DISSERTATIONES
MATHEMATICAE
UNIVERSITATIS
TARTUENSIS
103

TAUNO METSALU

Statistical analysis of multivariate data in bioinformatics

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
103

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
103

TAUNO METSALU

Statistical analysis of multivariate data in bioinformatics

UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (PhD) on January 19th, 2016 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor:

Prof. PhD. Jaak Vilo
University of Tartu
Tartu, Estonia

Opponent:

Dr. Simon Anders
Institute for Molecular Medicine Finland
Helsinki, Finland

The public defense will take place on March 4th, 2016 at 16:15 in Liivi 2-405.

The publication of this dissertation was financed by Institute of Computer Science, University of Tartu.

ISSN 1024-4212
ISBN 978-9949-77-043-4 (print)
ISBN 978-9949-77-044-1 (pdf)

University of Tartu
http://www.tyk.ee

Contents

List of publications **7**

Abstract **9**

Introduction **10**

1 Preliminaries **11**
1.1 Biological background 11
1.2 Statistical background 13

2 Sources of multivariate data **16**
2.1 Gene expression microarray 17
2.2 Tissue microarray 18
2.3 SNP array 20
2.4 RNA sequencing 20
2.5 Reverse transcription quantitative PCR 22
2.6 Summary 23

3 Analysis methods **25**
3.1 Exploratory analysis methods 26
3.1.1 Principal component analysis 26
3.1.2 Clustering 28
3.2 Confirmatory analysis methods 31
3.2.1 Differential expression 31
3.2.2 Gene set analysis 33
3.2.3 Support vector machine 34
3.2.4 Binomial test 35
3.3 Summary 35

4 Comparison of cancer models **36**
4.1 Overview of PREDECT project 36

4.2 Centralized data collection and analysis 38
4.3 Improved breast cancer xenograft model (paper I) 40
4.4 Tissue slices in different cultivation conditions (paper II) 42
4.5 ClustVis web tool for matrix visualization (paper III) 42

5 Imprinted and monoallelically expressed genes in the human placenta (paper IV) 45

6 Molecular mechanisms of atopic dermatitis (paper V) 48

7 MicroRNAs as diagnostic markers for endometriosis (paper VI) 49

Conclusions 50

Bibliography 51

Acknowledgements 59

Kokkuvõte (Summary in Estonian) 60

Publications 63

Curriculum vitae 187

PUBLICATIONS INCLUDED IN THIS THESIS

I G. Sflomos, V. Dormoy, T. Metsalu, R. Jeitziner, L. Battista, A. Treboux, M. Fiche, J. Delaloye, J. Vilo, A. Ayyanan, and C. Brisken. A Novel Preclinical Model for ERα Positive Breast Cancer Points to the Mammary Epithelial Microenvironment as a Critical Determinant of Luminal Phenotype and Hormone Response. Accepted for publication in Cancer Cell.

II E. J. Davies, M. Dong, M. Gutekunst, K. Närhi, H. J. A. A. van Zoggel, S. Blom, A. Nagaraj, T. Metsalu, E. Oswald, S. Erkens-Schulze, J. A. D. S. Martin, R. Turkki, S. R. Wedge, T. M. af Hällström, J. Schueler, W. M. van Weerden, E. W. Verschuren, S. T. Barry, H. van der Kuip, and J. A. Hickman. Capturing complex tumour biology in vitro: histological and molecular characterisation of precision cut slices. *Scientific Reports*, 5, 2015.

III T. Metsalu and J. Vilo. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, 43(W1):W566–W570, 2015.

IV T. Metsalu, T. Viltrop, A. Tiirats, B. Rajashekar, E. Reimann, S. Kõks, K. Rull, L. Milani, G. Acharya, P. Basnet, J. Vilo, R. Mägi, A. Metspalu, M. Peters, K. Haller-Kikkatalo, and A. Salumets. Using RNA sequencing for identifying gene imprinting and random monoallelic expression in human placenta. *Epigenetics*, 9(10):1397–1409, 2014.

V A. Rebane, M. Zimmermann, A. Aab, H. Baurecht, A. Koreck, M. Karelson, K. Abram, T. Metsalu, M. Pihlap, N. Meyer, R. Fölster-Holst, N. Nagy, L. Kemeny, K. Kingo, J. Vilo, T. Illig, M. Akdis, A. Franke, N. Novak, S. Weidinger, and C. A. Akdis. Mechanisms of IFN-γ-induced apoptosis of human skin keratinocytes in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology*, 129(5):1297–1306, 2012.

VI M. Saare, K. Rekker, T. Laisk-Podar, D. Sõritsa, A. M. Roost, J. Simm, A. Velthut-Meikas, K. Samuel, T. Metsalu, H. Karro, A. Sõritsa, A. Salumets, and M. Peters. High-Throughput Sequencing Approach Uncovers the miRNome of Peritoneal Endometriotic Lesions and Adjacent Healthy Tissues. *PLOS ONE*, 9(11):e112630, 2014.

My contribution in these articles was as follows:

I I analyzed gene expression data, helped to interpret the results and participated in writing the article.

II I did differential expression and principal component analysis, helped to interpret PCA plots and participated in writing the article.

III I created the web tool and wrote the article.

IV I implemented the pipeline for detecting ASE, made the plots about genes found, helped to interpret the results and participated in writing the article.

V I reanalyzed public gene expression data and wrote materials and methods for this part.

VI I created support vector machine classifier for clinical score formula and wrote methods part about it.

ABSTRACT

Proteins are one of the most important molecules of an organism. Their structure is coded in the DNA. By investigating the abundance of different proteins, it is possible to get information about the current state of the organism. Modern technologies allow to collect a large amount of data related to proteins in a short period of time. Analyzing high-throughput data needs different technical skills and has created a new field of science—bioinformatics.

The aim of the dissertation is to describe problems and solutions related to statistical analysis of multivariate data. It is shown that this type of data can be presented as a matrix. A pan-European consortium is described where data is collected from many partners, and it is also important to gather metadata in a structured way. Different studies are described where data analysis was needed. Web tools with a graphical user interface were created to reduce the amount of technical skills required for the analysis and make some types of analysis available to the people who do not necessarily have experience with data analysis.

Both biological and statistical background are introduced. An overview is given about different sources of multivariate data. Analysis methods are described which are divided into *exploratory* methods that are useful to discover general patterns, and *confirmatory* methods that aim to answer a particular research question.

It is shown how sources of data and analysis methods are used in practice. A pan-European project PREDECT is described. An overview is given about collecting metadata from multiple partners, and about web tools created for initial data analysis. An analysis concerning a novel breast cancer model is described, and a comparison of tissue slices in different cultivation conditions is made. A freely available web tool is introduced which allows to perform exploratory data analysis.

Next chapters describe data analysis in various projects. Multiple novel genes were found in the human placenta that have an allele-specific expression. Molecular mechanisms of atopic dermatitis are examined, more specifically the influence of the protein *IFN-*γ. MicroRNAs are found that can be used as markers for endometriosis, and a classifier is built to differentiate people with endometriosis from healthy people.

INTRODUCTION

The amount of data collected is larger than ever before. Many services have become electronic, needing well-structured databases to be established. The cost of gathering data from some biological experiments has even decreased faster than the cost of storing the data, driving the need for faster algorithms and better storage solutions. These trends, colloquially called information revolution, have also created the need for people who can analyze large amounts of data.

In this thesis, we describe the analysis of data originating from molecular biology, called bioinformatics. Most of these data sources can be converted into the multivariate form, presentable as a numeric table. We give basic knowledge about essential biological and statistical concepts and an overview of the data sources, ranging from different platforms for measuring gene expression to detecting genome variants and measuring protein expression. We introduce different analysis methods, covering both exploratory analysis methods that are used without a clear hypothesis in mind, and confirmatory analysis methods that give a score to each specific hypothesis.

In the following chapters, we give an overview of the articles included in the thesis. Three papers are covering PREDECT project about developing novel biological models for three frequent human cancers. The first article proposes a novel model for breast cancer, and the second article compares different tissue slice models. The third article describes a web tool for visualizing user-uploaded data using principal component analysis and heatmap.

Three remaining papers describe three topics regarding applied data analysis. First of them searches for novel imprinted genes, the second one describes molecular mechanisms of a disease called atopic dermatitis, and the last article describes how microRNA expression can be used to diagnose endometriosis.

The thesis has multiple goals.

- Describe data sources and analysis methods in bioinformatics. It is shown that any data source under consideration can be converted to a numeric matrix;
- Show concrete cases how these methods were applied to real datasets;
- Describe web tools that were developed for analyzing the data.

CHAPTER 1

PRELIMINARIES

1.1 Biological background

Hereditary information of living organisms is stored in the genome that consists of a long double-stranded molecule called *deoxyribonucleic acid* (DNA). Human DNA consists of 46 chromosomes. There are 22 different types of *autosomes*, all of them present as two copies and marked with sequential numbers 1, 2, . . . , 22. Two remaining chromosomes are called *allosomes* (or sex chromosomes) and are marked with X and Y. Males have one X and Y chromosome whereas females have two X chromosomes. Both mother and father give one chromosome from each pair to the offspring.

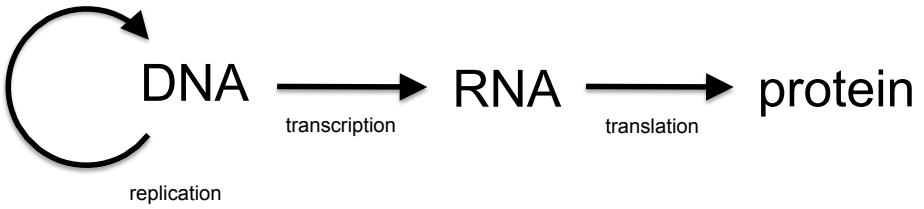

Figure 1.1: The main principles of transmitting hereditary information.

Information is coded in the DNA by four nucleic acids: adenine (A), cytosine (C), guanine (G) and thymine (T). The two strands are *complementary*: A is always paired with T and C with G. Therefore, only one strand of each chromosome is sufficient to determine the genome. The other strand can be synthesized using complementarity. This principle is the basis for the process called *DNA replication* where two identical copies of the initial DNA are created (see Figure 1.1). It starts with unwinding the strands and then synthesizing complementary sequence

for both strands.

A portion of the genome has known to have a highly important biological meaning. *Gene* is a region (or locus) of the genome that codes a functional biological object. Different variants of the same gene or part of the gene are called *alleles*. Gene products are produced during a synthesis process called *gene expression*. During *transcription*, the information from DNA is transferred into *messenger ribonucleic acid* (mRNA, see Figure 1.1).

The initial RNA (pre-mRNA) contains both coding regions–*exons*–and non-coding regions–*introns*. During *RNA splicing*, introns are cut out and mature mRNA is formed. Transcription is often followed by *translation* where mature mRNA is used for coding a *protein*. During translation, each nucleotide triplet encodes one amino acid—a building block for the protein.

Proteins are the most important gene products. They perform many of the functions within living organisms, including catalyzing chemical reactions, cell signaling, forming a solid structure and other. Another type of gene product is called a *microRNA* (miRNA, see Bartel (2004) for a review). These are small RNA molecules that act as post-transcriptional regulators of gene expression and are not translated into protein.

Genes are often categorized into groups based on function or general behavior. For example, a *genetic pathway* (or gene regulatory network) is a set of genes that regulate each other and other substances, dictating a specific gene expression pattern. Another way to classify genes is based on the origin of the expression. Usually, autosomal genes are expressed randomly from one or the other chromosome. A small number of genes express only from one of the alleles; this process is called *allele-specific expression* (ASE). A special type of this process where the expression is specific to the parent of origin is called *genomic imprinting*. The genes that behave according to this process are called *imprinted genes*.

To draw conclusions about biological experiments, measurements have to be performed. First, to know the exact structure of a DNA or an RNA, each nucleotide can be read and saved into computer memory using the process called *sequencing*. For few decades, *Sanger sequencing* was the main method to read DNA sequence (Sanger et al., 1977; Metzker, 2010). Recent development have led to widespread use of *next-generation sequencing* (NGS, also called *second-generation sequencing*) technologies that have dramatically reduced the cost. Sequencing machines have been created by different companies and using various technical solutions. Compared to Sanger sequencing, all of them can process multiple samples in parallel with a fully automated workflow, resulting in a greatly decreased sequencing time. Nowadays, sequencing-based methods are used in a broad range of applications (Metzker, 2010).

Individuals from the same species have large amount of the genome in com-

mon. Therefore, it is often reasonable to only measure the nucleotides that are known to differ the most among individuals—*single nucleotide polymorphisms* (SNPs). Usually, at each SNP position, two alleles can appear, identified by A and B. If the pair of nucleotides in one position, called genotype, is AA or BB, the SNP is called *homozygous*. In case AB or BA happens in a particular position, it is called *heterozygous*.

Besides genetic information, it is also important to know the abundance of gene products in the organism. The amount of proteins is often of great interest. Although there are recently developed large-scale methods to measure it directly (Vogel and Marcotte, 2012), there has long been a common practice to use mRNA expression as a proxy for protein expression. Measuring mRNA abundance is called *gene expression profiling*.

1.2 Statistical background

Data analysis often starts with *exploratory analysis*. The analyst takes a first look at the data by calculating summaries and making some general visualizations of the data. This step can discover problems with data quality and generate ideas for further analysis. Visualizations are suitable to get the first overview, but in many cases, it is quite subjective to draw final conclusions. For example, there can be borderline cases where point clouds on the plot are nearly overlapping, and it is hard to tell whether they are separate or not. In this case, a *confirmatory analysis* method such as a statistical test can help to make an objective decision.

Statistical testing uses concepts from probability theory. *Random variable* is a function that is defined on sample space (a set of all possible outcomes of an experiment) and has undetermined output. *Distribution* of a random variable describes the behavior of the output by assigning a *probability* to each possible event (the measurable subset of the sample space). If random variable R has distribution D, it is written as follows:

$$R \sim D.$$

The probability of an event A is denoted with $P(A)$.

If sample space is finite or countable, the random variable is called *discrete*. In this case, it can be described with *probability mass function* that assigns a probability to each outcome. If sample space is uncountable, the random variable is called *continuous*. A continuous variable can be described with a *probability density function* and non-zero probability can only be assigned to (unions of) intervals.

Statistical testing starts with formulating a pair of competing hypotheses. The *alternative hypothesis* proposes a potential novel finding whereas *null hypothesis*

	Null hypothesis rejected	Null hypothesis accepted
Null hypothesis invalid	Correct decision True positive	Type II error False negative
Null hypothesis valid	Type I error False positive	Correct decision True negative

Table 1.1: Possible outcomes of hypothesis testing.

states that there is no finding. The null hypothesis is often a simplified version of the alternative hypothesis where one or more parameters are omitted (set equal to zero). A *test statistic* is created, which is a formula that measures the deviation from the null hypothesis, given a sample. The test statistic is often defined in such a way that *null distribution*—the distribution of the test statistic if null hypothesis holds—has some well-known form. If this is not possible, null distribution can also be estimated from the data, for example by permuting sample labels.

Test statistic value is calculated from the data. We can then find how probable it is to observe the test statistic value or more extreme value assuming null distribution. This probability is called *p-value*. If extremality means both left and right tail of the distribution, the p-value (and the test itself) is *two-sided*. Otherwise, if only left or right tail is considered, the p-value is *one-sided*. If the p-value is smaller than a particular constant called *significance level*, the null hypothesis is rejected. The significance level is chosen before seeing the data; typical choices include 0.05 or 0.01.

The test can result in one of four possible outcomes (see Table 1.1). If the null hypothesis is valid and accepted or invalid and rejected, the decision is correct. If the null hypothesis is invalid but accepted, the decision is called *type II error*. The fourth possibility to reject valid null hypothesis is called *type I error*. When a single test is made, the significance level is equal to the probability of making type I error. Thus, statistical testing methodology sets the upper limit only to type I error, because this error suggests a potential new finding whereas there is no actual signal. This limitation should be kept in mind in the applications where type II error is also important.

In bioinformatics, hundreds or thousands of tests are made in parallel. Let us assume that we have a dataset where there is no real signal, and we use 5% significance level. Approximately 5% of the tests would end up rejecting the null hypothesis and, therefore, get a false positive result. It is important to take the number of tests we make into account. This operation is called *multiple testing correction*.

There are different ways to correct for multiple testing. A popular choice is called *false discovery rate* (FDR) which converts the p-values into corrected p-

values called *q-values*. The expected proportion of false positive findings among all positive findings (with the q-value less than the significance level) is not more than the significance level.

Besides statistical testing, there are other types of useful confirmatory data analysis methods. For example, one can build regression models for predicting a continuous outcome, or classification models (classifiers) for predicting a discrete outcome. These methods are called *supervised learning methods* because both input and output are given to the algorithm. *Unsupervised learning methods* are another type of methods that do not use the output for learning. These methods are useful for generating new hypotheses.

When building complex models, it should be made sure that the model does not *overfit*. This situation happens when the model has too many parameters and works well on the data where it is trained, but does not generalize well to the unseen data (gives poor performance). To avoid overfitting, a common strategy is to split the data into independent training and validation set. The first set is used for building multiple models and the second one for evaluating the performance and choosing the best model. For smaller datasets, the data can be re-used by doing multiple randomized splits into training and validation set and reporting the average performance. This method is called *cross-validation*.

CHAPTER 2

SOURCES OF MULTIVARIATE DATA

Multivariate data appears in our everyday life, although we might not directly notice it. Big companies are constantly collecting and storing lots of information needed for their business. This data can be stored in different forms. In this thesis, *multivariate data* means a dataset that can be presented as a numeric table with features (attributes) measured on multiple objects (e.g. individuals).

	Height (cm)
Alice	165

↓

	Height (cm)	Weight (kg)	Foot length (cm)	Arm span (cm)
Alice	165	67	25	167

↓

	Height (cm)	Weight (kg)	Foot length (cm)	Arm span (cm)
Alice	165	67	25	167
Bob	182	82	29	185
Charlie	185	87	31	190
Dave	178	95	27	179
Erin	162	61	24	165
Frank	181	97	31	183

Figure 2.1: Collecting multivariate data by performing single measurements in a structured way.

A convenient way to think about collecting multivariate data is to make multiple single measurements. For example, let us say we want to compare persons with each other based on the size of their body. First, we can measure the height of Alice and thus get one object with one feature (see Figure 2.1). If we now make some more measurements, e.g. measure weight, foot length, and arm span, we get multivariate data, but only about one object. By repeating this procedure for multiple individuals, we get a multivariate dataset (see Figure 2.1).

Conventionally, *features* are listed in columns and *objects* in rows. This way, when measuring new objects, we need to add new rows rather than columns. For practical reasons, especially when there are many more features than objects, it may be more convenient to present the data in a transposed form—it depends on the situation. One such example in bioinformatics where objects can be presented in columns and features in rows is measuring the expression of every gene in a small number of samples.

Both features and objects can be divided into groups based on some properties. For example, we can divide male and female individuals into two groups, or separate length measurements from weight measurements.

Collecting the data and transforming it into a suitable format can be quite challenging. In the next section, we will describe different ways to collect multivariate data from biological experiments.

2.1 Gene expression microarray

Measuring gene expression (mRNA abundance) is in practice easier than measuring protein levels. Since mRNA is used to create protein, gene expression is often used as a proxy for protein activity, although the correspondence is not perfect (de Sousa Abreu et al., 2009; Vogel et al., 2010). A platform called *gene expression microarray* allows to measure tens of thousands of genes simultaneously. Although this method was invented in the last decade and is nowadays often replaced by sequencing-based methods, it is still relatively popular because of its lower cost (Mantione et al., 2014).

The main working principles of microarrays have been described by Heller (2002) as follows. From each gene, some *probes* (short oligonucleotides) are taken and attached to a slide in multiple copies. The sample of interest is marked with a color, converted to *complementary DNA* (cDNA) and put on the slide so that it can hybridize with the probes. After washing, only pairs with full complementarity will stay attached. The intensity of the color is measured using a scanner and converted into numeric scale so that it will represent relative abundance of each probe in the sample.

These numeric values have to be normalized to take potential variability in the

background intensity into account. Often, probe level values are converted into *probeset* or gene values. From data analysis perspective, it should be kept in mind that microarray only measures the relative abundance of the mRNA. Comparing absolute values of samples from different microarray platforms is complicated if not impossible (Kawasaki, 2006).

2.2 Tissue microarray

Proteins are common targets of interest in molecular biology because of their importance in molecular machinery. A popular method for detecting proteins in a sample is called *immunohistochemistry* (IHC). A carefully chosen *antibody*—a special type of immune system protein—binds to a protein of interest and labels it with some color (Dabbs, 2013). Taking an image of the processed sample will show the places with lower and higher abundance of this protein (see Figure 2.2).

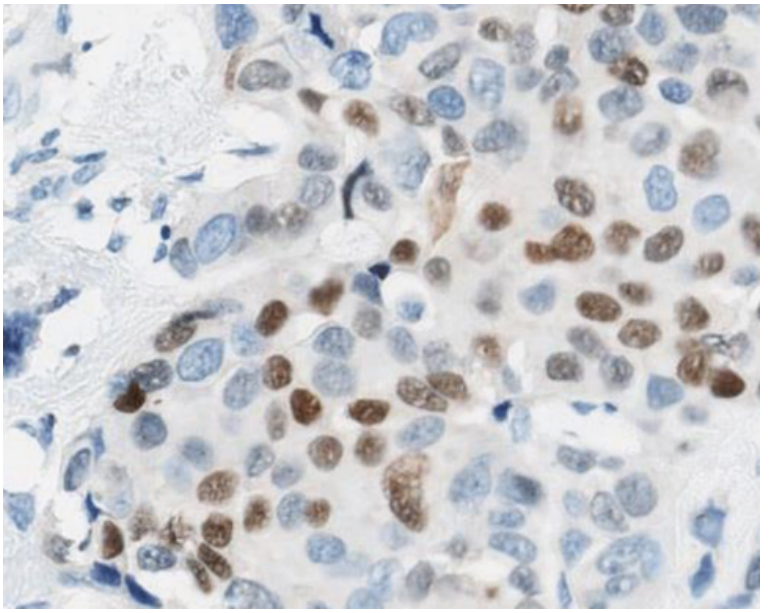

Figure 2.2: An example of antibody staining. Darker coloring shows the places with more abundance of the protein (Veta et al., 2014, Figure 1b, © 2014 IEEE).

Using this method on single samples is very time-consuming. Therefore a high-throughput technology called *tissue microarray* (TMA) has been developed (Kononen et al., 1998; Kallioniemi et al., 2001; Jawhar, 2009). Small cylindrical pieces of tissue are cut out and arranged to form a regular array called *TMA block*.

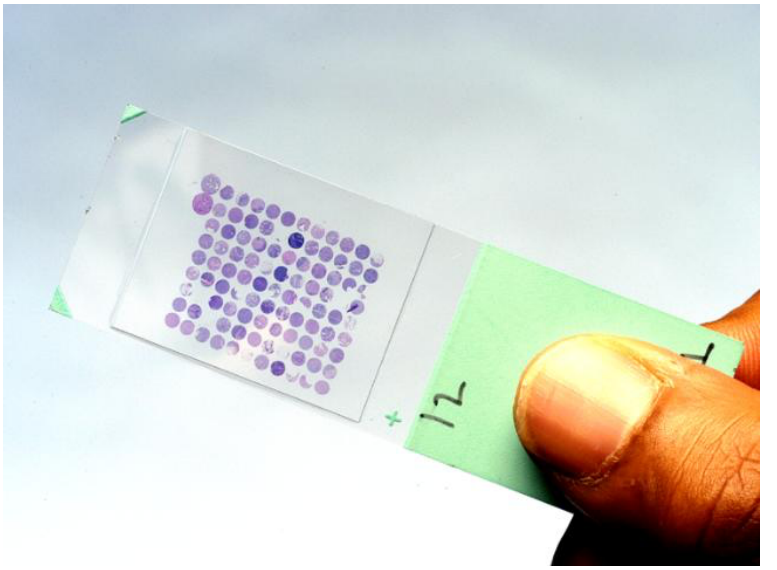

Figure 2.3: An example of a tissue microarray slide. Human thumb gives an intuition about the size of the tissue spots (Wikipedia, `https://en.wikipedia.org/wiki/Tissue_microarray`, January 12, 2016).

Multiple *TMA slides* can be constructed by cutting thin slices from the block (see Figure 2.3). Hundreds of samples can be placed on a single slide and analyzed in one run.

To convert an image of a stained TMA slide into a numeric matrix, some features of interest should be extracted (see Gurcan et al. (2009) for a review). Two popular features are as follows (see Figure 2.4).

- Percentage of positive cells—how many percent of all cells have staining intensity above a specified threshold;
- Staining intensity—the average staining intensity among all cells with the intensity above a threshold.

These two measures are complementary to each other. First of them characterizes the level of intensity globally, whereas the other measures the average intensity inside each cell.

The scores can be manually estimated by a professional pathologist or using some image analysis algorithm. The latter variant is often preferable to get the results faster and avoid human bias.

Repeating the same procedure with multiple proteins results in a full data matrix. Though, adding new proteins is not always straightforward. Some antibodies

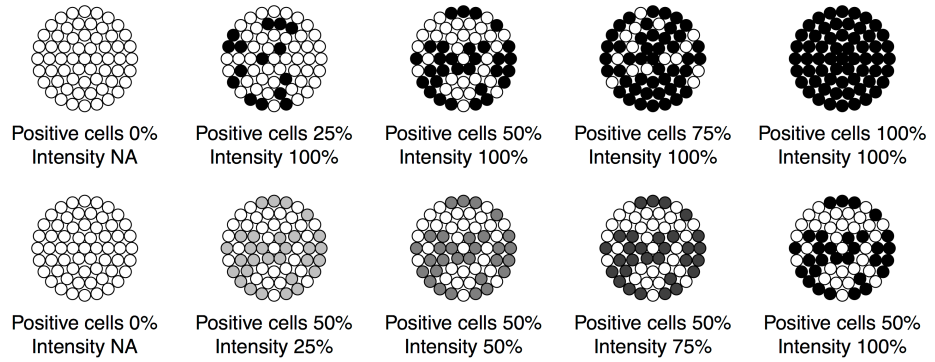

Figure 2.4: Illustration of the variation in percentage of positive cells (first row) and staining intensity (second row). If there are no positive cells, the intensity is not defined.

can nonintentionally interact with multiple proteins, causing unexpected results. Therefore, antibodies have to be carefully tested for cross-reactivity before using them in the study.

2.3 SNP array

Single nucleotide polymorphisms (SNPs) can also be detected using microarray technology (see LaFramboise (2009) for a review). It assumes that for all SNPs, possible alleles are known. This is often accomplished by sequencing a panel of individuals and finding more frequent changes before constructing the microarray. Usually, each SNP position has two possible alleles, identified by A and B.

SNP arrays are produced by different companies, but their technologies share the general principles. Probes complementary to genomic regions of interest are constructed by attaching both possible alleles separately to its flanking region. Samples are fragmented, labeled and hybridized with the array probes. Label brightness of A and B probes is converted into a numeric scale and perfectly matching fragments should have brighter labels. Each SNP from each sample can be assigned to a genotype call (AA, AB or BB) or no call (NC) if the signal is not strong enough.

2.4 RNA sequencing

RNA sequencing (RNA-seq) is a state-of-the-art method for large-scale gene expression profiling (see Wang et al. (2009) for a review). Instead of construct-

ing probes for detecting different sequences, mRNA is sequenced directly. This causes higher cost but allows much more precise measurement. RNA-seq also allows to conduct studies that were not possible using microarrays, for example predicting new genes or measuring allele-specific expression. One dataset can be re-used and converted to a different kind of numeric matrix by extracting other types of features. RNA-seq needs much more computational resources, and analysis methods are not yet so well matured compared to microarrays. Despite that, it seems probable that RNA-seq will replace microarrays in the future (Mantione et al., 2014).

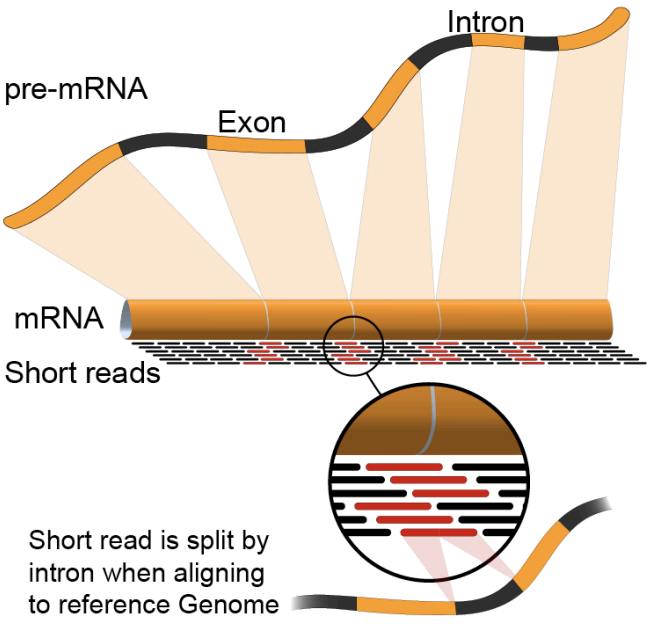

Figure 2.5: Principles of RNA-seq (Wikipedia, `https://en.wikipedia.org/wiki/RNA-Seq`, January 12, 2016).

The output of an RNA-seq experiment consists of short nucleotide sequences called *reads*. For almost any type of analysis, these reads should be *mapped* (or aligned) to the genome to get genomic coordinates and make sense of the data (see Figure 2.5). Many different mapping programs have been developed (Grant et al., 2011). Some analysis pipelines need mapping to multiple genomes. For example, when studying allele-specific expression, mapping to only reference genome would favor the reference allele and can produce artificial bias (Stevenson et al., 2013).

After mapping the reads to the genome, most types of analysis use additional

information such as gene coordinates to group the data according to known functional units. Various algorithms have been developed to count the reads and find differentially expressed genes (Rapaport et al., 2013). If studying allele-specific expression, reads can be counted for both alleles of each SNP. In any case, raw RNA-seq reads will be converted to a numeric matrix for further analysis.

2.5 Reverse transcription quantitative PCR

Reverse transcription quantitative polymerase chain reaction (RT-qPCR) is another method for measuring RNA abundance in the cell (Mullis et al., 1986; Kubista et al., 2006). It is a low-throughput method, measuring from few up to hundreds of genes. Since the technology is relatively robust, it is often considered a "gold standard" for measuring RNA expression and used to validate results from high-throughput experiments. It can be used to measure different types of RNAs, including mRNA and miRNA. The method relies on working principles of *polymerase chain reaction* (PCR).

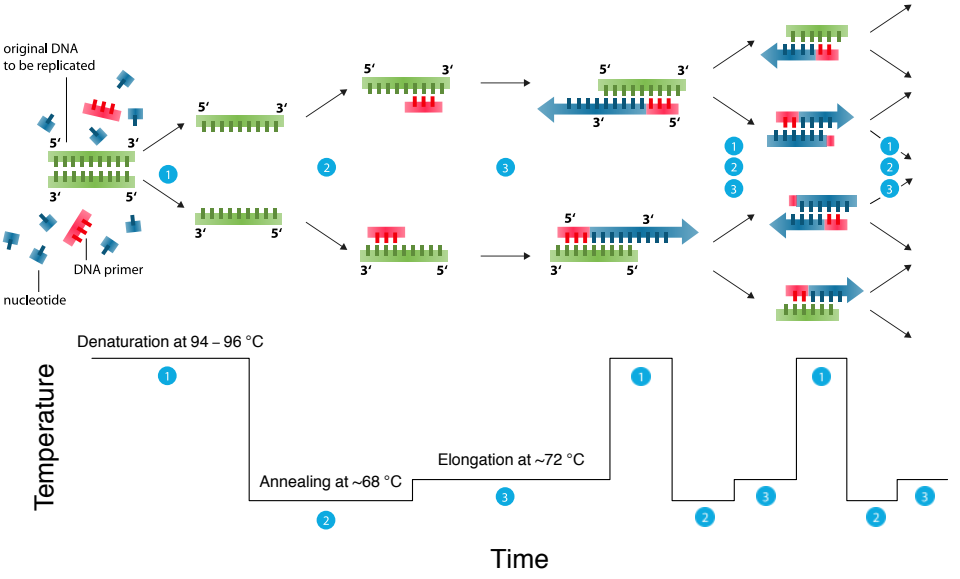

Figure 2.6: Principles of PCR (Wikipedia, `https://en.wikipedia.org/wiki/Polymerase_chain_reaction`, adapted, January 12, 2016).

PCR is a technology for amplifying fragments of DNA (Kubista et al., 2006). It is performed in a cyclic way by changing the temperature (see Figure 2.6). First, double-stranded DNA is *denaturated* (separated into single strands) by applying high temperature. Then, the temperature is lowered to allow *primers* (short

nucleotide sequences) to anneal (attach) to each of the sequences. Finally, the temperature is raised a bit to make optimal conditions for *DNA polymerase*—a DNA building enzyme. Polymerase attaches to the primer and starts building the second strand based on complementarity. When it is completed, we have two identical copies of the initial fragment of double stranded DNA. By repeating all the steps multiple times in a row, the number of identical DNA fragments grows exponentially, causing a chain reaction.

To use RNA as the input in this process, it should first be converted into *complementary DNA* (cDNA). This is performed using *reverse transcription*, lead by an enzyme called *reverse transcriptase*. To quantify the abundance, the input DNA should be labeled. Different techniques of fluorescent labeling can be used for this purpose (Kubista et al., 2006; Wong and Medrano, 2005). Specific dyes or probes are introduced into the reaction that emit fluorescence when attached to target DNA sequence. Signal and background DNA are labeled differently, and one can approximate the abundance of RNA by the difference between the intensities of the fluorescent dyes.

It may seem enough to measure the label intensity on one specific point in time. However, this can lead to serious bias since the process saturates at some point. The saturation happens due to exhaustion of some essential component needed for the reaction, for example, primers or nucleotide triphosphates that are used by the polymerase to build the strands (Kubista et al., 2006). Thus, the relationship between cycle number and time is not linear. To overcome this problem, quantitative PCR (which is also called real-time PCR) measures the intensity on each step of the cycle. A threshold is set, and the number of cycles needed to reach this threshold is found for both signal and background curves, denoted by CT. The relative expression can be obtained using the difference between CTs.

2.6 Summary

In this chapter, we have covered sources of multivariate data used in the thesis. Tissue microarray is the platform for measuring protein expression in PREDECT project (see Chapter 4). Multiple technologies for measuring gene expression were described. Each of them has its advantages (see Table 2.1). RT-qPCR is considered a gold standard for measuring gene expression, but it is a low-throughput method for up to hundreds of genes. Both gene expression microarray and RNA-seq are high-throughput methods. RNA-seq has many more applications besides simple gene expression analysis, but it is still more expensive than gene expression microarray.

Besides expression, we also described a platform for measuring genome variation (SNPs). As described before, SNP array internally produces a continuous

	Number of genes	Cost per gene	Reproducibility
RT-qPCR	$\sim$200	€€€	good
Microarray	$\sim$20000	€	average
RNA-seq	∞	€€	average

Table 2.1: Comparison of platforms for measuring gene expression.

scale of brightness, but in the further analysis, genotype calls are used. In this thesis, we use results from the SNP array to categorize SNPs into groups based on zygosity (whether an individual is homozygous or heterozygous). This information is used for finding allele-specific expression (see Chapter 5).

CHAPTER 3

ANALYSIS METHODS

All methods described in this thesis are applied to multivariate data (see Chapter 2). In general, there are two popular types of approaches how to analyze this kind of data.

- Apply a multivariate analysis method and report the results.
- Split the data, then apply some univariate analysis method on each piece and show all results or top ones in an ordered list.

Whether to choose one or the other strategy or a hybrid thereof depends very much on the situation.

A convenient way to present multivariate data is a table form called a *matrix*. Throughout this chapter, we denote the numeric dataset by a matrix $X : n \times p$ where rows represent n objects (e.g. samples), and columns represent p features (e.g. genes):

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}. \tag{3.1}$$

In many cases, it is also important to take sample grouping (annotations) into account. We denote this by a vector y:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}. \tag{3.2}$$

For convenience, we use condition on grouping to subset the data matrix. For example, $X_{y="normal"}$ means all rows from X with sample group "normal" and $X_{y="normal",1}$ means applying this condition to the first column.

3.1 Exploratory analysis methods

3.1.1 Principal component analysis

Principal component analysis (PCA) is a multivariate analysis method to reduce the dimensionality of the data (Pearson, 1901; James et al., 2014, p. 374–385). Geometrically, a p-dimensional point cloud is rotated until the largest variance is found. The projection of the data in this direction is called the first *principal component*. Next, the point cloud is rotated again to find the second largest variance, but keeping the orientation of the first component fixed. The projection on the axis with the second largest variance is called the second principal component. This process is repeated for all components so that the next component is found under the following conditions.

- The directions of all previous components are fixed;
- The new direction is perpendicular to all previous directions.

We can now reduce the dimensionality by removing some number of the last components which only explain a small amount of the variability. Since the shape is retained after finding the components, first few components can also be used to approximate the distances between objects. By plotting two first components on the scatterplot, we get the most informative two-dimensional view of the data, explaining as much variance as possible. On the other hand, there are situations where first components show uninformative variation such as batch effect (Leek et al., 2010). In such cases, it makes sense to look at further components that can explain more informative sources of variability.

Results of the PCA depend on the scale of the input variables. If one of the variables has much higher variance than all others, the direction of the first component will almost overlap with the axis of this variable. To avoid this problem, all variables can be *standardized* (mean subtracted and divided by standard deviation) before applying PCA. This procedure is equivalent to using correlation matrix instead of covariance matrix as input for PCA if a method called *eigenvalue decomposition* is used for finding the components. We prove it by the following lemma.

Lemma. *Correlation between two variables is equal to the covariance between respective standardized variables (mean subtracted and divided by standard deviation).*

Proof. Let EX and VX denote expected value and variance of the random variable X, respectively. Let A_1 and A_2 be random variables and B_1 and B_2 respective standardized variables:

$$B_i = \frac{A_i - EA_i}{\sqrt{VA_i}},\ i = 1, 2.$$

Using definition of correlation and covariance, we get

$$\begin{aligned} corr(A_1, A_2) &= \frac{cov(A_1, A_2)}{\sqrt{VA_1 VA_2}} \\ &= \frac{E[(A_1 - EA_1)(A_2 - EA_2)]}{\sqrt{VA_1 VA_2}} \\ &= E\left[\frac{A_1 - EA_1}{\sqrt{VA_1}} \cdot \frac{A_2 - EA_2}{\sqrt{VA_2}}\right] \\ &= E(B_1 \cdot B_2) \\ &= E[(B_1 - EB_1)(B_2 - EB_2)] \\ &= cov(B_1, B_2) \end{aligned}$$

since $EB_1 = EB_2 = 0$. □

Computationally, a more efficient way for finding the components is using a matrix factorization called *singular value decomposition* (SVD) (Jolliffe, 2002, p. 44–46). The following matrix operations are used:

- *Multiplication* of matrices $A : r \times s$ and $B : s \times t$ is defined as a $r \times t$ matrix calculated using the formula

$$AB := \begin{pmatrix} \sum_{k=1}^{s} a_{1,k} b_{k,1} & \cdots & \sum_{k=1}^{s} a_{1,k} b_{k,t} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{s} a_{r,k} b_{k,1} & \cdots & \sum_{k=1}^{s} a_{r,k} b_{k,t} \end{pmatrix} \tag{3.3}$$

  where a_{ij} and b_{ij} are the elements of row i and column j from matrices A and B, respectively;

- The *transpose* of a matrix A (created by reflecting A over its main diagonal) is denoted by A^T.

SVD divides the data matrix into three matrices:

$$X = ULA^T \tag{3.4}$$

where

- $U : n \times r$ and $A : p \times r$ are matrices with *orthonormal* columns (perpendicular to each other and having unit length);

- $L : r \times r$ is a diagonal matrix where the diagonal elements are square roots of the *eigenvalues* of $X^T X$, in descending order.

The eigenvalues of $X^T X$ are the values λ that satisfy the equation

$$X^T X v = \lambda v \tag{3.5}$$

for a non-zero vector v. These eigenvalues define the contribution of each principal component—the amount of variability explained.

3.1.2 Clustering

Multivariate data often has some objects or features that are highly similar to each other. Grouping them together will make the data more understandable and easier to interpret. This result can be achieved using *clustering* (or cluster analysis) that aims to form *clusters* (or groups) with high similarity using unsupervised learning approach.

Many clustering methods have been developed, for reviews see Aggarwal and Reddy (2013); Thalamuthu et al. (2006). In this thesis, we consider two most popular methods.

k-means clustering

A popular clustering method where the number of clusters is given as a parameter is called *k-means* (Lloyd, 1982; James et al., 2014, p. 386–390). Assuming that all observations belong to a cluster, cluster centers are calculated. Each observation is then re-assigned to the cluster with the closest center. This process is repeated until the cluster centers do not change anymore.

There are multiple ways for initialization; random selection is often involved (Celebi et al., 2013). Random initial choice can lead to poor quality clusters. Therefore, it is often recommended to run k-means repeatedly with different initial configuration and report the result with the best quality clusters according to some information criterion.

It is not guaranteed that the algorithm gives optimal division. It can converge to a local optimum that is not necessarily optimal globally. In practice, though, k-means has shown to give acceptable results and run in a reasonable amount of time (Aggarwal and Reddy, 2013).

Choosing an appropriate number of clusters is usually not trivial. Several strategies have been proposed (Yan, 2005; Kodinariya and Makwana, 2013). There are also extensions of the algorithm proposed which run k-means iteratively to find the optimal number of clusters (Likas et al., 2003; Pelleg and Moore, 2000; Ray and Turi, 1999).

Hierarchical clustering

Hierarchical clustering is a clustering method where a dendrogram (tree-like structure) is generated with original objects as leaves (Sneath, 1957; Ward Jr, 1963; James et al., 2014, p. 390–399). It can be built bottom-up (agglomerative) or top-down (divisive). We consider the first type.

First, all pairwise *distances* are calculated. Two objects with the smallest distance are merged, and distance from this pair to all other objects is re-calculated. This step is repeated until all objects belong to one large cluster.

Two important parameters determine the behavior of hierarchical clustering. We need to decide how to calculate the distance between clusters and which linkage method to use. We denote indicator function (resulting in 1 if the condition is met and 0 otherwise) by $I(\cdot)$, arbitrary vectors of length n by $a = (a_1, a_2, \ldots, a_n)^T$ and $b = (b_1, b_2, \ldots, b_n)^T$. Distance between these vectors is denoted by $d(a, b)$.

Some examples of the distance measures are as follows.

- correlation distance—Pearson correlation subtracted from 1.

$$d(a,b) = 1 - \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \tag{3.6}$$

  where

$$\bar{a} := \frac{1}{n}\sum_{i=1}^{n} a_i,$$

$$\bar{b} := \frac{1}{n}\sum_{i=1}^{n} b_i.$$

- Euclidean distance—the square root of the sum of square distances.

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{3.7}$$

- maximum distance—the greatest absolute difference between coordinates.

$$d(a,b) = \max_i |a_i - b_i| \tag{3.8}$$

- Manhattan distance—the sum of the absolute differences.

$$d(a,b) = \sum_{i=1}^{n} |a_i - b_i| \tag{3.9}$$

- Canberra distance—weighted Manhattan distance.

$$d(a,b) = \sum_{i=1}^{n} \frac{|a_i - b_i|}{|a_i| + |b_i|} \tag{3.10}$$

- binary distance (or Jaccard distance)—the number of a/b pairs that have exactly one non-zero divided by the number of pairs that have at least one non-zero.

$$d(a,b) = \frac{\sum_{i=1}^{n} I((a_i \neq 0 \wedge b_i = 0) \vee (a_i = 0 \wedge b_i \neq 0))}{\sum_{i=1}^{n} I(a_i \neq 0 \vee b_i \neq 0)} \tag{3.11}$$

  This distance is more meaningful for binary vectors that can be interpreted as sets. It is the ratio of the size of the symmetric difference to the union.

The *linkage method* defines how the distance between objects is generalized into the distance between clusters. Popular linkage methods include the following.

- Single linkage—using two closest objects from two clusters to be merged.
- Complete linkage—using two farthest objects.
- Average linkage (or UPGMA—unweighted pair group method with averaging)—the average distance of all possible pairs.
- McQuitty linkage (or WPGMA—weighted pair group method with averaging)—the average distance between both clusters to be merged and the cluster of interest.
- Median linkage (or WPGMC—weighted pair group method using centroids)—the median distance of all possible pairs.
- Centroid linkage (or UPGMC—unweighted pair group method using centroids)—the distance between cluster means.
- Ward linkage—using the sum of squared differences from points to centroids as the distance.

The choice of both distance and linkage method heavily influence the result. Euclidean and correlation distance are the most popular distance measures. Euclidean distance measures the absolute deviance between vectors whereas correlation compares the trends. The latter option may be more meaningful in gene expression measurements since the changing pattern of the expression is usually more informative than the absolute level.

The most often used linkage methods include complete and average linkage. Compared to the single linkage method, they tend to produce a more balanced clustering tree (James et al., 2014, p. 394–395).

Heatmap

The *heatmap* is a color coded data matrix that allows to easily detect patterns and outliers (Loua, 1873; Sneath, 1957). It is one of the most popular visualizations in bioinformatics (Wilkinson and Friendly, 2009). Optionally, rows and/or columns can be clustered using hierarchical clustering to group similar rows and columns together (see Figure 4.4 right for an example). If one of the dimensions is huge (this is common for gene expression datasets, measuring tens of thousands of genes), hierarchical clustering would be slow. In this case, k-means clustering can be used first to decrease the dimensionality, followed by hierarchical clustering on k-means cluster centers (Metsalu and Vilo, 2015; Kolde, 2015).

3.2 Confirmatory analysis methods

3.2.1 Differential expression

A frequent task in gene expression studies is to find genes that behave differently in two conditions. For example, we can study which genes are up-regulated (have higher expression) or down-regulated (have lower expression) among disease patients compared to healthy ones. A test can be made for each gene and the results ordered based on it. It should be noted that differential expression is always relative, there has to be some baseline expression to compare with.

Student's t-test

Student's t-test is a popular statistical test where a decision is made based on t-distribution (Student, 1908). It has multiple different versions; here we are describing *two-sample unpaired* t-test that is used in the Chapter 6. It is used to compare means of two sample groups by taking both mean and standard deviation into account, and assumes that both groups have equal variances.

For example, we may be interested whether men and women have different height on average. We can take a sample (e.g. ten people) from both groups and measure them. As a result, we may find out that average height in the sample is 180 cm for men and 170 cm for women. By looking at means only, it is hard to decide whether this difference holds in general or is caused just by the random fluctuation due to small sample size. In this situation, t-test can be used to find the probability to see such difference or a larger one in the sample, assuming that the null hypothesis of having no actual difference in the whole population holds.

Mathematically, for each $j = 1, \ldots, p$, we use the vectors

$$\begin{aligned} a &:= X_{y=\text{"}group1\text{"},j}, \\ b &:= X_{y=\text{"}group2\text{"},j}. \end{aligned} \tag{3.12}$$

We denote length of a and b by n_a and n_b, respectively, and calculate the t-statistic:

$$t = \frac{\overline{a} - \overline{b}}{\sqrt{\left(\frac{1}{n_a} + \frac{1}{n_b}\right) \cdot \frac{(n_a-1)s_a^2+(n_b-1)s_b^2}{n_a+n_b-2}}} \tag{3.13}$$

where

$$\begin{aligned} \overline{a} &= \frac{1}{n_a}\sum_{i=1}^{n_a} a_i, \\ \overline{b} &= \frac{1}{n_b}\sum_{i=1}^{n_b} b_i, \\ s_a^2 &= \frac{1}{n_a-1}\sum_{i=1}^{n_a}(a_i-\overline{a})^2, \\ s_b^2 &= \frac{1}{n_b-1}\sum_{i=1}^{n_b}(b_i-\overline{b})^2. \end{aligned} \tag{3.14}$$

Next, two-sided p-value is calculated based on the number of degrees of freedom:

$$p = P(T \geqslant |t| \vee T \leqslant -|t|) = 2 \cdot P(T \geqslant |t|) \tag{3.15}$$

where

$$T \sim t(n_a + n_b - 2). \tag{3.16}$$

This way, we can calculate a p-value for each feature and then rank them based on that.

Linear models

Linear models can be used to answer the same question as t-test, but they are applicable in a broader range of applications. They allow multi-factored design (multiple levels for grouping) that is not possible with a simple t-test. For example, it is possible to remove the batch effect by including the batch identifier in the model.

Linear models are implemented in the software package *limma* that is written using R statistics language (Smyth, 2004; Ritchie et al., 2015). It allows to fit various linear models to test for differential expression of genes and also perform supporting analysis steps. It was originally written for microarray data, but was extended with later developments to work with RNA-seq data as well.

An important improvement compared to simple linear models is the property that the models of different genes are not entirely independent. Limma borrows information across genes by modifying gene variances towards the global variance. This strategy gives more robust variance estimates resulting in less false positives and false negatives.

3.2.2 Gene set analysis

Some methods take a gene set as input. This set can have a biological meaning like a genetic pathway, or come as an output gene list from some other analysis step such as differential expression analysis. Compared to single gene analysis, taking gene set as a whole increases the statistical power and can make the results easier to interpret.

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) is a method for finding whether a given list of genes, such as a pathway or other annotated group, is significantly related to a binary phenotype of interest (Subramanian et al., 2005). For example, this method can reveal which pathways behave differently when comparing normal and tumor samples.

The method starts with calculating a statistic that shows how well the expression of each gene relates to the phenotype. Some options for this measure include t-test or linear models. Genes are ranked based on the statistic. Walking through the sorted list, the running sum is increased if the gene is in the list and decreased if it is not. The extent of the change is calculated based on the same measurements that were used for ordering the genes. An *enrichment score* (ES) is calculated as the maximum deviation of the running score from zero.

To estimate whether the deviation is significant, a permutation test is used. Labels of the samples are randomly shuffled, and ES is calculated. This step is repeated many times to generate null distribution. If multiple gene sets are tested, the enrichment scores are normalized, and results are corrected for multiple testing using FDR.

Hypergeometric test in g:Profiler

A set of tools for analyzing gene lists is called g:Profiler (Reimand et al., 2007, 2011). One part of the toolset called g:GOSt allows to perform functional enrichment analysis based on different types of biological evidence. It finds whether the given gene list is significantly overlapping with any functional category using *hypergeometric test*.

	In the gene list	Not in the gene list	Total
In the category	a_{11}	a_{10}	a_{1+}
Not in the category	a_{01}	a_{00}	a_{0+}
Total	a_{+1}	a_{+0}	a_{++}

Table 3.1: Input for hypergeometric test.

The test takes a 2×2 contingency table as input (see Table 3.1). Given fixed marginal frequencies (a_{1+}, a_{++}, a_{+1}), we formulate the null hypothesis that being in the gene list is independent of being in the functional category. In this case, the number of common genes in the category and in the gene list (denoted by A) has hypergeometric distribution:

$$A \sim HG(a_{1+}, a_{++}, a_{+1}). \tag{3.17}$$

We can calculate one-tailed p-value by using probability mass function of hypergeometric distribution:

$$p = P(A \geqslant a_{11}). \tag{3.18}$$

If many categories are tested, multiple testing correction is needed. In g:Profiler, a tailor-made algorithm called *g:SCS* is used by default for correcting p-values (Reimand et al., 2007). Other popular methods, such as Bonferroni correction and FDR, are also available.

3.2.3 Support vector machine

Support vector machine (SVM) is a supervised learning method to find optimal classifier (Vapnik and Lerner, 1963; Cortes and Vapnik, 1995; James et al., 2014, p. 337–372). The aim of the original SVM is to find the best linear separation between two groups by fitting an optimal hyperplane. There are several extensions for more than two groups and non-linear separation. We describe only the original linear SVM.

SVM takes data matrix X and annotation vector y as input. For simplicity, we assume that annotation groups are coded with $+1$ and -1. Linear SVM solves an optimization problem

$$\max_{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_n} M \tag{3.19}$$

subject to

$$\begin{gathered} \sum_{j=1}^{p} \beta_j^2 = 1, \\ y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \geqslant M(1 - \epsilon_i), \\ \epsilon_i \geqslant 0, \\ \sum_{i=1}^{n} \epsilon_i \leqslant C \end{gathered} \tag{3.20}$$

for $i = 1, \ldots, n$ where M is the width of the margin to be maximized, $\beta_0, \beta_1, \ldots, \beta_p$ are classifier coefficients, $\epsilon_1, \ldots, \epsilon_n$ measure how much each observation violates the margin, C is a constant that determines how many violations we tolerate. The constant C is typically chosen based on cross-validation by performing a grid search and choosing the constant that gives the lowest cross-validation error. The prediction for a new data vector $x^* = (x_1^*, \ldots, x_p^*)$ is given by

$$\hat{y}^* = \text{sign}(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^*). \tag{3.21}$$

3.2.4 Binomial test

Binomial test is a statistical test where the decision is made based on binomial distribution (Clopper and Pearson, 1934). For example, if we flip a coin 100 times, the total number of tails has a binomial distribution. If we get 55 tails, it is not clear whether the coin is biased (the proportion of tails is different from 50%). It is hard to intuitively draw the line between random fluctuation and clear difference. The test allows us to assign a statistical score.

Mathematically, let us assume that we are trying to detect allele-specific expression and for each gene and for each sample, we count the number of maternal reads n_m and paternal reads n_p. In case of biallelic expression, they should be similar. We can calculate the p-value using binomial distribution:

$$p = P(B \leqslant \min(n_m, n_p) \vee B \geqslant \max(n_m, n_p)) \tag{3.22}$$

where

$$B \sim B(n_m + n_p, 0.5). \tag{3.23}$$

3.3 Summary

As described, there are multiple exploratory and confirmatory analysis methods available. Which one to use in a given situation depends on the data type, the amount of data and other factors. In the next chapter, we describe how these methods were applied to multivariate datasets.

CHAPTER 4

COMPARISON OF CANCER MODELS

4.1 Overview of PREDECT project

I have been involved in the project about new models for *Preclinical Evaluation of Drug Efficacy in Common solid Tumours* (PREDECT). It is a collaborative project between academic partners, *small and medium-sized enterprises* (SMEs) and pharmaceutical companies from *European Federation of Pharmaceutical Industries and Associations* (EFPIA) funded by *Innovative Medicines Initiative* (IMI). The aim of the project is to develop novel cancer models for testing potential drugs before allowing them to be used on humans (see Figure 4.1). Three pathologies are considered: breast, prostate and lung cancer that form three separate work packages (WPs), respectively. The fourth work package (WP4) is responsible for managing and analyzing the material and data produced by the first three. The University of Tartu takes part in WP4 and is mainly involved in collecting metadata and performing the bioinformatic analysis.

In the early days of cancer research, the only way to study cancer and search for potential drugs was using hand-cut pieces from a human tumor. This method needs much patient material that is not always available. There emerged a need for imitating human cancer that would still be as similar to the original tumor as possible. Non-perfect resemblance leads to a major concern about drug development efficacy: the majority of potential new medicines passing preclinical studies do not succeed in the following clinical trials made on humans (Voskoglou-Nomikos et al., 2003; Kamb, 2005; Cook et al., 2012). To overcome this problem, PREDECT aims to develop novel models that better represent human cancer. Models imitate cancer in varying levels of complexity.

The simplest models are *cell lines* where cells originate from human cancer but are immortalized to keep proliferating. Researchers can grow them without the need for steady inflow of biological material. Cell lines can be cultured on a simple glass or plastic slide (called *2D models*) or in a more complex setting where

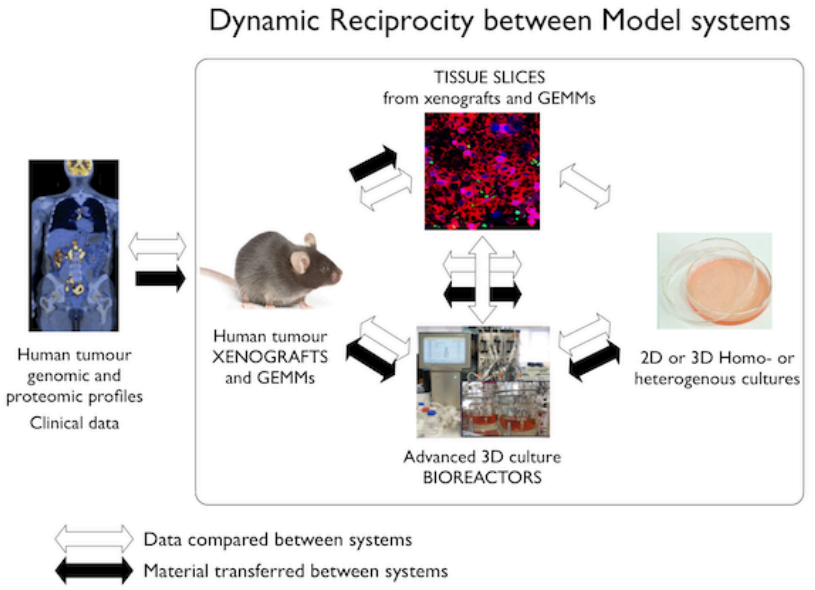

Figure 4.1: Model systems used in PREDECT project (PREDECT homepage, `http://www.predect.eu/about/project_overview/`, January 12, 2016).

the natural spatial shape of the tumor is retained (*3D models*). Specific control mechanisms can be added, which keep the biological conditions (e.g. nutrients, temperature, pressure) constant over time. This type of artificial environment is called *bioreactor*.

Models described above are called *ex vivo* or *in vitro* models because they grow outside the living organism, on an artificial platform ("in the tube"). An overview of such models is given in a review by Hickman et al. (2014). Another class of models is so called *in vivo* models where tumor cells grow in the living model organism (Frese and Tuveson, 2007). The most popular host is a mouse because of its fast breeding times and relatively good resemblance to human (de Jong and Maina, 2010). Therefore, *in vivo* models are often called *mouse models*. A mouse is clearly different from a human in many ways, and we should be careful when translating results from mouse studies into human (Wall and Shani, 2008; Barrett and Melenhorst, 2011).

There are several subclasses of *in vivo* models. *Genetically engineered mouse models* (GEMMs) are mice grown with a modified genome to grow cancer (Politi and Pao, 2011). Often, this is accomplished by making one or few tumor suppressor genes non-functional (gene knock-out). Another class is called *xenografts*

that are normal mice where a piece of cancer cell line or human tumor is injected (grafted) into a mouse. The tumor is then growing in a microenvironment similar to original location (Frese and Tuveson, 2007).

A separate class of models that complexity-vise lies between cell lines and mouse models is called *tissue slices*. These slices are cut from human tumor or mouse model and cultivated in a separate growth medium. The main advantage compared to the original tumor is a more economical use of the material. This improvement is following the widely adopted 3R's principles of animal experiments (Workman et al., 2010; Russell et al., 1959).

- Replace living organisms with non-living biological material where possible;
- Reduce the number of animals needed while keeping the required accuracy;
- Refine the procedures so that they cause less harm and pain for the animals.

4.2 Centralized data collection and analysis

The main platform for model comparison in PREDECT consortium is tissue microarray (see Section 2.2). Tissues are punched into small cylindrical pieces and arrayed into a paraffin block. This block can be sliced and put onto multiple glass slides. Slides are first stained with *hematoxylin and eosin* (H&E) which is a popular way to visualize the histological structure of a tissue. Additionally, slides are stained with different antibodies. Areas with higher expression of the target protein will appear with more intense stain (see Figure 2.2).

Both biological material and sample annotations are collected centrally. For the latter, University of Tartu has developed a web-based database system (MBase—Metadata Database) in co-operation with other PREDECT partners and Quretec Ltd. The model management and analysis process looks as follows.

- Laboratories participating in model development send their latest models to WP4 and annotate them in MBase.
- Samples are punched, and TMAs constructed centrally in WP4.
- H&E staining is made, and bad quality samples are reported to the partners to get them replaced.
- TMA slides are stained with selected antibodies.
- Stainings are converted into digital images.

- Images are converted into numeric data using image analysis methods or estimates from a professional pathologist.
- Image analysis results (scorings) are analyzed using heatmap and PCA plot to compare cancer models.

Figure 4.2: Spreadsheet form to facilitate sample submission in batch. Coloring refers to different data types. Where only a specified number of options is available, these are also shown to the user.

For data analysis, it is essential to collect good quality annotations about each sample. MBase uses *Qure Data Management Platform* (QDMP) as data input interface (Jäger et al., 2008). It has a secure web-based system to make it comfortable to use for all partners. Users can download and fill a form in a spreadsheet program. They can then easily upload annotations in batch to the database (see Figure 4.2). During the process, the data format is validated to make sure that the structure of the data is correct.

Building a unified annotation format for different types of models (cell lines, xenografts, GEMMs, human material, slices from different material) was not straightforward. It was evolving during the project as we faced new problems with some of the samples. Initially, we had a separate structure for samples from cell lines and tissues, but this division was not sufficient for collecting detailed and structured data about each platform. Therefore, the structure was later changed into platform-based classification (2D/3D/bioreactor, tissue slice, in vivo, breast human reference, prostate human reference, lung human reference).

MBase has multiple supporting tables besides sample annotations (see Figure 4.3). Samples are linked to TMAs, table of stainings is linked to both TMAs

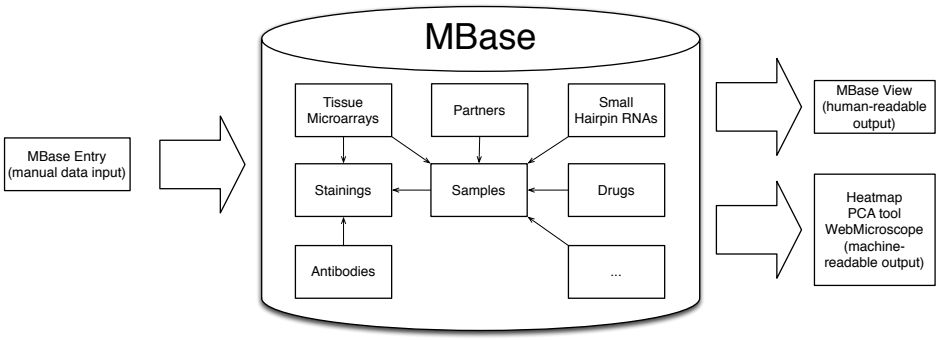

Figure 4.3: Structure and interfaces of MBase.

and antibodies. There are also multiple supporting tables that give input to samples, such as the list of partners, small hairpin RNAs, drugs, models, and model modifications. At the moment of writing (January 12, 2016), there are 2061 samples annotated.

MBase data is used in both human-readable form as tables and summary plots and machine-readable form for giving input to other PREDECT tools (see Figure 4.3). A custom heatmap tool (see Section 3.1.2) is developed where samples can be clustered based on different antibodies. The resulting plot is interactive; the user can click on it to visualize a row or column separately on violin plot or one specific cell of the heatmap on barplot. The user can further click on single bars on the barplot to go to *WebMicroscope*—a collection of digitized images about PREDECT samples. This tool provides access to images of digitized TMA spots stained with different antibodies.

The same data can be visualized on another tool using PCA plot (see Section 3.1.1). It is also interactive; sample groups can be removed separately, and each spot has a tooltip with additional information and link to WebMicroscope. Both heatmap tool and PCA tool can also be used to visualize gene expression microarray data collected from public sources and PREDECT partners.

4.3 Improved breast cancer xenograft model (paper I)

Breast cancer is a major cause of cancer-related deaths among women. Most patients with breast cancer are *estrogen receptor positive* (ER+), meaning that they have receptors for estrogen and can potentially respond to hormonal therapy. So far, there has been more success in growing *estrogen receptor negative* (ER-) xenografts (Zhang et al., 2013; Cottu et al., 2012). In this paper, the consortium developed a novel breast cancer ER+ xenograft model where tumor cells are in-

jected directly into mouse milk duct system as opposed to the fat pad injection that is commonly used.

First, six different breast cancer cell lines were injected into mouse milk ducts to confirm that appropriate microenvironment is provided. All except one grew without estrogen supplementation. Intraductally grafted xenografts showed a resemblance to the clinical tumor in terms of histology and tissue morphology. Cell line MCF7 had the highest ER expression and was chosen for further analysis.

Tumor appearance and progression in the MCF7 intraductal model were more similar with ER+ than in the fat pad model. The intraductal model also responded to endocrine therapy, showing decreased proliferation index after treatment.

Comparison of these two models was also done on the molecular level by measuring gene expression using microarray (see Section 2.1). We found many differentially expressed genes (see Section 3.2.1). First components from PCA (see Section 3.1.1) showed that intraductal model is more close to *luminal* subtype that tends to be ER+. This subtype is highly heterogeneous, consisting of different gene expression profiles, mutational repertoire and histological characteristics, with very different clinical outcomes and responses to systematic treatments. The name of the subtype comes from the fact that cancer has originated from luminal cells, one of the two epithelial cell types that the main tissue (parenchyma) of the breast is made of.

Pathway analysis with g:Profiler (see Section 3.2.2) confirmed the changes in the genes related to proliferation. The genes most extremely downregulated in the intraductal model were also related to *epithelial to mesenchymal transition* (EMT) as discovered by GSEA (see Section 3.2.2). This finding suggests that microenvironment plays an important role to keep luminal characteristics of the cells.

The suitability of the proposed model was confirmed by growing xenografts using nine patients with ER+ breast cancer. All patients retained luminal characteristics after grafting.

My contribution was to collect publicly available gene expression data about patient tumors (Guedj et al., 2012, ArrayExpress ID: E-MTAB-365) and cell lines (Neve et al., 2006, ArrayExpress ID: E-TABM-157) from ArrayExpress database (Brazma et al., 2003). I analyzed it together with private data using heatmap, PCA, g:Profiler and GSEA, helped to interpret the analysis results and participated in writing the article.

4.4 Tissue slices in different cultivation conditions (paper II)

Tissue slices are thin slices cut from mouse or human tumor. They allow us to get more study material from a single tumor, but can potentially lead to changes related to the response to stress and cultivation. In this paper, we investigated different culturing conditions on three pathologies (breast, prostate, and lung tumor).

Both IHC (see Section 2.2) and qPCR (see Section 2.5) were used to evaluate the culture induced changes from the parental tumor. Measurements were done immediately after cutting (day zero) and then after culturing on low or atmospheric oxygen. The slices were either floating inside the media or supported by the filter on the media. qPCR data was analyzed using PCA (see Section 3.1.1), heatmap (see Section 3.1.2) and linear models (see Section 3.2.1) showing the number of differentially expressed stress-related genes. This study showed that filter support with atmospheric oxygen is the optimal condition to keep the model as close as possible to the parent tumor.

Using immunohistochemical staining (see Section 2.2), it was also shown that the cells are not equally viable throughout the slice. Cells were found to be more viable the closer they were to the air interface of the slice. The viability gradient was coincident with $HIF1\alpha$ and $\gamma H2AX$, related to hypoxia and stress, respectively. Testing with two other types of filter systems showed similar results. We also tested a new technique called *incubation unit* that allows oxygen to access both sides of the tissue. It showed two-sided gradient going from both sides to the center, confirming the importance of oxygen supply.

My contribution was to perform differential expression analysis and PCA plots, help to interpret the PCA results and participate in writing the article.

4.5 ClustVis web tool for matrix visualization (paper III)

Considering the efforts needed to produce publication quality PCA plots (see Section 3.1.1) and heatmaps (see Section 3.1.2), we developed a separate web tool for custom data. This tool makes exploratory analysis easier for other scientists. Users can upload their data and interactively change the appearance of the plot using graphical control elements (widgets). The aim was to make it easy to use and not require any scripting experience. The tool is freely usable and downloadable at `http://biit.cs.ut.ee/clustvis/`.

Users have multiple options during the analysis.

1. Different data input options allow to upload custom data or choose one of the sample datasets. It is also possible to load one of the large gene

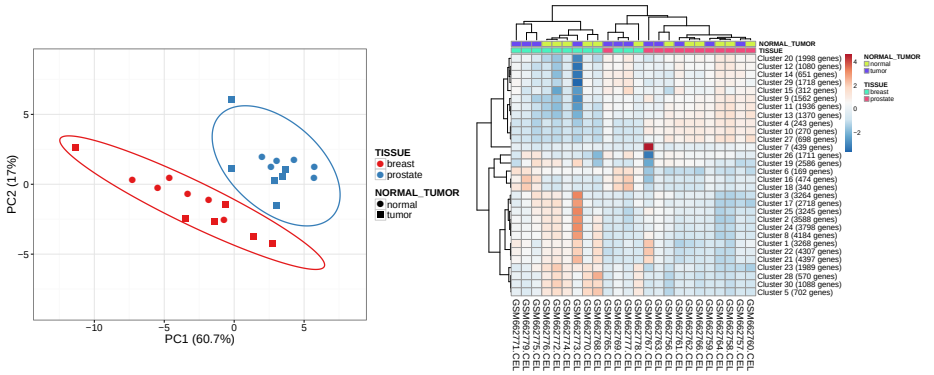

Figure 4.4: PCA plot and heatmap of stromal molecular signatures of breast and prostate samples. Ellipses and shapes on PCA plot and annotations on top of the heatmap show clustering of the samples.

expression microarray datasets from *Multi Experiment Matrix* (MEM) web tool (Adler et al., 2009).

2. Observations can be filtered and, in case of gene expression data, one pathway can be selected or all genes clustered into specified number of clusters.

3. Data transformations and method for calculating principal components can be chosen.

4. Appearance of the generated PCA plot and heatmap can be modified in many ways.

5. The plot can be exported in multiple file formats. The data and intermediate results can be exported as a text file.

To illustrate the output of ClustVis, we consider a gene expression microarray (see Section 2.1) dataset. It is about stromal molecular signatures of breast and prostate cancer samples (Planche et al., 2011, GEO ID: GSE26910) downloaded from *Gene Expression Omnibus* (GEO) database (Edgar et al., 2002). There are 54675 probesets measured in 24 samples, constituting equal groups of six samples from the breast tumor, normal breast, prostate tumor and normal prostate. We aggregate the probesets into 30 clusters using k-means clustering (see Section 3.1.2).

PCA plot of the two first components shows that ~78% of the variation is explained by them, leaving 22% for all other components (see Figure 4.4 left). This proportion is enough for drawing preliminary conclusions. The hidden variability can be explored by looking at further components.

When looking at the ellipses, we can see that breast and prostate tumor form separate clusters. We cannot differentiate normal from tumor samples inside these groups, but we can see that tumor samples vary more than normal samples. This property called tumor heterogeneity (Cusnir and Cavalcante, 2012) is common for cancer and is one of the leading causes why cancers are not easy to cure.

From the annotations of the heatmap, we can also recognize the clusters of breast and prostate samples (see Figure 4.4 right). Two samples on the heatmap (GSM662767 and GSM662773) seem different from others and are worth further investigation. In particular, cluster 7 consisting of 439 genes could be explored further to find out which genes are overexpressed in the sample GSM662767.

As described, ClustVis provides a comfortable way to perform exploratory analysis that can later be extended by confirmatory analysis using other tools. According to *Google Analytics* (`http://www.google.com/analytics/`, January 12, 2016), there have been 500 unique ClustVis users last month visiting the website. My contribution was to implement the web tool and write the article.

CHAPTER 5

IMPRINTED AND MONOALLELICALLY EXPRESSED GENES IN THE HUMAN PLACENTA (PAPER IV)

Each human starts the development in the uterus, connected to the mother through the placenta. This temporary organ supports the growth of the fetus by mediating nutrients, oxygen and waste products between the mother and the child. The development of the placenta is controlled by a specific gene expression pattern that is crucial for a normal pregnancy.

Most genes express equally from the maternal and paternal copy of the gene. There is a small group of genes that violate this general behavior. These constitute imprinted genes where expression depends on the heritage of the allele, and other monoallelically expressed genes that preferentially express from one allele but the parent of origin is not always the same. Dysregulation of such genes is a possible cause for fetal growth abnormalities and pregnancy complications (Piedrahita, 2011). The aim of this study was to find novel monoallelically expressed genes using a genome-wide approach.

We used placentas from ten family trios. The analysis pipeline was adapted from *AlleleSeq* (Rozowsky et al., 2011) and is shown in Figure 5.1. SNP array (see Section 2.3) was used on the DNA of all participants (mother, father, and child) to detect maternal and paternal alleles of the child. RNA of the placenta was sequenced (see Section 2.4). Sequence reads were mapped to constructed maternal and paternal genome, and the best mapping from the two was taken into account. Subsequently, we created a data matrix so that for each informative SNP, the number of maternal and paternal RNA-seq reads were counted. The counts were aggregated to gene level, and imbalance in the read count was tested using binomial test (see Section 3.2.4). The p-values were corrected using FDR (see Section 1.2), genes with significant corrected p-value ($q < 0.05$) were defined as

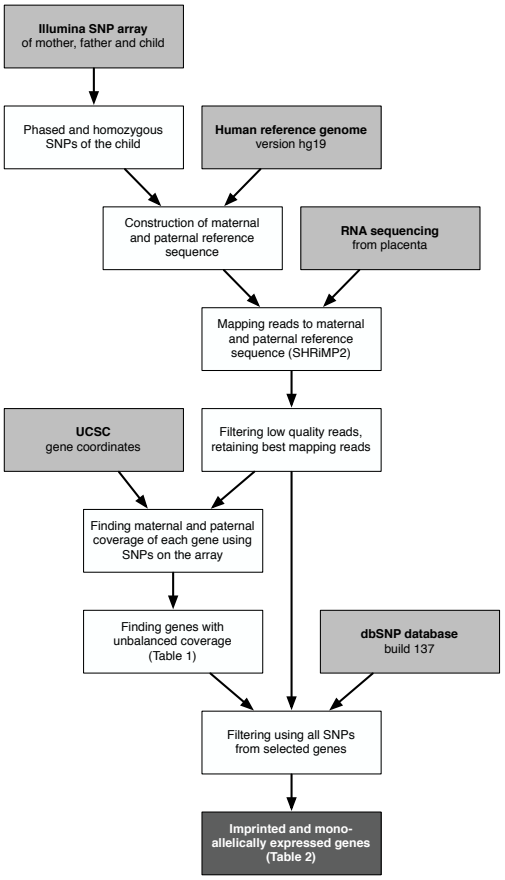

Figure 5.1: Pipeline for finding imprinted and monoallelically expressed genes (Metsalu et al., 2014, Figure 1).

having an allele-specific expression (ASE).

We used another filtering step to keep more confident genes. Besides SNPs on the microarray, we used all SNPs available in dbSNP database (Sherry et al., 2001). A similar pipeline was reused with the exception that SNPs were not aggregated to the gene level. The final list was made of genes where at least 75% of the SNPs showed significant ASE.

The list of 12 genes was further analyzed for differences between trios. Due to the low number of samples, a simple approach was used. The genes were considered having random monoallelic expression pattern if they had evidence for both maternal and paternal expression in at least one trio, and imprinted otherwise. Six genes (*LGALS14*, *SPTLC3*, *AIM1*, *PEG10*, *RHOBTB3*, and *ZFAT-AS1*) showed

the preference towards paternal expression, two (*PAPPA2*, *LGALS8*) towards maternal expression. Four genes (*ABP1*, *BCLAF1*, *IFI30*, and *ZFAT*) had random monoallelic expression pattern. The main functions of these genes include the following (see Metsalu et al. (2014, Table 3)).

- Mediating cellular apoptosis and tissue development;
- Regulating inflammation and immune system;
- Facilitating metabolic processes;
- Regulating cell cycle.

In general, results from high-throughput experiments should always be confirmed using some other method. Here, we chose three genes—*PEG10*, *RHOBTB3*, and *PAPPA2*—for further validation. Regions with multiple SNPs were found from each gene and sequenced using Sanger sequencing (see Section 1.1). All genes showed homozygous RNA expression where the DNA of the newborn was heterozygous, confirming the allele-specific expression. Other genes that we found but were not taken to the validation set need further confirmation by future studies.

My contribution was to implement the whole pipeline for finding imprinted and monoallelically expressed genes, help to interpret the results and participate in writing the article.

CHAPTER 6

MOLECULAR MECHANISMS OF ATOPIC DERMATITIS (PAPER V)

Atopic dermatitis (AD) is an inflammatory skin disease that causes a dry, red, scaly and itchy skin on distinct areas of the human body (Bieber, 2010). Most people with AD get the disease before the age of five, but in some cases, it can also start during adulthood. There is no known cure for AD, but there are treatment plans that can alleviate the symptoms and avoid AD from getting worse.

AD is related with *apoptosis of keratinocytes*—a type of skin cells dying in a controlled way. Enhanced apoptosis is a leading cause of eczema and spongiosis. In this article, we studied the molecular mechanisms of this process.

We showed that patients with AD have increased apoptosis of keratinocytes induced by *IFN*-γ, a characteristic cytokine for T cells. Also, several *IFN*-γ-inducible genes are upregulated in chronic AD lesional skin, and there are also differentially expressed genes related to apoptosis. Some genetic markers in the proximity of *IFN*-γ-inducible genes and apoptosis-related genes were also found. These results propose potential novel targets for developing new drugs against AD.

My contribution was to download and analyze a publicly available gene expression microarray dataset (Nograles et al., 2008, GEO ID: GSE12109) from GEO database (Edgar et al., 2002) for supporting the discussion. Primary keratinocytes treated with *IFN*-γ were compared with untreated keratinocytes using t-test (see Section 3.2.1). Resulting p-values were corrected for multiple testing using FDR (see Section 1.2). All probesets with q-value less than 0.05 were considered significantly differentially expressed.

CHAPTER 7

MICRORNAS AS DIAGNOSTIC MARKERS FOR ENDOMETRIOSIS (PAPER VI)

Endometriosis is a chronic gynecological disease where endometrium—a tissue that will normally be found only inside the uterus—will grow outside the uterus. Typical symptoms include pain and infertility, but the disease can also progress without symptoms. The biological mechanism is still obscure; different theories have been proposed for the cause of the disease (Hickey et al., 2014).

Multiple studies have shown that some microRNAs (miRNAs) have altered expression profile in endometriosis. However, most of these studies have analyzed a predefined subset of miRNAs using microarray or qPCR. In our paper, we used RNA sequencing (see Section 2.4) to study peritoneal endometriotic lesions. Eight potential new miRNA candidates were found, and two known miRNAs (*miR-34c*, *miR-449a*) showed significant differential expression in peritoneal endometriotic lesions compared to healthy controls.

These two differentially expressed miRNAs were selected for further experimental validation using quantitative PCR (see Section 2.5). Three more miRNAs (*miR-200a*, *miR-200b*, and *miR-141*) were added to the validation set since they showed a similar trend in one of the patients and have been found as differentially expressed in previous studies. Experimental validation was done using 22 endometriosis patients and 24 controls. For all five miRNAs, the differential expression was confirmed.

Based on validated results, my contribution was to create a *clinical score formula* (CSF) that can be used to evaluate biopsied samples without the need for healthy controls. Three miRNAs showing the best discrimination power were used to train a linear soft-margin SVM classifier (see Section 3.2.3). The CSF allows to classify the samples with more than 95% accuracy.

CONCLUSIONS

The fast developing field of bioinformatics offers us multiple sources of high-throughput data. Analyzing this data presents major challenges. In the first chapter of this thesis, we presented relevant biological and statistical background. In the next two chapters, we thoroughly described different data sources and analysis methods. Following chapters described the articles included in the thesis. My scientific contribution of them is, in general, twofold.

First, I was involved in data analysis of multiple projects to contribute to novel scientific discoveries. I analyzed data from a novel breast cancer model (paper I), compared tissue slices to find most optimal conditions (paper II), implemented and used a pipeline for finding imprinted and monoallelically expressed genes in the human placenta (paper IV), analyzed mechanisms of atopic dermatitis (paper V), and created a model for predicting endometriosis based on microRNAs (paper VI). All these projects presented different challenges for data preprocessing and analysis.

Second, I created a web tool with an intuitive user interface that can be used by other scientists without necessarily having programming skills. ClustVis was built and made publicly available that allows to upload any dataset and visualize the data using PCA plot and heatmap (paper III). This tool was inspired by analysis and tools built during PREDECT project.

Bibliography

P. Adler, R. Kolde, M. Kull, A. Tkachenko, H. Peterson, J. Reimand, and J. Vilo. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biology*, 10(12):R139, 2009.

C. C. Aggarwal and C. K. Reddy. *Data clustering: algorithms and applications*. CRC Press, 2013.

A. J. Barrett and J. J. Melenhorst. Is human cell therapy research caught in a mousetrap? *Molecular Therapy*, 19(2):224–227, 2011.

D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.

T. Bieber. Atopic dermatitis. *Annals of Dermatology*, 22(2):125–137, 2010.

A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, et al. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68–71, 2003.

M. E. Celebi, H. A. Kingravi, and P. A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.

C. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, pages 404–413, 1934.

N. Cook, D. I. Jodrell, and D. A. Tuveson. Predictive in vivo animal models and translation to clinical trials. *Drug Discovery Today*, 17(5):253–260, 2012.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

P. Cottu, E. Marangoni, F. Assayag, P. de Cremoux, A. Vincent-Salomon, C. Guyader, L. de Plater, C. Elbaz, N. Karboul, J. Fontaine, et al. Modeling

of response to endocrine therapy in a panel of human luminal breast cancer xenografts. *Breast Cancer Research and Treatment*, 133(2):595–606, 2012.

M. Cusnir and L. Cavalcante. Inter-tumor heterogeneity. *Human Vaccines & Immunotherapeutics*, 8(8):1143–1145, 2012.

D. J. Dabbs. *Diagnostic immunohistochemistry*. Elsevier Health Sciences, 2013.

E. J. Davies, M. Dong, M. Gutekunst, K. Närhi, H. J. A. A. van Zoggel, S. Blom, A. Nagaraj, T. Metsalu, E. Oswald, S. Erkens-Schulze, J. A. D. S. Martin, R. Turkki, S. R. Wedge, T. M. af Hällström, J. Schueler, W. M. van Weerden, E. W. Verschuren, S. T. Barry, H. van der Kuip, and J. A. Hickman. Capturing complex tumour biology in vitro: histological and molecular characterisation of precision cut slices. *Scientific Reports*, 5, 2015.

M. de Jong and T. Maina. Of mice and humans: are they the same?–Implications in cancer translational research. *Journal of Nuclear Medicine*, 51(4):501–504, 2010.

R. de Sousa Abreu, L. O. Penalva, E. M. Marcotte, and C. Vogel. Global signatures of protein and mRNA expression levels. *Molecular Biosystems*, 5(12): 1512–1526, 2009.

R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30 (1):207–210, 2002.

K. K. Frese and D. A. Tuveson. Maximizing mouse cancer models. *Nature Reviews Cancer*, 7(9):654–658, 2007.

G. R. Grant, M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, C. J. Stoeckert, J. B. Hogenesch, and E. A. Pierce. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, 2011.

M. Guedj, L. Marisa, A. De Reynies, B. Orsetti, R. Schiappa, F. Bibeau, G. Macgrogan, F. Lerebours, P. Finetti, M. Longy, et al. A refined molecular taxonomy of breast cancer. *Oncogene*, 31(9):1196–1206, 2012.

M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

M. J. Heller. DNA microarray technology: devices, systems, and applications. *Annual Review of Biomedical Engineering*, 4(1):129–153, 2002.

M. Hickey, K. Ballard, and C. Farquhar. Endometriosis. *British Medical Journal*, page g1752, 2014.

J. A. Hickman, R. Graeser, R. de Hoogt, S. Vidic, C. Brito, M. Gutekunst, H. van der Kuip, et al. Three-dimensional models of cancer for pharmacology and cancer cell biology: Capturing tumor complexity in vitro/ex vivo. *Biotechnology Journal*, 9(9):1115–1128, 2014.

M. Jäger, L. Kamm, D. Krushevskaja, H.-A. Talvik, J. Veldemann, A. Vilgota, and J. Vilo. Flexible Database Platform for Biomedical Research with Multiple User Interfaces and a Universal Query Engine. In *DB&IS*, pages 301–310, 2008.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2014.

N. M. Jawhar. Tissue microarray: a rapidly evolving diagnostic and research tool. *Annals of Saudi Medicine*, 29(2):123, 2009.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

O.-P. Kallioniemi, U. Wagner, J. Kononen, and G. Sauter. Tissue microarray technology for high-throughput molecular profiling of cancer. *Human Molecular Genetics*, 10(7):657–662, 2001.

A. Kamb. What's wrong with our cancer models? *Nature Reviews Drug Discovery*, 4(2):161–165, 2005.

E. S. Kawasaki. The end of the microarray Tower of Babel: will universal standards lead the way? *Journal of Biomolecular Techniques: JBT*, 17(3):200, 2006.

T. M. Kodinariya and P. R. Makwana. Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), 2013.

R. Kolde. *pheatmap: Pretty Heatmaps*, 2015. R package version 1.0.7.

J. Kononen, L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O.-P. Kallionimeni. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4(7):844–847, 1998.

M. Kubista, J. M. Andrade, M. Bengtsson, A. Forootan, J. Jonák, K. Lind, R. Sindelka, R. Sjöback, B. Sjögreen, L. Strömbom, et al. The real-time polymerase chain reaction. *Molecular Aspects of Medicine*, 27(2):95–125, 2006.

T. LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, page gkp552, 2009.

J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11 (10):733–739, 2010.

A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

T. Loua. *Atlas statistique de la population de Paris*. J. Dejey & cie, 1873.

K. J. Mantione, R. M. Kream, H. Kuzelova, R. Ptacek, J. Raboch, J. M. Samuel, and G. B. Stefano. Comparing bioinformatic gene expression profiling methods: Microarray and RNA-Seq. *Medical Science Monitor Basic Research*, 20: 138, 2014.

T. Metsalu and J. Vilo. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, 43(W1):W566–W570, 2015.

T. Metsalu, T. Viltrop, A. Tiirats, B. Rajashekar, E. Reimann, S. Kõks, K. Rull, L. Milani, G. Acharya, P. Basnet, J. Vilo, R. Mägi, A. Metspalu, M. Peters, K. Haller-Kikkatalo, and A. Salumets. Using RNA sequencing for identifying gene imprinting and random monoallelic expression in human placenta. *Epigenetics*, 9(10):1397–1409, 2014.

M. L. Metzker. Sequencing technologies–the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposium in Quantitative Biology*, 51:263–273, 1986.

R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J.-P. Coppe, F. Tong, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6):515–527, 2006.

K. Nograles, L. Zaba, E. Guttman-Yassky, J. Fuentes-Duculan, M. Suárez-Fariñas, I. Cardinale, A. Khatcherian, J. Gonzalez, K. Pierson, T. White, et al. Th17 cytokines interleukin (IL)-17 and IL-22 modulate distinct inflammatory and keratinocyte-response pathways. *British Journal of Dermatology*, 159(5): 1092–1102, 2008.

K. Pearson. On Lines and Planes of Closest Fit to System of Points in Space. *Philosophical Magazine*, 2:559–572, 1901.

D. Pelleg and A. W. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *ICML*, pages 727–734, 2000.

J. A. Piedrahita. The role of imprinted genes in fetal growth abnormalities. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 91(8):682–692, 2011.

A. Planche, M. Bacac, P. Provero, C. Fusco, M. Delorenzi, J.-C. Stehle, and I. Stamenkovic. Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. *PLOS ONE*, 6(5):e18640, 2011.

K. Politi and W. Pao. How genetically engineered mouse tumor models provide insights into human cancers. *Journal of Clinical Oncology*, 29(16):2273–2281, 2011.

F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):R95, 2013.

S. Ray and R. H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143, 1999.

A. Rebane, M. Zimmermann, A. Aab, H. Baurecht, A. Koreck, M. Karelson, K. Abram, T. Metsalu, M. Pihlap, N. Meyer, R. Fölster-Holst, N. Nagy, L. Kemeny, K. Kingo, J. Vilo, T. Illig, M. Akdis, A. Franke, N. Novak, S. Weidinger, and C. A. Akdis. Mechanisms of IFN-γ-induced apoptosis of human skin keratinocytes in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology*, 129(5):1297–1306, 2012.

J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, 35(suppl 2):W193–W200, 2007.

J. Reimand, T. Arak, and J. Vilo. g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, 39(suppl 2): W307–W315, 2011.

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*, page gkv007, 2015.

J. Rozowsky, A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, R. Bjornson, Y. Kong, N. Kitabayashi, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, 7(1), 2011.

W. M. S. Russell, R. L. Burch, and C. W. Hume. The principles of humane experimental technique. 1959.

M. Saare, K. Rekker, T. Laisk-Podar, D. Sõritsa, A. M. Roost, J. Simm, A. Velthut-Meikas, K. Samuel, T. Metsalu, H. Karro, A. Sõritsa, A. Salumets, and M. Peters. High-Throughput Sequencing Approach Uncovers the miRNome of Peritoneal Endometriotic Lesions and Adjacent Healthy Tissues. *PLOS ONE*, 9 (11):e112630, 2014.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74 (12):5463–5467, 1977.

G. Sflomos, V. Dormoy, T. Metsalu, R. Jeitziner, L. Battista, A. Treboux, M. Fiche, J. Delaloye, J. Vilo, A. Ayyanan, and C. Brisken. A Novel Preclinical Model for ERα Positive Breast Cancer Points to the Mammary Epithelial Microenvironment as a Critical Determinant of Luminal Phenotype and Hormone Response. Accepted for publication in Cancer Cell.

S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

P. H. Sneath. The application of computers to taxonomy. *Journal of General Microbiology*, 17(1):201–226, 1957.

K. R. Stevenson, J. D. Coolon, and P. J. Wittkopp. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, 14(1):536, 2013.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412, 2006.

V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and remote control*, 24:774–780, 1963.

M. Veta, J. P. Pluim, P. J. van Diest, M. Viergever, et al. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, 2014.

C. Vogel and E. M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4): 227–232, 2012.

C. Vogel, R. de Sousa Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, 6(1), 2010.

T. Voskoglou-Nomikos, J. L. Pater, and L. Seymour. Clinical predictive value of the in vitro cell line, human xenograft, and mouse allograft preclinical cancer models. *Clinical Cancer Research*, 9(11):4227–4239, 2003.

R. Wall and M. Shani. Are animal models as good as we think? *Theriogenology*, 69(1):2–9, 2008.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2), 2009.

M. L. Wong and J. F. Medrano. Real-time PCR for mRNA quantitation. *Biotechniques*, 39(1):75, 2005.

P. Workman, E. Aboagye, F. Balkwill, A. Balmain, G. Bruder, D. Chaplin, J. Double, J. Everitt, D. Farningham, M. Glennie, et al. Guidelines for the welfare and use of animals in cancer research. *British Journal of Cancer*, 102(11):1555–1577, 2010.

M. Yan. *Methods of determining the number of clusters in a data set and a new clustering criterion*. PhD thesis, Virginia Tech, 2005.

X. Zhang, S. Claerhout, A. Prat, L. E. Dobrolecki, I. Petrovic, Q. Lai, M. D. Landis, L. Wiechmann, R. Schiff, M. Giuliano, et al. A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer Research*, 73(15):4885–4897, 2013.

ACKNOWLEDGEMENTS

I thank my supervisor Prof. Jaak Vilo for endless ideas and support, and for creating a perfect working environment. My scientific development has been supported by discussions with many BIIT members and co-authors of the papers, too many to mention all of them here. Special thanks go to Raivo Kolde, who introduced me bioinformatics. Thank you, Konstantin Tretyakov, Meelis Kull, Anna Leontjeva, Priit Adler, Emma Davies and Georgios Sflomos for giving valuable feedback about my thesis.

My doctoral studies have been supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115188, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and European Federation of Pharmaceutical Industries and Associations (EFPIA) companies' in kind contribution; the European Regional Development Fund through the Estonian Center of Excellence in Computer Science (EXCS); Estonian Research Council (grant IUT34-4); Tiger University program of the Information Technology Foundation for Education; European Social Fund's Doctoral Studies and Internationalisation Programme DoRa; IT Academy and EMT-Elion scholarship.

I thank my parents and, last and most importantly, my wife Mailis, who supports me in all of my activities.

KOKKUVÕTE
(SUMMARY IN ESTONIAN)

MITMEMÕÕTMELISTE ANDMETE STATISTILINE ANALÜÜS BIOINFORMAATIKAS

Organismi ühed tähtsaimad molekulid on valgud, mille ehitus on kodeeritud pärilikkusaines – DNAs. Erinevate valkude kogust uurides on võimalik saada infot organismi seisundi kohta. Tänapäevased seadmed võimaldavad valkudega seotud andmeid koguda lühikese aja jooksul väga suurel hulgal. Suuremahuliste andmete analüüs vajab erinevaid tehnilisi oskusi ja sellega seoses on tekkinud uus teadusharu – bioinformaatika.

Dissertatsiooni eesmärgiks on kirjeldada mitmemõõtmeliste andmete statistilise analüüsiga seotud probleeme ja nende lahendusi. Näidatakse, et sellised andmed on esitatavad maatriksina. Kirjeldatakse üleeuroopalist konsortsiumit, kus andmeid kogutakse paljudelt parnteritelt ja väga tähtis on koguda ka metaandmed struktureeritud kujul. Vaadeldakse erinevaid uuringuid, kus on tekkinud vajadus andmeid analüüsida. Luuakse graafilise kasutajaliidesega veebitööristad, et vähendada andmete analüüsiks vajalikku tehniliste oskuste hulka ja teha mõned tüüpanalüüsid kättesaadavaks ka nendele, kes pole andmete analüüsiga varem kokku puutunud.

Esimeses peatükis tuuakse lühidalt välja vajalikud taustateadmised. Bioloogilises osas kirjeldatakse inimese genoomi ülesehitust ja selle funktsioone. Tutvustatakse väga lühidalt päriliku info masinloetavale kujule viimise põhimõtteid. Statistilises osas tehakse ülevaade andmeanalüüsi etappidest ja statistilise testimise metoodikast. Muuhulgas kirjeldatakse bioinformaatikas tihti esinevaid probleeme nagu mitmene testimine ja ülesobitamine.

Teises peatükis tehakse ülevaade erinevatest andmeallikatest. *Koe mikrokiibid*

võimaldavad mõõta korraga ühe valgu ekspressiooni paljudes proovides. Geeniekspressiooni on võimalik mõõta erinevate platvormidega nagu *geeniekspressiooni mikrokiip*, *RNA sekveneerimine* ja *pöördtranskriptsiooniga kvantitatiivne polümeraasi ahelreaktsioon*. Kõigil neist on oma eelised ja puudused ning platvormi valikul tuleb arvesse võtta, mitut geeni soovitakse analüüsida, kui palju on analüüsiks võimalik raha kulutada ning kui hästi peavad tulemused olema korratavad. Kirjeldatakse ka *ühenukleotiidiliste erinevuste* tuvastamiseks mõeldud mikrokiipi.

Kolmandas peatükis tutvustatakse statistilisi analüüsimeetodeid. Need jagatakse *kirjeldavateks*, mille abil on võimalik saada andmetest ülevaade ilma konkreetseid hüpoteese püstitamata, ja *kinnitavateks*, mille korral otsitakse vastust konkreetsele küsimusele.

Kirjeldavate meetodite hulka kuulub *peakomponentide analüüs*, mis võimaldab teisendada andmeid nii, et võimalikult suur hulk varieeruvusest oleks kirjeldatud väikese arvu esimeste komponentide abil. *Klasterdamine* võimaldab objekte sarnasuse põhjal rühmitada. Erinevatest klasterdusmeetoditest vaadeldakse *k-keskmiste klasterdamist*, kus klastrite arv peab olema varasemalt teada, ja *hierarhilist klasterdamist*, kus luuakse objektidest sarnasuse põhjal puukujuline struktuur. Populaarne viis klasterduse joonisel näitamiseks on *soojuskaart*, kus arve näidatakse erinevate värvitugevuste abil ning read ja/või veerud on tavaliselt klasterdatud.

Kinnitavatest meetoditest vaadeldakse *diferentsiaalse ekspressiooni* meetodeid, mis aitavad leida, kas ekspressioon on konkreetse geeni korral rühmades erinev. Kahe rühma võrdlemiseks sobib *t-test*, suurema arvu rühmade jaoks *lineaarsed mudelid*. *Binoomtestiga* saab samuti leida, kas ekspressioon jaguneb kahe grupi vahel võrdselt, kuid see eeldab, et suudame seda täisarvuliselt loendada. *Tugivektormasinaga* on võimalik leida teatud mõttes parim klassifitseerija, mis võimaldab kahe rühma objekte eristada. Üksikute geenide asemel on võimalik analüüsida ka geenigruppe. *Geenigrupi rikastatuse analüüs* leiab, kas geenigrupp käitub kahes rühmas erinevalt. *Hüpergeomeetriline test* võimaldab leida, kas konkreetne geenigrupp kattub oluliselt mõne varem teada oleva grupiga.

Neljandas peatükis tutvustatakse üleeuroopalist vähiuuringute projekti PREDECT. Tehakse ülevaade metaandmete kogumisest ja andmete esmaseks analüüsiks loodud tööriistadest. Kirjeldatakse uue rinnavähi mudeliga seotud analüüsi ning samuti koelõikude mudelite võrdlust erinevates laboritingimustes. Tutvustatakse vabalt kasutatavat veebitööriista, mille abil on peakomponentide analüüsi joonise ja soojuskaardi tegemine lihtsam.

Järgmised peatükid kirjeldavad andmeanalüüsi erinevates projektides. Viiendas peatükis tuvastatakse mitmeid uusi geene, mille korral esineb inimese platsentas alleelispetsiifiline ekspressioon. Kuuendas peatükis vaadeldakse atoopili-

se dermatiidi molekulaarseid mehhanisme, täpsemalt valgu *IFN-*γ mõju. Seitsmendas peatükis leitakse mikroRNAd, mis sobivad endometrioosi markeriteks, ja luuakse klassifitseerija endometrioosihaigete eristamiseks tervetest.

CURRICULUM VITAE

Personal data

Name	Tauno Metsalu
Birth	May 22nd 1986 Tallinn, Estonia
Citizenship	Estonian
Marital Status	Married
Languages	Estonian, English, German
Address	Puiestee 24-2, 50303, Tartu, Estonia
Contact	+372 56669250 tauno.metsalu@eesti.ee

Education

2011–	University of Tartu, computer science PhD student
2009–2011	University of Tartu, MSc in mathematical statistics (*cum laude*)
2008–2009	Military service
2005–2008	University of Tartu, BSc in mathematical statistics (*cum laude*)
1993–2005	Tallinn German Gymnasium of Kadriorg, secondary education (with silver medal)

Employment

2011–	University of Tartu, Institute of Computer Science, programmer
2010–2011	Quretec OÜ, scientist
2007–2010	Swedbank AS, junior specialist

Scholarships

2013	Tiger University program scholarship
2012	IT Academy scholarship
2012	DoRa travel scholarship for attending ECCB conference
2012	EMT-Elion scholarship
2011	Videomat scholarship
2010	DoRa travel scholarship for attending Erice summer school
2007	Hansapank scholarship

Publications

1. G. Sflomos, V. Dormoy, T. Metsalu, R. Jeitziner, L. Battista, A. Treboux, M. Fiche, J. Delaloye, J. Vilo, A. Ayyanan, and C. Brisken. A Novel Preclinical Model for ERα Positive Breast Cancer Points to the Mammary Epithelial Microenvironment as a Critical Determinant of Luminal Phenotype and Hormone Response. Accepted for publication in Cancer Cell.

2. E. J. Davies, M. Dong, M. Gutekunst, K. Närhi, H. J. A. A. van Zoggel, S. Blom, A. Nagaraj, T. Metsalu, E. Oswald, S. Erkens-Schulze, J. A. D. S. Martin, R. Turkki, S. R. Wedge, T. M. af Hällström, J. Schueler, W. M. van Weerden, E. W. Verschuren, S. T. Barry, H. van der Kuip, and J. A. Hickman. Capturing complex tumour biology in vitro: histological and molecular characterisation of precision cut slices. *Scientific Reports*, 5, 2015.

3. T. Metsalu and J. Vilo. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, 43(W1):W566–W570, 2015.

4. M. Saare, K. Rekker, T. Laisk-Podar, D. Sõritsa, A. M. Roost, J. Simm, A. Velthut-Meikas, K. Samuel, T. Metsalu, H. Karro, A. Sõritsa, A. Salumets, and M. Peters. High-Throughput Sequencing Approach Uncovers the miRNome of Peritoneal Endometriotic Lesions and Adjacent Healthy Tissues. *PLOS ONE*, 9(11):e112630, 2014.

5. T. Metsalu, T. Viltrop, A. Tiirats, B. Rajashekar, E. Reimann, S. Kõks, K. Rull, L. Milani, G. Acharya, P. Basnet, J. Vilo, R. Mägi, A. Metspalu, M. Peters, K. Haller-Kikkatalo, and A. Salumets. Using RNA sequencing for identifying gene imprinting and random monoallelic expression in human placenta. *Epigenetics*, 9(10):1397–1409, 2014.

6. A. Rebane, M. Zimmermann, A. Aab, H. Baurecht, A. Koreck, M. Karelson, K. Abram, T. Metsalu, M. Pihlap, N. Meyer, R. Fölster-Holst, N. Nagy, L. Kemeny, K. Kingo, J. Vilo, T. Illig, M. Akdis, A. Franke, N. Novak, S. Weidinger, and C. A. Akdis. Mechanisms of IFN-γ-induced apoptosis of human skin keratinocytes in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology*, 129(5):1297–1306, 2012.

ELULOOKIRJELDUS

Isikuandmed

Nimi	Tauno Metsalu
Sünniaeg ja -koht	22. mai 1986 Tallinn, Eesti
Kodakondsus	Eesti
Perekonnaseis	abielus
Keelteoskus	eesti, inglise, saksa
Aadress	Puiestee 24-2, 50303, Tartu, Eesti
Kontaktandmed	+372 56669250 tauno.metsalu@eesti.ee

Haridustee

2011–	Tartu Ülikool, informaatika doktorant
2009–2011	Tartu Ülikool, MSc matemaatilises statistikas (*cum laude*)
2008–2009	Ajateenistus
2005–2008	Tartu Ülikool, BSc matemaatilises statistikas (*cum laude*)
1993–2005	Tallinna Kadrioru Saksa Gümnaasium, keskharidus (hõbemedaliga)

Teenistuskäik

2011–	Tartu Ülikool, Arvutiteaduse instituut, programmeerija
2010–2011	Quretec OÜ, teadur
2007–2010	Swedbank AS, nooremspetsialist

Stipendiumid

2013	Tiigriülikooli stipendium
2012	IT Akadeemia stipendium
2012	DoRa välislähetuse toetus ECCB konverentsil osalemiseks
2012	EMT-Elioni stipendium
2011	Videomati stipendium
2010	DoRa välislähetuse toetus Erice suvekoolis osalemiseks
2007	Hansapanga majandusstipendium

Publikatsioonid

1. G. Sflomos, V. Dormoy, T. Metsalu, R. Jeitziner, L. Battista, A. Treboux, M. Fiche, J. Delaloye, J. Vilo, A. Ayyanan, and C. Brisken. A Novel Preclinical Model for ERα Positive Breast Cancer Points to the Mammary Epithelial Microenvironment as a Critical Determinant of Luminal Phenotype and Hormone Response. Accepted for publication in Cancer Cell.

2. E. J. Davies, M. Dong, M. Gutekunst, K. Närhi, H. J. A. A. van Zoggel, S. Blom, A. Nagaraj, T. Metsalu, E. Oswald, S. Erkens-Schulze, J. A. D. S. Martin, R. Turkki, S. R. Wedge, T. M. af Hällström, J. Schueler, W. M. van Weerden, E. W. Verschuren, S. T. Barry, H. van der Kuip, and J. A. Hickman. Capturing complex tumour biology in vitro: histological and molecular characterisation of precision cut slices. *Scientific Reports*, 5, 2015.

3. T. Metsalu and J. Vilo. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, 43(W1):W566–W570, 2015.

4. M. Saare, K. Rekker, T. Laisk-Podar, D. Sõritsa, A. M. Roost, J. Simm, A. Velthut-Meikas, K. Samuel, T. Metsalu, H. Karro, A. Sõritsa, A. Salumets, and M. Peters. High-Throughput Sequencing Approach Uncovers the miRNome of Peritoneal Endometriotic Lesions and Adjacent Healthy Tissues. *PLOS ONE*, 9(11):e112630, 2014.

5. T. Metsalu, T. Viltrop, A. Tiirats, B. Rajashekar, E. Reimann, S. Kõks, K. Rull, L. Milani, G. Acharya, P. Basnet, J. Vilo, R. Mägi, A. Metspalu, M. Peters, K. Haller-Kikkatalo, and A. Salumets. Using RNA sequencing for identifying gene imprinting and random monoallelic expression in human placenta. *Epigenetics*, 9(10):1397–1409, 2014.

6. A. Rebane, M. Zimmermann, A. Aab, H. Baurecht, A. Koreck, M. Karelson, K. Abram, T. Metsalu, M. Pihlap, N. Meyer, R. Fölster-Holst, N. Nagy, L. Kemeny, K. Kingo, J. Vilo, T. Illig, M. Akdis, A. Franke, N. Novak, S. Weidinger, and C. A. Akdis. Mechanisms of IFN-γ-induced apoptosis of human skin keratinocytes in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology*, 129(5):1297–1306, 2012.

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. **Eno Tõnisson.** Solving of expession manipulation exercises in computer algebra systems. Tartu, 2002, 92 p.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.

43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.

64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.
74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
82. **Vitali Retšnoi.** Vector fields and Lie group representations. Tartu 2012, 108 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
85. **Erge Ideon**. Rational spline collocation for boundary value problems. Tartu, 2013, 111 p.
86. **Esta Kägo**. Natural vibrations of elastic stepped plates with cracks. Tartu, 2013, 114 p.

87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
88. **Boriss Vlassov.** Optimization of stepped plates in the case of smooth yield surfaces. Tartu, 2013, 104 p.
89. **Elina Safiulina.** Parallel and semiparallel space-like submanifolds of low dimension in pseudo-Euclidean space. Tartu, 2013, 85 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
93. **Kerli Orav-Puurand.** Central Part Interpolation Schemes for Weakly Singular Integral Equations. Tartu, 2014, 109 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
95. **Kaido Lätt.** Singular fractional differential equations and cordial Volterra integral operators. Tartu, 2015, 93 p.
96. **Oleg Košik.** Categorical equivalence in algebra. Tartu, 2015, 84 p.
97. **Kati Ain.** Compactness and null sequences defined by ℓ_p spaces. Tartu, 2015, 90 p.
98. **Helle Hallik.** Rational spline histopolation. Tartu, 2015, 100 p.
99. **Johann Langemets.** Geometrical structure in diameter 2 Banach spaces. Tartu, 2015, 132 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.