliminate these systematic sources of variation, some sort of pre-processing or normalisation of the data should be considered.

As explained in Section 3.3, each platform has a specific limit of quantification, under which true signals can not be separated from the background noise. Thus, low-abundant lipid species are not often detected at all, and other species may include a set of censored values, that are low-level concentrations considered to be too imprecise to be reported as a single number. These so called left-censored values are a commonplace phenomena for proteomic, metabolomic, and lipidomic data from mass spectrometry platforms. Often, in the final dataset, left-censored values are notated by "< LLOD", where LLOD is some positive real number. This notation describes the property that left-censored values are known to be somewhere between zero and LLOD. Finding a proper way to account for left-censored values in the statistical analyses is crucial as the simple exclusion of them produces an upward bias in subsequent measures of location, such as means and medians.

4 Types of omics studies

4.1 Differential expression

Both NGS and lipidomics data are typically collected under an experimental design, where one is interested in comparing the expression or concentration levels of genes or lipids between two groups of samples exposed to different conditions, to understand molecular basis of phenotypic variation in biology. For example, to understand an effect a certain drug is having, it is interesting to compare diseased and control groups, consisting typically of minimum three replicates, which genes or lipids are up regulated (increased in expression/concentration) or down regulated (decreased in expression/concentration). This is called differential expression [99, 36, 112].

For continuous expression or concentration measurements, such as DNA microarray or lipidomics data, the group expression or concentration level can be summarised by the mean expression level of the replicates. Thus, the problem is fundamentally comparison of the means. With two comparison groups, differential expression problem can be solved by conducting a t-test, if one can assume that the data is normally distributed, or a non-parametric Mann-Whitney U-test [31]. With more than two comparison groups, one can conduct a variance analysis or the non-parametric equivalent, Kruskall-Wallis' test.

On a whole genome- or lipidome-wide scale, the goal of the differential expression analysis is to generate a list of all genes or lipids that are differentially expressed. Thus, thousands of hypotheses are tested simultaneously, causing a problem of multiple comparisons. To avoid numerous false positive discoveries, some sort of multiple comparison correction, such as Bonferroni correction, is usually applied. An other approach is to control the expected proportion of falsely rejected null hypotheses. This so-called false discovery rate (FDR), first introduced by Benjamini and Hochberg [14], provides a less rigorous control of type I error in comparison to multiple comparison correction. The theory of FDR relies on null hypothesis tail areas (p-values), and as such, is an extension of conventional frequentist hypothesis testing to simultaneous inference.

Identifying differentially expressed genes from RNA-seq data requires methods beyond elementary statistics due to high-dimensionality, different sequencing depths, count format and non-normal distributional features of the data. Also, the efficiency of an elementary one-gene-at-a-time analysis is questionable in cases where the number of replicates remains small. A thorough review of differential expression for RNA-seq data is given by [131]. A number of software packages have been developed to conduct differential expression analysis for RNA-seq data [122, 123, 8].

4.2 Biomarker discovery

High-throughput studies of biological systems are accumulating omics-scale data at an unprecedented rate, and the datasets are expanding both in the number of variables measures as samples analysed, as for example some bacteria samples collected in hospitals is already being sequenced routinely. These data, when collected from different groups of samples under different biological conditions, provide new understanding on how diseases should be managed and how new drugs and tests could be developed and used, and thus, enable the identification of genetic and molecular biomarkers for disease processes [110].

A biomarker is a biological characteristic that can be used as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [17]. In context of omics data, these features can be, for example, individual genes or molecules, a bigger set of genes and/or molecules, a relation of two lipid concentrations, or a whole lipidomic profile. One example of a successful genomic biomarker discovery and further development into a multiparameter gene test is given by [113]. The test helps to determine which early stage breast cancer patients are at higher risk of recurrence and thus may be more likely to benefit from chemotherapy. Thus it allows women at lower risk to safely forgo chemotherapy, avoiding toxicities, cost, and quality-of-life issues associated with treatment.

Simultaneously, the deluge of data has complicated the extraction of meaningful molecular fingerprints of biological processes from these complex datasets. From the statistical point of view, a quest for

biomarker discovery is a model selection or a binary classification problem. Due to the high dimensionality of the data, and relatively low number of samples, the use of conventional statistical models is not possible, as there are too many parameters to estimate, with not enough information to do it. Instead, analysis approaches reducing the dimensionality of the data, or performing variable selection and regularisation in order to identify only a small fraction of molecules giving the best prediction accuracy, could be used.

4.3 Interaction studies and pathway discovery

High-throughput experimental methods produce a great volume of complex, interconnected data. Often the initial goal in many omics studies is to find a set of co-regulated genes or a set of molecules that share a related expression pattern in a certain phenotype, disease model or human disease, or in response to a drug treatment. Thus, it is often more meaningful to consider all genes/molecules simultaneously, than compare expression levels of individual genes/molecules between two or more biological conditions. Visualisation of these complex systems as pathways (graphs that show overall changes in state) or networks (graphs that do not necessarily show state changes but describe the association and co-expression structure) has been found useful in creating understanding of biological systems [51].

Biological association and interaction networks provide information about the essential processes behind different conditions, and help to recognise the important distinguishing molecules, for example for therapeutic purposes. Here, between-molecule association describes the similarity of the concentration levels of two molecules and how they change together. The core of a network analysis is a connectivity score that represents the strength of the association between two particles. At its simplest, the connectivity can be represented with a correlation coefficient. Further, differential network analysis provides a formal statistical method capable of inferential analysis to examine differences in network structures under two or more biological conditions [53]. It also guides in identifying potential relationships requiring further biological investigation.

5 Setup and aims

The classical statistical methods rely mostly on complete data vectors measured on all samples. However, this assumption is not often met due to reasons varying from technical obstacles or limitations to nonresponse. Missing data may have a significant effect on the conclusions drawn from the data. Thus proper handling of the remaining data is crucial. By selecting only the samples with fully observed variable profiles can lead to a great reduction of the sample size and hence to a serious loss of precision. Thus, one of the main themes carried through the research conducted in this thesis is to provide efficient methodology for various types of missing or censored omics data.

In omics data, expression or concentration values can be missing for various reasons and in various amounts. In DNA microarray data, typically $1-10\%$ of measurements are missing, affecting up to 95% of the genes [21]. Gene expression values in microarray assays can be filtered out due to low spot pixel frequencies, occurrence of technical errors during the hybridization, low fluorescent intensities, or due to presence of dust, scratches, and systematic errors on the slides [21]. Due to the nature of next generation technologies, RNA-seq data does not include missing count values. If no read is aligned on a specific gene, and gene expression "is missing", the value is recorded as a zero. Data generated on mass-spectrometry platforms contains typically a large number of missing values accounting for $10-40\%$ of data and affecting up to 80% of all variables [71, 55]. Also, most of the missing values do not occur randomly but rather as a function of signal intensity. Some values are missing as the relating molecule is absent in a given sample, whereas some molecules are detected, but in such a low abundances, that their concentrations fall below a set lower limit of detection [22], and thus can possibly be mixed with noise. These values are known as left-censored concentrations or non-detects. In addition, values may not be measured properly owing to a technical problem. Depending on their origins, missing values should be considered differently and dealt with in suitable ways. For example, if data containing left-censored values is analysed using only the completely observed data, the means of the concentrations would be overestimated and the standard deviations would be

underestimated. Consequently, any related test statistic or estimate would be biased.

To review how widely missing data analysis methods are acknowledged in published omics studies, a Web of Knowledge topic search with the following search phrase was conducted:

```
TS=genomics AND
(TS="missing data analysis" OR
TS="missing value imputation" OR
TS="multiple imputation" OR
TS="imputation of missing values" OR
TS="imputing missing values" OR
TS="missing values").
```

The word "genomics" was replaced with proteomics and lipidomics in the subsequent searches. The results show, that after 15 years since the first applications of imputation methods to gene expression data, the rate of publications on missing data analysis on omics data is not slowing down (see Figure 4). New and improved methods are published at a steady rate. The processing and efficient use of missing values provides a rich source of appealing research questions, especially when considering application-specific modifications exploiting information sources relevant to the missing data problem.

The need for proper handling of missing data has previously been recognised in the analysis of DNA microarray data [148], in gel-based proteomics data (a method to separate proteins prior to mass spectrometric analysis) [6], and in metabolomics data [71]. Studies have been conducted to evaluate how missing values effect the estimation of statistical parameters [148], how they influence univariate data analysis [127] and multivariate [115] data analysis, and to give recommendations on optimal methods for their imputation [80, 149, 9]. Web of Knowledge search for missing data analysis methods specifically for lipidomics data refers only to two publications included in this thesis, [P1] and [P3]. Thus, there is still demand for development of solid statistical analysis procedures combining missing or censored values with high-dimensional data.

With the recent accumulation of high throughput data, the analysis of

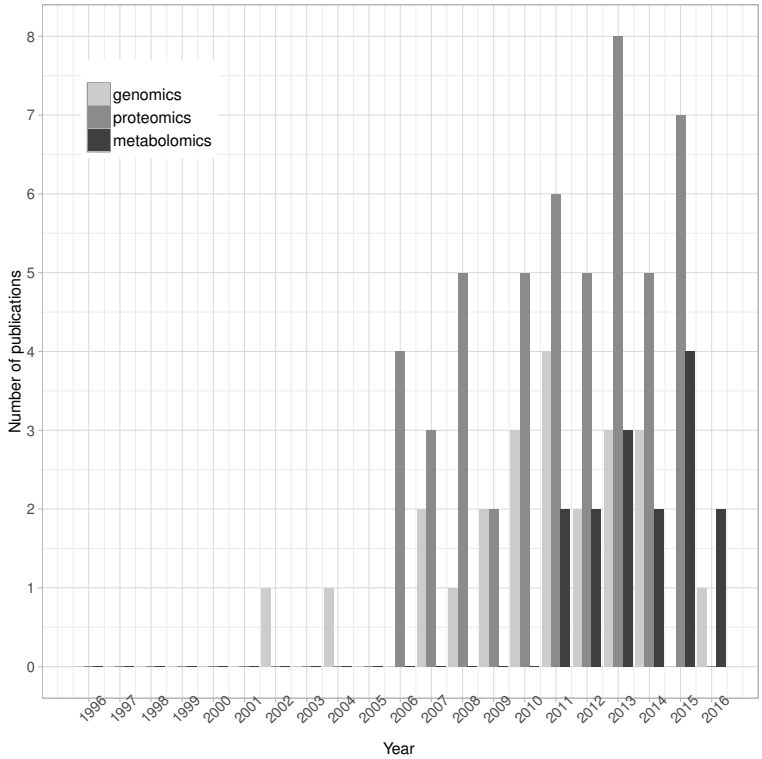

Figure 4: Number of missing data analysis related publications during the past two decades in the fields of genomics, proteomics, and metabolomics found by Web of Knowledge topic search. For 2016, the number of publications indicates papers published before March 19, 2016.

biological networks has gained significant interest. Biological association or interaction networks provide information about the essential processes behind different conditions, such as healthy and diseased statuses, and help to recognise the important distinguishing features. Existing network reconstruction methods are primarily developed for continuous and complete data. Thus, the research conducted in this thesis, aims at providing methodology to augment network reconstruction on incomplete and count data. Besides learning the association structures within a condition, an interesting problem is to compare the network structures between different conditions. For the most parts, previous work on this so-called differential network analysis has been based on complete case analysis, that is, including

only those samples for which all measurements have been detected. Thus, in this thesis, also the differential network analysis is expanded such that it can be implemented on left-censored multiple imputed data.

CHAPTER 2

Regression analyses on incomplete omics data

1 On the missing data terminology

Missing data are a common occurrence in omics data and thus understanding the reasons behind missing values is necessary when analysing the remaining data. Simply omitting the missing values in the analyses can lead to a severe loss of the effective sample size, may cause bias and a loss of precision. According to a classification originally presented by Rubin [124], missing values can be divided into three subgroups. These are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

The partition of the missing values into these categories is based on the probability of an observation being missing. If this probability does not depend on observed or unobserved measurements then the observation is MCAR and the analyses performed on the complete data are unbiased, although some information is usually lost. If the missingness mechanism can be expressed solely using the observed data, the values are MAR. This is the most general condition under which the data can be analysed using the observed data only and no information about the missing value mechanism is needed to be incorporated in the analysis. Thus, in the likelihood setting, a

term ignobility is often used to refer to MAR mechanism. However, it is important to note that only the mechanism is ignorable, not the missing values themselves. This is due to the property, that the analysis methods based on likelihood are valid under MAR, whereas non-likelihood methods, such as the ones based on generalised estimating equations, will lead to biased results unless adjusted externally. Such proposed adjustments include for example multiple imputation before data analysis, or inverse probability weighting. To summarise, with likelihood-based methods in their standard form, inference based on both complete data and missing data mechanism models would be the same if inference was based on complete data only.

Finally, an observation is MNAR if, even accounting for all the available observed information, the probability of a value being missing depends on the value itself or other unobserved reasons. The phenomenon is also known as non-ignorable nonresponse. In general, this is a very challenging to handle, but in particular cases, valid inference is achieved using a joint model of both data and the missingness mechanism: with censored data, a subset of values are unobservable due to some censoring mechanism. If this mechanism is known, it can be modelled and hence the censored observations can be used to improve the inference. These kind of informative missing values were present in Publications [P1], [P2], and [P3], where analysis methods were adjusted to efficiently incorporate left-censored values (non-detects). Due to the fact that censored values are a specific trait of the omics data, two censored data inference methods are reviewed in detail in the following sections.

2 Censored data analysis approaches

Hewett and Ganser [65] divide censored data analysis methods into four categories: substitution methods, log-probit regression, maximum likelihood (ML) estimation methods, and non-parametric methods. None of the methods has been recommended to be the ideal solution in all different scenarios. The recommendation depends on the sample size, the divergence from log-normal distribution or the degree of censoring. However, due to its many desirable statistical properties, ML estimation is often considered the gold

standard provided the data is well-described by some parametric probability distribution [62, 83, 172]. In Publications [P1] and [P3], left-censored and MAR missing values were multiple imputed, while in Publication [P2], ML estimation techniques were employed. Under identical assumptions, both approaches produce estimates that are consistent, efficient and asymptotically normal.

2.1 Maximum likelihood based approach

The ultimate aim of the missing data analysis is never to predict the exact values themselves, but rather to facilitate revealing the most important findings from an incompletely observed data at hand as precisely as possible. Thus, methods that can handle missing values without any pre-processing procedures, are ideal. An approach that fulfils this property, is to analyse an incomplete data set using ML estimation. As ML estimation can be used to estimate the parameters of a statistical model given data, it requires an assumption on the distributional properties of given variables. This method does not impute any data, but rather uses observed data to compute ML estimates, that are particular values of the parameters that make the observed data the most probable given the model.

With or without missing data, the first step of the ML estimation is to construct a general likelihood function. Let n be the number of samples, $i = 1, \ldots, n$ and p the number of variables $x_{i1}, \ldots, x_{ip}$ whose expression levels or concentrations are measured. The likelihood function is then defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_{i1}, \ldots, x_{ip}; \theta),$$

where f is the probability density function of the assumed distribution relating to variables $x_{i1}, \ldots, x_{ip}$, and θ is a set of parameters to be estimated. To achieve the ML estimates, this likelihood function is maximised. Often, the likelihood is presented in the mathematically convenient log-scale as a sum of the logarithms of likelihood contributions of each sample,

$$l(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(x_{i1}, \ldots, x_{ip}; \theta) = \sum_{i=1}^{n} \log f(x_{i1}, \ldots, x_{ip}; \theta).$$

If for sample i, say, the values of two first variables x_1 and x_2 are missing, and the missing data mechanism is assumed to be ignorable, the likelihood contribution of this particular sample is

$$l_i(\theta) = \int \int \log f(x_1, \ldots, x_p; \theta) d\, x_1 d\, x_2.$$

Essentially, ML approach to incorporate missing values in the analysis is done by integrating over all possible values for samples including missing data. When a sample has left-censored values, the integrals are taken over the observation space that is not observeble and that is defined by LLODs. If a sample includes both MAR or MCAR values and left-censored values, the likelihood contribution can be partitioned further, as was done in Publication [P2].

2.2 Multiple imputation approach

One commonly used approach is to substitute the left-censored values with a suitable constant and then analyse the resulting data as it was complete. Potential substitution values include the sample mean or median of the uncensored values for the corresponding variable, zero, LLOD/2, or a minimum of the observed values. These alternative approaches have been investigated in previous studies [43, 40, 141]. All of them are more or less biased, but they are still used despite the criticism [63, 64].

For metabolomics and microarray data, more advanced substitution method has been recommended: the substitution value is computed by finding k metabolites or genes most similar in terms of their intensity profiles across all samples, identified based on the Euclidean distance similarity measure. The substitution value is then estimated as a weighted average of the k metabolites, weights given by their similarity [138, 148]. Also, [108] suggested an imputation approach consisting of three stages, principal component regression, Bayesian estimation and the expectation-maximisation repetitive algorithm. The missing values were then replaced with the expectation of the estimated posterior distribution. However, the two latter approaches assume that missing values occur randomly and independently of other features, which does not hold for the left-censored values.

No matter how advanced the substitution method, single imputation does not reflect the full uncertainty created by missing data. This issue has served as a motivation for multiple imputation (MI), a statistical technique reflecting the uncertainty that arises when data remains unobserved. In contrast to ML approach where everything is done under a single model, MI approach requires separate models for imputation and analysis. The key idea of MI is to use the conditional distribution on the observed data to generate a set of plausible imputations for the missing data. In practice, the draws are based on an appropriate posterior distribution [25]. Imputations are repeated M times, creating multiple data sets, which are analysed individually as if they were complete, resulting in a set of parameter estimates. Finally, the results are combined across all multiple imputed data sets by averaging them, and the standard errors of the estimates are computed as a combination of within-imputation and between-imputation variances, by so-called Rubin's rules [125]. These rules incorporate the imputation related uncertainty into the analysis.

MI is widely used with various omics datasets [80, 133, 2, 88, 16]. The origin of missing values can be caused by different reasons and depending on these origins missing values should be considered differently and dealt with in different ways. Especially, left-censored values should be multiple imputed with caution. The MI methods for left-censored data are appealing due to their relatively simple computational algorithms. The literature includes applications in univariate [12, 73], bivariate [26] and multivariate settings [69, 27].

MI was implemented in Publications [P1] and [P3] with a technique called MI by chained equations [120, 150]. In this approach, the imputation model, specifying the dependence of the conditional distribution of the missing data on the observed data, is constructed through a set of univariate conditional regressions, once for each incomplete variable. The choice of the model is flexible depending on the type of the variable to be imputed, for example, linear regression for the continuous variables, and logistic regression for the binary variables. In practice, the imputation is carried out using an acceptance-rejection sampling principle. For the left-censored values, draws from the conditional distribution are accepted only if they fall below the observed LLOD. If a candidate value does not

meet this condition, it is rejected, and a new candidate is drawn sufficiently many times until acceptance. For missing values originating for other reasons than left-censoring, all draws are accepted.

2.3 ML or MI?

In an ideal situation, ML approach is simple to implement, as everything is done under a single model which produces a deterministic result, and the approach has optimal statistical properties, if the underlying assumptions are met. In contrast, MI method gives a different result every time it is run due to the random draws as a part of the imputation process. However, in some situations proper substitution of the censored values, using for example multiple imputation, is computationally more feasible. For example in Publication [P2], the evaluation of likelihood function to be maximised is computationally very demanding, partly due to numerical integration of the multivariate normal cumulative distributions, and to be usable in practice, an approximation of the ML results is provided. While combining the results achieved from multiple imputed data sets takes some effort, the imputation of a high dimensional data multiple times is usually faster than solving the high dimensional optimisation problem in ML approach. At the moment, the software available for implementing ML approach are rather limited, whereas a big attraction of MI is that once the imputed data sets are generated, any chosen software or method can be used to analyse the datasets.

3 Regularised regression

High throughput omics data includes often large number of variables p measured in relatively small number of patients n. A single data set may contain expression profiles for over 20,000 genes, measured over a range of time points and experimental conditions, so that determining which genes are potentially relevant to the studied problem requires an extensive search through a large amount of often noisy, multivariate data. In general, common statistical techniques cannot be employed in such situations without very specific hypotheses about important variables to be included in the models. An attempt to fit an ordinary least squares regression model on a data set with more variables than observations would lead to a saturated

model. Therefore there has been interest in applying methods that automatically select important variables in some fashion.

One possible solution could be stepwise regression where the choice of predictive variables is carried out by an automatic procedure [67]. The stepwise model selection has received severe criticism, stating, for example, that the p-values are too low due to multiple comparisons and are difficult to correct, the standard errors of the parameter estimates are too small, and the parameter estimates are unstable when the number of variables is relatively large and variables are highly correlated [59].

Several alternatives to classical variable selection techniques have been suggested where regression coefficients are constrained in some manner, for example, by setting L_1- or L_2-penalties, or a combination of both. Such methods are generally called regularised regression models and their use in analysing omics data has become a common procedure [143]. Regularised regression techniques can be used, for example, to identify the variables giving a best classification in biomarker discovery problems as was done in Publication [P1] or to select edges when reconstructing association networks, as was briefly tested in Publication [P3].

A general log-likelihood penalised by L_2-norm, also known as *ridge penalty* [68] or Tikhonov regularisation, is formulated as

$$l(\theta) - \lambda \sum_{j=1}^{p} \theta_j^2,$$

where λ is a tuning parameter controlling the effect of the penalisation. Ridge penalty shrinks all directions, but sets a larger shrinkage on low-variance directions. Ideally, λ is large enough to shrink the parameter estimates relating to unimportant variables close to zero, but keeping the important ones non-zeros.

The least absolute shrinkage and selection operator (*lasso*) penalty [146] maximises the objective function

$$l(\theta) - \lambda \sum_{j=1}^{p} |\theta_j|.$$

An attractive property of lasso is that it shrinks parameter estimates relating to unimportant variables exactly to zero, making the resulting models easy to interpret. However, lasso is able to select at most n observations before it saturates, a property that is not ideal in $p >> n$ -situations. Also, in the presence of high degree of collinearity between the predictive variables, lasso occasionally produces poor results where as ridge performs better [146].

Elastic net regularisation [174] combines both of the ridge and lasso methods, and thus maximises the objective function

$$l(\theta) - \lambda_1 \sum_{j=1}^{p} \theta_j^2 - \lambda_2 \sum_{j=1}^{p} |\theta_j|.$$

By blending the two penalties, elastic net performs at worst as well as lasso or ridge, an in certain mentioned conditions, outperforms both.

These penalised regression methods can be implemented on missing data using both approaches presented in Sections 2.1 and 2.2. The penalisation term can substracted from the partitioned likelihood functions presented in Section 2.1 or then penalised regression models can be fitted on multiple imputed datasets. However, the latter approach can result in divergent sets of selected variables between imputed datasets. In Publication [P1], this problem was solved building on the ideas presented by [160, 151]. The imputed datasets were stacked and then a selected regularised regression model was fitted to the resulting large dataset with weights proportional to the number of observed values on each sample.

Alternatively, regularised regression can be implemented in MI data using an approach called MI-lasso, suggested by Chen and Wang [27]. Their method treats the parameters relating to a particular variable across all imputed datasets as a group and applies a group lasso penalty [166]. As a results, the parameter estimates of the same covariate are either all zero or nonzero leading to consistent variable selection across MI datasets.

4 Dimension reduction techniques

Various dimension reduction methods are also useful tools when analysing high-dimensional, complex data. In these approaches, the data matrix is linearly transformed to a set of derived variables whose number is smaller than or equal to the rank of the data matrix. The derived variables can then be used in a chosen regression model instead of the original variables. Examples of such derived variables are principal components (PCs) and partial least squares (PLS) components.

Principal component analysis (PCA) converts, using an orthogonal transformation, a set of possibly mutually correlated variables into a set of not linearly dependent PCs. This transformation is defined so that the first PC captures as much of the variability in the data as possible and each of the subsequent components has the highest variance possible conditioned on being orthogonal to the preceding components. In short, PCA selects the M largest varying directions and discards the rest. The number of PCs M is usually smaller than the number of original variables, making it attempting to use especially in context of high-dimensional data. PCA is used mostly for exploratory data analysis and predictive modelling (PC regression) [79]. Often, it can be thought as method to reveal the internal structure of the data in a way that best explains the variance in the data. One disadvantage of PCA is that the PCs are weighted combination of all original variables. Sparse PCA [175] defeats this limitation by shrinking some of the weights to zeros, and thus producing PCs being a combination of only a subset of the original variables. This eases the interpretation of the PCs, especially in extremely high-dimensional cases, and provides a way to identify the most interesting variables for further experiments.

Similar to PCA, PLS components are constructed as a set of linear combinations of the original variables which compress the information into a lower dimension. While PCA seeks directions with the highest variation, PLS seeks directions having both high variance and high correlation with the response, in other words, maximises the covariance between the two sets of variables. PLS has been found to be a versatile tool for the analysis of high-dimensional omics data, employed for example in tumour classification from transcriptome

data, identification of relevant genes, survival analysis and modelling of association networks and transcription factor activities [20]. Also, a sparse version of PLS has been suggested to conduct variable selection for high-dimensional datasets [100].

PCA is also related to canonical correlation analysis (CCA). CCA is a general procedure for exploring the linear relationships between two sets of variables measured on the same individual. While PCA defines new orthogonal components to describe variance in a single dataset, CCA defines components that optimally describe the covariance structure across two datasets. In other words, CCA defines what is common amongst the two sets. In the context of omics data, CCA could be used to explore, for example, relationships between lipidomic profiles and various phenotypic measurements. As for PCA and PLS, also a sparse version of CCA has been proposed to increase the biological plausibility and interpretability of the results achieved from the analysis of high dimensional data [159].

As the multivariate statistical methods discussed above often rely on a sample covariance matrix, estimation of the sample covariance matrix in the presence of left-censored values provides an interesting challenge. Conventional estimators of the covariance matrix require complete data matrix, leading to either filtering out the observations or variables with incomplete measurements, or to impute the missing values. Either of these approaches changes the structure of the data and the resulting inferences from the used multivariate method. There are guidelines to conduct PCA in the presence of missing data [75, 50], but only recently PCA was considered in the presence of left-censored data [154]. Extensive literature exists regarding univariate and bivariate ML-based methods for estimating the measures of centrality and variability in the presence of the left-censored values, such as [94, 95, 158]. The extensions of ML-based estimation techniques to the multivariate setting are fewer. In this thesis, Publication [P2] addresses the same issue, but provides a more general solution by estimating the covariance matrix, which can then be used as a basis for any statistical method, such as PCA. The resulting lower-dimensional representations can then be used for visualisation, clustering, or classification of the data.

5 Models for non-continuous responses

So far, all the models considered in this thesis have had continuous responses. However, often the responses in the omics data problems are either discrete numeric such as count data generated on NGS platforms, or nominal, for example, when predicting the class generating the data. Generalised linear models provide a flexible framework for analysing various types of data.

5.1 Models for count data

When analysing NGS data, assuming that the read counts, or more specifically, the expected number of read counts within a given region of the transcriptome from a given experiment, follow a negative binomial distribution has been shown to satisfactorily capture both biological and technical variability [8, 123, 36]. The negative binomial distribution is especially useful for discrete data consisting of positive integers, whose sample variance exceeds the sample mean. The choice of the negative binomial distribution can be justified by the following. The process, where one takes an RNA transcriptome and chooses a location at random to produce a read, is a Poisson process. But when the selected depth of the sequence is considered, the process will be Poisson distributed. However, the inability of Poisson distribution to model unequal mean and variance for the read counts makes it a poor model for explaining the variability between biological replicates. The dispersion parameter of the negative binomial distribution allows us to model this variation. When there is no biological variation between the replicates, the negative binomial distribution reduces to Poisson. Thus, in some NGS applications, technical variation can be treated as Poisson, on top of which the biological variation is represented by the overdispersion parameter of the negative binomial distribution.

Negative binomial regression is a generalised linear model where the dependent variable, say Y is a count of the number of times an event occurs. A convenient parametrisation of the negative binomial distribution is given as

$$P(Y=y)=\frac{\Gamma(y+1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}\Big(\frac{1}{1+\alpha\mu}\Big)^{1/\alpha}\Big(\frac{\alpha\mu}{1+\alpha\mu}\Big)^{y},$$

where $\mu > 0$ is the mean of Y and $\alpha > 0$ is the heterogeneity parameter. In generalised linear models, the relationship between the mean (or some other parameter of interest) and the predictive variables $x_1, \ldots, x_p$ is defined via link function. In the case of negative binomial model, the link function is natural logarithm, and the negative binomial regression model is designated as

$$\log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ are the regression coefficients to be estimated. The corresponding likelihood function is then defined as a product of the probabilities above with μ replaced by the exponent of the linear predictor.

In Publication [P4] the negative binomial regression model was employed in estimating the associations between large number of genes. There, each gene at time were set as a dependent variable y_i, and the dependent variables x_i were PLS-components constructed on the remaining genes. It is worth noting, that also negative binomial regression models could be fitted via penalised maximum likelihood resulting some of the constructed association scores shrinking to zero.

5.2 Regularised generalised linear models

In biomarker discovery, the designs are often either prospective, matched case-control designs or sets of cases and controls without matching. Generalised linear models, such as logistic regression and conditional logistic regression, are useful in analysing such data. However, fitting a full logistic regression model resulting in a large vector of regression coefficients is not very easy to interpret or to adopt to clinical practice as a risk indicator. A researcher would almost always prefer a considerably reduced model with the most relevant variables related to a given disease. In addition, fitting a full model to high-dimensional data is obviously not always possible.

In a binary classification problem, the study design postulates the likelihood. In a prospective design, if for each sample i, $i = 1, \ldots, n$, one observes a dependent variable y_i and a set of predictive variables

$x_i = (x_{i1}, \ldots, x_{ip})$, the likelihood function of the logistic regression model is

$$l(\beta) = \sum_{i=1}^{n} y_i \log p_i + (1 - y_i) \log(1 - p_i),$$

where

$$p_i = P(y_i = 1) = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}.$$

Variable selection can be performed by regularising this likelihood function with a chosen penalty term. In a case-control matched study, the stratification of the patients needs to be taken into account. This leads to conditional logistic regression model and gives a log-likelihood

$$l(\beta) = \sum_{i=1}^{m} \log\Big(\frac{1}{1 + e^{-z_i \beta}}\Big),$$

where m is the number of samples in a case group and $z_i = x_{i,\text{case}} - x_{i,\text{control}}$ is a vector of the differences between a case-control pair. Such regularised generalised linear models were employed in Publicaton [P1] when exploring a combination of lipid concentrations giving a best prediction of case/control statuses of cardiovascular disease patients.

CHAPTER 3

Association networks and differential network analysis

1 From differential expression to differential networking

Thousands of biomolecules with different chemical structures and functions have an impact on biological regulatory systems. These complex systems are often represented as networks, where the components interact with each other directly or indirectly through other components [13]. Examples of such networks include metabolic networks [77], protein-interaction networks [136] and transcriptional regulatory networks [11]. Despite the fact that networks are just hypothetical representations of regulatory systems, they enable the studying of the systems as whole. For example, human metabolism is a vital cellular process, that is determined by genome, environment, and diet. Understanding metabolic genotype-phenotype associations requires combining information from genome, protein and metabolite levels. A metabolic network can then be used in visualising the observed patterns, showing interactions between enzymes and metabolites, and test how removing one node effects the rest of the network. Computational tools are needed to model and understand these complex and dynamic life processes and to gain to deeper

understanding of a working cell. Networks provide understanding about the whole system instead of just reporting a list of individual parts. In statistical terms, differential expression analysis aims at identifying changes in first-order moments, means, whereas differential co-expression explores changes in second-order moments, covariances.

Interest in the study of networks has increased in pursuance of the advances in measuring technologies, bioinformatics and biostatistics, as well as high-throughput data available in public databases [147]. Genomic co-expression studies have been conducted already for over fifteen years [24, 139]. Despite the important findings achieved by differential expression studies, much of the information contained in omics data is ignored when genes or other components are considered individually. For example, known pathogen genes are often not differentially expressed in between diseased and healthy samples as mutations in the coding region of can affect the function of the gene without affecting its expression level [33]. Moreover, a variety of post-translational modifications can effect regulatory functions of a gene product regardless of its expression level. The ability to examine biological systems on a genome, proteome, and metabolome scale has revolutionised the study of diseases by allowing to consider the effect and associations of thousands of genes/lipids/proteins simultaneously. The identification of disease related biomolecules requires the studying of these molecules as a part of the regulatory systems, not as individual factors [82].

2 Extracting meaning from network structures

The concepts and properties from graph theory are useful for describing, inferring and visualising relationships between biomolecules as part of larger biological systems [72]. Formally, a network consists of collection of nodes, representing the components of the network, such as genes, proteins or lipids, and their interactions, given by a set of edges (Figure 1). Biological networks have often a modular structure where components belonging to different clusters have a weak or no connection between them, while within a cluster components are connected by short paths with strong connections. Components sharing similar structures or functions can be hypothesised to belong

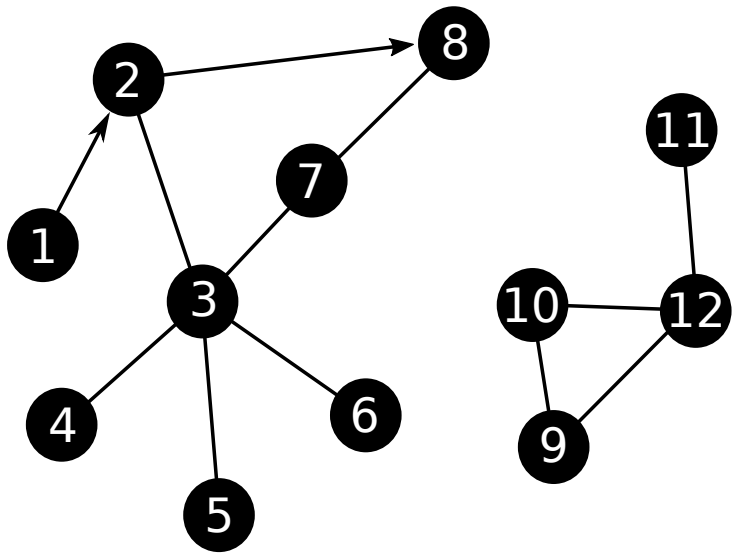

Figure 1: A hypothetical network consisting of twelve nodes (black circles) and twelve edges between them. Two of the edges are directed (black arrows). The network is consisted of two modules: the one on the left including nodes 1-8 and the one on the right including nodes 9-12. The node 3 is so-called hub node, that is densely connected to its surroundings.

to a same module [109]. In an unsupervised study, one goal of the network analysis is to identify such modular structures [60].

The edges of a network represent functional, causal or physical interactions between the nodes. Edges can be directed (arrows in Figure 1), indicating an effect heading from a source node to a target node, or undirected (straight lines in Figure 1), indicating symmetrical interaction. Here, between-component interaction (or association) describes the similarity of the expression or concentration levels of two components and how they change together. Edges can also be weighted to reflect the strength of an association.

One of the most general observations on biological networks is the tendency to form scale-free structures. A scale-free network has an overall sparse connectivity, where few nodes tend to be densely connected whereas the majority of the nodes have only a few connections [7]. Basic network topology properties, such as network hubs and

modules can reveal information about the biological significance of network components. Hub components are highly connected nodes that are often central to the network structure [13, 3] and are believed to be biologically important as they represent tightly regulated processes. In proteomics, it has been shown that hub proteins tend to be essential for survival in lower organisms, such as yeast [4, 78]. Some have argued that it is critical to rather focus on intramodular hubs instead of whole network hubs, at least in applications concerning co-expression networks [168]. In Publication [P3], one specific lipid belonging to class of ceramics had an important role in connecting different sphingolipid species together, thus it was recognised as an hub lipid. Earlier studies had identified the same lipid as a key metabolite for increased cardiovascular outcome risk.

3 Reconstruction of the association network

The goal of all network reconstruction and inference based on the network is the detection of the unobserved association structure of a given set of biological components based on the measured expression or concentration levels. Various strategies for recovering the underlying structure of the system have been proposed, and below, short reviews of the most common ones are provided. Often, the performance of the construction algorithms is assessed by simulations, as no complete biological interaction database exists to objectively evaluate the results obtained for the real data.

3.1 Correlation networks

The core of a network analysis is a defined connectivity score that represents the strength of the association or interaction between two components. At its simplest, the connectivity can be represented with a correlation coefficient. Thus, the most straightforward approach to network reconstruction is to estimate a sample covariance matrix using Pearson's correlation for continuous omics data and Spearman's correlation for count data. As a downside, correlation approach considers only direct pairwise association between all possible pairs of components. Such pairwise analyses of association are prone to Simpson's paradox [130] which can lead to inclusion of false associations. To avoid this, conventional estimates of correlation

could thus be replaced with partial correlations, that measures the association between two components, with the effect of a selected set of controlling components removed. However, a multiple comparison problem is encountered, when the inclusion or exclusion of each potential association is based on testing the significance of pairwise correlation coefficients. The inevitable correction for this problem decreases the statistical power to detect the true associations as the number of components in a network increases. The inclusion criteria for associations can be also done without testing, using so-called hard thresholding, referring to selection of a single threshold correlation value, such as 0.5, above which all associations are included in the network.

To avoid the multiple comparison problem and the coarseness of hard thresholding, [168] introduced weighted gene co-expression network analysis (WGCNA) that also uses a sample correlation matrix, but selects the associations of the network by soft thresholding. The absolute values of the pairwise correlations are raised to a power of β, where β is an integer, selected based on a scale-free topology criteria [168]. As β increases, smaller correlations, corresponding to "no interaction present" -cases, approach zero. Thus, pairwise associations corresponding to correlations approaching zero are excluded from the network. Various correlation-based co-expression network reconstruction have been widely used in the context of omics data [147, 170, 137, 165, 74, 82].

3.2 Bayesian networks

Bayesian framework can also be used to as a basis for generating networks [114] and have also been applied to gene expression data [49]. Bayesian network, or belief network, is a graphical representation of the joint probability distribution of the nodes that are considered as random variables. This structure is commonly referred to as directed acyclic graph (DAG), in which the lack of an edge between node 1 and node 2 within the graph denotes a conditional independence of node 1 from node 2 given the nodes that do have a connection to the node 1. Bayesian networks have the benefit of clear interpretation of the type of connection (dependence/independence) and can be flexibly applied to different stochastic observation models.

The construction of the Bayesian network is generally an optimisation problem, and considering all the possible connections between large number of nodes, the problem is computationally demanding. However, efficient search algorithms have been constructed and are widely available as ready implementations [97, 48].

3.3 Gaussian graphical models

Gaussian graphical models are a group of models using graphs to present dependencies among variables, while assuming that the variables follow a multivariate normal distribution with a particular structure of the inverse of the covariance matrix [167]. For Gaussian graphical models, it is usually assumed that the patterns of variation in expression or concentration for a given component can be predicted by those of a small subset of other components, leading to sparsity in the inverse covariance matrix.

Graphical lasso (glasso) is the most widely used Gaussian graphical model that estimates, using a lasso penalty applied to the inverse covariance matrix, a graph from a given data [47, 102]. The elements of the estimated inverse covariance matrix can be interpreted, similar to partial correlations, as pairwise measures of association in the presence of all other components. If two components are conditionally independent (conditional on all other components present), the corresponding element of the inverse covariance matrix is shrunk to zero.

In general, a challenging problem in sparse estimation of the graphical models is the selection of the regularisation parameter. Two commonly used criteria are highly efficient rotation information selection criterion [171] and a stability approach to regularization selection (stars) which selects the optimal graph by variability of subsamplings [92]. After selecting the optimal value for the regularisation parameter, the associations corresponding to the non-zero elements of the estimated inverse covariance matrix, are included in the network.

Graphical lasso, along other gaussian graphical models, has been used to model the sparse network structure especially in the context of genomics data [126, 128, 89, 116].

3.4 Model based approaches

Estimating the between-component association by a statistical model provides, not only a way to estimate the association in the presence of all other components, but enables also to adjusting of the associations for additional covariates, such as age, smoking status or use of statins. However, as noted in Chapter 2, the high-dimensionality sets some requirements on the model fitting on omics data. One convenient way to avoid the curse of dimensionality is to transform the variables to a smaller set of linear combinations of original variables. Using this property, [32] showed that in the context of microarray data PLS regression is a powerful tool for exploring relationships which also translate into biologically meaningful associations. Later, [117] proposed a more systematic approach to the PLS-based network reconstruction and showed that PLS based networks outperformed those reconstructed with simple correlations or partial correlations.

In practice, a chosen regression model is fitted for each variable as a response at a time, with the remaining $p-1$ variables and additional covariates as predictors. When using PLS-regression, first the PLS components are estimated, and then used predictors in a subsequent model fitting step. The score measuring the association strength between components j and k is then computed in a symmetrised form revising the roles of j and k as the average of the regression coefficients from the respective models. For PLS-regression, the association score between j and k is computed as a sum over a chosen number of PLS components of products of respective PLS component coefficient (from the first estimation step) and regression coefficient (from the second estimation step). If the models are adjusted for additional covariates, the regression coefficients relating to the additional covariates are not used in computing the connectivity score. Publication [P4] generalises the model PLS-approach for count data generated on NGS platforms, whereas Publication [P3] uses the approach suggested by [117] for multiple imputed lipidomic data.

Also, PCs could be used instead of PLS-components. Then, the network reconstruction for left-censored omics data could be based for a covariance matrix estimated as suggested in Publication [P2]. An attempting approach would also be the use a regularised regression

model, such as lasso or elastic net. Regularised regression would result in a sparse vector of estimated regression coefficients, and consequently a portion of the computed association scores would be zeros or close to zero and thus would result in a sparse network structure without additional decision making steps.

3.5 Edge selection

One challenge especially in model-based network reconstruction is establishing a threshold for each edge to be included in the network. Formally, we need to test multiple hypotheses

$$H_0 : s_{jk} = 0,\ H_1 : s_{jk} \neq 0, \quad j = 1,\ldots,p, \quad k = 1,\ldots,p, \quad j \neq k$$

where s_{jk} are the computed population association scores. The multiple hypothesis problem is even more severe than in the context of differential expression setting, as network includes $p(p-1)/2$ edges to test.

The local false discovery rate methodology [39], a continuum for Benjamini-Hockberg's FDR, operates edge selection with a minimum of modelling assumptions. Intuitively, tail area-based FDR is simply a P-value corrected for multiplicity, whereas local FDR is a corresponding probability value. In Efron's formulation, it is assumed that an association score $\hat{s}_{jk}$ comes from a mixture distribution

$$f(s) = p_0 f_0(s) + p_1 f_1(s),$$

where p_0 and p_1 are the mixing proportions, $f_0(s)$ is the density corresponding to the "no association present" -scores (null distribution), and $f_1(s)$ is the density corresponding to the "association present" -scores. The ratio $f_0(\hat{s}_{jk})/f(\hat{s}_{jk})$ represents an upper bound on the posterior probability of $\hat{s}_{jk}$ coming from the distribution $f_0(s)$ (no association between genes j and k). In practice, the mixture density $f(s)$ can be empirically estimated by fitting a smooth density curve to the histogram of observed $\hat{s}_{jk}$. As for $f_0(s)$, one can parametrically estimate the distribution by assuming normal distribution with the estimate of the expected value ave$(\hat{s}_{jk})$ and sample variance var$(\hat{s}_{jk})$. Thus, by evaluating the value of the ratio for each $\hat{s}_{jk}$, we get a likelihood for $\hat{s}_{jk}$ coming from the null distribution, and by comparing it to the selected value q (the maximum false discovery rate that we

are ready to accept), the network structure can be inferred. If $\hat{s}_{jk} < q$, an edge between genes j and k is added to the network.

In Publication [P4], the edges of the genomic networks were selected using local FDR. For the empirical local false discovery rate to perform satisfactorily, it is assumed that the mixing proportion p_0 is close to one (majority of the scores represent no association present). Also, the distribution of the observed scores should be approximately normal. For the PLS-based scores, used in Publications [P3] and [P4], the latter holds true while the sample size is relatively large. When these conditions are not met, filtering out some non-interesting or unreliable genes (such as genes with only little variation or genes including zero read counts) prior analysis can help to improve the behaviour of the associations scores and reduce the false positive rate. In Publication [P3], where the main interest was to compare network structures under two biological conditions, the modular structures of the networks (and thus inclusion or exclusion of the edges) were identified as a part of the statistical test using a connectivity threshold parameter ε.

Also, other modifications of FDR based inference have been suggested. In the contact of differential expression analysis, Genovese et al. [52] and more recently, in the context of genomic associaltion network construction, Gui et al. [56] assigned weights, representing the strength of existing evidence, for each p-value and then conducted the Benjamini-Hochberg FDR as usual. Futhermore, Li et al. [90] proposed a new method for estimating FDR for RNA sequencing data, based on a novel permutation plug-in approach.

4 Differential network analysis

Whereas differential expression describes a state where the mean expression levels of a given component is significantly different between two biological conditions (such as healthy and diseased states), differential co-expression (or differential network structure) denotes that the association between the expression levels of a given set of components is significantly different between two biological conditions. Comparing the structures of two or more networks provides insight into condition specific alternations in the regulatory

systems underlying the reconstructed association patterns. Exploring changes in connectivity patterns can help to identify condition-related nodes and thus potentially improve diagnosis and prognosis of condition outcome.

Kostka and Spang [85] proposed a first method to investigate differentially co-expressed groups of genes using a score based on an additive model followed by a heuristic algorithm to find high-scoring sets of genes displaying the characteristic pattern under examination. The proposed score measures the mutual correlation of a group of genes, but in contrast to standard correlation, one can explore how including or excluding a single candidate gene will affect the score without refitting the model. An ideal target set of genes would be such with low scores expressed in the samples belonging to group A, but not in samples belonging to group B.

Since then, many other approached have been suggested. Changes in the connectivity patterns can also be studied by comparing differences in topological properties across sparse, group-specific networks, summarised using network concepts, such as whole network connectivity, intramodular connectivity, topological overlap, and the clustering coefficient [169, 70, 38]. Alternatively, weighted condition-specific networks, for example such reconstructed by WGCNA, could be compared by computing a dissimilarity measure (function of the edge-specific weight differences) and then implementing a chosen clustering technique to identify modules [145]. This method groups together genes, whose correlations to the same sets of genes change between different conditions. Instead of working with edge-specific differential co-expression, one can also focus on exploring sets of genes and identifying which covariance patterns differ between conditions [121, 153].

An approach implemented in Publication [P3], with adjustments to fit the left-censored data, is the one proposed by Gill et al. [53]. Unlike previous approaches, they construct formal statistical tests on differential connectivities and modular structures based on the connectivity scores collected in the adjacency matrix. The formulated tests are based on permutations and thus provide a robust way to test differential network structures without any distributional assumptions. In their approach, three different interesting scenarios can be

tested: 1) whether the modular structures of two or more networks are different, 2) whether the connectivity of a single selected lipid is different between two or more networks, and 3) whether the connectivities of a selected interesting set of genes is different between two or more networks. While the tests are not tied to any specific type of association score, they are applicable on any omics data, with an association score best fitting to a data set analysed.

CHAPTER 4

Conclusions

The research conducted in this thesis contributes on developing statistical analysis methods for modern omics data, concentrating especially on lipidomics and NGS data. Two common themes are carried through the research papers: to efficiently incorporate missing or censored values in the analyses and to customise and develop analysis methods suitable for high-dimensional omics data measured either on continuous or count scale.

In addition to high dimensionality, one major challenge in performing valid inference on omics data is the presence of numerous missing and censored values. The research presented in this thesis solves the problem by relying both on MI of the censored values as well as on ML approach. Both approaches are operable when data includes left-censored values. Often the main criticism against MI of the left-censored values is that the censored values are not MCAR or MAR. However, the left-censoring mechanism is informative and can be implemented in the imputation model. Even though considered as the golden standard of censored data analysis methods, ML approaches can be computationally demanding. This was also noted in Publication [P2], where an approximation for the proposed ML estimator was derived. This approach succeeded in decreasing the computation times substantially and still produced accurate estim-

ates. Similar solutions were implemented in the literature, where ML function were approximated for example using pseudo-likelihoods.

Another statistically efficient and computationally fast alternative could be to regularise the sample covariance matrix or its inverse by making it sparse by setting some of the covariance elements to zero. This shrinkage estimation of the covariance matrix is closely related to lasso and ridge regressions. The shrinkage approach is proposed mainly for inferring large-scale covariance matrices in high-dimensions, such as clustering genes using data from a microarray experiment or building association networks. Combining shrinkage with ML estimation could provide an efficient estimator of the covariance matrix also in the presence of left-censored data. One could add a penalty term to the likelihood function and force some of the covariance elements to zero or close to zero producing interpretable estimates.

A major challenge of high-throughput omics data is to detect interactions from large-scale observations. By identifying co-varying components and significant relationships, it is possible to display conditional dependencies among the considered particles and discover the underlying network structure representing biological pathways. Many existing methods, especially in context of metabolomics and NGS data, were based on conventional correlation matrices. However, expressions of thousands of genes, for example, are often measured from relatively few samples. In small sample sizes, the sample correlation coefficient is a fairly unstable and inaccurate estimate of the true correlation. Thus, the association network reconstruction methods implemented both in Publication [P3] and [P4] were based on modelling the pairwise associations in the presence of all other variables. In this approach, the models can also be adjusted for additional covariates. As an alternative to the PLS regression implemented in Publications [P3] and [P4], the network construction could be employed using penalised regression models such as elastic net. This approach was only briefly tested in Publication [P3] and could be employed further in future studies.

Summaries of the Original Publications

P1 In this paper, our objective was to identify an efficient statistical methodology for the analysis of lipidomics data - especially in finding interpretable and predictive biomarkers useful for clinical practice. In two case studies, the need for data preprocessing was addressed prior to fitting a regression model of a binary response. The preprocessing steps include 1) a normalisation step, in order to remove experimental variability, and 2) a multiple imputation step, to make the full use of the incompletely observed data with potentially informative missingness. We then present and suggest the use of a permutation based test for a global test of association between full lipid profile and the outcome. Finally, by cross-validation, we compare stepwise variable selection to L_1- and L_2-penalised regression models on stacked multiple imputed data sets.

P2 The conventional estimators of a covariance matrix rely on complete data vectors on all subjects - an assumption that is rarely met. For example, data generated on mass spectrometry platforms is filtered to separate the actual signals from background noise. These left-censored values are considered too imprecise to be reported by a single number but known to exist somewhere between zero and a set lower limit of detection. We consider a maximum likelihood based estimator that handles the left-censored values without any pre-processing of the data. The presented estimator efficiently uses all the information available and thus, based on our simulation studies, produces

the least biased estimates compared to often used competing estimators. As the estimation problem can be solved fast only in low dimensions, we also suggest an element-wise approximation which then adjusted to nearest positive-definite matrix to meet the properties of a covariance matrix. It is shown by a simulation study that the suggested estimator and its approximation accomplish in decreasing the computation times substantially while still producing accurate estimates.

P3 Following the recent advances in mass spectrometry techniques for lipid profiling and increasing amounts of data available, the analysis of lipidomic networks, enabling studying the biological systems as a whole, has gained significant interest. Differential network analysis provides a formal statistical method for inferential analysis when the goal is to examine differences in lipid network structures under two biological conditions, to identify lipids and lipid classes that interact with each other, and to recognise the most important differentially expressed lipids between two subgroups. In this paper, we provided a recipe to combine differential network analysis with state of the art multiple imputation techniques, particularly paying attention to the left-censored values typical for a wide range of data sets in life sciences. As a case study, we analyse lipid profiles from two groups of coronary artery disease patients rom the LUdwigshafen RIsk and Cardiovascular Health study.

P4 A new algorithm to construct an association network for high-dimensional count data, called cPLS, is presented. The suggested approach is applicable to the raw counts data, without requiring any further pre- processing steps. The predictions for the associations are estimated using PLS regression model based approach. In the first step of the algorithm, the variation of the count data is compressed into a small number of PLS-components and in the second step, these components are used as predictors in a negative binomial model. In the settings investigated, the cPLS-algorithm outperformed the two comparative methods, glasso and WGCNA.

References

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409 (6822):860–921, 2001.

[2] T. Aittokallio. Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, 11(2):253–264, 2009.

[3] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[4] R. Albert, H. Jeong, and A. L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, Taylor & Francis Group, 2008.

[6] D. Albrecht, O. Kniemeyer, A. A. Brakhage, and R. Guthke. Missing values in gell based proteomics. *Proteomics*, 10(6):1202–1211, 2010.

[7] A. B. and. Scale-free networks: A decade and beyond. *Science*, 325:412–413, 2009.

[8] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11, 2010.

[9] E. G. Armitage, J. Godzien, V. Alonso-Herranz, A. Lopez-Gonzalvez, and C. Barbas. Missing value imputation strategies

for metabolomics data. *Electrophoresis*, 36(24):3050–3060, 2015.

[10] P. Auer and R. W. Doerge. Statistical design and analysis of rna sequencing data. *Genetics*, 185:405–416, 2010.

[11] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, 14 (3):283–291, 2004.

[12] A. Baccarelli, R. Pfeiffer, D. Consonni, A. C. Pesatori, M. Bonzini, D. G. J. Patterson, P. A. Bertazzi, and M. T. Landi. Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the seveso chloracne study. *Chemosphere*, 60(7):898–906, 2005.

[13] A. L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5: 101–115, 2004.

[14] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society B*, 57(1):289–300, 1995.

[15] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, and M. R. Pratt. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

[16] S. Bijlsma, I. Bobeldijk, E. R. Verheij, R. Ramaker, S. Kochhar, I. A. Macdonald, B. van Ommen, and A. K. Smilde. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry*, 78(2):567–574, 2006.

[17] Biomarkers Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69:89–95, 2001.

[18] S. J. Blanksby and T. W. Mitchell. Advances in mass spectrometry for lipidomics. *Annual Review of Analytical Chemistry*, 3:433–465, 2010.

[19] Y.-C. Bor, J. Swartz, Y. Li, J. Coyle, D. Rekosh, and M.-L. Hammarskjöld. Northern blot analysis of mrna from mammalian polyribosomes. *Nature Protocols*, 2006.

[20] A.-L. Boulesteix and K. Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44, 2006.

[21] A. Brevern, S. Hazout, and A. Malpertuy. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, 5(114), 2004.

[22] R. W. Browne and B. W. Whitcomb. Procedures for determination of detection limits. application to high-performance liquid chromatography analysis of fat-soluble vitamins in human serum. *Epidemiology*, 21(4):S4–S9, 2010.

[23] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11 (94), 2010.

[24] A. Butte, P. Tamayo, D. Slonim, T. Golub, and I. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97:12182–12186, 2000.

[25] J. R. Carpenter and M. G. Kenward. *Multiple Imputation and its Application*. Wiley, 2013.

[26] H. Chen, S. A. Quandt, J. G. Grzywacz, and T. A. Arcury. A distibution-based multiple imputation method for handling bivariate pesticide data with values below the limit of detection. *Environmental Health Perspectives*, 119:351–356, 2011.

[27] Q. Chen and . S. Wang. Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, 32:3646–3659, 2013.

[28] M. Clamp, B. Fry, M. Kamal, X. Xie, J. Cuff, M. F. Lin, M. Kellis, K. Lindblad-Toh, and E. S. Lander. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 104:19428–19433, 2007.

[29] F. H. C. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, XII:139–163, 1956.

[30] F. H. C. Crick. Central dogma of molecular biology. *Nature*, 227 (5258):561–3, 1970.

[31] X. Cui and G. A. Churchill. Statistical tests for differential expression in cdna microarray experiments. *Genome Biology*, 4 (210), 2003.

[32] S. Datta. Exploring relationships in gene expressions: a partial least squares approach. *Gene Expression*, 9(6):249–255, 2001.

[33] A. de la Fuente. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26:326–333, 2010.

[34] E. A. Dennis. Lipidomics joins the omics evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 106(7):2089–2090, 2009.

[35] K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26(1):51–78, 2007.

[36] Y. Di, D. W. Schafer, J. S. Cumbie, and J. H. Chang. The nbp negative binomial model for assessing differential gene expression from rna-seq. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.

[37] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keimea, G. Marot, D. Castel, J. Estelle,

G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. L. Gall, B. Schaeffer, S. L. Crom, M. Guedj, and F. Jaffrézic. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 14 (6):671–683, 2013.

[38] J. Dong and S. Horvath. Understanding network concepts in modules. *BMC Systems Biology*, 1(24), 2007.

[39] B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99:96–104, 2004.

[40] A. H. El-Shaarawi and S. R. Esterby. Replacement of censored observations by a constant: an evaluation. *Water Research*, 26 (6):835–844, 1992.

[41] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23 (22):5866–5878, 2014.

[42] E. Fahy, S. Subramaniam, R. Murphy, M. Nishijima, C. Raetz, T. Shimizu, F. Spener, G. van Meer, M. Wakelam, and E. A. Dennis. Update of the lipid maps comprehensive classification system for lipids. *Journal of lipid research*, 50:S9–S14, 2009.

[43] I. M. Farnham, A. K. Singh, K. J. Stetzenbach, and K. H. Johannesson. Treatment of nondetects in multivariate analysis of groundwater geochemistry data. *Chemometrics and Intelligent Laboratory Systems*, 60:265–281, 2002.

[44] F. Finotello and B. D. Camillo. Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, 14(2):130–142, 2015.

[45] R. A. Fisher. *The design of experiments*. Oliver and Boyd Ltd, 1951.

[46] W. M. Freeman, S. J. Walker, and K. E. Vrana. Quantitative rt-pcr: pitfalls and potential. *Biotechniques*, 26:112–125, 1999.

[47] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9: 432–441, 2008.

[48] N. Friedman, I. Nachman, and D. Pe'er. Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm. *Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*, pages 196–205, 1999.

[49] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

[50] P. J. Garcia-Laencia, J. Sancho-Gomez, and A. R. Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computational Applications*, 19:263–282, 2009.

[51] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A.-C. Gavin. Visualization of omics data for systems biology. *Nature Methods*, 7(30):S56–S68, 2010.

[52] C. R. Genovese, K. Roeder, and L. Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[53] R. Gill, S. Datta, and S. Datta. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11(95), 2010.

[54] F. D. Girolamo, I. Lante, M. Muraca, and L. Putignani. The role of mass spectrometry in the "omics" era. *Current organic chemistry*, 17(23):2891–2905, 2013.

[55] J. Godzien, V. Alonso-Herranz, C. Barbas, and E. G. Armitage. Controlling the quality of metabolomics data: new strategies to get the best out of the qc sample. *Metabolomics*, 11:518–528, 2015.

[56] J. Gui, C. S. Greene, C. Sullivan, W. Taylor, J. H. Moore, and C. Kim. Testing multiple hypotheses through imp weighted

fdr based on a genetic functional network with application to a new zebrafish transcriptome study. *BioData Mining*, 8(17), 2015.

[57] X. Han. *Lipidomics: Comprehensive Mass Spectrometry of Lipids*. John Wiley & Sons Inc, Hoboken, New Jersey, 2016.

[58] X. Han and R. W. Gross. Shotgun lipidomics: Electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples. *Mass spectrometry reviews*, 24(3):367–412, 2005.

[59] F. E. Harrell. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis,*. Springer-Verlag, New York, 2001.

[60] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:47–52, 1999.

[61] M. J. Heller. Dna microarray technology: Devices, systems, and applications. *Annual Review of Biomedical Engineering*, 4 (129-153), 2002.

[62] D. R. Helsel. Less than obvious: Statistical treatment of data below the reporting limit. *Environmental Science and Technology*, 24(12):1766–1774, 1990.

[63] D. R. Helsel. *Nondetects and data analysis*. John Wiley & Sons Inc, 2005.

[64] D. R. Helsel. Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*, 65:2434–2439, 2006.

[65] P. Hewett and G. H. Ganser. A comparison of several methods for analysing censored data. *Annals of Occupational Hygiene*, 51(7):611–632, 2007.

[66] H. Hinterwirth, C. Stegemann, and M. Mayr. Quest for molecular lipid biomarkers in cardiovascular disease. *Circulation: Cardiovascular Genetics*, 7:941–954, 2014.

[67] R. R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[68] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[69] P. K. Hopke, C. Liu, and D. B. Rubin. Multiple imputation for multivariate data with missing and below-threshold mmeasurement: Time series concentrations of pollutants in the arctic. *Biometrics*, 57:22–33, 2001.

[70] S. Horvath and J. Dong. Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology*, 4 (8), 2008.

[71] O. Hrydziuszko and M. R. Viant. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, 8:S161–S174, 2012.

[72] W. Huber, V. J. Carey, L. Long, S. Falcon, and R. Gentleman. Graphs in molecular biology. *Bioinformatics*, 8:S8, 2007.

[73] T. Huybrechts, O. Thas, J. Dewulf, and H. V. Langenhov. How to estimate moments and quantiles of environmental data sets with nondetected observations? a case study on volatile organic compounds in marine water samples. *Journal of Chromatography A*, 975(1):123–133, 2002.

[74] O. D. Iancu, S. Kawane, D. Bottomly, R. Searles, R. Hitzemann, and S. McWeeney. Utilizing rna-seq data for de novo coexpression network inference. *Bioinformatics*, 28(12), 2012.

[75] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, 11:1957–2000, 2010.

[76] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, 2004.

[77] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[78] H. Jeong, S. Mason, A. Barabasi, and Z. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41, 2001.

[79] I. Jolliffe. *Principal component analysis*. John Wiley & Sons, 2002.

[80] R. Jornsten, H. Wang, W. Welsh, and M. Ouyang. Dna microarray data imputation and significance analysis of differential expression. *Bioinformatics*, 21:4155–4161, 2005.

[81] M. B. Khalil, W. Hou, H. Zhou, F. Elisma, L. A. Swayne, A. P. Blanchard, Z. Yao, S. A. Bennett, and D. Figeys. Lipidomics era: acomplishments and challenges. *Mass spectrometry reviews*, 29:877–929, 2010.

[82] L. J. A. Kogelman, S. Cirera, D. V. Zhernakova, M. Fredholm, L. Franke, and H. N. Kadarmideen. Identification of co-expression gene networks, regulatory ggene and pathways for obesity based adipose tissue rna sequencing in a porcine model. *BMC Medical Genomics*, 7(57), 2014.

[83] J. W. Koo, F. Parham, M. C. Kohn, S. A. Masten, J. W. Brock, L. L. Needham, and C. J. Portier. The association between biomarker-based exposure estimates for phthalates and demographic factors in a human reference population. *Environmental Health Perspectives*, 110(4):405–410, 2002.

[84] D. E. Koshland. The seven pillars of life. *Science*, 295(5563): 2215–2216, 2002.

[85] D. Kostka and R. Spang. Finding disease specific alternations in the co-expression of genes. *Bioinformatics*, 20(Suppl. 1): i194–199, 2004.

[86] R. Laaksonen and K. Ekroos. Lipidomics: a new asset to the clinical lipidology tool-kit. *Clinical lipidology*, 6(1):21–23, 2011.

[87] M. Lappas, P. A. Mundra, G. Wong, K. Huynh, D. Jinks, H. M. Georgioua, M. Permezela, and P. J. Meikle. The prediction of type 2 diabetes in women with previous gestational diabetes mellitus using lipidomics. *Diabetologia*, 58(7):1436–1442, 2015.

[88] M. Lee, L. Kong, and l Weissfeld. Multiple imputation for left-censored biomarker data based on gibbs sampling method. *Statistics in Medicine*, 31(17):1838–1848, 2012.

[89] H. Li and J. Gui. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2006.

[90] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

[91] M. Li, L. Yang, Y. Bai, and H. Liu. Analytical methods in lipidomics and their applications. *Analytical Chemistry*, 86(1):161–175, 2014.

[92] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.

[93] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405:827–836, 2000.

[94] R. H. Lyles, J. K. Williams, and R. Chuachoowong. Correlating two viral load assays with known detection limits. *Biometrics*, 57:1238–1244, 2001.

[95] H. S. Lynn. Maximum likelihood inference for left-censored hiv rna data. *Statistics in Medicine*, 20:33–45, 2001.

[96] E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008.

[97] D. Margaritis. *Learning Bayesian Network Model Structure from Data*. PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 2003.

[98] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanzaand, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[99] D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012.

[100] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62–69, 2012.

[101] P. J. Meikle, G. Wong, C. K. Barlow, and B. A. Kingwell. Lipidomics: Potential role in risk prediction and therapeutic monitoring for diabetes and cardiovascular disease. *Pharmacology and Therapeutics*, 143:12–23, 2014.

[102] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

[103] H. K. Min, S. Lim, B. C. Chung, and M. H. Moon. Shotgun lipidomics for candidate biomarkers of urinary phospholipids in prostate cancer. *Analytical and bioanalytical chemistry*, 399 (2):823–830, 2011.

[104] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5:621–628, 2008.

[105] O. G. Mouritsen. *Life - as a matter of fat. The emerging science of lipidomics*. Heidelberg Germany: Springer-Verlag, 2005.

[106] P. S. Niemelä, S. Castillo, M. Sysi-aho, and M. Oresic. Bioinformatics and computational methods for lipidomics. *Journal of Chromatography*, 877:2855–2862, 2009.

[107] T. Nolan, R. E. Hands, and S. A. Bustin. Quantification of mrna using real-time rt-pcr. *Nature Protocols*, 1(3):1559–1582, 2006.

[108] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

[109] S. Oliver. Guilt-by-association goes global. *Nature*, 403:601–603, 2000.

[110] G. Ommen, C. D. DeAngelis, D. L. DeMets, T. R. Fleming, G. Geller, J. Gray, D. F. Hayes, I. C. Henderson, L. Kessler, S. Lapidus, D. Leonard, H. L. Moses, W. Pao, R. D. Pentz, N. D. Price, J. Quackenbush, E. Railey, D. Ransohoff, E. A. Reece, and D. M. Witten. Evolution of translational omics: lessons learned and the path forward. *Institute of Medicine National Academies of Sciences Press, Washington DC*, 2012.

[111] A. Oshlack and M. J. Wakefield. Transcript length bias in rna-seq data confounds systems biology. *Biology Direct*, 4(14), 2009.

[112] A. Oshlack, M. D. Robinson, and M. D. Young. From rna-seq reads to differential expression results. *Genome Biology*, 11 (220), 2010.

[113] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004.

[114] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, California, 1988.

[115] R. Pedreschi, M. Hertog, S. Carpentier, J. Lammertyn, J. Robben, J. P. Noben, B. Panis, R. Swennen, and B. M. Nicolai. Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. *Proteomics*, 8:1371–1383, 2008.

[116] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of American Statistical Association*, 104(486):735–746, 2009.

[117] V. Pihur, S. Datta, and S. Datta. Reconstruction of genetic association networks from microarray data: a partial least squares approach. *Bioinformatics*, 24(4):561–568, 2008.

[118] P. Proitsi, M. Kim, L. Whiley, M. Pritchard, R. Leung, H. Soininen, I. Kloszewska, P. Mecocci, M. Tsolaki, B. Vellas, P. Sham, S. Lovestone, J. F. Powell, R. J. B. Dobson, and C. Legido-Quigley. Plasma lipidomics analysis finds long chain cholesteryl esters to be associated with alzheimer's disease. *Translational Psychiatry*, 5(1), 2015.

[119] O. Quehenberger, A. M. Armando, A. H. Brown, S. B. Milne, D. S. Myers, A. H. Merrill, S. Bandyopadhyay, K. N. Jones, S. Kelly, R. L. Shaner, C. M. Sullards, E. Wang, R. C. Murphy, R. M. Barkley, T. J. Leiker, C. R. Raetz, Z. Guan, G. M. Laird, D. A. Six, D. W. Russell, J. G. McDonald, S. Subramaniam, E. Fahy, and E. A. Dennis. Lipidomics reveals a remarkable diversity of lipids in human plasma. *Journal of lipid research*, 51(11):3299–3305, 2010.

[120] T. E. Raghunathan, J. M. Lepkowksi, J. V. Hoewyk, and P. Solenbeger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27:85–95, 2001.

[121] Y. Rahmatallah, F. Emmert-Streib, and G. Glazko. Gene sets net correlations analysis (gsnca): a multivariate differential coexpression test for gene sets. *Bioinformatics*, 30(3):360–368, 2013.

[122] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[123] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[124] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[125] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York USA: John Wiley and Sons, 2004.

[126] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:article 32, 2005.

[127] I. Scheel, M. Aldrin, I. Glad, R. Sorum, H. Lyng, and A. Frigessi. The influence of missing values imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, 21:4272–4279, 2005.

[128] E. Segal, N. Friedman, N. Kaminski, A. Regev, and D. Koller. From signatures to models: understanding cancer using microarrays. *Nature Genetetics*, 37:S38–45, 2005.

[129] A. Shevchenko and K. Simons. Lipidomics: coming to grips with lipid diversity. *Nature Reviews Molecular Cell Biology*, 11: 593–598, 2010.

[130] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society B*, 13:238–241, 1951.

[131] C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14(91), 2013.

[132] L. J. Sparvero, A. A. Amoscato, P. M. Kochanek, B. R. Pitt, V. E. Kagan, and H. Bayir. Mass-spectrometry based oxidative lipidomics and lipid imaging: applications in traumatic brain injury. *Journal of Neurochemistry*, 115:1322–1336, 2010.

[133] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig. pcamethods - a bioconductor package providing pca methods for incomplete data. *Bioinformatics Applications note*, 23(9): 1164–1167, 2007.

[134] C. Stegemann, I. Drozdov, J. Shalhoub, J. Humphries, C. Ladroue, A. Didangelos, M. Baumert, M. Allen, A. H. Davies, C. Monaco, A. Smith, Q. Xu, and M. Mayr. Comparative lipidomics profiling of human atherosclerotic plaques. *Circulation*, 4: 232–242, 2011.

[135] C. Stegemann, R. Pechlaner, P. Willeit, S. Langley, M. Mangino, U. Mayr, C. Menni, A. Moayyeri, P. Santer, G. Rungger, T. D. Spector, J. Willeit, S. Kiechl, and M. Mayr. Lipidomics profiling and risk of cardiovascular disease in the prospective population-based bruneck study. *Circulation*, 129(18):1821–1831, 2014.

[136] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[137] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19:1019–1026, 2003.

[138] R. Steuer, K. Morgenthal, W. Weckwerth, and J. Selbig. A gentle guide to the analysis of metabolomic data. *Methods in Molecular Biology*, 358:105–126, 2007.

[139] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255, 2003.

[140] S. Subramaniam, E. Fahy, S. Gupta, M. Sud, R. W. Byrnes, D. Cotter, A. R. Dinasarapu, and M. R. Maurya. Bioinformatics and systems biology of the lipidome. *Chemical Reviews*, 111(10): 6452–6490, 2011.

[141] P. A. Succop, S. Clark, M. Chen, and W. Galke. Imputation of data values that less than a detection limit. *Journal of Occupational and Environmental Hygiene*, 1(7):436–441, 2004.

[142] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, J. A. H. Merril, R. C. Murphy, C. R. H. Raetz, D. W. Russel, and S. Subramaniam. Lmsd: Lipid maps structure database. *Nucleic Acids Research*, 35:D527–D532, 2007.

[143] J. Sung, Y. Wang, S. Chandrasekaran, D. M. Witten, and n D Price. Molecular signatures from omics data: From chaos to consensus. *Biotechnology Journal*, 7:946–957, 2012.

[144] F. C. Taylor, M. Huffman, and S. Ebrahim. Statin therapy for primary prevention of cardiovascular disease. *The Journal of the American Medical Association*, 310(22):2451–2452, 2013.

[145] B. M. Tesson, R. Breitling, and R. C. Jansen. (2010) diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC bioinformatics*, 11(497), 2010.

[146] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.

[147] D. Toubiana, A. R. Fernie, Z. Nikoloski, and A. Fait. Network analysis: tackling complex data to study plant metabolism. *Trends in Biotechnology*, 31(1):29–36, 2013.

[148] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17: 520–525, 2001.

[149] J. Tuikkala, L. L. Elo, O. S. Nevalainen, and T. Aittokallio. Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics*, 9(202), 2008.

[150] S. VanBuuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242, 2007.

[151] Y. Wan, S. Datta, D. J. Conklin, and M. Kong. Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation*, 85(9):1902–1916, 2015.

[152] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[153] M. Watson. Coxpress: differential co-expression in gene expression data. *BMC bioinformatics*, 7(509), 2006.

[154] B.-J. M. Webb-Robertson, M. M. Matzke, T. O. Metz, J. E. McDermott, H. Walker, K. D. Rodland, J. G. Pounds, and K. M. Waters. Sequential projection pursuit principal component analysis - dealing with missing data associated with new -omics technologies. *Biotechniques*, 54:165–168, 2013.

[155] J. N. Weinstein. Fishing expeditions. *Science*, 282(5389):627, 1998.

[156] M. R. Wenk. The emerging field of lipidomics. *Nature Reviews, Drug Discovery*, 4:594–610, 2005.

[157] M. R. Wenk. Lipidomics: New tools and applications. *Cell*, 143: 888–895, 2010. doi: 10.1016/j.cell.2010.11.033.

[158] M. S. Williams and E. D. Ebel. Estimating the correlation between concentrations of two species of bacteria with censored microbial testing data. *International Journal of Food Microbiology*, 175:1–5, 2014.

[159] D. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3): 515–534, 2009.

[160] A. M. Wood, I. R. White, and P. Royston. How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27:3227–3246, 2008.

[161] P. L. Wood. Lipidomics of alzheimer's disease: current status. *Alzheimer's Research and Therapy*, 4, 2012.

[162] L. Yang, X. Cui, N. Zhang, M. Li, Y. Bai, X. Han, Y. Shi, and H. Liu. Comprehensive lipid profiling of plasma in patients with benign breast tumor and breast cancer reveals novel biomarkers. *Analytical and bioanalytical chemistry*, 407:5065–5077, 2015.

[163] L. Yang, M. Li, Y. Shan, S. Shen, Y. Bai, and H. Liu. Recent advances in lipidomics for disease research. *Journal of separation science*, 39:38–50, 2016.

[164] J. R. Yates, C. I. Ruse, and A. Nakorchevsky. Proteomics by mass spectrometry: Approaches, advances, and applications. *Annual Review of Biomedical Engineering*, 11:49–79, 2009.

[165] L. Yetukuri, M. Katajamaa, G. Medina-Gomez, T. Seppänen-Laakso, A. Vidal-Puig, and M. Oresic. Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. *BMC Systems Biology*, 1(12), 2007.

[166] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68:49–67, 2006.

[167] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[168] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[169] B. Zhang, H. Li, R. B. Riggins, M. Zhan, J. Xuan, Z. Zhang, E. P. Hoffman, R. Clarke, and Y. Wang. Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics*, 25:526–532, 2009.

[170] J. Zhang, Y. Xiang, L. Ding, K. Keen-Circle, T. B. Borlawsky, H. G. Ozer, R. Jin, P. Payne, and K. Huang. Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. *Bioinformatics*, 11(Suppl 9):S5, 2010.

[171] T. Zhao and H. Liu. The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13:1059–1062, 2012.

[172] Y. Zhao and H. C. Frey. Uncertainty for data with non-detects: Air toxic emissions from combustion. *Human and Ecological Risk Assessment: An international Journal*, 12(6):1171–1191, 2006.

[173] X. Zhou, J. Mao, J. Ai, Y. Deng, M. R. Roth, C. Pound, J. Henegar, R. Welti, and S. A. Bigler. Identification of plasma lipid biomarkers for prostate cancer by lipidomics and bioinformatics. *PLoS One*, 7(11), 2012.

[174] H. Zou and T. Hastie. Regression shrinkage and selection via the elastic net, with applications to microarrays. *Stanford University, Department of Statistics. Technical Report*, 2003.

[175] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):262–286, 2006.