Computational methods for the identification and characterization of tissues and cells

Dissertation zur Erlangung des akademischen Grades des Doktors der Naturwissenschaften *(Dr. rer. nat.)*

vorgelegt von

Khadija El Amrani

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Berlin 2017

Erstgutachter:	Prof. Dr. Peter N. Robinson
Zweitgutachter:	Prof. Dr. Miguel Andrade
Drittgutachter:	Dr. Andreas Kurtz

Tag der Disputation: 12.02.2018

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorgelegte Dissertation selbstständig angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind.
Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen in- oder ausländischen Hochschule oder bei einem anderen Fachbereich beantragt. Die Bestimmungen der Promotionsordnung sind mir bekannt.

Berlin, 18.09.2017 ______________________

Khadija El Amrani

Acknowledgements

First and foremost, I would like to thank Dr. Andreas Kurtz, for giving me the opportunity to complete my PhD thesis in his group, for allowing me the freedom to pursue my own ideas, for helpful input and discussions, and especially for supporting my two months as a visiting scientist at Dr. Wataru Fujibuchi's lab in the Center for iPS Cell Research and Application (Kyoto, Japan).

I especially want to thank my supervisor Prof. Peter Robinson for his assistance and help.

Further, I would like to thank Prof. Miguel Andrade for his valuable advice and support in pursuance of this work.

I am deeply grateful to Dr. Nancy Mah for all her guidance, ideas, several interesting discussions, the careful proof-reading of this thesis, and for sharing her knowledge and experience with me. She has basically been my most important advisor throughout my PhD project. I also want to thank Vindi Jurinovic for proof-reading some parts of this thesis.

Thanks also to all the former and current members in the Dr. Kurtz's lab.

I also want to thank Dr. Junko Yamane for helping me during my stay in Kyoto, and for assistance with the primer design, during her visit to the BCRT. Thanks also to Dr. Kunie Sakurai and all the wonderful people that I had the opportunity to meet in Japan.

Further, I would like to acknowledge all researchers who deposit their experimental data in public databases, without them the validation of the presented tools in this work could not have been realized.

My deepest heartfelt appreciation goes to all my family and friends, particularly I want to express deep gratitude to my parents, my sisters and my brothers for their steady love, support, and encouragement. Words cannot express how thankful I am to have you all.

Last but not least, I am especially grateful and thankful to my husband for believing in me and being there with me every step of the way.

Finally, big thank you to everyone that I couldn't mention here, who have helped and encouraged me one way or the other.

Abstract

Identification of marker genes associated with a specific tissue or cell type, and discrimination between different classes of samples such as different cell types or tissues using gene expression profiles are important problems in cell research. Comparing the gene expression profiles of different types of samples is of major importance for understanding differentiation, development and disease.

In this thesis, I present new bioinformatics tools to detect marker genes and classify samples using gene expression profiles. These contributions can be divided into three sub-projects: First, I optimized and extended the marker tool `MGFM` (Marker Gene Finder in Microarray gene expression data) to support the detection of marker genes from RNA-seq data. For this purpose, I implemented an R package called `MGFR` (Marker Gene Finder in RNA-seq data). Furthermore, I present a comparison study between microarrays and RNA-seq. I identify robust marker genes (predicted by both `MGFM` and `MGFR`) for a set of 16 human tissues, and suggest novel candidate marker genes for each of the examined tissues. Next, I compare the set of predicted marker genes to a gold-standard list of marker genes obtained from the Tissue-specific Gene Expression and Regulation (TiGER) database. In addition, I validated the expression of top ranked marker genes by reverse transcriptase-polymerase chain reaction (RT-PCR) for a set of five tissues.

Second, I developed `sampleClassifier`, a novel computational method, which uses a simple algorithm called "Shared Marker Genes" (SMG) to classify samples based on their gene expression profiles. As the name suggests, the number of shared marker genes between a reference and a query sample is used as a similarity measure. I demonstrate the utility and effectiveness of the proposed approach by the classification of different tissues using public microarray and RNA-seq datasets. Furthermore, I compared my tool to a Support Vector Machines (SVMs) classifier. My approach performed comparably or better than SVMs. The SMG algorithm is implemented as an R package, which is available from the Bioconductor repository (http://www.bioconductor.org).

Finally, I apply `MGFM` and `sampleClassifier` to publicly available biopsy-based microarray gene expression data from eight diverse kidney diseases. I identify marker genes for each of the examined diseases, and demonstrate the performance of the classification tool in distinguishing between normal and disease samples, as well as between different types of renal diseases.

List of abbreviations

CKD	Chronic Kidney Disease
DNA	Deoxyribonucleic Acid
ESCs	Embryonic Stem Cells
ESRD	End-Stage Renal Disease
EST	Expressed Sequence Tag
FPKM	Fragments Per Kilobase of exon model per Million mapped reads
GEO	Gene Expression Omnibus database
iPSCs	Induced Pluripotent Stem Cells
MGFM	Marker Gene Finder in Microarray gene expression data
MGFR	Marker Gene Finder in RNA-seq data
NGS	Next Generation Sequencing
RMA	Robust Multiarray Average
RNA	Ribonucleic Acid
RNA-seq	RNA deep-sequencing
RT-PCR	Reverse Transcriptase-Polymerase Chain Reaction
SMG	Shared Marker Genes
SVMs	Support Vector Machines
TiGER	Tissue-specific Gene Expression and Regulation database

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

As the basic structural and functional unit of life, cells play a role in multiple aspects of human physiology, including development and disease. Cells associate to form tissues that carry out particular physiological functions, which in turn organize in a specific manner to form functional structures or organs. Although each somatic cell within an organism contains the same deoxyribonucleic acid, or DNA (a few exceptions include red blood cells and some immune system cells), the gene expression is specific for each cell type and it determines cell structure and functions. The identification of specific genes selectively expressed in various cells, tissues and organs, including various disease states, is of considerable interest for understanding the gene function, the molecular mechanisms underlying complex diseases, and may lead to the development of new therapeutic targets. Throughout this thesis, I will refer to the genes that are highly expressed in specific cells or tissues as marker genes.

The use of high-throughput technologies such as microarrays or RNA-seq for gene expression profiling has revolutionized genetic and biomedical research. Consequently, vast amounts of data of gene expression have accumulated in many public repositories, such as Gene Expression Omnibus (GEO) (Barrett et al., 2013) and ArrayExpress (Rustici et al., 2013). Hence, computational methods to make use of these data are strongly in demand. In this work I introduce computational methods to address two important issues: i) the problem of finding marker genes to distinguish and compare between different cell types and ii) sample class prediction or classification using gene expression data from both normal tissues and disease states. I focus on genome-wide expression data from microarray and RNA-seq technologies.

1.2 Biological background

Deoxyribonucleic acid (DNA) is a molecule that encodes the genetic instructions used in the growth, development, function and reproduction of all known living organisms and many viruses. DNA was first isolated by Friedrich Miescher in 1869. Miescher called the novel substance "nuclein", as it was located in the nucleus of each cell. The term "nuclein" was later changed to "nucleic acid" and eventually to "deoxyribonucleic acid" or "DNA". The molecular structure of DNA was identified by James Watson and Francis Crick in 1953, based on physical and chemical data generated by other labs (Watson and Crick, 1953). The major contributors to the model were Rosalind Franklin and Maurice Wilkins and their X-ray diffraction data. In 1962, Watson, Crick and Wilkins received the Nobel Prize in Physiology or Medicine for this discovery.

DNA is a double stranded helix (Figure 1.1.a) composed of nucleotides. Each nucleotide consists of a deoxyribose (5-carbon sugar), a phosphate group, and a nitrogenous base. Attached to each sugar is one of four bases: adenine (A), cytosine (C), guanine (G), or thymine (T). A and G are classified as purines. The primary structure of a purine consists of two carbon-nitrogen rings. C and T are classified as pyrimidines which have a single carbon-nitrogen ring as their primary structure. The strands are held together by formation of base pairs: A is paired with T through two hydrogen bonds; G is paired with C through three hydrogen bonds (Figure 1.1.b). The two strands are anti-parallel in nature; that is, one strand will have the 3’ carbon of the sugar in the "upward" position, whereas the other strand will have the 5’ carbon in the upward position.

The second major nucleic acid present in cells is ribonucleic acid or RNA. RNA molecules are also linear polymers, but are much smaller than genomic DNA. In most forms of RNA molecule there are also just four bases, three being the same as in DNA, but thymine is replaced by the pyrimidine uracil (U). In contrast to DNA, almost all RNA molecules in living systems are single stranded. One major type of RNA molecule, called messenger RNA (mRNA), provides the information for the ribosome to catalyze protein synthesis in a process called translation. The processes of transcription and translation are collectively referred to as gene expression. A gene is defined as the union of genomic sequences encoding a coherent set of potentially overlapping functional products (Gerstein et al., 2007). Genes are organized and packaged in units called "chromosomes". The sum of all the genes and intergenic DNA on all the different chromosomes of a cell is referred to as the cellular genome. Gene expression is the synthesis of a functional gene product from the information that is encoded in the gene. Gene products are often proteins, however non-protein coding genes can encode functional RNA (e.g., ribosomal RNA (rRNA), transfer RNA (tRNA)). Proteins are large biomolecules, or macromolecules, consisting of one or more long chains of amino acid residues. Proteins are essential to the cell and serve many different functions. Many proteins are enzymes that catalyze almost all biological reactions in a living organism. Other proteins perform a structural role for the cell - either in the cell wall, the cell membrane or in the cytoplasm.

The flow of genetic information from DNA to mRNA to protein is described by the central dogma (Figure 1.2). The central dogma of molecular biology postulated by Francis Crick (Crick Mc, 1970) states that genes specify the sequences of mRNAs, which in turn specify the sequences of proteins.

The regulation of gene expression is a crucial aspect of proper cell function, so that different cell types express different subsets of genes. Measuring mRNA levels can provide a detailed molecular view of the subset of genes expressed in different cell and tissue types under different conditions.

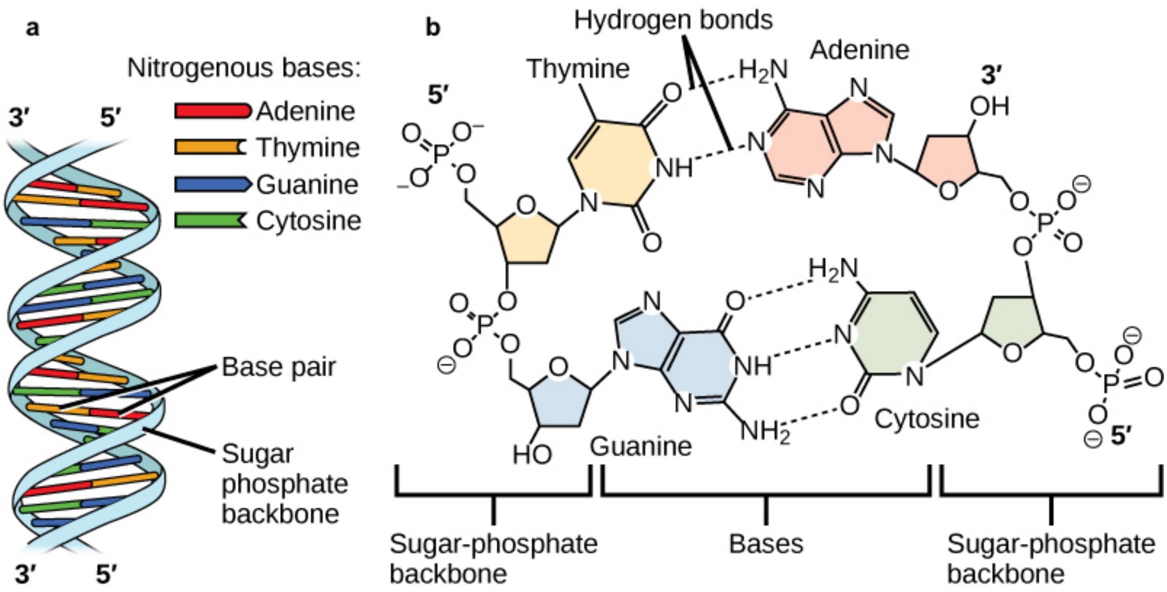


Figure 1.1: Structure of DNA. (a) DNA forms a double stranded helix, and (b) adenine pairs with thymine and cytosine pairs with guanine. *Figure Source:* (https://opentextbc.ca/biology/chapter/9-1-the-structure-of-dna/)

1.3 Microarray technology

The term microarray was first introduced by Schena et al. (Schena et al., 1995) and the first genome of an eukaryotic species completely investigated (Saccharomyces cerevisiae) by a microarray was published in 1997 (Lashkari et al., 1997). A typical microarray consists of

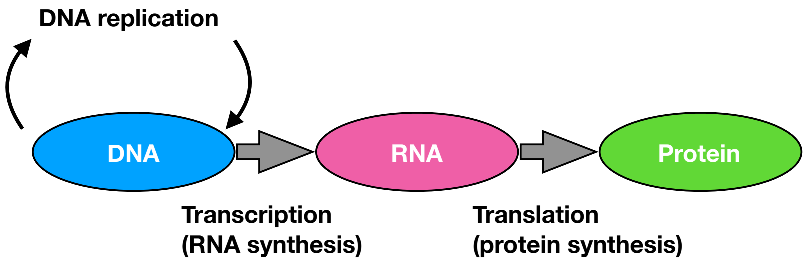


Figure 1.2: Central dogma of molecular biology. DNA is first transcribed into messenger RNA (mRNA), which is then translated into proteins. This process is called gene expression.

known biological molecules, probes, affixed to a solid support, which can be a glass slide, a custom surface, or a membrane. These probes bind their labeled targets, and the resulting signal is analyzed computationally. The key to microarray technology is that a probe is detected at a level that is proportional in a predictable way to the amount of its target present in the labeled extract (Wheelan et al., 2008). There are many types of microarrays, based on the biological materials immobilized on the solid substrate and the purpose of the microarray, including: i) DNA microarrays, such as oligonucleotide microarrays, single nucleotide polymorphism (SNP) arrays, and methylation microarrays; ii) protein microarrays for detailed analysis or optimization of protein-protein interactions; iii) transfection microarrays; iv) tissue microarrays. The focus of this thesis is on gene expression or DNA microarrays that are typically used for measuring relative mRNA expression abundances.

DNA microarrays

DNA microarrays are a high-throughput technology used to simultaneously measure the expression level of thousands of genes within a particular mRNA sample (Schena et al., 1995). The technology takes advantage of the ability of the complementary single-stranded sequences of nucleic acids (DNA or RNA) to form double stranded hybrids. The microscopic spots on a microarray contain single stranded DNA oligonucleotides called probes. Each of these spots contain DNA which is of a complementary sequence to the specific DNA molecule that corresponds to the gene that it is targeting.

The first step in microarray experiments is target preparation, in which RNA from cells or tissue of interest is extracted and either, labeled directly, converted to a labeled cDNA or converted to a T7 RNA promoter tailed cDNA which is further converted to cRNA. A variety of methods have been developed for labeling of the cDNA or cRNA including: incorporation of fluorescently labeled nucleotides during the synthesis, incorporation of biotin labeled nucleotide which is subsequently stained fluorescently labeled streptavidin, incorporation of a modified reactive nucleotide to which a fluorescent tag is added later, and a variety of signal amplification methods (Bumgarner, 2013). The two most frequently used methods are the incorporation of fluorescently labeled nucleotides in the cRNA or cDNA synthesis step or the incorporation of a biotin labeled nucleotide in the cRNA synthesis step (as is done by Affymetrix).

The labeled cRNA or cDNA are then hybridized to the microarray. In this way, the amount of hybridization that has taken place can be measured by the level of fluorescence at each spot, which is detected by a scanner. This scanner then outputs a text file for each array, which contains the relevant data pertaining to that array, such as the intensities of each spot and the level of background noise. These text files are further computationally analyzed. The aim is to calculate the intensity of each spot. In theory, the fluorescence intensity of each spot on the array is proportional to the concentration of target bound to that spot, which in turn is proportional to the amount of target in the original solution.

DNA microarrays can be manufactured in different ways, depending on the number of probes under examination, costs, customization requirements, manufacturers, etc. There are several microarray manufacturers, the most prominent ones are Affymetrix and Illumina. There are mainly two different types of DNA microarrays, namely cDNA microarrays and oligonucleotide microarrays.

The general process in microarray experiments is depicted in Figure 1.3. In cDNA microarrays (Figure 1.3.a), cDNAs prepared from two samples of interest (e.g. diseased and healthy tissue), are labeled with fluorescent dyes of different color (usually red Cy3 dye and green Cy5 dye), and hybridized to a single chip. The relative level of gene expression in the two samples is then measured as the logarithmic ratio between the intensities of the dyes.

In oligonucleotide microarrays (Figure 1.3.b), only one sample is hybridized per chip, and estimations of the absolute levels of gene expression are given. In Affymetrix arrays, each gene is typically represented by a set of 11-20 pairs of oligonucleotides, each 25 bases long, referred to as probeset. To improve the quantification accuracy, each pair of probes consist of perfect match probes (PM), which are perfect matching 25-mer oligos to the target transcripts, and corresponding mismatch probes (MM), which contain sequences with the 13th position of the corresponding PM sequence being modified to the complement nucleotide. The expression level for a gene is a summary of the data from the entire probeset.

Pre-processing and normalization are essential steps in data analysis of gene expression microarray data, to obtain reliable estimates of relative abundances for each gene. The purpose of normalization is to adjust the effects that arise from variations in the microarray technology rather than from biological differences between the RNA samples or between the printed probes. Normalization is an essential step to compare measurements from different array hybridizations due to many diverse sources of variation. These include different efficiencies of reverse transcription, labeling, or hybridization reactions, physical problems with the arrays, reagent batch effects and laboratory conditions. Without a proper normalization, the comparison of data across arrays can provide misleading results. However, normalization procedures do not adjust the data for batch effects. Here, the term "batch effects" refers to experimental variations of datasets generated by different labs. When combining batches of data (particularly batches that contain large batch-to-batch variation), normalization is not sufficient for adjusting for batch effects and other procedures must be applied (Johnson et al., 2007).

1.4 RNA sequencing

RNA sequencing (RNA-seq) is a high-throughput technology, which uses the capabilities of next generation sequencing (NGS) methods for comprehensive transcriptome study. RNA-seq enables precise quantification of gene and transcript levels compared to other methods. RNA-seq has a wide variety of applications, such as differential expression, novel transcripts detection, splice junction analysis, de novo assembly, and SNP analysis. There are various high-throughput sequencing platforms such as Illumina, Roche 454 Life Science, and Sequencing by Oligonucleotide Ligation Detection (SOLiD). Advantages of RNA-seq over microarrays include greater dynamic range, higher sensitivity, and the ability to characterize RNA sequences without prior genomic information. The latter makes RNA-seq particularly attractive for transcriptome profiling in non-model organisms without a reference genome. A typical RNA-seq procedure is depicted in Figure 1.4. Briefly, a population of RNA (total or fractionated, such as poly(A)+) is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing). The reads can range from 30–400 bases, depending on the DNA-sequencing technology used. In principle, any high-throughput sequencing technology can be used for RNA-seq (Wang et al., 2009).

After sequencing, the first step in the bioinformatics analysis of RNA-seq data is the mapping of the short reads from RNA-seq to the reference genome (if a genome sequence is available for the studied organism), or assembling reads de novo into contigs and then mapping them onto the transcriptome. There are several programs for mapping reads to the genome, such as Bowtie (Langmead et al., 2009), BWA (Li and Durbin, 2010), SOAP (Li et al., 2008b), MAQ (Li et al., 2008a), RMAP (Smith et al., 2008), and TopHat (Trapnell et al., 2009). In this thesis the RNA-seq data were processed using the set of open source software programs of the Tuxedo suite: TopHat and Cufflinks (Trapnell et al., 2010). TopHat (http://ccb.jhu.edu/software/tophat/index.shtml) aligns reads to the genome and discovers transcript splice sites. These alignments are used during downstream analysis in several ways. Cufflinks (http://cufflinks.cbcb.umd.edu/) uses this map against the genome to assemble the reads into transcripts (Trapnell et al., 2012). Cuffnorm, a part of the Cufflinks package, takes the aligned reads from two or more conditions and generates tables of expression values that are properly normalized for library size.

1.5 Classification

In machine learning and statistics, classification or class prediction is the problem of identifying to which of a set of classes a new observation belongs, based on a training set of vectors whose classification is known *a priori*. In general, class prediction can deal with a two-class (binary) or multi-class classification problem. In gene expression experiments, classification of data is a crucial step for the prediction of phenotype of cells. In classification applications of gene expression data, the classes are predefined (e.g. different tissue or cell types) and the aim is to build a "classifier" able to distinguish between these classes based on the gene expression profiles of the samples. There are several approaches that can be used for the purpose of classification such as nearest neighbor (Li et al., 2001), random forests (Breiman, 2001) and support vector machines (SVMs (Cortes and Vapnik, 1995)). In this work, an algorithm was developed to classify gene expression data (microarray and RNA-seq) based on the number of marker genes shared between a query and a reference sample.

Support vector machines

In machine learning, SVMs (Cortes and Vapnik, 1995) are supervised learning models, which were first introduced by Vladimir Vapnik in early 90s. Briefly, SVMs build a classifier based on a training set, and seek for an optimal separating hyperplane between two classes by maximizing the margin between the classes' closest points (Figure 1.5). Consider a set of labeled training examples:

$$\{\mathbf{x}_i, y_i\} \;\; i = 1, \ldots, l, \;\; x_i \in \mathbb{R}^d, \text{ and } y_i \in \{-1, 1\} \tag{1.1}$$

The training set (1.1) is called separable by the hyperplane $H = \{\mathbf{x}_i \in \mathbb{R}^d \,|\, \langle w, x \rangle + b = 0, w \in \mathbb{R}^d, b \in \mathbb{R}\}$ if there exist both a unit vector $w(\|w\| = 1)$ and a constant b such that the following inequality holds:

$$y_i(\langle w, x_i \rangle + b) \geq 0, \ i = 1, \ldots, l \tag{1.2}$$

where $\langle ., . \rangle$ denotes the inner product in $\mathbb{R}^d$. The hyperplane H defined by w and b is called a "separating hyperplane". The optimal hyperplane is the unique one which separates the training data with a maximal margin (Cortes and Vapnik, 1995). It is the solution of the optimization problem:

$$\begin{cases} \text{minimize } \frac{1}{2}||w||^2 \\ \text{subject to } y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \text{ for all } i \leq l \end{cases} \tag{1.3}$$

This optimization problem can be solved by finding the saddle point of the primal Lagrangian (Schölkopf and Smola, 2002):

$$L_P(w, b, \alpha) = \frac{||w||^2}{2} - \sum_{i=1}^{l} \alpha_i y_i(\langle w, x_i \rangle + b) + \sum_{i=1}^{l} \alpha_i, \tag{1.4}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers. The Lagrangian L_P has to be minimized with respect to the primal variables w and b and maximized with respect to the dual variables α_i. The dual problem is to find multipliers α_i which solve the problem.

$$\begin{cases} \text{maximize } L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{subject to } \alpha_i \geq 0 \text{ for all } i = 1, \ldots, l \text{ and } \sum_{i=1}^{l} \alpha_i y_i = 0, \end{cases} \tag{1.5}$$

The label of a new test point x can be predicted by the decision function::

$$f(x) = \text{sgn}(\sum_{i=1}^{l} y_i \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b) \tag{1.6}$$

Often, real-life data are not linearly separable in the original space. In this case, the separability constraints (1.2) are relaxed by introducing misclassification penalties ξ_i $(i = 1, \ldots, l)$:

$$y_i(\langle w, x_i \rangle + b + \xi_i) \geq 0, \ i = 1, \ldots, l \tag{1.7}$$

And the optimal separating hyperplane can be found by minimizing:

$$\begin{cases} \text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ \text{subject to } y_i(\langle w, x_i\rangle + b) - 1 + \xi_i \geq 0 \text{ and } \xi_i \geq 0 \end{cases} \tag{1.8}$$

where C is a regularization parameter used to decide a trade-off between the training error and the margin. This formulation is called the *soft-margin* SVM (Cortes and Vapnik, 1995). By the use of kernel functions the pattern vectors $x_i \in \mathbb{R}^d$ are mapped to a high dimensional space $\mathscr{H}$ and separated there by a linear classifier. This results in a classifier nonlinear in $\mathbb{R}^d$. Given a mapping $\phi : \mathbb{R}^d \to \mathscr{H}$ from input space $\mathbb{R}^d$ to an (inner product) feature space $\mathscr{H}$, the function $k : \mathbb{R}^d \to \mathbb{R}^d$ is called a kernel function, if for all $x_i, x_j \in \mathbb{R}^d$:

$$k(x_i, x_j) = \left\langle \phi(x_i), \phi(x_j) \right\rangle_{\mathscr{H}} \tag{1.9}$$

The kernel trick is to replace the inner product $\langle .,. \rangle$ with the kernel function $k(.,.)$ in 1.5. The optimization problem can then be represented as:

$$\begin{cases} \text{maximize } L_D = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to } 0 < \alpha_i \leq C \text{ for all } i = 1, \ldots, l \text{ and } \sum_{i=1}^{l}\alpha_i y_i = 0, \end{cases} \tag{1.10}$$

Three commonly used kernel functions are defined by:

$$K(x_i, x_j) = (\left\langle x_i, x_j \right\rangle + 1)^d \tag{1.11}$$

$$K(x_i, x_j) = exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \tag{1.12}$$

$$K(x_i, x_j) = tanh(\kappa \left\langle x_i, x_j \right\rangle + \theta) \tag{1.13}$$

Equation (1.11) is the polynomial kernel function with degree d. Equation (1.12) is the Gaussian radial basic function, where $\sigma > 0$ is a parameter that controls the width of the Gaussian. Equation (1.13) is the sigmoid kernel, where $\kappa > 0$ and $\theta < 0$.

SVMs were basically developed for binary classification. Several extensions have been

suggested to allow for multi-class classification. The most popular strategies are *one-against-all* and *one-against-one*, which decomposes the multi-class problem into a predefined set of binary problems. In *one-against-all* approach, every class is separated by an SVM from the pooled data points of all other $l-1$ classes. To test a new point, one calculates the distance to all l hyperplanes and assigns it to the class for which it achieves maximum distance. In *one-against-one* approach, an SVM classifier is built for each pair of classes. More details about SVMs can be found for example in the two introductory books (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002).
As software to run SVMs, I used in this thesis the R package `e1071` (Meyer et al., 2017).

1.6 R and Bioconductor

R (R Development Core Team, 2016) is an open-source programming language and software environment for statistical computing and graphics. It provides a variety of statistical and graphical techniques, and is highly extensible through the use of packages. These are libraries for specific functions or specific areas of study, frequently created by R users and distributed under suitable open-source licenses. A large number of packages are available at the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org or at the Bioconductor (Huber et al., 2015) repository at http://www.bioconductor.org. Bioconductor is an open-source and open-development software project primarily based on R, but it also contains contributions in other programming languages, and it provides tools for the analysis and comprehension of genomic data.
The tools developed in this work are implemented in the programming language R, and available as R packages from the Bioconductor website.

1.7 Thesis outline

This thesis introduces new bioinformatics tools to detect marker genes of cells and tissues and to classify samples using gene expression profiles. It is divided in five chapters. The first chapter provides a brief introduction to the microarray and RNA-seq technologies. In addition, classification and SVMs are illustrated. In chapter 2, I introduce `MGFR`, a Bioconductor R package for marker gene detection from RNA-seq data. This tool is an extension of the tool `MGFM`, which I have developed to detect marker genes from microarray data. Next, I show results of the experimental validation of top ranked marker genes by reverse transcriptase-polymerase chain reaction (RT-PCR) for a set of five human tissues. Furthermore, I present a comparison study between microarrays and RNA-seq. I compare the overlap of marker genes obtained using a public microarray and an RNA-seq dataset for the same set of 16 human tissues. I identify robust marker genes (predicted by both `MGFM` and `MGFR`), and suggest novel candidate marker genes for each of the examined tissues. Finally, I compare the set of predicted marker genes for ten tissues to a gold-standard list of marker genes obtained from the Tissue-specific Gene Expression and Regulation (TiGER) database. A part of this work (concerning the marker tool `MGFM` and the experimental validation by RT-PCR) was published in (El Amrani et al., 2015):

> **El Amrani, K.**, Stachelscheid, H., Lekschas, F., Kurtz, A., and Andrade-Navarro, M. A. (2015). MGFM: a novel tool for detection of tissue and cell specific marker genes from microarray gene expression data. *BMC Genomics*, 16(1):645

The main change in `MGFR` compared to `MGFM` is the mapping of gene identifiers to gene symbols and Entrez Gene IDs, and the use of a cutoff value of 1 FPKM (fragments per kilobase of exon model per million mapped reads) as a cutoff for marker gene expression. I created a new package, because `MGFM` was already almost two years in active use by the community in the time `MGFR` was submitted to Bioconductor. Since I use the algorithm developed for `MGFM` for marker gene detection in `MGFR`, I could have modified the `MGFM` package. But this would have resulted in modifying and adding more input parameters to the functions, which may cause problems and inconveniences for users, who are using the tool or its output in their own analysis procedures.

In chapter 3 I introduce the classification tool `sampleClassifier`, which is designed for the classification of samples using their gene expression profiles. The package supports the classification of microarray and RNA-seq gene expression profiles. I demonstrate its performance using public microarray and RNA-seq data and compare it to SVMs.

In chapter 4 I introduce an application of the previously described approaches (`MGFM` and `sampleClassifier`) to publicly available biopsy-based microarray data from eight diverse kidney diseases. I identify marker genes, and demonstrate the performance of the classification tool in distinguishing between normal and disease samples as well as between different types of renal diseases.

Finally, in chapter 5, I summarize the contributions of my work and give an outlook.

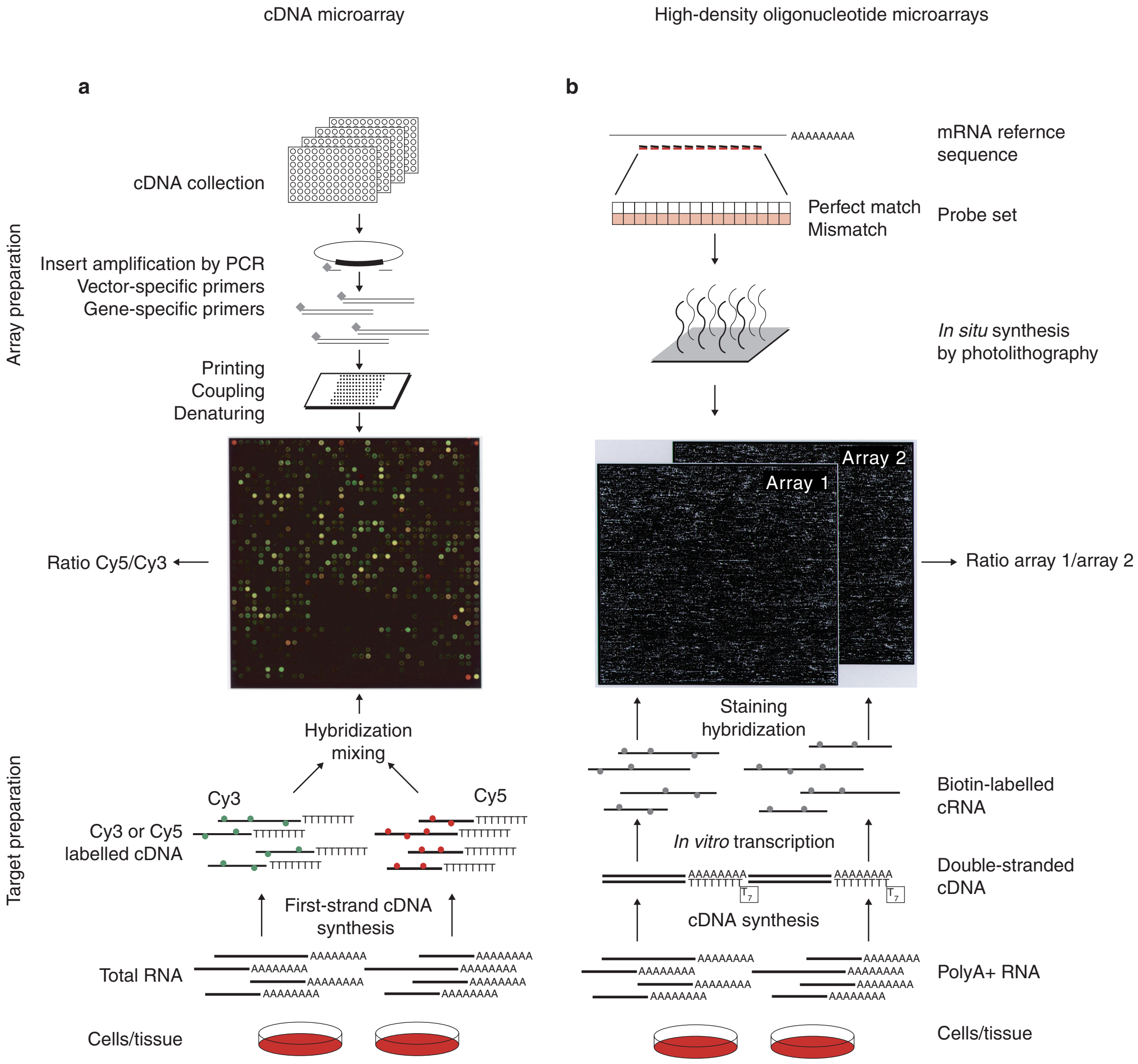


Figure 1.3: Schematic overview of probe array and target preparation for a) spotted cDNA microarrays and b) high-density oligonucleotide microarrays. Figure reproduced from (Schulze and Downward, 2001)

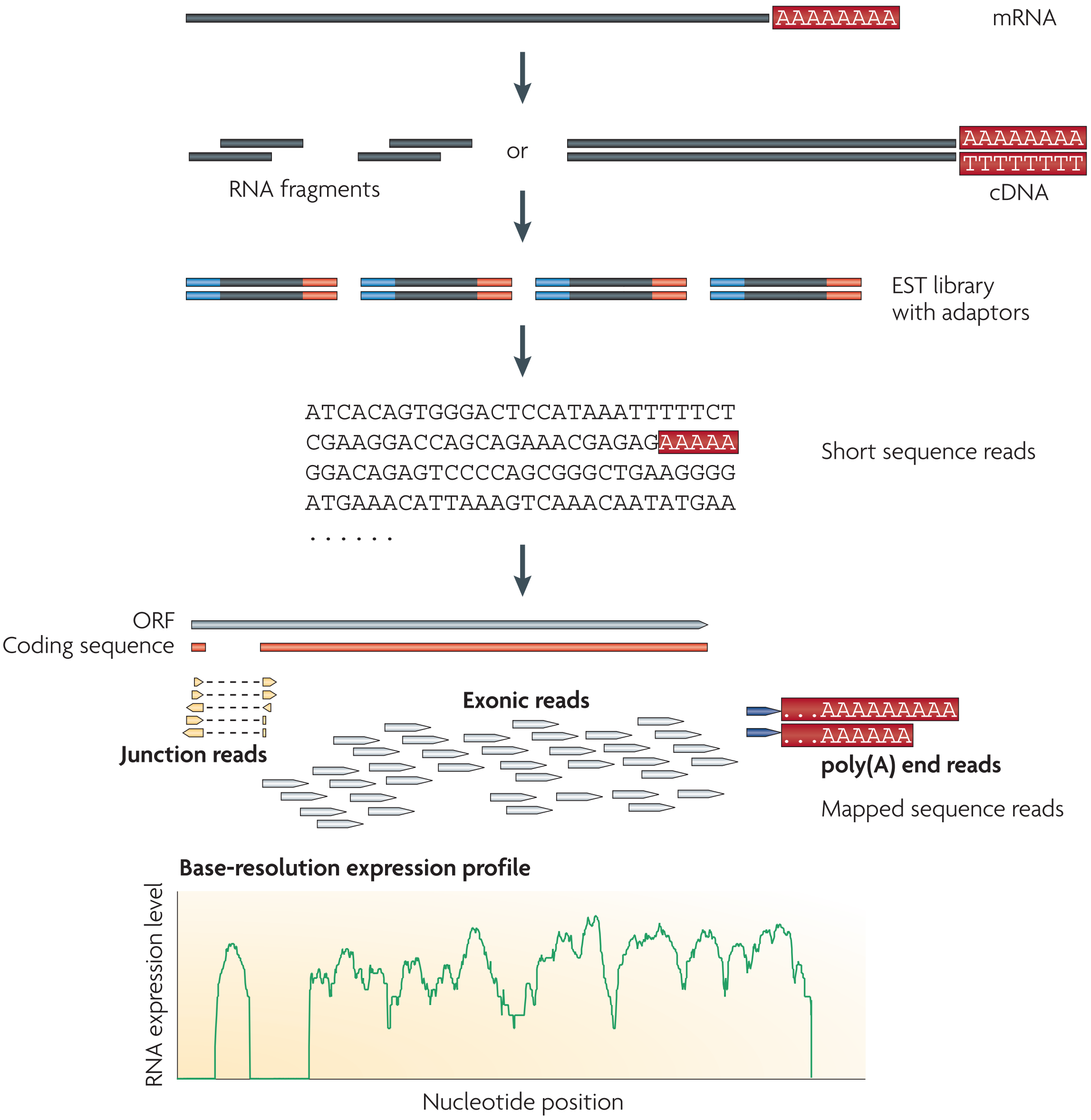


Figure 1.4: A typical RNA-seq experiment. mRNA is converted into a library of cDNA fragments typically following RNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene (bottom). Figure reproduced from (Wang et al., 2009).

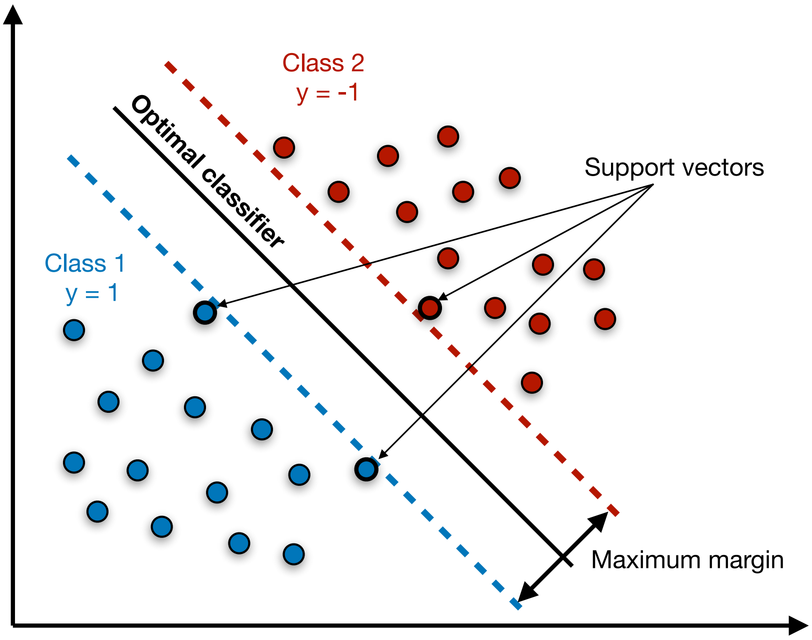


Figure 1.5: Binary support vector machines

Chapter 2

Marker gene detection

Genes may be divided into two categories: 1) genes that are constitutively expressed in all tissues and cell types. These genes are required for the maintenance of basic cellular function and are often referred to as housekeeping genes; 2) tissue-specific genes, whose expression is preferred in specific tissues or cell types. From here on forward, I define marker genes of a tissue or cell type as the tissue-specific genes whose expression pattern distinguishes the tissue or cell type from other tissues or cell types. These genes are important for tissue function and cellular determination. Since disease-associated genes are more likely to show tissue specific expression (Reverter et al., 2008), marker genes could be used to understand the molecular mechanisms underlying complex diseases. Genes involved in diseases might be identified by comparing gene expression between tissues from healthy and diseased individuals. Moreover, marker genes are useful to determine the tissue identity and to characterize cells grown *in vitro*. I developed a marker tool as an R package named `MGFM` (Marker Gene Finder in Microarray gene expression data) to detect marker genes from microarray data. Within the scope of this thesis, this tool was further optimized and updated to support the detection of marker genes from RNA-seq data. For this purpose, I implemented an R package called `MGFR` (Marker Gene Finder in RNA-seq data). In addition, the mRNA expression of top-ranked marker genes was experimentally validated by reverse transcriptase-polymerase chain reaction (RT-PCR). The main change in `MGFR` compared to `MGFM` is the mapping of gene identifiers to gene symbols and Entrez Gene IDs, and the use of a cutoff value of 1 FPKM (fragments per kilobase of exon model per million mapped reads) as a cutoff for marker expression. Both marker tools `MGFM` and `MGFR` are available from the Bioconductor website (http://www.bioconductor.org/packages/release/bioc/html/MGFM.html) or (http://www.bioconductor.org/packages/release/bioc/html/MGFR.html), respectively. In this chapter, I introduce the marker tools and explain the algorithm to identify marker genes. Then, I apply both `MGFM` and `MGFR` to a public microarray and an RNA-seq dataset profiling the same set of 16 human tissues, respectively. I assess the performance of the tools using tissue-specific genes taken from the Tissue-specific Gene Expression and Regulation (TiGER) database (Liu et al., 2008). TiGER is a database for generating comprehensive information about human tissue-specific gene regulation, including both expression and regulatory data. The database contains tissue-specific gene expression profiles or expressed sequence tag (EST) data, cis-regulatory module (CRM) data, and combinatorial gene regulation data. Next, I compare the predicted marker gene lists for each of the examined tissues. Finally, I suggest robust marker genes (i.e. predicted by both tools `MGFM` and `MGFR`) that were not previously associated with the 16 examined tissues as novel candidate marker genes.

2.1 Materials and methods

2.1.1 Data sources and pre-processing

The microarray dataset for this analysis is publicly available from GEO with the series number GSE3526 (Roth et al., 2006b). The primary RNA-seq data (reads) are available through the ArrayExpress Archive (www.ebi.ac.uk/arrayexpress/) under the accession number: E-MTAB-1733 (Fagerberg et al., 2014). I selected 16 tissues that were common to both datasets: adrenal, bone marrow, brain, colon, endometrium, esophagus, heart, kidney, liver, lung, lymph node, prostate, salivary gland, spleen, testis, and thyroid. The IDs of samples used in the microarray and RNA-seq dataset are shown in Tables A.1 and A.2 (Appendix A), respectively.

The microarray data were normalized using YuGene (Lê Cao et al., 2014). YuGene uses the cumulative proportion transform. Let P_i denote the expression of the probesets on the chip, and $P_{(i)}$ the expression of these same probesets but in decreasing order, from the highest to the lowest values ($i = 1, \dots, n$).

$$Y_{(i)} = 1 - \frac{\sum_{j=1}^{i} P_{(j)}}{\sum_{j=1}^{n} P_{(j)}} = \frac{\sum_{j=1}^{n} P_{(j)} - \sum_{j=1}^{i} P_{(j)}}{\sum_{j=1}^{n} P_{(j)}} = \frac{\sum_{j=i+1}^{n} P_{(j)}}{\sum_{j=1}^{n} P_{(j)}}, \text{ for } i = 1, \dots, p-1$$

$$Y_{(n)} = 0,$$

where $Y_{(i)}$ is the YuGene transformed value for probeset (i), $P_{(i)}$ is the pre-processed raw value for probeset (i), and n is the total number of probesets on the array. The output for each probeset $Y_{(i)}$ is a value between zero (lowest expression) and close to one (highest expression). When equivalent values occur in the raw data, for example $P_{(i)} = P_{(i+1)}$, the same YuGene value is assigned to each probeset such that:

$$Y_{(i)} = Y_{(i+1)} = 1 - \frac{\sum_{j=1}^{i} P_{(j)}}{\sum_{j=1}^{n} P_{(j)}}$$

For the RNA-seq dataset, the reads from the study E-MTAB-1733 were mapped to the GRCh37 version of the human genome with TopHat v2.1.0 (Trapnell et al., 2009). Normalized FPKM (fragments per kilobase of exon model per million mapped reads) values were calculated using cuffquant and cuffnorm from the Cufflinks package v2.2.1 (Trapnell et al., 2010). Cufflinks estimates the abundance of isoforms or transcripts by probabilistically assigning reads to the isoforms. The probability that a fragment originates from transcript t and the probability of selecting a fragment from transcript t are denoted by β_g and γ_t, respectively. These parameters are estimated from a likelihood function and the abundance of a transcript $t \in$ gene g is given in FPKM units:

$$\text{FPKM} = \frac{10^6 \cdot 10^3 \cdot \beta_g \cdot \gamma_t}{\tilde{l}(t)}$$

where $\tilde{l}(t)$ is an adjusted length of transcript t (Trapnell et al., 2010).

The analyzed RNA-seq data were extracted after the calculation of FPKM values of all samples and averaging across technical replicates.

For the comparison of MGFM and MGFR, the method Jetset v3.3.0 (Li et al., 2011) was used to select the optimal probeset for each gene in the microarray dataset. Then, genes common to both the microarray and RNA-seq dataset were matched using the Entrez Gene IDs. Only common genes to both datasets (a total of 18415 genes) were considered for the analysis.

Jetset gives each Affymetrix gene probe a score based on specificity, splice isoform coverage, and robustness against transcript degradation. Using these scores the Jetset method selects a single representative probeset for each gene, thus creating a simple one-to-one mapping between gene and probeset.

2.1.2 Marker gene identification

The tools MGFM and MGFR require a normalized reference matrix with replicates for each sample type.

Marker genes are identified following the steps below:

Sorting expression values for each gene: In this step the expression values are sorted in decreasing order.

Marker selection: To analyze the sorted distribution of expression values of a gene to define if it is a potential candidate marker I define cut-points as those that segregate samples of different types. A sorted distribution can have multiple cut-points; a cut-point can segregate one sample type from the others, or it can segregate multiple sample types from multiple sample types. Figure 2.1 illustrates the procedure of marker selection for the gene *CGNL1* (cingulin like 1). For simplicity, only 6 sample types are shown. In this example, the distribution has two cut-points (cut-point 1 and cut-point 2), the first cut-point segregates kidney samples from the rest, and the second cut-point segregates liver and salivary gland samples from the rest. Each cut-point is defined by the ratio of the expression averages of the groups of samples adjacent to it. That is, given a distribution with n cut-points and n+1 segregated groups, cut-point i receives a score that is the ratio of the average expression of samples in the group i+1 (following the cut-point) divided by that of group i (preceding the cut-point). This value is < 1 because the values are sorted in decreasing order. The closer the values, the closer the score to 1 and therefore the smaller is the gap between expression values at the cut-point. For marker identification only the first two cut-points are required. A gene is considered as marker if it has a cut-point that segregates one tissue at high expression from the rest (as in Figure 2.1 for kidney). I define the score associated with the first cut-point as specificity score for the gene. This score has a value between 0 and 1. Values near 0 would indicate high specificity and large values closer to 1 would indicate low specificity. I disregard negative markers (segregating samples from one tissue at low expression) or multiple tissue markers (segregating samples from more than one tissue from other multiple tissues).

Mapping of probeset IDs or gene IDs to gene symbols or Entrez Gene IDs: For mapping between microarray probeset IDs and gene symbols or Entrez Gene IDs, `MGFM` uses annotation packages from Bioconductor that contain annotation data about particular microarray platforms (ChipDb). The mapping of gene identifiers to gene symbols and Entrez Gene IDs is done in `MGFR` using the R package biomaRt (Durinck et al., 2005). `MGFR` supports Ensembl (ENSG), RefSeq, and UCSC identifiers.

2.1.3 Implementation

Both tools `MGFM` and `MGFR` are implemented in the programming language R. Marker genes can be detected using the function `getMarkerGenes()` from `MGFM` in the case of microarray data or `getMarkerGenes.rnaseq()` from `MGFR` for RNA-seq data. Both functions require a normalized reference matrix. In addition, the function `getHtmlpage()` from `MGFM` or `getMarkerGenes.rnaseq.html()` from `MGFR` can be used to show the marker genes in HTML tables with links to various online annotation sources (Ensembl, GenBank and

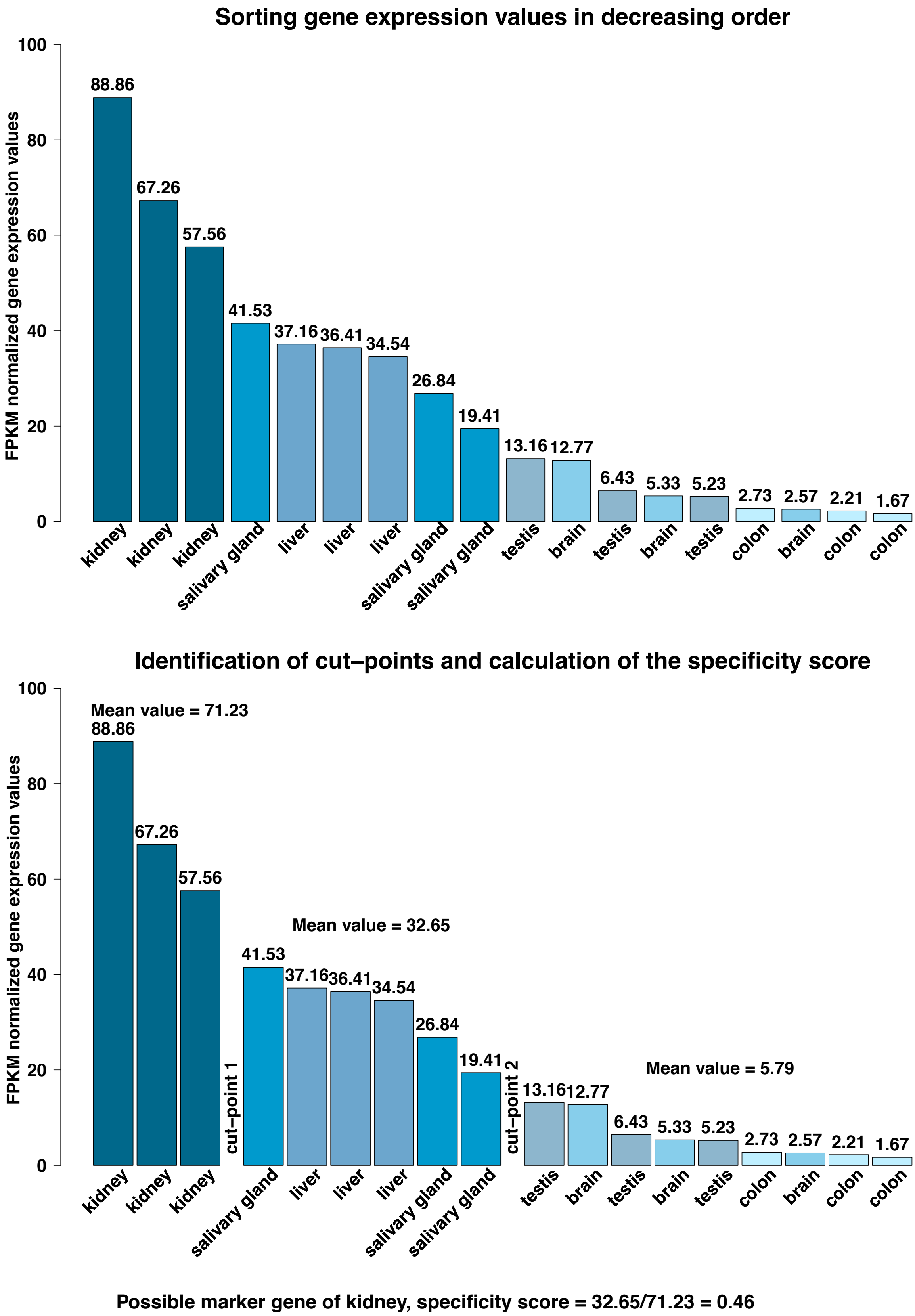


Figure 2.1: An example showing how marker genes are identified by MGFR. The expression values correspond to the gene *CGNL1* (cingulin like 1).

EntrezGene repositories). More technical details about the provided functions and their input parameters can be found in the Vignettes of the R packages.

2.1.4 Output

The functions `getMarkerGenes()` and `getMarkerGenes.rnaseq()` return as output a list with the predicted marker genes for each sample type. Each entry in the output list is a data frame, in which the marker genes are sorted according to their specificity score. For each marker the probeset or gene ID, the gene symbol, the Entrez Gene ID, and the specificity score are shown, if the input parameter *annotate* is set to TRUE. Otherwise, only the probeset or gene ID, and the specificity score are shown. The `getHtmlpage()` and `getMarkerGenes.rnaseq.html()` create HTML tables for each sample type to show the predicted marker genes with links to various online annotation sources (Ensembl, GenBank and EntrezGene repositories).

2.1.5 Ethics statement

Human kidney tissue was provided from leftover diagnostic biopsies from the Department of Nephrology at Charité Universitätsmedizin Berlin. RNA from heart and lung tissues was provided by the German Heart Center Berlin, and RNA from liver from the Department of Experimental Surgery at Charité Universitätsmedizin Berlin. All tissue donors were consented and ethics approval obtained from the responsible ethics Committee at Charité (Nr. EA1/110/10) and the German Heart Center (Nr. EA4/028/12).

2.1.6 cDNA synthesis and RT-PCR analysis

Human total RNA was isolated from liver, lung, heart and kidney with TRIzol reagent (Invitrogen) according to the manufacturer's protocol. Human RNA from brain was purchased from Clontech Laboratories (Mountain View, CA, USA). RNA was reverse transcribed into cDNA with random primers using SuperScript III First-Strand Synthesis System (Invitrogen) according to the manufacturer's protocol. 5 µg of total RNA was used for cDNA synthesis. The PCR reaction consisted of 1 µl of cDNA, 0.5 µl of 10 mM deoxynucleoside triphosphate mix (dNTP), 5 µl of 5X Crimson Taq (Mg-free) Reaction Buffer, 1.5 µl of 25 mM $MgCl_2$, 0.5 µl of each 10 µM forward and reverse primers, 0.125 µl of Crimson Taq DNA polymerase, and nuclease-free water up to 25 µl. The cycling conditions were performed as following: 95 °C for 2 min, followed by 30 cycles of 95 °C for 30s, temperature specific annealing for 30s, and 72 °C for 45s with a final elongation at 72 °C for 7 min. A 1% agarose gel was used to check PCR amplification. The housekeeping gene beta-actin was used as positive control. All primers used are listed in Appendix A (Table A.3).

2.2 Results

I developed the marker tool MGFM to compare small sets of samples using microarray gene expression data. Within the scope of this thesis I modified this tool to reduce the computation time and implemented it as a Bioconductor R package. In order to assess the accuracy of MGFM, I verified top ranked marker genes predicted for a set of five human tissues by RT-PCR. Furthermore, I modified the optimized version of MGFM to enable the detection of marker genes from RNA-seq. For simplicity and to avoid changing the original tool, I created a new R package named MGFR. Both packages are open-source, and available in the Bioconductor repository (http://www.bioconductor.org/packages/release/bioc/) since September 2014 or July 2016, respectively, and are accepted and used by the Bioconductor community. Figures 2.2 and 2.3 visualize the download statistics provided by the Bioconductor team for the package MGFM and MGFR, respectively. In this section I will show that the optimization of MGFM reduced the running time significantly. In addition, I show results of the experimental validation of top ranked marker genes by RT-PCR for a set of five human tissues. Next, I present the results of the application of both marker tools to publicly available gene expression data. First, I test the performance of MGFM and MGFR independently in finding markers with a gold-standard obtained from the TiGER database for ten of the examined human tissues (bone marrow, brain, heart, kidney, liver, lung, lymph node, prostate, spleen, and testis) (Sections 2.2.4 and 2.2.5). Next, to facilitate the comparison of performance of MGFM and MGFR, eliminating the array-specific limits in detecting genes, I repeat the analysis using the datasets with the common genes to both platforms (18415 genes) and using the set of TiGER genes that both could potentially detect (2373 genes) for validation (Section 2.2.6). Finally, I suggest robust marker genes (i.e. predicted by both tools MGFM and MGFR), as well as novel candidate marker genes that were not previously associated with the 16 examined tissues.

2.2.1 Optimization of MGFM

In order to enable a fast running time, MGFM was modified and optimized by adding the following main changes: i) replacing for loops by `apply` functions; ii) the identification of cut-points was limited to the first two cut-points, since only two are required for marker identification (more details about cut-points and marker identification are given in Section 2.1.2); iii) modification of the input parameter such as the use of a new input parameter

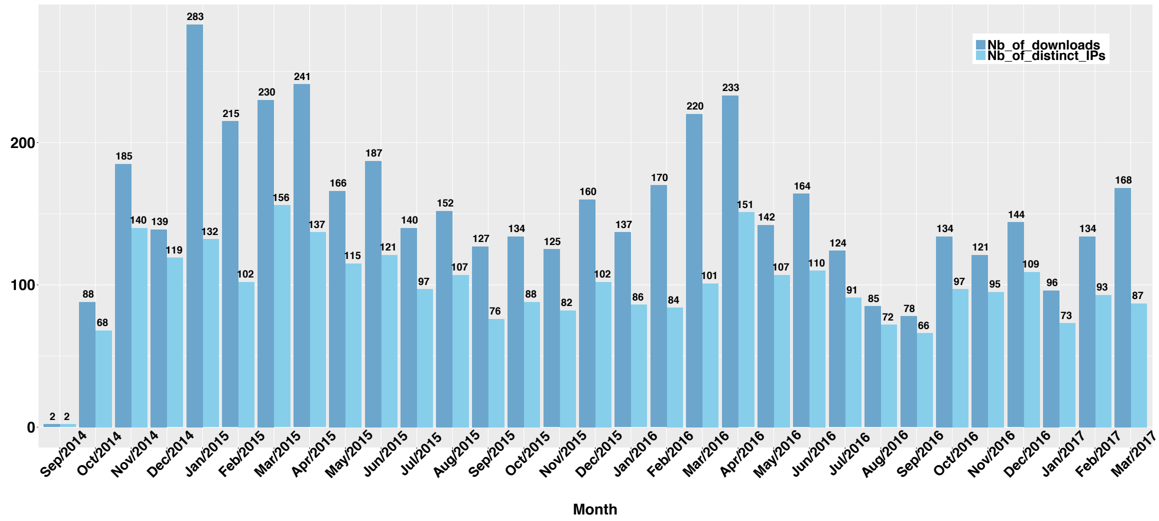


Figure 2.2: Download statistics for the software package `MGFM` (http://bioconductor.org/packages/stats/bioc/MGFM/).

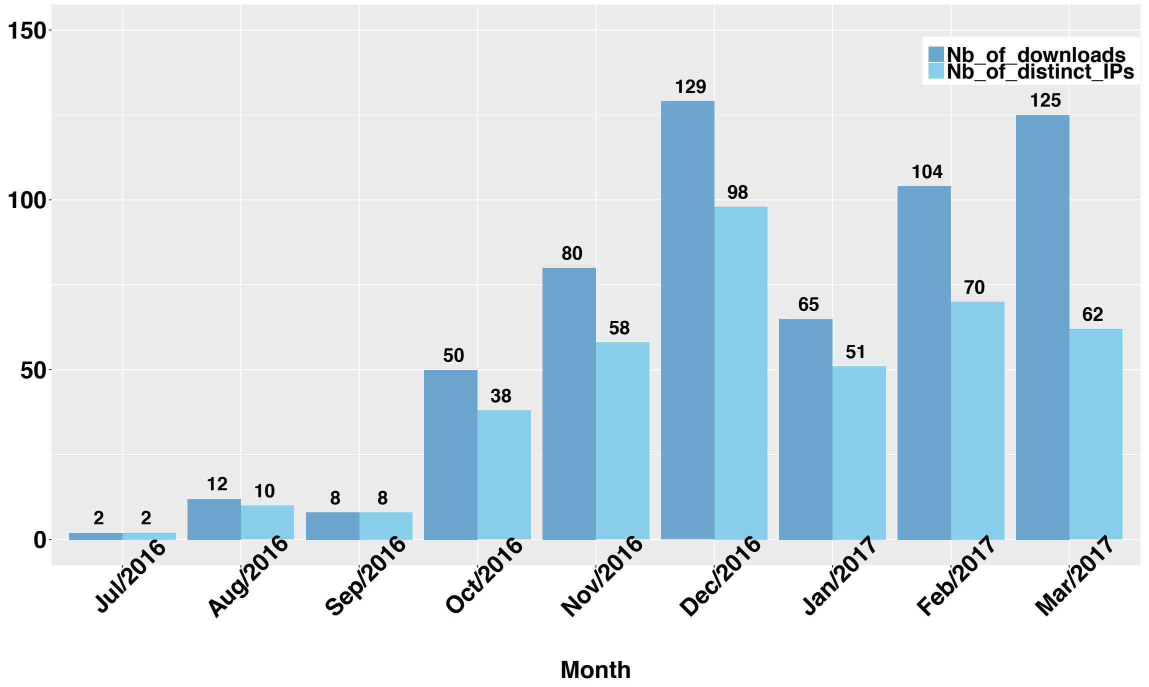


Figure 2.3: Download statistics for the software package `MGFR` (http://bioconductor.org/packages/stats/bioc/MGFR/).

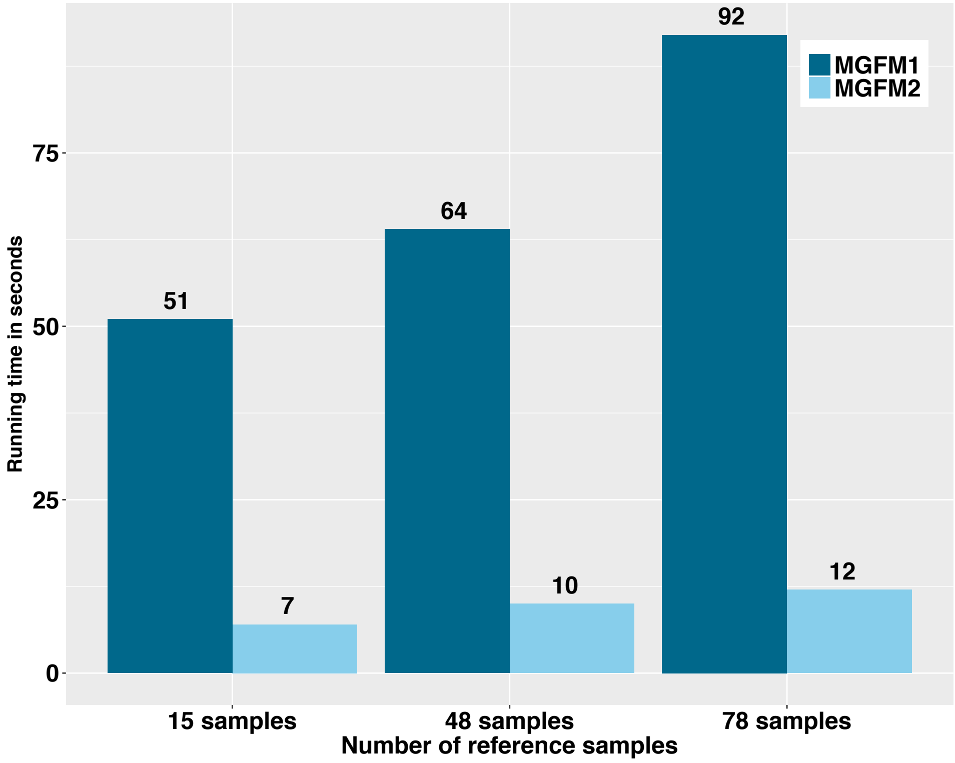


Figure 2.4: Comparison of the runnig time of the first version of `MGFM` (`MGFM1`) and the optimized version (`MGFM2`) on microarray datasets (Affymetrix Human Genome U133 Plus 2.0 Array) with a total of 54675 probesets and different numbers of reference samples. The tools were run on an Apple Macbook with a 2.9 GHz Intel Core i7 processor, and 8 GB 1600 MHz DDR3 memory (OS X 10.12.1).

samples2compare to enable the comparison of particular sample types in the reference matrix; iv) implementation of functions to show the predicted marker genes in HTML pages. In addition the tool was implemented as a Bioconductor R package. Figure 2.4 illustrates the comparison of the running time of the first version of `MGFM` (`MGFM1`) and after optimization (`MGFM2`) using three microarray datasets with different numbers of samples (N=15, 48, and 78). Obviously, the running time was significantly reduced after the optimization. Based on the performed examples, the running time is accelerated by an average factor of 7. The tools were run on an Apple Macbook with a 2.9 GHz Intel Core i7 processor, and 8 GB 1600 MHz DDR3 memory (OS X 10.12.1).

2.2.2 Verification by RT-PCR

I investigated top ranked marker genes predicted by `MGFM` using two microarray expression datasets derived from GEO. The first data set consisted of 15 samples and is derived from five human tissues (heart atrium, kidney cortex, liver, lung, and midbrain). The microarray data set is publicly available from GEO with the series number GSE3526 (Roth et al., 2006a). The second dataset was derived from five human tissues (brain, heart, kidney, liver, and lung) from two GEO Series GSE1133 (Su et al., 2004) and GSE2361 (Ge et al., 2005). To verify the tissue-specific expression of top-ranked marker genes, I examined these genes by RT-PCR. A total of 11 marker genes were selected for liver and 12 genes for each of the tissues: brain, heart, kidney, and lung. The resulting gel electrophoresis images are shown in Appendix A (Figures A.1, A.2, A.3, A.4, A.5, and A.6). In addition, the RT-PCR results are summarized in Table 2.1 using + or - for present or absent, respectively. As shown in Table 2.1, all genes, predicted as markers of a tissue, were indeed detected in that tissue except *GAP43* in the brain, and the four genes *SLC12A1*, *SLC3A1*, *FXYD2*, and *CA12* predicted as markers of kidney. All identified marker genes are shown in Appendix A (Tables A.4, A.5, A.6, A.7, and A.8) and descriptions of their functions in normal or disease states provided if available. This work was published in (El Amrani et al., 2015):

> **El Amrani, K.**, Stachelscheid, H., Lekschas, F., Kurtz, A., and Andrade-Navarro, M. A. (2015). MGFM: a novel tool for detection of tissue and cell specific marker genes from microarray gene expression data. *BMC Genomics*, 16(1):645

2.2.3 Marker tool in CellFinder

The marker tool `MGFM` was integrated into the CellFinder platform (http://cellfinder.org/analysis/marker) by Fritz Lekschas. CellFinder (Stachelscheid et al., 2014) is a comprehensive on-line repository for diverse data characterizing mammalian cells in different tissues and in different development stages. It is built from carefully selected datasets stemming from other curated databases and the biomedical literature. `MGFM` can be used within CellFinder using three public microarray datasets to generate lists of marker genes for a set of tissues (Figure 2.5). Using the marker tool within CellFinder, users can: i) calculate the potential marker genes at the gene level (using JetSet (Li et al., 2011) to associate genes to probesets), ii) display and rank the list of marker genes associated with each sample type according to the specificity, and iii) download the list of all found markers for further use. Moreover, probesets are linked to CellFinder's gene view which allows for an immediate evaluation of potential marker genes utilizing expression values from the RNA Seq Atlas (Krupp et al., 2012). Also, gene ontology annotations (Ashburner et al., 2000) are included for better understanding of functional properties of genes.

Table 2.1: RT-PCR results

Predicted marker genes for brain

Gene	Liver	Lung	Heart	Brain	Kidney	Gene	Liver	Lung	Heart	Brain	Kidney
GAP43	-	-	-	-	-	*MBP*	-	+	+	+	+
GFAP	-	-	-	+	-	*GRIA2*	-	-	-	+	-
TMEFF1	-	-	-	+	-	*KIF5C*	-	-	-	+	-
FUT9	-	-	-	+	+	*STMN2*	-	-	-	+	-
SYT1	-	-	-	+	-	*NEFM*	-	-	-	+	-
SNAP25	-	+	+	+	-	*GABBR2*	-	-	-	+	-

Predicted marker genes for heart

Gene	Liver	Lung	Heart	Brain	Kidney	Gene	Liver	Lung	Heart	Brain	Kidney
MYOZ2	-	-	+	-	-	*PLN*	-	+	+	-	+
TNNI3	-	+	+	-	-	*MB*	-	-	+	-	-
SYNPO2L	-	+	+	-	-	*TTN*	-	+	+	-	+
MYH6	-	-	+	-	-	*MYL7*	-	-	+	-	-
CSRP3	-	-	+	-	-	*MYH7*	-	-	+	-	-
CKM	-	-	+	-	-	*TPM1*	+	+	+	-	+

Predicted marker genes for kidney

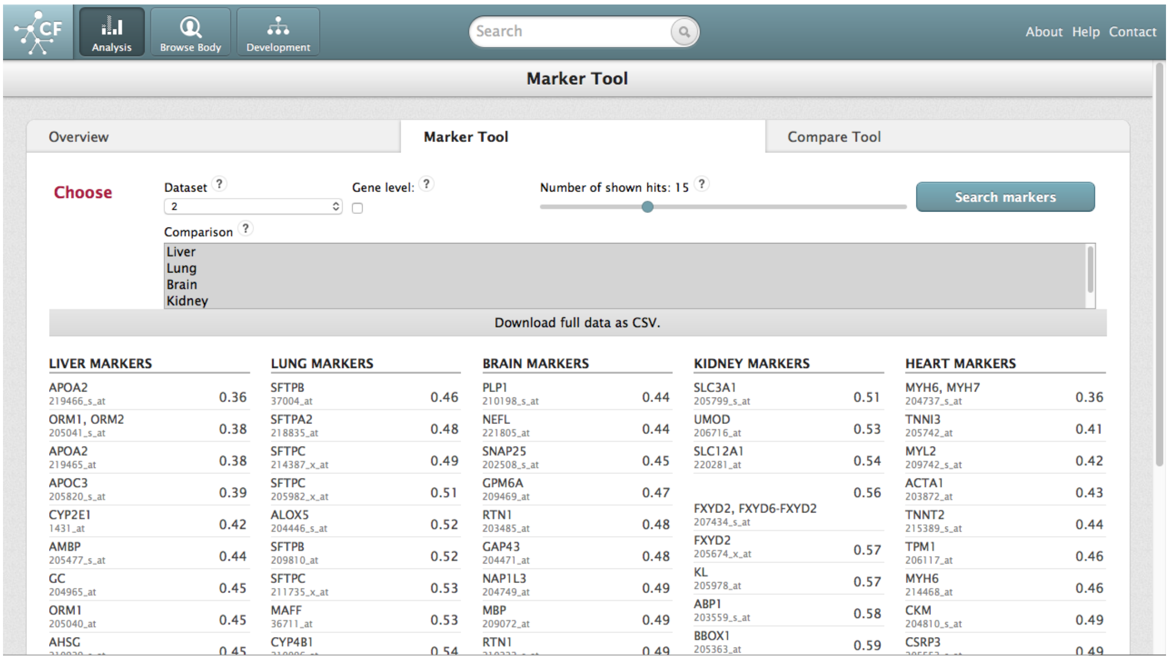
Gene	Liver	Lung	Heart	Brain	Kidney	Gene	Liver	Lung	Heart	Brain	Kidney
SLC12A1	-	-	-	-	-	*CA12*	-	-	-	-	-
SLC3A1	-	-	-	-	-	*PDZK1IP1*	-	-	-	-	+
UMOD	-	-	-	-	+	*FXYD2*	-	-	-	-	-
AOC1	-	-	-	-	+	*CDH16*	-	-	-	-	+
CD24	-	+	-	-	+	*SLC22A8*	-	-	-	-	+
HSD11B2	-	+	-	-	+	*CLDN8*	-	-	-	-	+

Predicted marker genes for liver

Gene	Liver	Lung	Heart	Brain	Kidney	Gene	Liver	Lung	Heart	Brain	Kidney
AKR1D1	+	+	-	-	-	*CYP2E1*	+	-	-	-	-
FGG	+	-	+	-	-	*APOC3*	+	-	-	-	-
APOA2	+	+	-	-	-	*SERPINC1*	+	-	-	-	-
CYP2C8	+	-	-	-	-	*AHSG*	+	-	-	-	-
GC	+	-	-	-	-	*AMBP*	+	-	-	-	-
CPS1	+	-	-	-	-						

Predicted marker genes for lung

Gene	Liver	Lung	Heart	Brain	Kidney	Gene	Liver	Lung	Heart	Brain	Kidney
CLDN18	-	+	-	-	-	*LAMP3*	-	+	+	-	-
NKX2-1	-	+	+	-	-	*AGER*	-	+	-	-	-
SCGB1A1	-	+	+	-	-	*LYZ*	+	+	+	-	-
SFTPB	-	+	-	-	-	*SFTPD*	-	+	-	-	-
CYP4B1	-	+	+	-	-	*SFTPC*	-	+	-	-	-
CD52	-	+	+	-	-	*SLC34A2*	-	+	-	-	+



Choose

Dataset ? 2

Gene level: ? □

Number of shown hits: 15 ?

Comparison ?
Liver
Lung
Brain
Kidney

LIVER MARKERS		LUNG MARKERS		BRAIN MARKERS		KIDNEY MARKERS		HEART MARKERS	
APOA2 219466_s_at	0.36	SFTPB 37004_at	0.46	PLP1 210198_s_at	0.44	SLC3A1 205799_s_at	0.51	MYH6, MYH7 204737_s_at	0.36
ORM1, ORM2 205041_s_at	0.38	SFTPA2 218835_at	0.48	NEFL 221805_at	0.44	UMOD 206716_at	0.53	TNNI3 205742_at	0.41
APOA2 219465_at	0.38	SFTPC 214387_x_at	0.49	SNAP25 202508_s_at	0.45	SLC12A1 220281_at	0.54	MYL2 209742_s_at	0.42
APOC3 205820_s_at	0.39	SFTPC 205982_x_at	0.51	GPM6A 209469_at	0.47		0.56	ACTA1 203872_at	0.43
CYP2E1 1431_at	0.42	ALOX5 204446_s_at	0.52	RTN1 203485_at	0.48	FXYD2, FXYD6-FXYD2 207434_s_at		TNNT2 215389_s_at	0.44
AMBP 205477_s_at	0.44	SFTPB 209810_at	0.52	GAP43 204471_at	0.48	FXYD2 205674_x_at	0.57	TPM1 206117_at	0.46
GC 204965_at	0.45	SFTPC 211735_x_at	0.53	NAP1L3 204749_at	0.49	KL 205978_at	0.57	MYH6 214468_at	0.46
ORM1 205040_at	0.45	MAFF 36711_at	0.53	MBP 209072_at	0.49	ABP1 203559_s_at	0.58	CKM 204810_s_at	0.49
AHSG	0.45	CYP4B1	0.54	RTN1	0.49	BBOX1 205363_at	0.59	CSRP3	0.49

Figure 2.5: Marker tool MGFM in the CellFinder platform

2.2.4 Detection of marker genes using MGFM

MGFM predicted a total of 10 368 out of 54 675 probesets (or 19% of all probesets on the microarray) as markers for the 16 examined tissues at a score cutoff of 0.9. These correspond to 7096 unique genes. Figure 2.6 shows the number of detected marker probesets and the corresponding unique genes for each of the examined tissues. More marker genes were predicted for the four tissues: testis, midbrain, liver, and bone marrow, compared to the other tissues.

A benchmark of MGFM

To validate the set of predicted marker genes, I used as a gold-standard tissue-specific genes from the TiGER database for ten of the examined human tissues (bone marrow, brain, heart, kidney, liver, lung, lymph node, prostate, spleen, and testis). RefSeq IDs for the TiGER tissue-specific genes were downloaded from the TiGER website (http://bioinfo.wilmer.jhu.edu/tiger/), and mapped to Entrez Gene IDs using the biomaRt (Durinck et al., 2005) R package. For validation of the potential marker sets, only gold-standard marker genes that were also found on the microarray were considered for the validation. This corresponded to a total of 2394 marker genes for the ten human tissues. MGFM identified 904 of the gold-standard marker genes (or 37.8%) at a score cutoff of 0.9 (Figure 2.7). The best performance is achieved for testis, where 74.8% of the gold-standard marker genes were correctly identified.

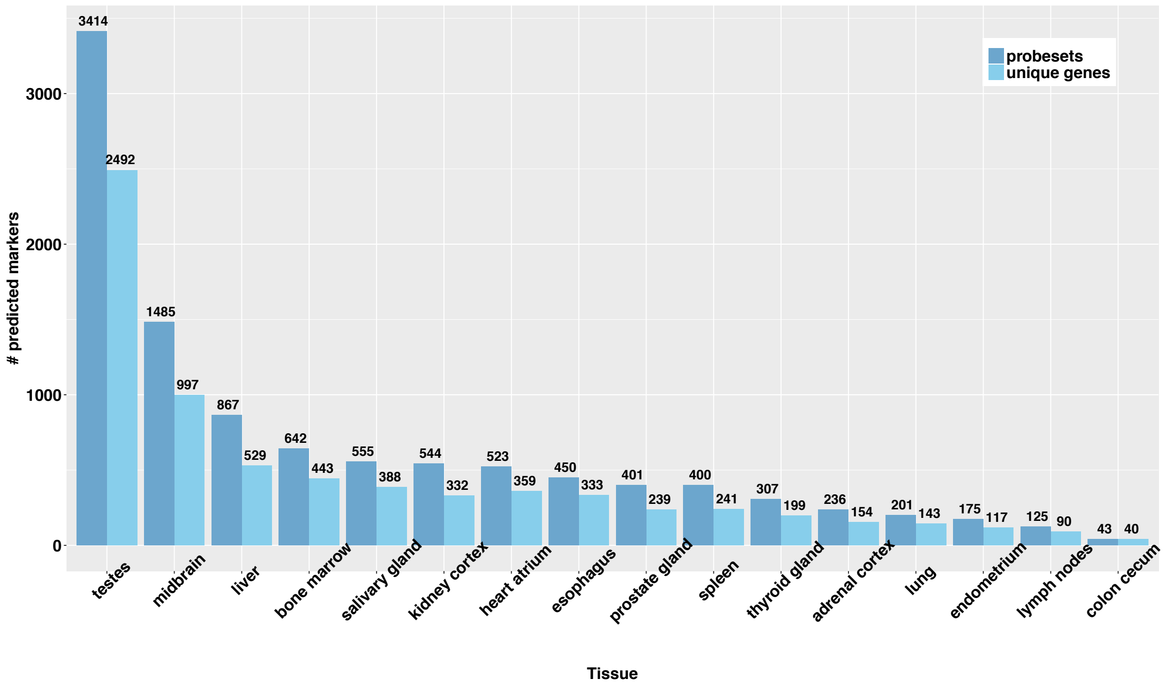


Figure 2.6: Number of marker probesets and the corresponding unique genes predicted by MGFM at a score cutoff of 0.9 for each of the examined tissues.

The lowest accuracy was obtained for lymph node, for which only 9 from the 306 TiGER genes (or 3%) were identified. Increasing the score cutoff from 0.9 to 1, MGFM identified 968 of the 2394 gold-standard marker genes (or 40.4%).

2.2.5 Detection of marker genes using MGFR

MGFR predicted a total of 8783 out of 43 039 genes (or 20.4% of all genes) as markers for the 16 examined tissues at a score cutoff of 0.9. These correspond to 7159 unique genes. Figure 2.8 shows the number of detected marker genes and the unique genes for each of the examined tissues. More marker genes were predicted for the four tissues: testis, bone marrow, liver, and brain, compared to the other tissues.

A benchmark of MGFR

Similar to the validation of MGFM, I used the TiGER genes to evaluate the performance of MGFR. MGFR identified 999 of the 2512 gold-standard marker genes (or 39.8%) at a score cutoff of 0.9 (Figure 2.9). The best performance is achieved for liver, where 69% of the gold-standard marker genes were correctly identified. The lowest accuracy was obtained

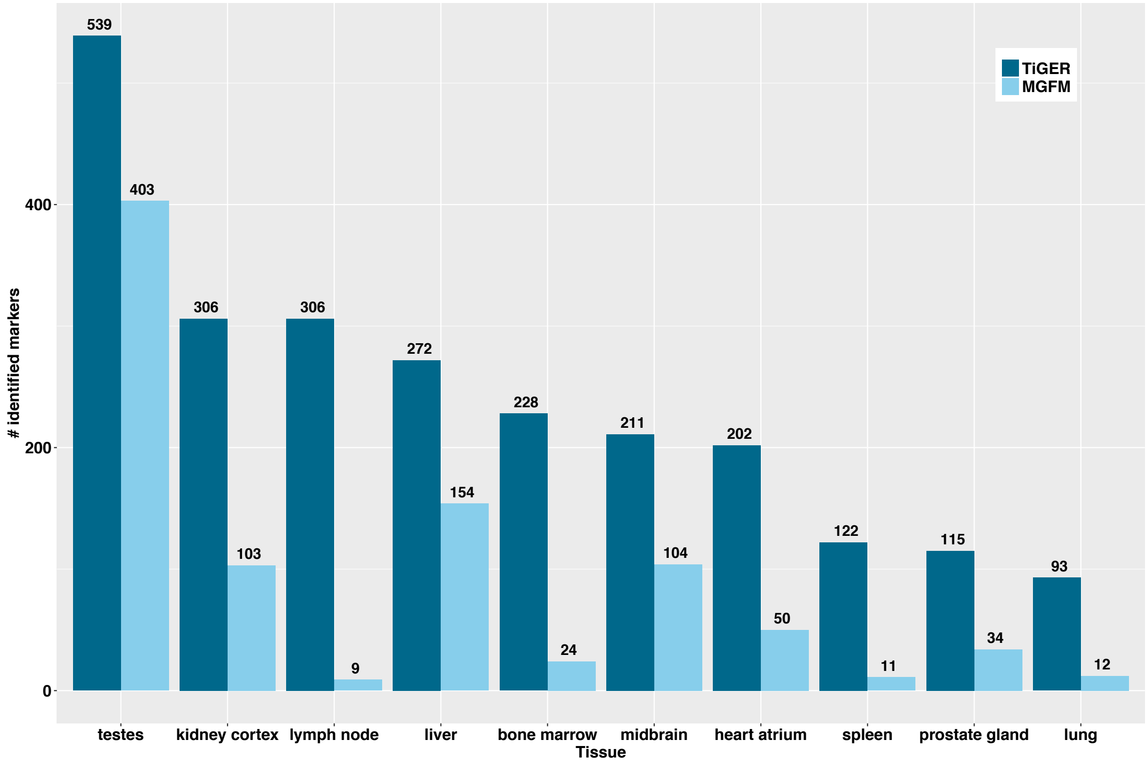


Figure 2.7: Number of obtained gold-standard marker genes from the TiGER database and the number of correctly identified marker genes by `MGFM` at a score cutoff of 0.9 for each of the examined tissues.

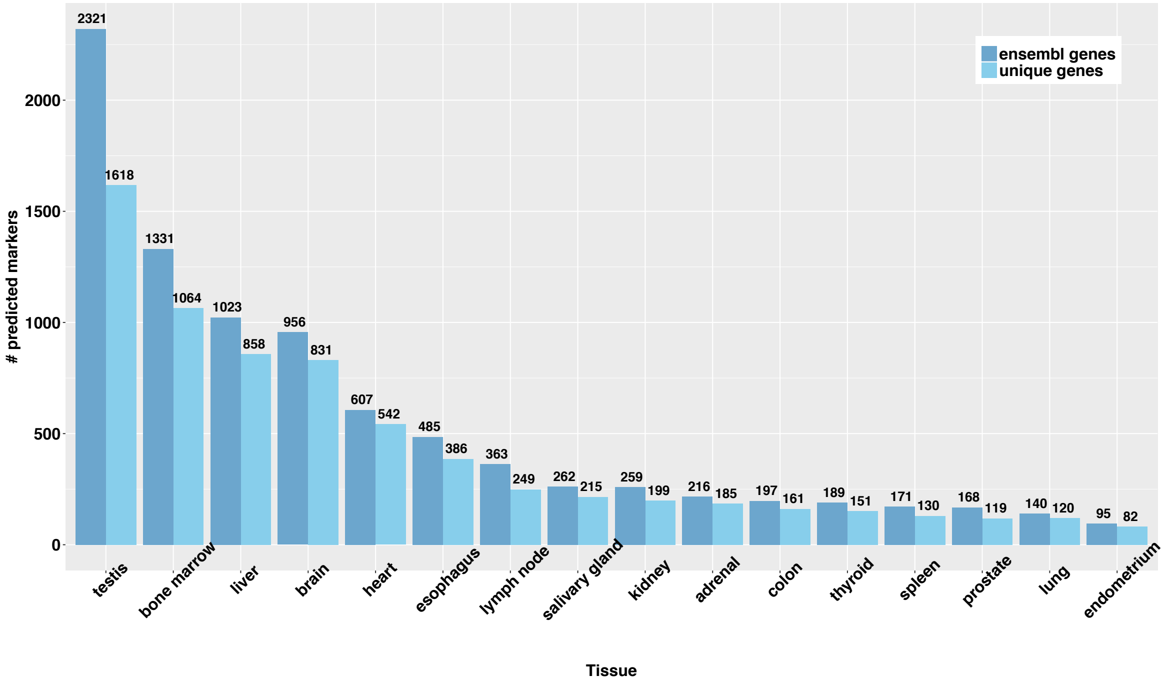


Figure 2.8: Number of marker genes and unique genes predicted by MGFR at a score cutoff of 0.9 for each of the examined tissues.

for spleen, for which only 13 from the 129 TiGER genes (or 10%) were identified. The performance of MGFR does not improve when increasing the score cutoff from 0.9 to 1.

2.2.6 A comparison of MGFM and MGFR

After the application of the marker tools to the whole datasets, in this section I reduce the datasets and consider only genes that are common to both the microarray and RNA-seq dataset. For the microarray data the method Jetset (Li et al., 2011) was used to select the optimal probeset for each gene. Then, genes common to both the microarray and RNA-seq dataset were matched using the Entrez Gene IDs. Only genes common to both datasets (a total of 18415 genes) were considered for the analysis.

Marker selection

Using MGFM a total of 5703 out of 18415 genes (or 31% of the common genes) were selected as markers. Using MGFR a total of 6222 genes from the 18415 genes (or 34% of the genes) were selected as markers. Figure 2.10 shows the number of detected marker genes for each of the examined tissues by MGFM and MGFR, and the number of common genes. Compared to

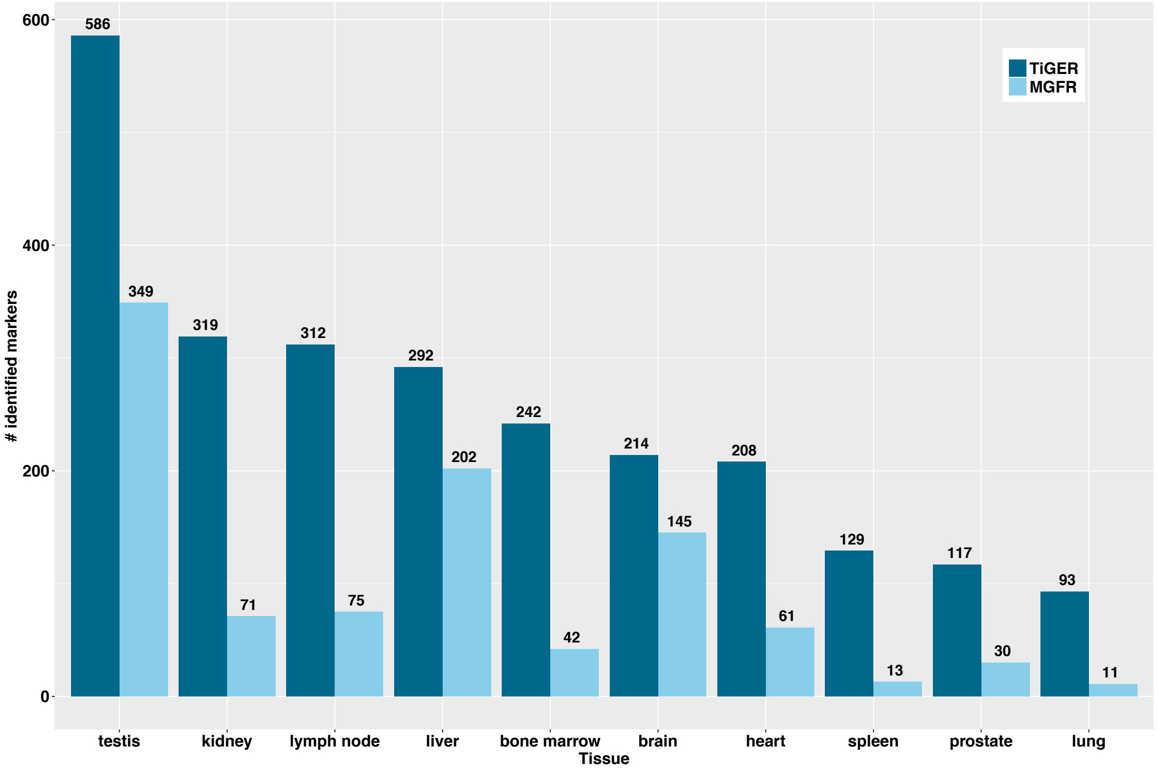


Figure 2.9: Number of obtained gold-standard marker genes from the TiGER database and the number of correctly identified marker genes by MGFR for each of the examined tissues.

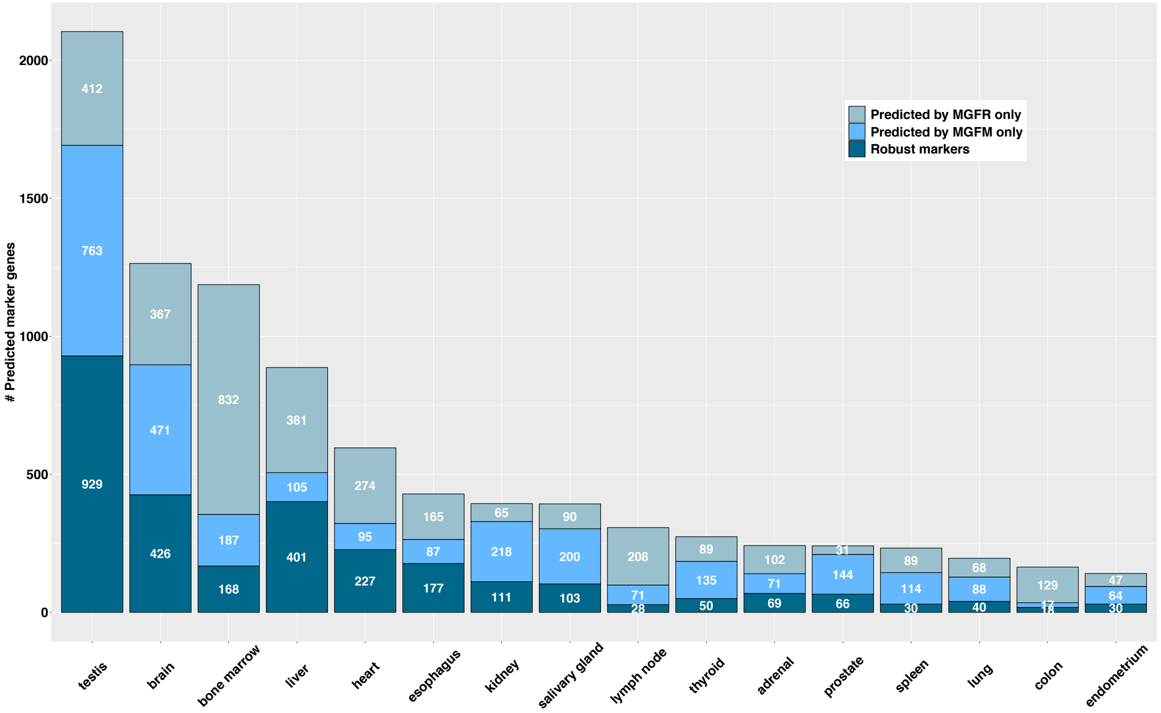


Figure 2.10: Number of marker genes detected for each tissue using MGFM and MGFR, and the number of robust markers (i.e. predicted by both MGFM and MGFR).

other tissues, a greater number of marker genes were predicted for these four tissues: testis, brain, bone marrow, and liver.

Performance analysis

Again, I validated the performance of the tools MGFM and MGFR on the reduced datasets with the common genes to both platforms, focusing on the set of TiGER genes that both could potentially detect (2373 genes). MGFR and MGFM identified 965 genes (40.7%) and 898 genes (37.8%) of the 2373 gold-standard marker genes, respectively, (Figure 2.11). The overlap of TiGER genes identified by both tools contained 726 genes, corresponding to 30.6% of the 2373 TiGER genes.

Next, I investigated the specificity scores of the correctly identified marker genes by MGFM and MGFR. The marker genes identified by MGFR show higher tissue specificity (i.e. lower specificity scores) compared to the markers identified by MGFM (Figure 2.12). This is due to the fact that RNA-seq is more sensitive in detecting genes with low expression compared to microarrays. In addition, RNA-seq has a much wider dynamic range than microarrays.

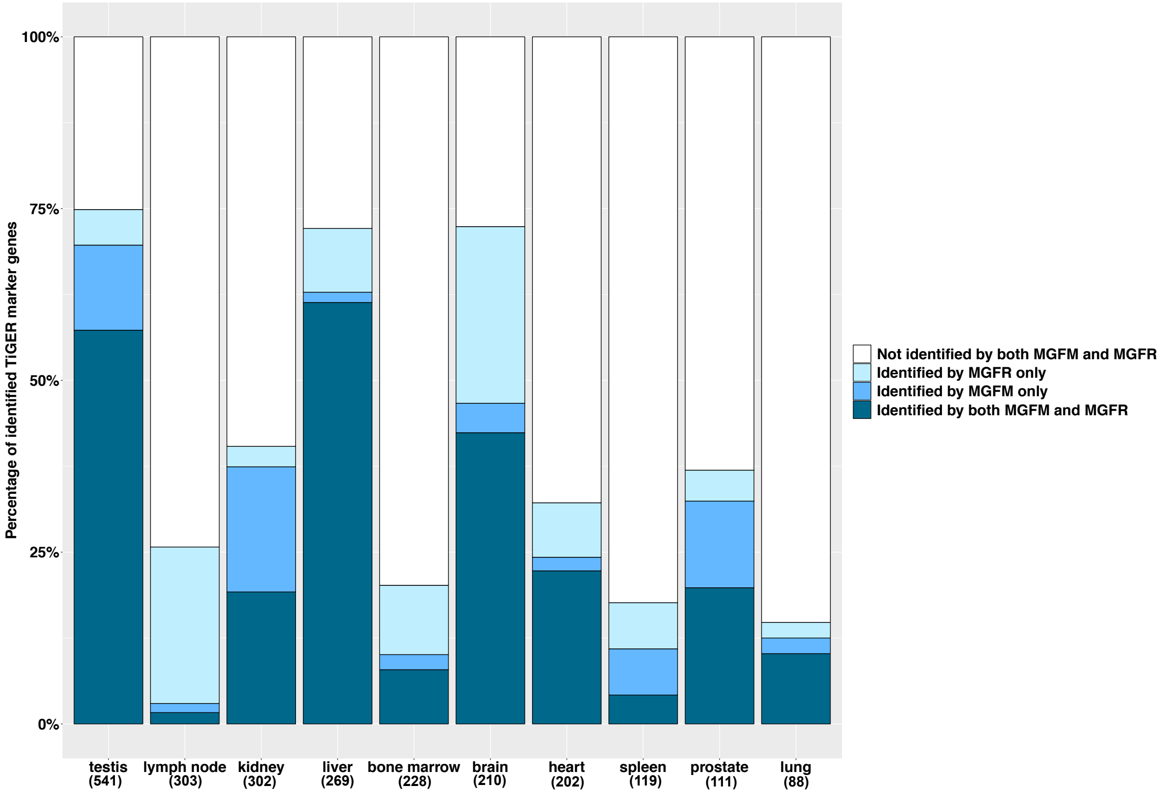


Figure 2.11: Number of gold-standard marker genes obtained from the TiGER database (shown in parentheses), percentage of correctly identified marker genes by `MGFM` and `MGFR` for ten of the examined tissues, and the overlap of identified markers by both tools.

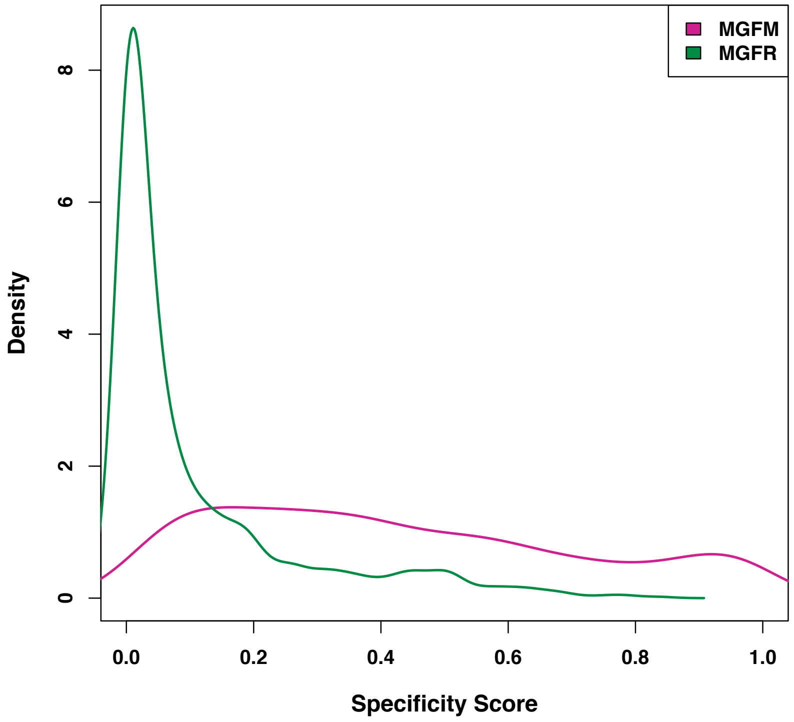


Figure 2.12: Specificity score distribution of correctly identified marker genes by `MGFM` and `MGFR`.

Robust marker genes: overlap of sets of predicted marker genes

For each of the examined tissues, I compared the lists of predicted marker genes by MGFM and MGFR using the reduced datasets with genes common to both the microarray and RNA-seq dataset (Figures 2.10 and 2.13.a). I define the marker genes of a tissue that are predicted by both MGFM and MGFR as robust marker genes, and the genes predicted for two different tissues as conflicting markers. The largest number of conflicting markers was obtained for bone marrow, for which 40 of the marker genes predicted by MGFR were predicted by MGFM for testis. To evaluate the consistency of predicted markers by MGFM and MGFR for a tissue, the ratio of the number of conflicting markers to the number of robust markers was calculated (Figure 2.13.b). For all tissues the number of robust markers is higher than the number of conflicting markers, except for lymph node, for which 28 robust markers were predicted and 33 markers were conflicting markers. The lowest ratio of conflicting markers to robust markers (0.08) was obtained for heart, testis and liver.
The marker genes predicted by each tool as well as the set of robust marker genes are available on github at https://github.com/khadija-a/Marker-genes/blob/master/predicted-marker-genes.xlsx.

Gene Ontology enrichment analysis

The Gene Ontology (GO) (Ashburner et al., 2000) is a comprehensive resource, which provides a dynamic structure of biological knowledge using a controlled vocabulary consisting of GO terms. The GO describes function with respect to three aspects: molecular function (the biochemical activities performed by gene products), biological process (biological objective to which the gene or gene product contributes), and cellular component (the place in the cell where a gene product is active) (Ashburner et al., 2000). To assess whether the sets of robust marker genes (i.e. predicted by both MGFM and MGFR) show significant over-representation of biological characteristics related to their corresponding tissues, GO enrichment analysis was performed. The GO enrichment (molecular function and biological process) was calculated for the robust marker genes associated with ten of the examined tissues (bone marrow, brain, heart, kidney, liver, lung, lymph node, prostate, spleen, and testis). Gene Ontology enrichment analysis was assessed with the hypergeometric statistic as implemented in the R package GOstats (Falcon and Gentleman, 2007) (Version: 2.38.1), with all genes common to both the microarray and RNA-seq dataset (18415 genes) as background. The cutoff for p-values was 0.01. For each tissue, the significantly enriched top GO terms that do not overlap more than 80% are displayed (Tables 2.2 and 2.3). The top enriched GO terms demonstrate that many of the predicted marker genes for the examined tissues have functions

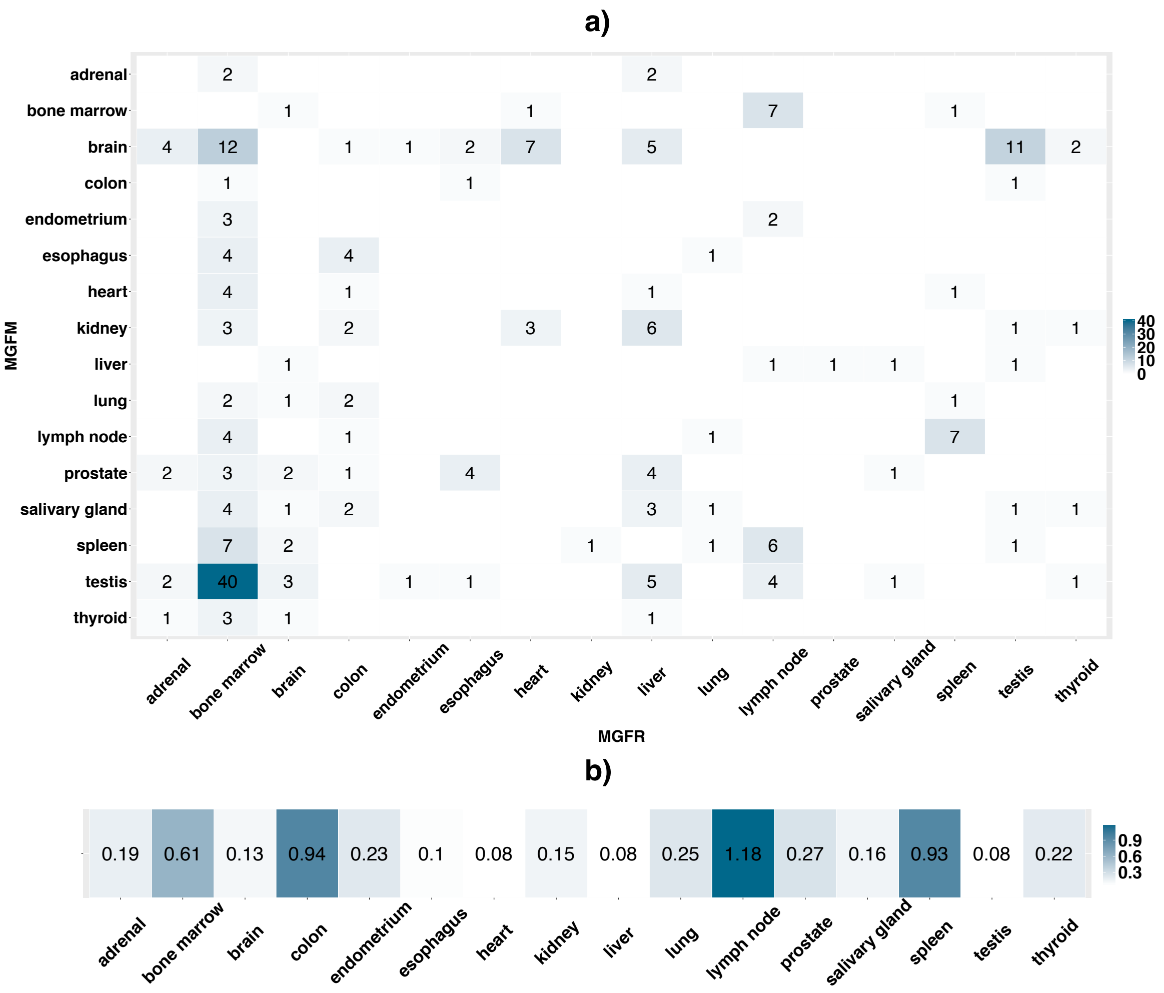


Figure 2.13: a) Number of conflicting marker genes for the examined tissues, that is the number of marker genes predicted by MGFM and MGFR for two different tissues. b) Ratio of the number of conflicting markers to number of robust markers for each of the examined tissues.

consistent with these tissues. For example, the predicted marker genes for brain were enriched in *syntaxin binding* and *glutamate receptor activity*, those for heart were enriched in *actin binding* and *titin binding*, kidney enriched terms related to transporter activity such as *carboxylic acid transmembrane transporter activity* and *solute:sodium symporter activity*, and liver predictions were associated with *oxidoreductase activity* and *alcohol binding*.

Detection of novel marker genes

The set of robust marker genes contained known marker genes, but also genes that have been reported in recent studies as novel marker genes, such as *RTP3* (also known as *TMEM7*), *SRHC*, *TTC36* (also known as *HBP21*), *TNFAIP8L1*, and *ETNPPL* in liver, and *RTKN2* in lung, and *TMEM72* in kidney. Wrzesiński et al. (Wrzesiński et al., 2015) reported a downregulation of *TMEM72* in clear cell renal cell carcinoma (ccRCC). In recent studies, *RTP3* (Zhou et al., 2007), *SRHC* (Zheng et al., 2015), *TTC36* (Jiang et al., 2015), *TNFAIP8L1* (Zhang et al., 2015), and *ETNPPL* (Ding et al., 2016) were reported to be downregulated in hepatocellular carcinoma (HCC). *RTKN2* was reported as novel candidate marker gene of Idiopathic Interstitial Pneumonias (Steele et al., 2015a). The expression of these genes was found to be downregulated in diseased tissues as compared to normal tissues. Hence, I hypothesize that these disease-implicated genes in tissue-specific disease may play important roles in the function of normal tissues.

The set of robust marker genes also included genes that were not previously associated at all with the 16 examined tissues. Searching for the gene symbols of these marker genes in association with the corresponding tissue in NCBI PubMed (http://www.ncbi.nlm.nih.gov/pubmed, all publications until March 29, 2017) yielded no publications. I suggest these genes as novel candidate marker genes for further investigation (https://github.com/khadija-a/Marker-genes/blob/master/Novel-candidate-marker-genes.xlsx). Figure 2.14 shows heatmaps of the suggested novel candidate marker genes for each of the examined tissues using each of the two datasets. Using both the microarray and the RNA-seq dataset, these novel candidate marker genes show tissue-specific expression.

2.3 Discussion

I developed a marker tool named `MGFM` (Marker Gene Finder in Microarray gene expression data) as an R package to detect marker genes from microarray data. Within the scope of this thesis, this tool was further optimized and updated to support the detection of marker genes from RNA-seq data. For this purpose, an R package called `MGFR` (Marker Gene Finder in RNA-seq data) was implemented. Marker genes are a group of genes whose expression is

Table 2.2: Gene Ontology enrichment (Molecular Function) of predicted marker genes for ten of the examined tissues. Column labels are as follows: GO ID is the GO identifier; GO is the description of the GO term; p-value is the hypergeometric p-value for over-representation of each GO term; Odds Ratio is an indicator of the level of enrichment in genes within the list as against the universe; Expected/Gene Count are the expected and actual gene counts; and Size is the number of genes within each GO term.

GO ID	GO	p-value	Odds Ratio	Expected Count	Gene Count	Size
Bone marrow						
GO:0050786	RAGE receptor binding	0.0001	42.26	0	3	10
GO:0030246	carbohydrate binding	0.0001	4.55	2	10	234
GO:0016744	transferase activity, transferring aldehyde or ketonic groups	0.0006	97.96	0	2	4
GO:0008329	signaling pattern recognition receptor activity	0.0006	21.12	0	3	17
GO:0008201	heparin binding	0.0008	5.02	2	7	147
Brain						
GO:0019905	syntaxin binding	4.14×10^{-10}	10.09	2	15	77
GO:0016917	GABA receptor activity	1.11×10^{-08}	27.33	0	8	20
GO:0030276	clathrin binding	7.45×10^{-06}	7.85	1	9	56
GO:0008066	glutamate receptor activity	3.82×10^{-05}	11.64	1	6	27
GO:0097109	neuroligin family protein binding	0.0001	60.66	0	3	5
Heart						
GO:0003779	actin binding	7.36×10^{-09}	5	5	22	374
GO:0008307	structural constituent of muscle	5.19×10^{-08}	18.94	1	8	40
GO:0031433	telethonin binding	9.68×10^{-06}	221.95	0	3	4
GO:0031432	titin binding	1.48×10^{-05}	37.17	0	4	12
GO:0005523	tropomyosin binding	2.12×10^{-05}	33.03	0	4	13
Kidney						
GO:0046943	carboxylic acid transmembrane transporter activity	4.44×10^{-09}	15.74	1	10	116
GO:0015370	solute:sodium symporter activity	5.33×10^{-07}	24.02	0	6	46
GO:0015296	anion:cation symporter activity	6.08×10^{-07}	23.44	0	6	47
GO:0015301	anion:anion antiporter activity	1.18×10^{-06}	33.04	0	5	29
GO:0015108	chloride transmembrane transporter activity	0.0002	10	1	5	84
Liver						
GO:0016491	oxidoreductase activity	1.53×10^{-27}	5.84	16	72	649
GO:0017171	serine hydrolase activity	4.87×10^{-15}	8.75	4	26	150
GO:0004857	enzyme inhibitor activity	1.33×10^{-12}	4.81	9	35	343
GO:0043168	anion binding	1.16×10^{-07}	1.93	61	100	2436
GO:0043178	alcohol binding	5.68×10^{-07}	6.469	2	13	94
Lung						
GO:0015114	phosphate ion transmembrane transporter activity	0.0003	97.72	0	2	13
GO:0019905	syntaxin binding	0.0005	22.53	0	3	77
GO:00509981	nitric-oxide synthase binding	0.0005	71.64	0	2	17
GO:0017075	syntaxin-1 binding	0.0006	67.16	0	2	18
Lymph node						
GO:0042608	T cell receptor binding	4.23×10^{-05}	315.7	0	2	6
GO:0004896	cytokine receptor activity	4.46×10^{-04}	23.17	0	3	88
GO:0030159	receptor signaling complex scaffold activity	7.01×10^{-04}	60.06	0	2	23
GO:0048020	CCR chemokine receptor binding	7.01×10^{-04}	60.06	0	2	23
Prostate						
GO:0003700	transcription factor activity, sequence-specific DNA binding	5.04×10^{-05}	4.43	4	13	1078
GO:0044212	transcription regulatory region DNA binding	9.05×10^{-04}	4.08	3	9	749
GO:0004252	serine-type endopeptidase activity	9.07×10^{-04}	10.25	0	4	126
Spleen						
GO:0008889	glycerophosphodiester phosphodiesterase activity	5.92×10^{-05}	252.5	0.01	2	7
GO:0042288	MHC class I protein binding	3.35×10^{-04}	90.14	0	2	16
GO:0060089	molecular transducer activity	5.15×10^{-04}	5.44	2	8	1164
Testis						
GO:0003796	lysozyme activity	0.001	20.19	0	3	7
GO:0004004	ATP-dependent RNA helicase activity	0.001	4.09	2	8	61
GO:0004175	endopeptidase activity	0.002	2.01	13	24	351
GO:0004802	transketolase activity	0.004	53.74	0	2	3
GO:0034584	piRNA binding	0.004	53.74	0	2	3

Table 2.3: Gene Ontology enrichment (Biological Process) of predicted marker genes for ten of the examined tissues. Column labels are as follows: GO ID is the GO identifier; GO is the description of the GO term; p-value is the hypergeometric p-value for over-representation of each GO term; Odds Ratio is an indicator of the level of enrichment in genes within the list as against the universe; Expected/Gene Count are the expected and actual gene counts; and Size is the number of genes within each GO term.

GO ID	GO	p-value	Odds Ratio	Expected Count	Gene Count	Size
Bone marrow						
GO:0034101	erythrocyte homeostasis	7.39×10^{-09}	12.5	1	11	102
GO:0002446	neutrophil mediated immunity	1.03×10^{-08}	33.66	0	7	28
GO:0001817	regulation of cytokine production	1.36×10^{-07}	4.4	6	21	537
GO:0045321	leukocyte activation	4.24×10^{-07}	3.81	7	23	679
GO:0060326	cell chemotaxis	1.035×10^{-05}	5.7	2	11	209
Brain						
GO:0045664	regulation of neuron differentiation	5.92×10^{-22}	5.66	13	57	517
GO:0007409	axonogenesis	1.71×10^{-16}	4.05	18	59	718
GO:0097479	synaptic vesicle localization	5.44×10^{-16}	11.96	3	23	104
GO:0050803	regulation of synapse structure or activity	1.02×10^{-15}	7.05	6	32	227
GO:0007269	neurotransmitter secretion	1.62×10^{-14}	9.29	3	24	133
Heart						
GO:0060048	cardiac muscle contraction	5.49×10^{-19}	20.11	1	21	104
GO:0055007	cardiac muscle cell differentiation	1.14×10^{-17}	21.05	1	19	90
GO:0002027	regulation of heart rate	3.22×10^{-13}	17.24	1	15	82
GO:0030048	actin filament-based movement	1.55×10^{-12}	13.44	1	16	108
GO:0055008	cardiac muscle tissue morphogenesis	8.8×10^{-12}	21.18	1	12	55
Kidney						
GO:0098656	anion transmembrane transport	9.72×10^{-18}	19.04	1	20	214
GO:0006812	cation transport	1.72×10^{-05}	3.48	6	19	963
GO:0072015	glomerular visceral epithelial cell development	3.36×10^{-05}	65.68	0	3	10
GO:0072017	distal tubule development	4.6×10^{-05}	57.47	0	3	11
GO:0072044	collecting duct development	6.1×10^{-05}	51.08	0	3	12
Liver						
GO:0019752	carboxylic acid metabolic process	1.03×10^{-53}	7.71	25	123	974
GO:0055114	oxidationreduction process	1.63×10^{-30}	5.06	25	95	993
GO:0006066	alcohol metabolic process	3.45×10^{-19}	6.02	9	46	370
GO:0015721	bile acid and bile salt transport	2.27×10^{-15}	36.82	1	14	29
GO:0017144	drug metabolic process	1.35×10^{-13}	30.08	1	13	30
Lung						
GO:0007585	respiratory gaseous exchange	1.44×10^{-05}	32.04	0	4	64
GO:0010193	response to ozone	0.0001	180.9	0	2	7
GO:0072593	reactive oxygen species metabolic process	0.00019	10.77	1	5	233
GO:0043129	surfactant homeostasis	0.0003	90.42	0	2	12
GO:0051384	response to glucocorticoid	0.0004	12.65	0	4	155
Lymph node						
GO:0050863	regulation of T cell activation	2.33×10^{-16}	54.71	0	13	273
GO:0030217	T cell differentiation	4.18×10^{-16}	62.81	0	12	207
GO:0050851	antigen receptor-mediated signaling pathway	5.17×10^{-13}	49.67	0	10	190
GO:0001816	cytokine production	6.5×10^{-06}	10.45	1	8	599
GO:0071345	cellular response to cytokine stimulus	1.81×10^{-05}	9.01	1	8	689
Prostate						
GO:0030518	intracellular steroid hormone receptor signaling pathway	4.14×10^{-05}	14.83	0	5	115
GO:0060740	prostate gland epithelium morphogenesis	1.24×10^{-04}	36.21	0	3	29
GO:0043401	steroid hormone mediated signaling pathway	2.34×10^{-04}	10.1	1	5	166
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	3.46×10^{-04}	4.02	3	11	982
GO:0002067	glandular epithelial cell differentiation	5.29×10^{-04}	21.37	0	3	47
Spleen						
GO:0002449	lymphocyte mediated immunity	2.07×10^{-05}	18.11	0	5	193
GO:0072643	interferon-gamma secretion	4.13×10^{-04}	80.39	0	2	17
GO:0046636	negative regulation of alpha-beta T cell activation	8.32×10^{-04}	54.78	0	2	24
Testis						
GO:0007283	spermatogenesis	1.72×10^{-94}	17.46	14	133	426
GO:0007281	germ cell development	3.81×10^{-40}	13.54	7	60	203
GO:0007126	meiotic nuclear division	6.02×10^{-28}	12.17	5	43	153
GO:0035036	sperm-egg recognition	3.54×10^{-19}	28.37	1	20	41
GO:0043046	DNA methylation involved in gamete generation	2.97×10^{-14}	58.64	1	12	18

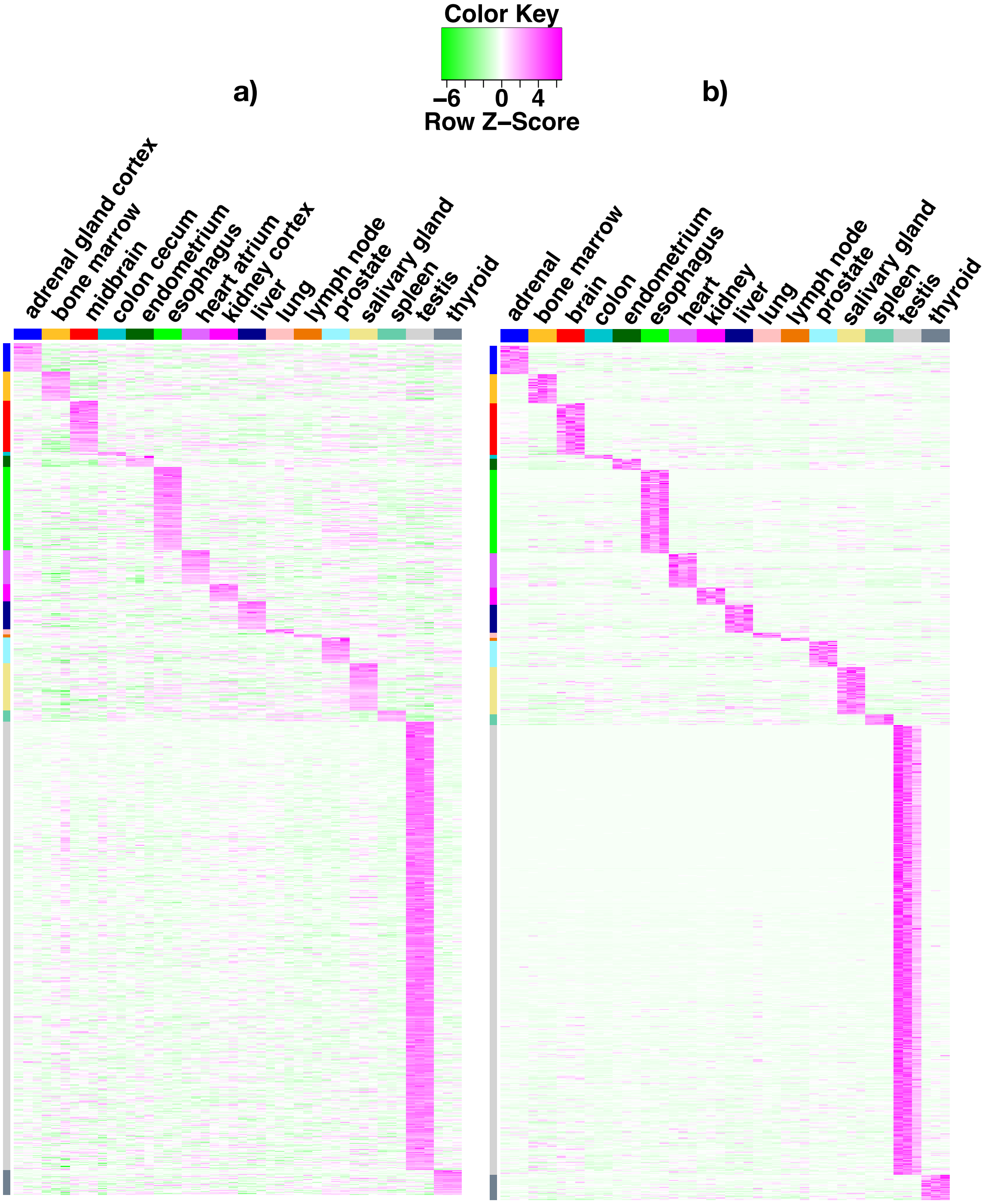


Figure 2.14: Expression of the suggested novel candidate marker genes in each of the examined tissues using a) the microarray dataset and b) the RNA-seq dataset. These marker genes are selected from the set of robust marker genes (i.e. predicted by both MGFM and MGFR). The scaled expression of each gene, denoted as the row Z-score, is plotted in green–magenta colour scale with magenta indicating high expression and green indicating low expression.

elevated in one or several tissues or cell types. Identification of these genes helps to better understand tissue-specific gene function and the molecular mechanisms underlying complex diseases. Moreover, marker genes are of great importance to determine tissue identity and to characterize cells grown *in vitro*. Both MGFM and MGFR require a reference dataset with replicates for each sample type. It is worth noting that the list of predicted marker genes for a sample type expectedly depends on the sample types included in the reference dataset, and this may differ by adding or removing a sample type from the reference dataset. A possible method to identify marker gene candidates is to identify genes that are differentially expressed between two experimental groups using a statistical test such as a *t*-test. Genes associated with each sample type could be used as markers. While this procedure enables the comparison of two sample types, my tools enable the comparison of multiple sample types in one run. In contrast to very comprehensive but static databases of tissue-specific genes such as TiGER (Liu et al., 2008) or PaGenBase (Pan et al., 2013), my tools enable users to easily modify and adapt the sample types in the reference to their set of interest.

I applied my tools MGFM and MGFR to a microarray and an RNA-seq dataset profiling 16 human tissues (adrenal, bone marrow, brain, colon, endometrium, esophagus, heart, kidney, liver, lung, lymph node, prostate, salivary gland, spleen, testis, and thyroid). In agreement with another study (Uhlén et al., 2015), both MGFM and MGFR predicted more marker genes for testis, brain, and liver. To evaluate the performance of my tools, I compared their results with tissue-specific genes obtained from the TiGER database for a set of ten human tissues (bone marrow, brain, heart, kidney, liver, lung, lymph node, prostate, spleen, and testis). Importantly, the gold-standard is independent from the used microarray or RNA-seq gene expression data as the gold-standard marker genes from TiGER were calculated based on EST (expressed sequence tag) counts. The use of EST counts to quantify gene expression levels is less sensitive than microarray technology. Therefore, the used gold-standard marker gene lists from TiGER are not comprehensive, and marker genes predicted with my tools may not be contained in the lists from TiGER. First, I tested each tool separately on the whole microarray or RNA-seq dataset. MGFR identified 999 of 2512 known marker genes (39.8%), slightly more than MGFM, which identified 968, but could only potentially detect 2394 TiGER markers due to the absence of probes on the microarray for some genes and thus performed with similar efficiency (40.4%). Together, MGFM and MGFR covered 1220 TiGER genes. To facilitate the comparison in the detection of markers between MGFM and MGFR, eliminating the array specific limits in detecting genes, I repeated the analysis focusing on the set of TiGER genes that both methods could potentially detect (2373 genes). MGFR identified 965 marker genes (or 40.7%) of the 2373 gold-standard marker genes, whereas MGFM identified 898 marker genes (or 37.8%) of the gold-standard marker genes.

To assess whether the sets of robust marker genes (i.e. genes predicted by both MGFM and MGFR) show significant over-representation of biological characteristics related to their corresponding tissues, the GO enrichment (molecular function and biological process) was calculated for the robust marker genes associated with ten of the examined tissues (bone marrow, brain, heart, kidney, liver, lung, lymph node, prostate, spleen, and testis). The top enriched GO terms demonstrate that many of the predicted marker genes for the examined tissues have functions consistent with these tissues (Tables 2.2 and 2.3).

The set of predicted marker genes for the 16 examined tissues contained known marker genes as well as genes not previously associated with the corresponding tissues. Searching for the gene symbols of these markers in association with the corresponding tissue in NCBI PubMed (http://www.ncbi.nlm.nih.gov/pubmed, all publications until March 29, 2017) returned no results. I suggest these genes as novel candidate marker genes for further investigation. Finally, I investigated the expression of top marker genes predicted by MGFM in another study for a set of five human tissues (brain, heart, kidney, liver, and lung). I was able to test the marker genes experimentally by RT-PCR in all five tissues. While not all marker genes were unambiguous markers, and some were not detected, the vast majority (92%) was experimentally confirmed (Table 2.1).

Chapter 3

Gene expression profile-based sample classification

Discrimination between different classes of samples such as different cell types or tissues using gene expression profiles is of great importance in cell research. It has several implications and can contribute to our understanding of cell phenotype differences and will allow precise identification of various cell types and tissues. I developed a bioinformatics tool for the classification of samples based on gene expression profiles. The tool requires a training and a test dataset (Figure 3.1), and uses a simple algorithm called "Shared Marker Genes" (SMG). As the name suggests, the number of shared marker genes between a reference and a query sample is used as a similarity measure. Marker genes are detected using the tool `MGFM` for microarray data and `MGFR` for RNA-seq data, which have been described in Chapter 2 and are available as Bioconductor R packages. I demonstrate the utility and effectiveness of the proposed approach by the classification of different tissues using public microarray and RNA-seq datasets. I verified my tool using 186 test samples from four human tissues (heart, kidney, liver and lung), from the NCBI's Gene Expression Omnibus public repository. My approach accurately classified 99% of these 186 test samples. Furthermore, I compared my tool to Support Vector Machines (SVMs). My approach performed comparably or better than SVMs. The tool is implemented as an R package named `sampleClassifier`, which is available from the Bioconductor website (https://bioconductor.org/packages/release/bioc/html/sampleClassifier.html). `sampleClassifier` can be applied: i) to evaluate the similarity of experimentally derived cells with their desired target cell type; ii) to compare *in vitro* derived organoids (e.g. kidney organoids) to their *in vivo* counterparts; iii) to classify different types of diseases.
To facilitate the use of `sampleClassifier`, a data package called `sampleClassifierData` (https://bioconductor.org/packages/release/data/experiment/html/sampleClassifierData.html) was implemented, which contains a collection of publicly available microarray and RNA-

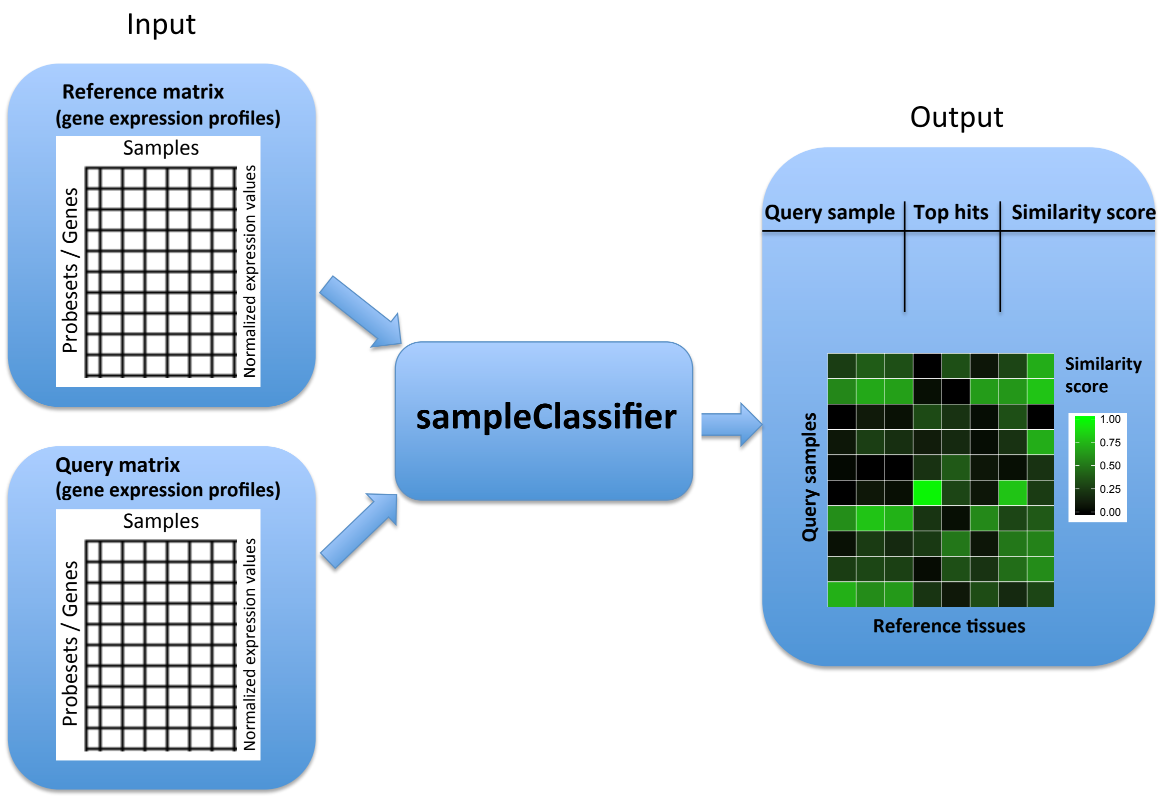


Figure 3.1: Overview of `sampleClassifier`

seq datasets that have been pre-processed for use with the `sampleClassifier` package. These pre-processed datasets can be used as reference matrices for gene expression profile classification using `sampleClassifier`.

3.1 Materials and methods

3.1.1 Data sources and pre-processing

The reference matrix for microarray data was derived from 78 samples from 26 tissues from the study GSE3526 (Roth et al., 2006a) with three replicates each (Figure 3.2). The sample accession numbers and tissue types are listed in Table B.1 (Appendix B). To test the performance of the `sampleClassifier` algorithm, microarray data from independent studies of four human tissues: heart, kidney, liver, and lung (Table 3.1) were used as test samples. In addition, the tool was tested on 16 samples from another independent study, GSE2361 (Ge et al., 2005).

The reference matrix for RNA-seq data was derived from 71 samples from 24 tissues (Figure

Table 3.1: Microarray test samples

Tissues	**Study ID**	**Number of Samples**
Heart	GSE29819	38
Kidney	GSE22459	25
Liver	GSE12720	63
Lung	GSE33356	60
Bone marrow, cerebellum, heart, kidney, liver, lung, ovary, pituitary gland, prostate, skeletal muscle, spinal cord, spleen, testis, and thalamus	GSE2361	16

3.2) from the study E-MTAB-1733 (Fagerberg et al., 2014), which is available from the ArrayExpress database. Each tissue was represented by three replicates, except ovary, which was represented by two replicates (Table B.2). For both the microarray and RNA-seq reference dataset, I performed hierarchical clustering with an average linkage and an Euclidean distance metric, and selected three replicates for each tissue that showed the highest similarity to each other.

The RNA-seq test samples were derived from the study E-MTAB-513 (Illumina Body Map). The fetal reference dataset used in Section 3.2.3 which contains transcriptional signatures of 13 fetal tissues from the first trimester was downloaded from the `KeyGenes` website (http://www.keygenes.nl). The test RNA-seq dataset (analyzed in Section 3.2.3) profiling kidney organoid differentiation from 4 time points (day 0, 3, 11 and 18) is available from GEO under the accession number GSE70101 (Takasato et al., 2015).

Two methods were used for normalization of the microarray datasets, the Robust Multiarray Averaging (Irizarry et al., 2003) (`RMA`) and `YuGene` (Lê Cao et al., 2014). `RMA` consists of the following three particular processing steps:

1. background correction to adjust raw perfect match (PM) probe intensities using a model based on observed intensity being the sum of signal and noise.

2. quantile normalization (Bolstad et al., 2003) of corrected PM probes. The aim of quantile normalization is: to make the distribution of probe intensities the same for every chip, and to average each quantile across chips.

3. Summarization: in this step the multiple probe intensities for each probeset are combined to produce an expression value using median polish. A robust multichip linear

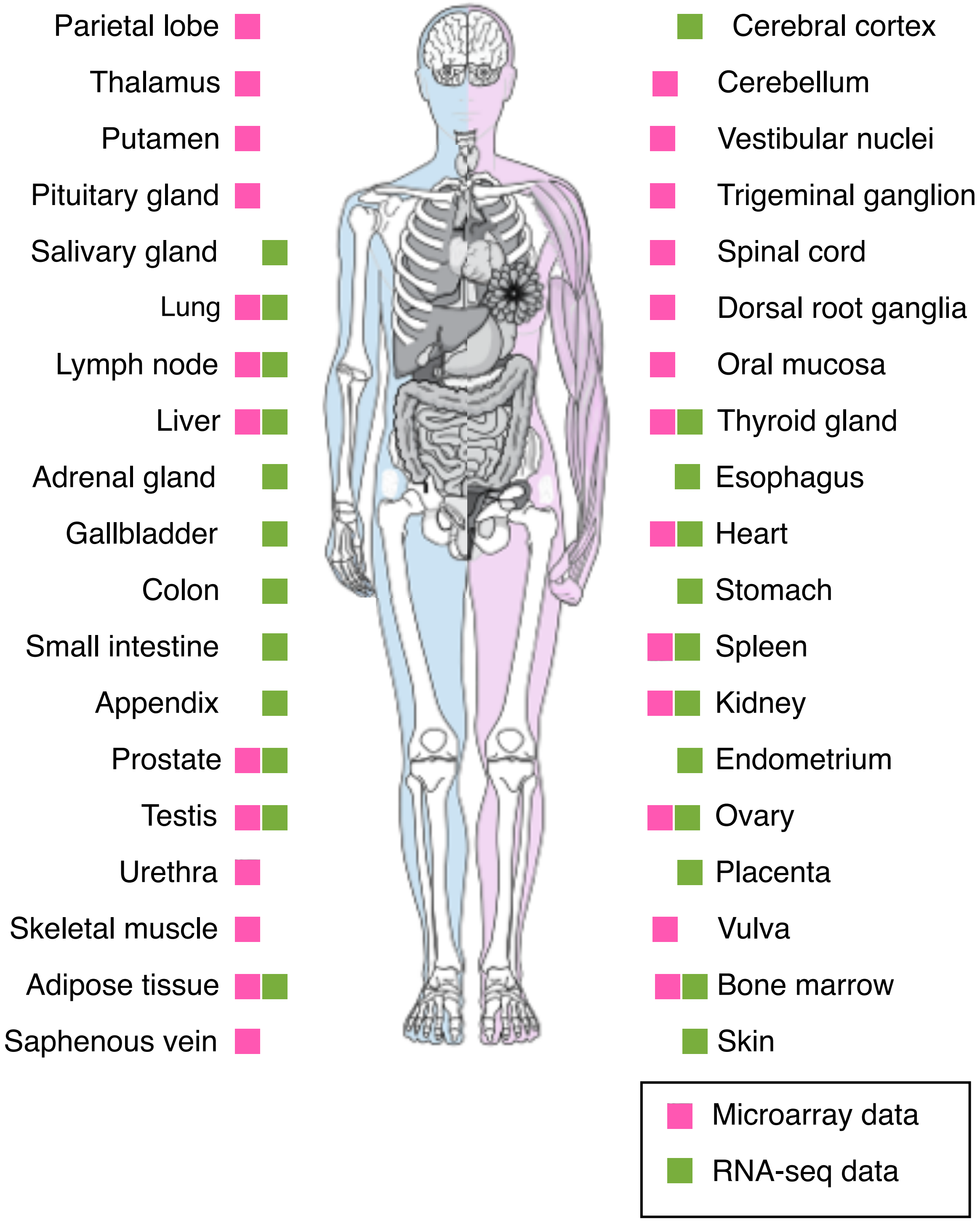


Figure 3.2: The human tissues and organs represented in the microarray (magenta) and RNA-seq (green) reference datasets.

model is fitted to the log of the preprocessed PM probes for a particular probeset. In particular for a probeset k with $i = 1, ..., I_k$ probes and data from $j = 1, ..., J$ arrays, the following model is fitted:

$$\log_2(PM_{ij}^k) = \alpha_i^k + \beta_j^k + \varepsilon_{ij}^k$$

where α_i is the probe effect and β_j is the $\log_2$ expression value.

YuGene is briefly explained in Section 2.1.

Prior to YuGene normalization, each dataset was background corrected, and log2 transformed. To classify samples from a different platform than the reference, multiple probesets that mapped to the same gene were summarized using their mean expression value. This step is implemented in the function `classifyProfile()`.

The reads from the study E-MTAB-1733 and E-MTAB-513 were mapped to the GRCh37 version of the human genome with TopHat v2.1.0 (Trapnell et al., 2009). FPKM (fragments per kilobase of exon model per million mapped reads) values were calculated using cuffquant and cuffnorm from the Cufflinks package v2.2.1 (Trapnell et al., 2010). The data from the study E-MTAB-1733 were extracted after the calculation of FPKM values for all samples and averaging across technical replicates.

KeyGenes was run with the fetal training and test set (analyzed in Section 3.2.3) based on raw read counts. For the classification with `sampleClassifier`, the raw counts from both the fetal training and test set were converted to CPM (counts per million) values using the `cpm` function in `edgeR` (version: 3.14.0) (Robinson et al., 2010). In addition, the gene symbols from the test dataset were converted to Ensembl Gene IDs, and matched to the fetal dataset, only common genes to both datasets were considered.

3.1.2 Implementation

The classification tool `sampleClassifier` is implemented in the programming language R. It provides functions to classify microarray and RNA-seq gene expression profiles. Microarray data can be classified using the function `classifyProfile()` and RNA-seq data can be classified using the function `classifyProfile.rnaseq()`. Both functions expect a normalized reference and a query matrix. One query profile or multiple query profiles can be classified in one run. In principle, each query profile is compared to each sample type in the reference and a similarity score is calculated. The class of the sample type with the highest similarity score is predicted as class of the query profile. To visualize the classification results in a heatmap, the function `get.heatmap()` can be used with the output list generated by `classifyProfile()` or `classifyProfile.rnaseq()` as input. In order to compare the

classification results to SVMs, I implemented the functions classifyProfile.svm() and classifyProfile.rnaseq.svm() for microarray and RNA-seq data, respectively. These functions are based on the functions svm() and predict() from the R package e1071 (Meyer et al., 2017). The input reference matrix is reduced to the marker probesets or genes that are used by my tool for classification and used as a training matrix for SVMs. Similarly, the test matrix is reduced to contain marker probesets or genes. More details about the input parameters of the functions provided in sampleClassifier can be found in the Vignette of the R package.
To facilitate the use of sampleClassifier, a data package called sampleClassifierData (https://bioconductor.org/packages/release/data/experiment/html/sampleClassifierData.html) was implemented, which contains pre-processed microarray and RNA-seq datasets to be used as reference for gene expression profile classification using sampleClassifier.

3.1.3 Output

The functions classifyProfile() and classifyProfile.rnaseq() return as output a list with a data frame for each query profile. Each data frame contains a comparison of the query profile to each sample type in the reference, with the hits sorted according to their similarity to the query profile. For each comparison the similarity score, which is the ratio of the number of shared marker genes to the number of marker genes used for the classification is provided (see Section 3.1.4 for more details). In addition, the output can be written to a file by setting the parameter *write2File* to TRUE. In this case the markers shared between each query and each sample type in the reference are also written to the file. The function get.heatmap() displays the classification results in a heatmap. The functions classifyProfile.svm() and classifyProfile.rnaseq.svm() return a data frame with the predicted class for each query profile using SVMs.

3.1.4 sampleClassifier algorithm details

The algorithm used in sampleClassifier is a simple algorithm called "Shared Marker Genes" (SMG). As the name indicates, the number of shared marker genes between a query and a reference is used as similarity measure. The steps of the algorithm are summarized in Figure 3.3. The tool requires a reference matrix with replicates for each sample type. This matrix is used for marker gene detection using MGFM for microarray data or MGFR for RNA-seq data. Since the number of detected markers differs depending on the sample types, I filter the list of marker genes of each sample type. Using the complete list of markers of each sample type will result in a bias towards the sample type with the largest number of marker genes. For example, if testis is the tissue with the largest number of marker genes, using all marker genes for classification will result in classifying query samples often as testis.

Let X be an $n \times m$ normalized gene expression data matrix:

$$X = \begin{array}{c} \\ G_1 \\ G_2 \\ \vdots \\ G_n \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & \cdots & S_m \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix} \end{array}$$

where $x_{i,j}$ is the gene expression level of the i^{th} gene in the j^{th} sample, for $i = 1,...,n$, and $j = 1,...,m$. Let $L_i = \{G_1,...,G_{l_i}\}$ denote the list of predicted marker genes associated with a reference sample type S_i. I filter the list of marker genes L_i as follows:

$$L_i' = \begin{cases} L_i, & |L_i| \leq \text{median}(|L_1|,....,|L_s|) \\ \left\{G_1,...,G_{\lceil \text{median}(|L_1|,....,|L_s|) \rceil}\right\}, & otherwise \end{cases}$$

where s is the number of unique sample types in the reference matrix X. L_i' is the filtered list of marker genes used for classification. For example if the reference matrix contains four tissues: liver, lung, kidney, and testis, and $v = (16, 20, 100, 500)$ is the vector of lengths of predicted marker genes for these tissues. The filtering is based on the median number of marker genes, in this case $\text{median}(v) = 60$. If the number of predicted markers for a tissue $\leq$ 60, then all markers are used for classification. If the number of predicted markers for a tissue $>$ 60, then only the top 60 marker genes will be used for classification. Since the marker genes are sorted according to their specificity, the most specific genes are selected. After the filtering step, each query sample will be compared to all sample types in the reference and the number of marker genes shared between the query and each sample type in the reference is calculated. Let Q be the gene expression profile of a query sample S_q to be classified:

$$Q = \begin{array}{c} \\ G_1 \\ G_2 \\ \vdots \\ G_n \end{array} \begin{array}{c} S_q \\ \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{pmatrix} \end{array}$$

To assess the similarity of S_q to a reference sample S_i, I calculate the number of marker genes of S_i (from the filtered list L_i'=$\{G_1,...,G_p\}$) that are shared with the query sample. Let $X_{G,k}$=$\{x_{k,1},...,x_{k,m}\}$ denote the vector of expression values of the marker gene $G_k \in L_i'$ in the reference matrix X sorted in decreasing order, and r denote the number of replicates of the sample type S_i in the reference matrix. By definition of a marker gene (see Section 2.1.2), the first highest r expression values in $X_{G,k}$ correspond to the sample type S_i. A query sample S_q shares a marker gene G_k with a reference sample S_i if:

$$q_k \geq x_{k,r+1}$$

where q_k is the expression value of the marker gene G_k in the query sample S_q. I define a score $S_{(G_k,S_q)}$ for each marker gene G_k in the filtered list of marker genes of the sample type S_i as:

$$S_{(G_k,S_q)} = \begin{cases} 1, & q_k \geq x_{k,r+1} \\ 0, & otherwise \end{cases}$$

$S_{(G_k,S_q)}$=1 means that the query sample S_q shares the marker gene G_k with the reference sample S_i. Finally, the similarity score of the query sample S_q to the reference sample type S_i is defined as:

$$S_{(S_q,S_i)} = \frac{\sum_{k=1}^{p} S_{(G_k,S_q)}}{p}$$

where $k \in \{1,...,p\}$ and $p = |L_i'|$. $S_{(S_q,S_i)}$ is the ratio of the number of shared marker genes and the total number of markers used for classification and has a value in $[0,1]$. A value of 1 means that the query S_q shares all marker genes with the reference sample S_i, and a value of 0 means that no marker genes are shared between the query and the reference sample. For each query sample, the hits are sorted according to this score. The class of the first top hit is predicted as a class for the query sample.

3.2 Results

I applied `sampleClassifier` to classify gene expression data from public repositories. Here, I will report the performance of my tool on microarray and RNA-seq data. For the microarray data, I tested two normalization methods, namely `RMA` and `YuGene`. I chose `RMA` because it is commonly used in analyzing Affymetrix microarray data. The choice of `YuGene` was motivated by the desire to test the impact of technical batch effects on the classification of test samples from different studies or platforms than the reference.

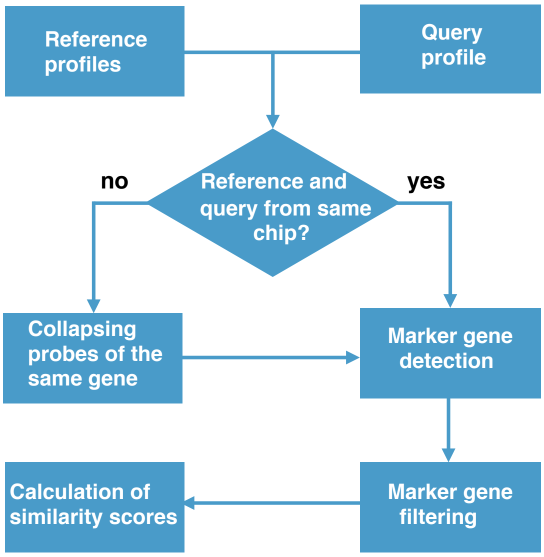

Reference profiles
Query profile
Reference and query from same chip?
no
yes
Collapsing probes of the same gene
Marker gene detection
Calculation of similarity scores
Marker gene filtering

Figure 3.3: Workflow of `sampleClassifier`

3.2.1 Classification of tissue types based on microarray data

First, I tested my tool on microarray test samples from the same study as the reference (GSE3526). Samples that were not chosen as reference samples (Table B.1, Appendix B) were used as test, that is 61 samples from 21 tissues. Using `YuGene` or `RMA` for normalization, `sampleClassifier` classified 55 (90%) or 54 samples (89%) of the 61 test samples correctly with a mean similarity score of 0.63 and 0.64, respectively, (Appendix B, Figures B.1 and B.2). Three samples representing lymph nodes, pituitary gland, thalamus, and two samples representing vestibular nuclei superior were misclassified using both `YuGene` and `RMA`. Since most of the samples from the same tissue type were classified correctly, a possible reason for misclassification might be variation in gene expression due to low tissue quality. Using `YuGene`, the second best scores by three of the six misclassified samples pointed to the correct tissue, and using `RMA` the second best scores by four of the seven misclassified samples pointed to the correct tissue.

Next, I tested my tool on samples from the same platform as the reference, Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), but from different studies. I used a total of 186 test samples from four human tissues (heart, kidney, liver and lung), from GEO (Table 3.1). Using `YuGene`, `sampleClassifier` classified all test samples correctly except one lung sample, which was misclassified into lymph node and one kidney sample, which was misclassified into skeletal muscle (Appendix B, Figures B.3.a, B.4.a, B.5.a, and B.6.a). The correctly classified heart, kidney, liver and lung samples had a mean similarity score of 0.66, 0.74, 0.78 and 0.63, respectively. Using `RMA`, `sampleClassifier` classified all test samples correctly except two lung samples, which were misclassified into lymph node or testis, and one kidney sample, which was misclassified into skeletal muscle. The correctly classified heart, kidney, liver and lung samples had a mean similarity score of 0.76, 0.96, 0.94 and 0.9, respectively, (Appendix B, Figures B.3.b, B.4.b, B.5.b, and B.6.b). The lung sample (GSM494675) was misclassified using `YuGene` and `RMA` into lymph node with a similarity score of 0.43 or 0.82, respectively. The kidney sample (GSM557865) was misclassified using `YuGene` and `RMA` into skeletal muscle with a similarity score of 0.7 or 0.98, respectively. Since all other lung and kidney samples were correctly classified with high similarity scores, this may indicate that the samples GSM557865 and GSM494675 may not be in fact from lung or kidney, respectively. The misclassified lung sample (GSM494657) using `RMA` was correctly classified using `YuGene` with a low similarity score of 0.33.

Figure 3.4 shows a heatmap with the top 10 marker genes used for classification for each of the four test tissues (heart, kidney, liver and lung). Hierarchical clustering of genes was based on Pearson's correlation. This set of top 10 classifier genes contained genes that are verified as tissue-specific in previous publications or are said to contain tissue-related functions (marked with an asterisk (*)) (In heart: *FGF12* (Hennessey et al., 2013), *CORIN* (Pang et al., 2015), *TBX5* (Waldron et al., 2016), and *PKP2* (Ramond et al., 2017). In kidney: *SLC13A1* (Markovich, 2014), and *SLC12A1* (also known as *NKCC2*) (Igarashi et al., 1995). In liver: *ADH4* (Wei et al., 2012), and *AKR1D1* (Chaudhry et al., 2013). In lung: *SLC6A14* (Corvol et al., 2015), *CLDN18* (Shimobaba et al., 2016)). In addition, two genes RTKN2 (Steele et al., 2015b) and MTUS2 (Du Puy et al., 2009) (also known as CAZIP) were reported in recent publications as novel markers for lung or heart, respectively.

Finally, to validate the performance of my tool on samples from a different platform than the reference, I tested it on 16 samples from tissues represented in the reference dataset from the study GSE2361 (Affymetrix Human Genome U133A Array (GPL96)). Using `YuGene`, all samples were classified correctly, except one spleen sample, which was misclassified as bone marrow with a similarity score of 0.28 (Figure 3.5.a), perhaps not surprising, as both are hematopoietic organs. However, the second best score of 0.26 pointed to spleen. In addition, the fetal liver sample was assigned to liver and bone marrow with the same similarity score of 0.47. Using `RMA`, 6 samples were misclassified as testis (Figure 3.5.b). Using `YuGene`, more samples were classified correctly with a mean similarity score of 0.51, whereas using `RMA` the correctly classified samples had a mean similarity score of 0.96. In contrast to `YuGene`, the similarity scores of the top hits obtained using `RMA` are close to each other.

In order to test if the performance of my tool would improve after removing batch effects, I applied ComBat (Leek and Storey, 2007) from the R-package sva (Leek et al., 2016) to the

RMA normalized data. After adjusting the batch effects, all samples were classified correctly with a mean similarity score of 0.63 (Figure 3.6). The lowest score was obtained for the fetal lung sample (GSM44705), which was identified as lung with a similarity score of 0.17.

3.2.2 Classification of tissue types based on RNA-seq data

First, I tested my tool on RNA-seq test samples from the same study as the reference (E-MTAB-1733). Samples that were not included in the reference and from tissues represented in the reference were used as test, that is 18 samples from 12 tissues. sampleClassifier classified all 18 samples correctly with a mean similarity score of 0.85 (Figure 3.7). Next, I applied my tool to 12 samples from the study E-MTAB-513. My tool classified 9 of the 12 test samples correctly with a mean similarity score of 0.49 (Figure 3.8).

3.2.3 Application to kidney organoids

Human induced pluripotent stem cells (iPSCs) can give rise to multiple cell or tissue types and are attractive sources of cells for regenerative medicine and disease-modeling. A challenging task in stem cell research is to determine the similarity of iPSCs differentiated derivatives to their target cell or tissue types. To test sampleClassifier in this aspect, I applied it to an RNA-seq dataset representing kidney organoids from 4 time points (day 0, 3, 11 and 18 after aggregation) obtained from iPSCs (Takasato et al., 2015). As a reference for fetal development, I took the fetal training set from the KeyGenes publication (Roost et al., 2015). KeyGenes is an algorithm, which predicts the identity of a test tissue from its transcriptional profile based on deep serial analysis of gene expression (DeepSAGE) data of 21 human fetal and extra-embryonic tissues from the first and second trimester of development. DeepSAGE (Nielsen et al., 2006) is a technique that detects short tags of 21-22 base pairs at the most 3' end of a transcript. For this reason the data complexity is lower compared to RNA-seq. In order to compare my tool to KeyGenes, I applied it using a reference with 13 fetal tissues from the first trimester (Figure 3.9). Using sampleClassifier, all the transcriptional profiles obtained at day 0, 3, 11 and 18 showed similarity to kidney with a mean similarity score of 0.36, 0.41, 0.41, 0.39, respectively. Using the KeyGenes algorithm, the transcriptional profiles obtained at day 0 showed similarity to gonad with a mean identity score of 0.26. At day 3, one organoid showed similarity to kidney with an identity score of 0.24, whereas the other two replicates showed similarity to gonad with a mean score of 0.24. This is not unexpected, given the common embryologic origin of the kidneys and gonads from intermediate mesoderm. Finally, all profiles from day 11 and 18 showed similarity to kidney with mean identity scores of 0.89 and 0.62, respectively.

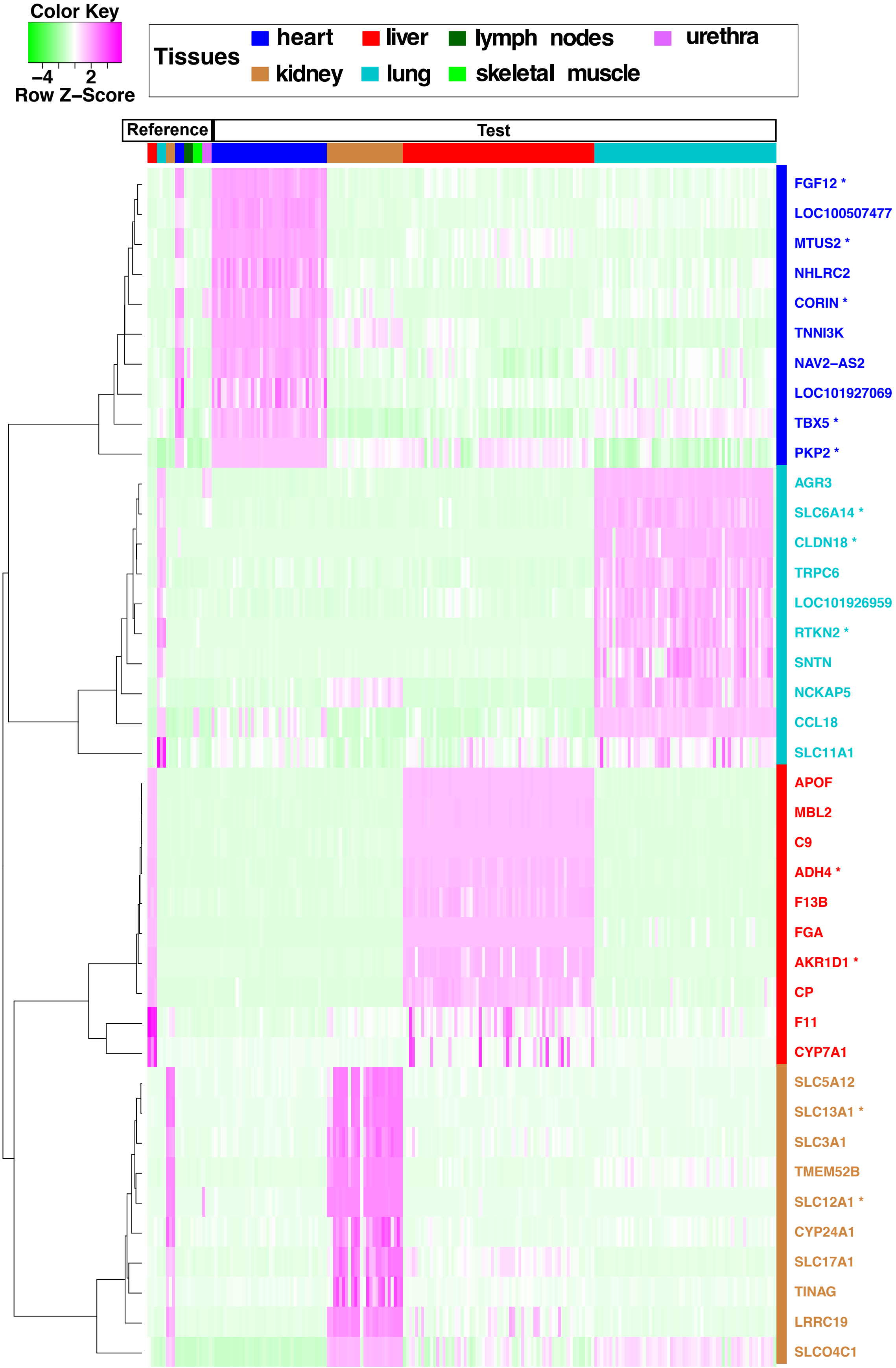


Figure 3.4: Top 10 marker genes used for classification for each of the tissues: heart, kidney, liver, and lung. Genes marked with an asterisk (*) are verified as tisssue-specific in previous publications. The scaled expression of each gene, denoted as the row Z-score, is plotted in green–magenta colour scale with magenta indicating high expression and green indicating low expression.

3.2.4 Comparison to SVMs

To evaluate my tool's classification performance, I compared `sampleClassifier` with Support Vector Machines (SVMs). SVMs (Cortes and Vapnik, 1995) are supervised learning models, which were first introduced by Vapnik in early 90s, and were developed for binary classification. SVMs build a classifier based on a training set, and seek for an optimal separating hyperplane between two classes by maximizing the margin between the classes' closest points. The points lying on the boundaries are called support vectors. The linear SVMs can be extended to nonlinear ones by transforming the problem into a feature space using a set of nonlinear basis functions.

I applied SVMs to the same test samples used to validate `sampleClassifier`. I used the R package `e1071` (Version: 1.6-8) (Meyer et al., 2017) to run SVMs. For each test I tested different kernels. The best results were obtained with the sigmoid kernel, except for the classification of samples from the study GSE2361 normalized with `YuGene`, for which SVMs performed best with the linear kernel (13 from 16 samples were correctly classified using linear kernel, and 12 samples were correctly classified using sigmoid kernel). It is worth noting that SVMs performed slightly better or similarly when using the sigmoid kernel compared to the linear kernel, whereas the worst performance was obtained with the polynomial kernel, and radial basis function kernel. Hence, for the following classification the sigmoid kernel was used, except for the classification of samples from the study GSE2361 normalized with `YuGene`, for which I show the results obtained with the linear kernel. First, I applied SVMs to classify 61 samples from the study GSE3526, which were not included in the reference. Using all probesets and `YuGene` or `RMA` for normalization, SVMs classified 49 samples (80%) and 42 (69%) correctly, respectively. Reducing the training and test datasets to the marker probesets used by my tool for classification (instead of using all probesets on the microarray) improves the accuracy of SVMs to 89% using `YuGene` and 85% using `RMA` (Table 3.2). In contrast, `sampleClassifier` classified 54 out of 61 samples (89%) correctly using `RMA`, and 55 (90%) using `YuGene`. The misclassified samples by SVMs are shown in Appendix B, Table B.3. In all the following tests I consider marker probesets that were used by my tool for classification (instead of using all probesets on the microarray). Next, I applied SVMs to classify 186 test samples from four human tissues (heart, kidney, liver and lung) taken from different studies, but run on the same platform as the reference (Table 3.1). SVMs performed similarly to `sampleClassifier`, except for lung, for which `sampleClassifier` misclassified two samples (out of 60 samples) using `RMA`, and SVMs misclassified one sample (Table 3.3). Importantly the lung sample GSM494675 and the kidney sample GSM557865 were misclassified by both SVMs and `sampleClassifier` into lymph nodes or skeletal muscle, respectively.

To evaluate the performance of SVMs on samples from a different platform than the reference, I applied them to classify 16 samples from the study GSE2361. SVMs classified 5 samples correctly using RMA, and 13 samples using YuGene. sampleClassifier classified 10 samples correctly using RMA and 15 samples correctly using YuGene (Table 3.4). The misclassified samples by SVMs are shown in Appendix B, Table B.4.

In order to test if the performance of SVMs will improve after removing batch effects, I applied ComBat from the R-package sva to the RMA normalized data. After adjusting the batch effects, SVMs classified 14 of the 16 test samples correctly. The prostate sample (GSM44678) and fetal lung sample (GSM44705) were misclassified into urethra or bone marrow, respectively. In contrast to SVMs, sampleClassifier classified all samples correctly after adjusting the batch effects. sampleClassifier seems to be less susceptible to batch effects compared to SVMs.

To test the performance for RNA-seq data, I applied SVMs to classify 18 RNA-seq samples from the study E-MTAB-1733. SVMs classified 17 samples correctly, whereas sampleClassifier classified all 18 test samples correctly (Table 3.5). Finally, I applied SVMs to classify 12 RNA-seq samples from a different study than the reference (E-MTAB-513). SVMs classified 7 samples correctly, whereas sampleClassifier classified 9 samples correctly (Table 3.5). The 3 samples misclassified by sampleClassifier were also misclassified by SVMs (Appendix B, Table B.5).

Table 3.2: Classification results of the microarray test data using sampleClassifier and SVMs

Method	RMA		YuGene	
	Correctly classified	Misclassified	Correctly classified	Misclassified
sampleClassifier	**54 (89%)**	7	**55 (90%)**	6
SVMs (using marker probesets only)	52 (85%)	9	54 (89%)	7
SVMs (using all probesets)	42 (69%)	19	49 (80%)	12

3.3 Discussion

Gene expression profile-based sample classification is a central problem in cell research. In addition, the differentiation of ESCs or iPSCs to different cell types is of great medical interest. A challenging task in stem cell research is to determine the similarity of iPSCs differentiated derivatives to their target cell or tissue types. Hence, there is a huge demand of

Table 3.3: Classification results of the microarray test samples from the study GSE2361 using `sampleClassifier` and SVMs

Method	**Tissue**	**RMA**		**YuGene**	
		Correctly classified	**Misclassified**	**Correctly classified**	**Misclassified**
sampleClassifier	heart	38	0	38	0
SVMs	heart	38	0	38	0
sampleClassifier	kidney	24	1	24	1
SVMs	kidney	24	1	24	1
sampleClassifier	liver	63	0	63	0
SVMs	liver	63	0	63	0
sampleClassifier	lung	58	2	59	1
SVMs	lung	**59**	1	59	1

bioinformatics tools to assess the identity of differentiated or reprogrammed cells. Several tools have been described for this purpose. For example, PluriTest (Müller et al., 2011) assesses the resemblance of cell samples to embryonic stem cells, based on gene expression profile comparison. CellNet (Cahan et al., 2014) assesses similarity of *in vitro* generated cells to 20 cell and tissue types, and KeyGenes (Roost et al., 2015) compares profiles of stem cell derivatives with those of fetal tissues. Here, I introduce a novel tool to classify samples based on their gene expression profiles. The tool is implemented as an R package called `sampleClassifier` and is available from the Bioconductor website. The tool supports the classification of microarray and RNA-seq gene expression profiles. It requires a reference and a test dataset, and uses a simple algorithm called "Shared Marker Genes" (SMG). As its name indicates, the number of shared marker genes between a reference and a query sample is used as a similarity measure. To facilitate the use of `sampleClassifier`, a data package called `sampleClassifierData` was implemented, which contains a collection of publicly available microarray and RNA-seq datasets that have been pre-processed for use with the `sampleClassifier` package. The microarray and RNA-seq datasets contain samples from 26 or 24 tissue types, respectively. These pre-processed datasets can be used as reference matrices for gene expression profile classification using `sampleClassifier`.

I evaluated the performance of `sampleClassifier` by classifying tissues using public microarray and RNA-seq data. These included the following: (1) 186 microarray test

Table 3.4: Classification results of the microarray test data using `sampleClassifier` and SVMs

Method	**RMA**		**YuGene**	
	Correctly classified	**Misclassified**	**Correctly classified**	**Misclassified**
sampleClassifier	**10**	6	**15**	1
SVMs	5	11	13	3

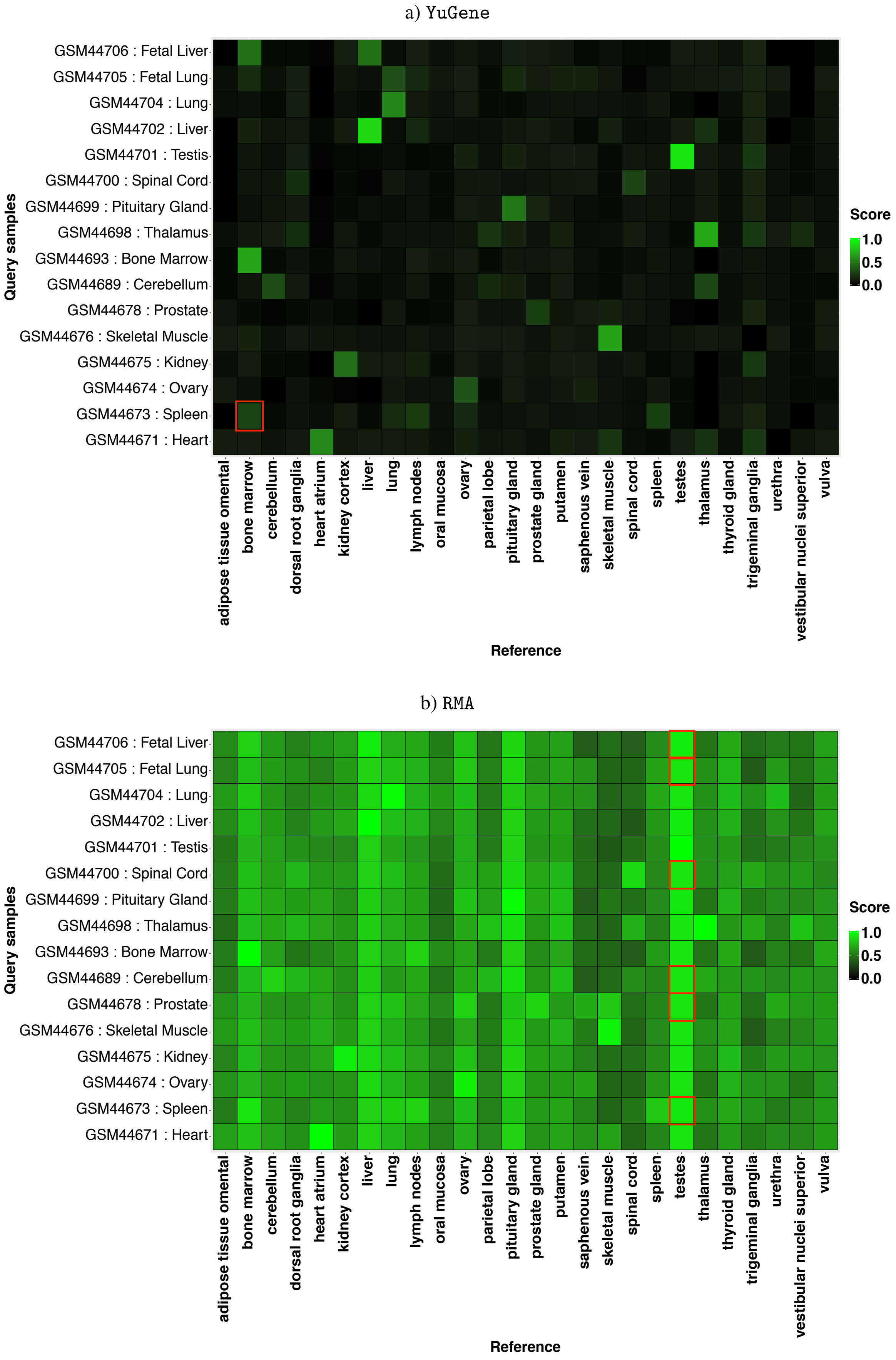
Table 3.5: Classification results of the RNA-seq test data using `sampleClassifier` and SVMs

Study ID	**Number of Samples**	**sampleClassifier**		**SVMs**	
		Correctly classified	**Misclassified**	**Correctly classified**	**Misclassified**
E-MTAB-1733	18	**18**	0	17	1
E-MTAB-513	12	**9**	3	7	5

samples from four tissues (kidney, heart, liver, and lung); (2) 16 microarray test samples, representing 16 tissues from a different study and platform than the reference (GSE2361); (3) 18 RNA-seq samples (E-MTAB-1733); (4) 12 RNA-seq test samples from a different study than the reference (E-MTAB-513). In addition, I tested the performance using two normalization methods for microarray data, namely `RMA` and `YuGene`. Furthermore, I compared my tool to the popular classification tool SVMs. For the classification of microarray test samples, `sampleClassifier` performed similar to SVMs using both `RMA` and `YuGene`, except for the case where the test samples were from a different platform than the reference, and `RMA` was used for normalization, where `sampleClassifier` outperformed SVMs. `sampleClassifier` classified 10 out of 16 samples correctly, whereas SVMs classified only 5 of the 16 test samples. `YuGene` was developed to enable the comparison of samples from different studies or platforms, whereas `RMA` was not built for this purpose. In order to test if the performance of my tool and SVMs will improve by removing batch effects, the method ComBat was used. After adjusting the batch effects, SVMs classified 14 of the 16 test samples correctly, whereas `sampleClassifier` classified all 16 test samples correctly. Since `sampleClassifier` performed better than SVMs without batch effects correction, this suggests that my tool is more robust to the batch effects arising from the comparison of samples from different platforms.

For the RNA-seq data, `sampleClassifier` performed slightly better compared to SVMs

(Table 3.5). Nevertheless, `sampleClassifier` has the following advantages compared to SVMs: first, `sampleClassifier` compares the similarity of each query to all sample types in the reference and calculates a similarity score. Secondly, the similarity score provides information about the marker genes shared between the query and a reference profile. Thirdly, the algorithm used for classification (SMG) is easy to understand and use. Finally, `sampleClassifier` can be applied: i) to assess the similarity of experimentally derived cells to their *in vivo* target cell types; ii) to compare *in vitro* derived organoids to their *in vivo* counterparts; iii) to classify different types of diseases.



a) YuGene

b) RMA

Figure 3.5: Classification heatmaps of 16 samples from the study GSE2361 using `sampleClassifier`, and a) `YuGene` or b) `RMA` for normalization. The misclassified samples are marked in red.

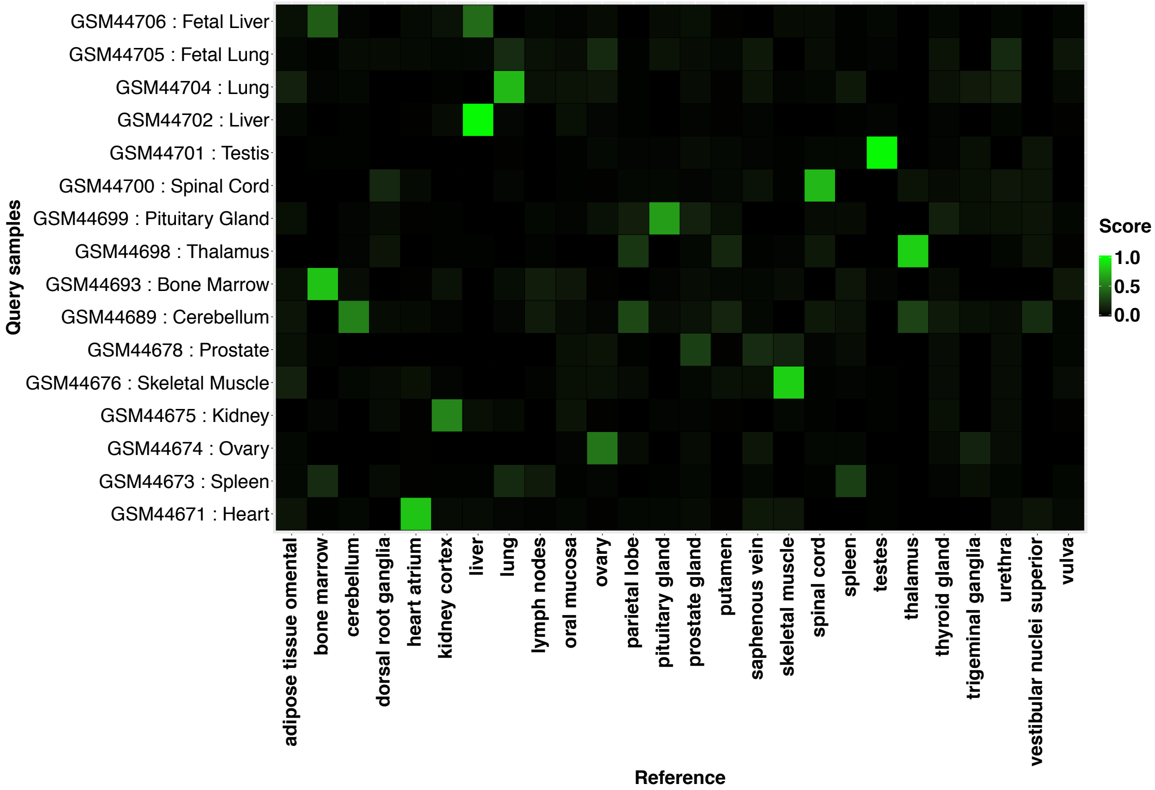


Figure 3.6: Classification heatmap of 16 samples from the study GSE2361 using `sampleClassifier`, and `RMA` and `ComBat` for normalization.

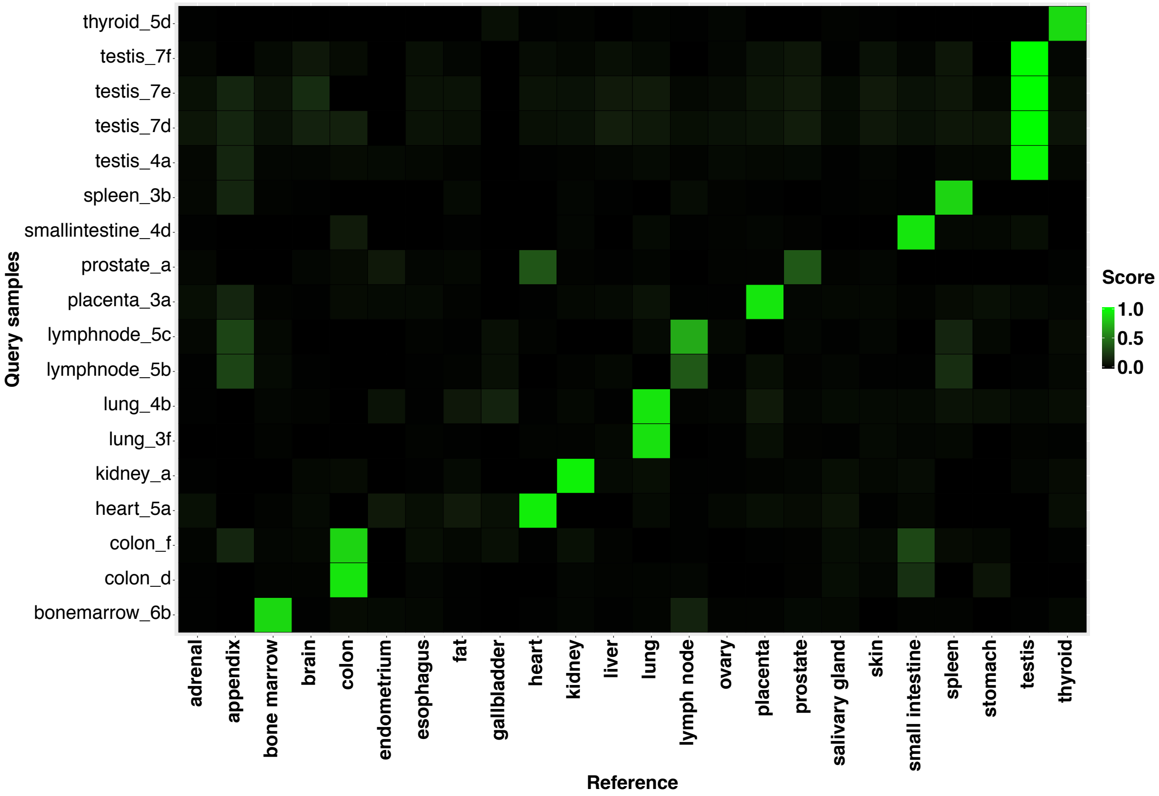


Figure 3.7: Classification heatmap of 18 samples from the study E-MTAB-1733 using `sampleClassifier`.

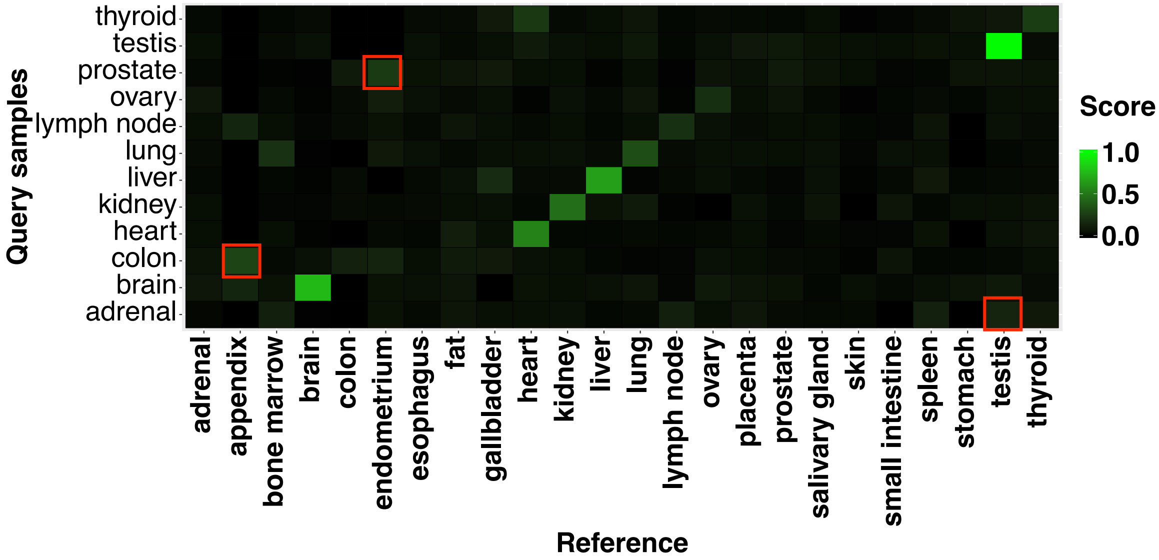


Figure 3.8: Classification heatmap of 12 samples from the study E-MTAB-513 using `sampleClassifier`.

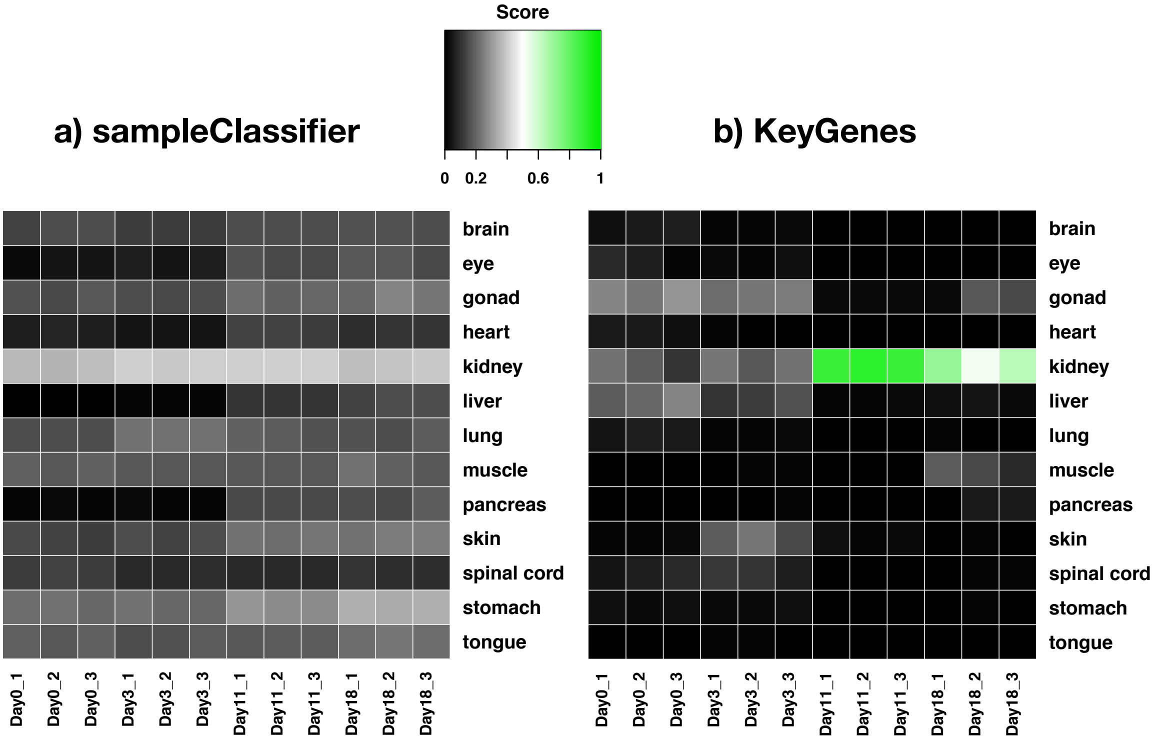


Figure 3.9: Classification heatmap of kidney organoids obtained using a) `sampleClassifier` and b) `KeyGenes`, and a reference with 13 fetal tissues from the first trimester. RNA-seq was performed on whole kidney organoids from 4 time points (day 0, 3, 11 and 18 after aggregation) with 3 individual organoids from 1 experiment per time point (Takasato et al., 2015).

Chapter 4

Classification of kidney diseases

Large amounts of microarray experimental data from numerous studies of many diseases are available in public repositories. Using this resource in order to gain insight into the molecular pathology of diseases is a fundamental challenge in biomedical research. Kidney diseases are a major health problem with high morbidity and mortality rates, and their prevalence is increasing. Chronic kidney disease (CKD) affects more than 10% of the population in many countries worldwide (Eckardt et al., 2013; James et al., 2010). Obesity, as well as type 2 diabetes mellitus, and hypertension are major risk factors for CKD (Haroun et al., 2003; Kramer et al., 2005; Narkiewicz, 2006). Untreated or poorly treated, CKD often progresses to kidney failure (also referred to as end-stage renal disease, ESRD) which means regular hemodialysis treatment or a kidney transplant is needed to survive. Kidney transplantation is considered the best treatment for many people with severe CKD as it not only offers freedom from dialysis but improves survival, provides better quality of life and is more cost effective. Unfortunately, the growing disparity between the expanding annual numbers of patients being added to the kidney transplant wait list compared to the available pool of donor organs means that kidney transplantation can be offered to an increasingly smaller proportion of the ESRD population. Improving our understanding of the molecular mechanisms underlying the diverse renal diseases holds promise for the development of new diagnostic tests and therapies that directly target the pathophysiologic processes underlying these diseases. I curated publicly available biopsy-based microarray data from eight diverse kidney diseases (diabetic nephropathy, focal and segmental glomerulosclerosis, hypertensive nephropathy, IgA nephropathy, lupus nephritis, membranous glomerulonephritis, minimal change disease, and thin membrane disease). In this chapter, I will apply the classification tool `sampleClassifier` and evaluate its performance on the classification of the different kidney diseases in terms of classification accuracy. Finally, using the marker tool (`MGFM`), I extract lists of robust marker genes associated with each disease type.

By comparing gene expression profiles between different kidney disease types and healthy normal controls, we may be able to identify genes involved in the development of the different kidney diseases, or new candidate marker genes for each of these disease types, a result that may reveal new therapeutic targets.

The kidneys are the central organs for homeostasis of the body's extracellular fluids (Figure 4.1.a). Each day, the kidneys filter around 180 liters of blood, accounting for around 20% of cardiac output. Filtering removes metabolic waste products, and kidney action adjusts water, salt and pH to maintain the homeostatic balance of tissue fluids. The kidneys also regulate blood pressure through the renin-angiotensin-aldosterone system, erythrocyte production through production of erythropoietin, and circulating calcium and phosphate levels, in part through the activation of vitamin D (McMahon, 2016). The functional unit of the kidney is the nephron (Figure 4.1.b). The adult human kidney contains approximately one million nephrons. Each nephron is composed of one glomerulus (renal corpuscle) and one double hairpin-shaped tubule that drains the filtrate into the renal pelvis. The glomeruli located in the kidney cortex are bordered by the Bowman's capsule (Figure 4.1.c). They are lined with parietal epithelial cells and contain the mesangium with many capillaries to filter the blood (Kurts et al., 2013). The space between the tubuli and glomeruli, which contains capillaries, fibroblasts and dendritic cells, is called tubulointerstitium.

Disease classification using gene expression data poses a challenge because of the heterogeneity across different data sources and the high variability between individuals. Since `sampleClassifier` requires reference samples for classification, I was interested to test how its performance (in terms of classification accuracy) changes by varying the reference sample set. To this end, I performed 100 runs, and in each run I selected three random samples. The overall mean classification accuracy for glomeruli samples over the 100 runs was 62%, and the highest classification accuracy was 76%. To test if the performance of my tool would improve by using a different strategy for reference selection, I constructed so-called pseudo samples by combining the gene expression values from different samples and used them as reference. `sampleClassifier` reached an overall mean classification accuracy of 71% for the glomeruli test samples over the 100 runs, and the best classification accuracy of 83%.

4.1 Description of the kidney diseases for classification

From a pathological and pathogenetic point of view kidney diseases can broadly be divided into three groups (Matovinović, 2009):

1. Nonproliferative (without cell proliferation) glomerular diseases without glomerular inflammation and without deposition of immunoglobulins such as:

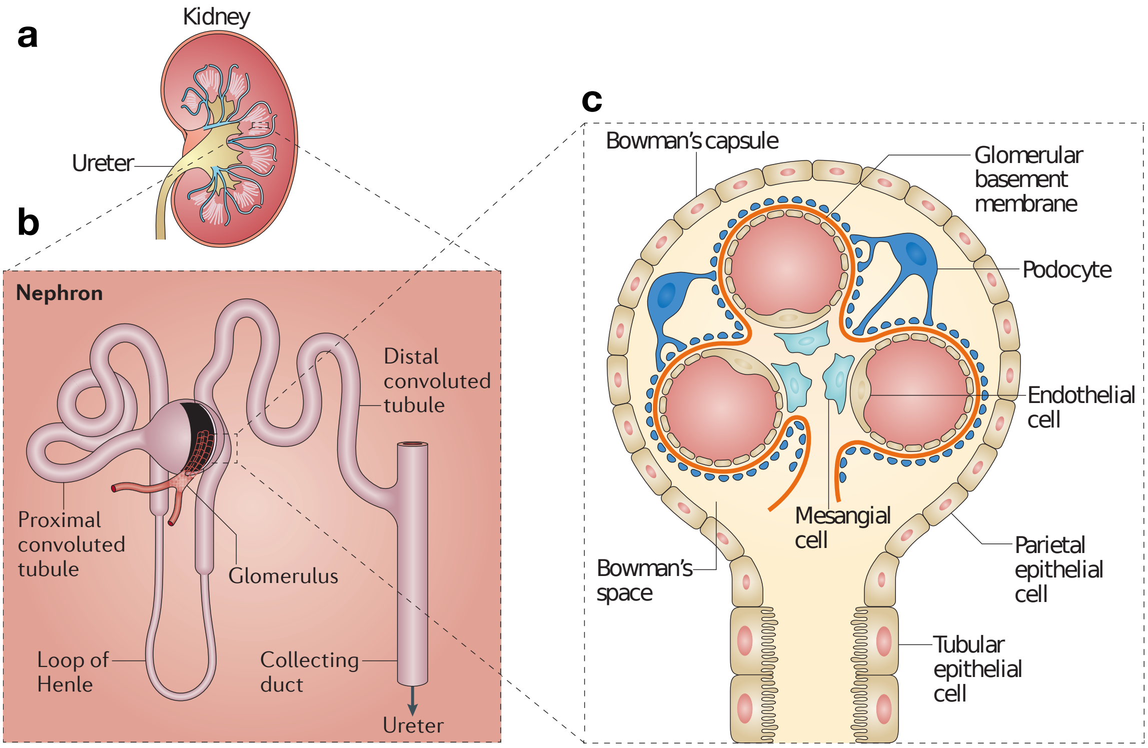


Figure 4.1: Structure of a kidney. Figure reproduced from (Kurts et al., 2013)

- **Minimal change disease (MCD):** also known as nil disease or lipoid nephrosis. As its name implies there is little or no change in the glomerular capillaries by light microscopy or immunofluorescence, but there is clear evidence of disruption of podocyte architecture by high power electron microscopy.
- **Focal and segmental glomerulosclerosis (FSGS):** a glomerular disease characterized by primary podocyte injury, and a lesion that occurs secondarily in any type of chronic kidney disease (Fogo, 2015). Early in the disease course, glomerulosclerosis is both focal, involving a minority of glomeruli, and segmental, affecting a portion of the glomerular globe. With progression, more widespread and global glomerulosclerosis develops (D'Agati et al., 2011).
- **Thin membrane disease (TMD):** is characterized clinically by persistent hematuria, minimal proteinuria, normal renal function, and a uniform thinning of the glomerular basement membrane.

Nonproliferative glomerular diseases with deposition of immunoglobulins, but without glomerular inflammation, most likely because of subepithelial localization of immunoglobulins, like:

- **Membranous glomerulonephritis (MG):** is characterized by formation of subepithelial deposits of immunoglobulins, resulting in alteration to the glomerular basement membrane, and leading to proteinuria.

2. Proliferative glomerular diseases with deposition of immunoglobulins leading to increased cellularity such as:

   - **IgA nephropathy (IN):** also known as Berger's disease, occurs when IgA (Immunoglobulin A) deposits build up in the kidneys glomeruli.
   - **Lupus nephritis (LN):** one of the most important clinical complications of systemic lupus erythematosus (SLE), a chronic autoimmune disease that can involve any organ system.

3. Heterogenous group of kidney diseases in systemic diseases like:

   - **Diabetic nephropathy (DN):** a progressive kidney disease caused by damage to the capillaries in the glomeruli. It is characterized by nephrotic syndrome and diffuse scarring of the kidney glomeruli. It is the main microvascular complication of diabetic disease.
   - **Hypertensive nephropathy (HN):** is defined as histological lesions in renal arteries, arterioles and interstitium that occur due to long-term primary arterial hypertension (Kubiak et al., 2014).

4.2 Materials and methods

4.2.1 Data sources and pre-processing

The microarray expression data were accessed from the NCBI Gene Expression Omnibus (GEO) database. I curated gene expression profiles from human glomeruli and tubulointerstitium obtained using Affymetrix HG-U133A arrays (GPL96). I collected a total of 168 samples for glomeruli and 164 samples for tubulointerstitium (Table 4.1) from the following studies: GSE21785 (Lindenmeyer et al., 2010), GSE32591, GSE37455, GSE37460 (Berthier et al., 2012), GSE35487 (Reich et al., 2010), GSE47183 (Martini et al., 2014), and GSE47184 (Ju et al., 2013).

The microarray data were normalized using YuGene (Lê Cao et al., 2014).

4.2.2 Classification procedure

In contrast to the classification of normal tissue or cell types, disease classification has a variety of problems to deal with, such as disease heterogeneity and the high variability between individuals. In order to evaluate the performance of `sampleClassifier` on the classification of different kidney diseases, I separated glomeruli and tubulointerstitium samples. For each disease class, the samples were partitioned randomly into an initial reference set consisting of half of the samples and a test set containing the second half. The splitting was repeated 100 times and 100 independent runs were performed. In addition, I tested two strategies for the selection of the final reference samples for each disease type. In the first strategy, I selected randomly three samples from the initial reference set of each disease type und used these samples as reference. In the second strategy, I constructed so-called pseudo samples from the initial reference set of each disease type. I call the samples used in the second classification strategy pseudo samples because the gene expression values are combined from different samples. To construct the pseudo samples for a disease type, I sort the vector of expression values of each gene in decreasing order. Let n be the number of samples in the initial reference set of disease type i, and $V_k = (v_1, v_2, ..., v_n)$ the sorted vector of gene expression values of a gene G_k. The pseudo samples are constructed as follows:
i) if n is even: for a gene G_k, the first pseudo sample will be assigned the value $v_{\frac{n}{2}}$, which is the gene expression value at position $P_1 = \frac{n}{2}$ in the vector V_k. The second pseudo sample will be assigned the value $mean(v_{\frac{n}{2}}, v_{\frac{n+2}{2}})$, which is mean of gene expression value at position $P_1 = \frac{n}{2}$ and gene expression value at position $P_2 = \frac{n+2}{2}$ in the vector V_k. Finally, the third pseudo sample will be assigned the value $v_{\frac{n+2}{2}}$.
ii) if n is odd: for a gene G_k, the first pseudo sample will be assigned the value $v_{\frac{n-1}{2}}$, which is the gene expression value at position $P_1 = \frac{n-1}{2}$ in the sorted vector V_k. The second pseudo sample will be assigned the value $v_{\frac{n+1}{2}}$, which is the gene expression value at position $P_2 = \frac{n+1}{2}$ in the vector V_k. Finally, the third pseudo sample will be assigned the value $v_{\frac{n+3}{2}}$, which is the gene expression value at position $P_3 = \frac{n+3}{2}$ in the vector V_k. Figure 4.2 illustrates how pseudo samples are calculated using two genes as an example.
In order to compare the two strategies, I used the same set of samples for testing in each run. The remaining samples from the initial reference set after selecting three random samples as reference were not further considered in the analysis. The schematic view of the entire classification procedure is shown in Figure 4.3.

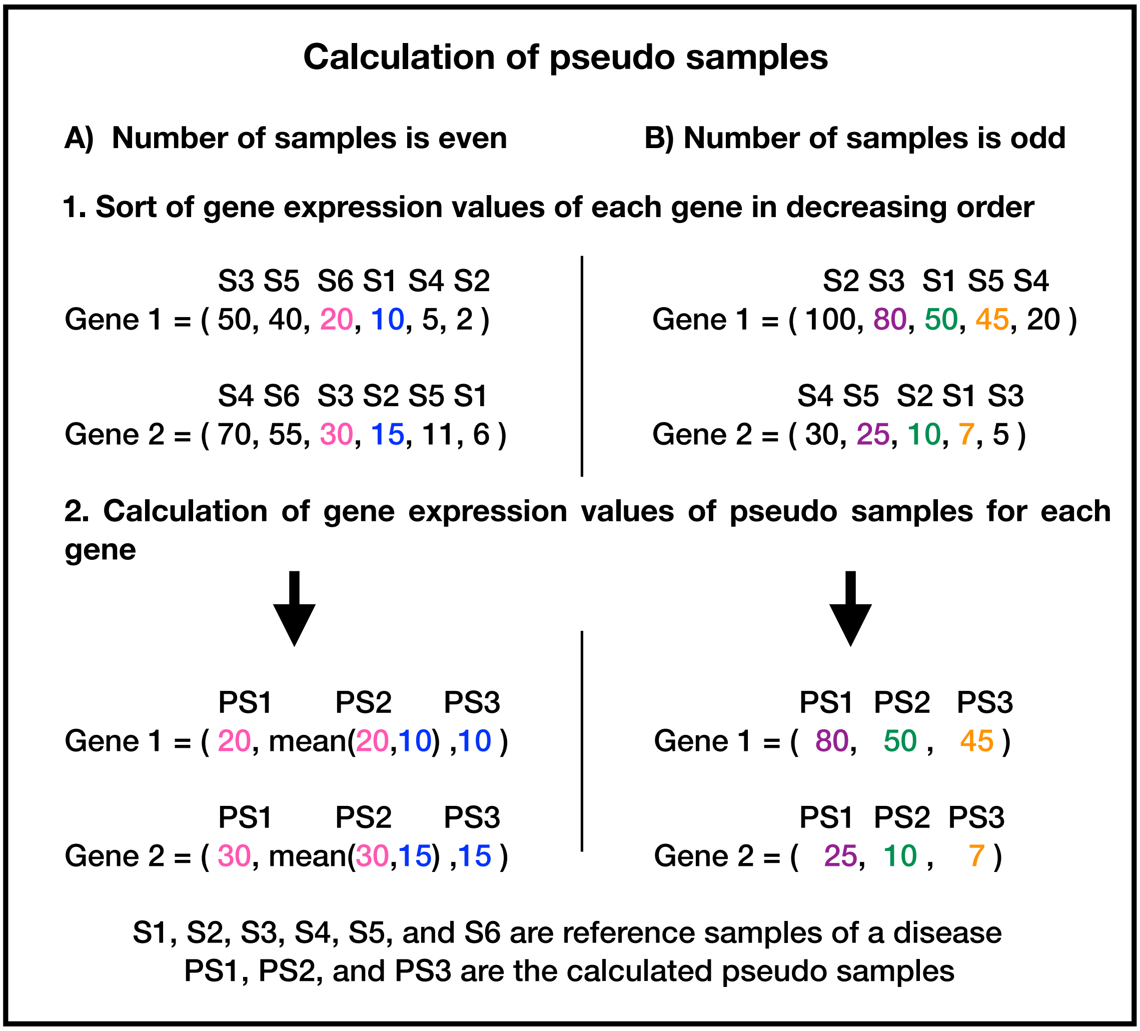

Calculation of pseudo samples
A) Number of samples is even
B) Number of samples is odd
1. Sort of gene expression values of each gene in decreasing order
S3 S5 S6 S1 S4 S2
Gene 1 = (50, 40, 20, 10, 5, 2)
S4 S6 S3 S2 S5 S1
Gene 2 = (70, 55, 30, 15, 11, 6)
S2 S3 S1 S5 S4
Gene 1 = (100, 80, 50, 45, 20)
S4 S5 S2 S1 S3
Gene 2 = (30, 25, 10, 7, 5)
2. Calculation of gene expression values of pseudo samples for each gene
PS1 PS2 PS3
Gene 1 = (20, mean(20,10) ,10)
PS1 PS2 PS3
Gene 2 = (30, mean(30,15) ,15)
PS1 PS2 PS3
Gene 1 = (80, 50 , 45)
PS1 PS2 PS3
Gene 2 = (25, 10 , 7)
S1, S2, S3, S4, S5, and S6 are reference samples of a disease
PS1, PS2, and PS3 are the calculated pseudo samples

Figure 4.2: An example showing how the expression values of pseudo samples are calculated

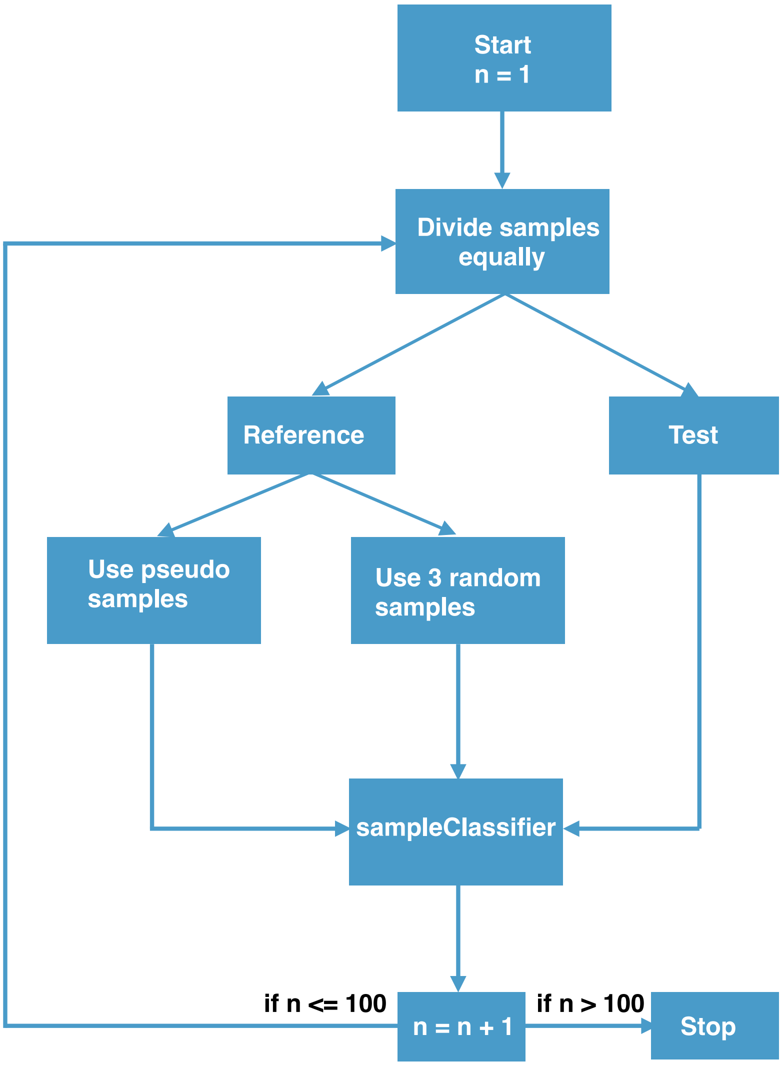


Figure 4.3: Schematic view of the entire classification procedure

Table 4.1: Number of samples for each disease type

Disease	**Number of samples for Glomeruli**	**Number of samples for Tubulointerstitium**
diabetic nephropathy	7	11
focal and segmental glomerulosclerosis	13	10
healthy normal	29	30
hypertensive nephropathy	15	21
IgA nephropathy	27	26
lupus nephritis	32	32
membranous glomerulonephritis	21	18
minimal change disease	10	10
thin membrane disease	0	6
unaffected tumor control	14	0

4.3 Results

In this section I report the performance of `sampleClassifier` in the discrimination between samples from glomeruli and tubulointerstitium. Next, I present the results of the classification of the different kidney diseases using two strategies for the selection of the reference samples. Finally, I provide marker genes for each of the examined kidney diseases.

4.3.1 Classification of glomeruli and tubulointerstitium samples

First, I investigated whether `sampleClassifier` can discriminate glomeruli and tubulointerstitium samples. I challenged `sampleClassifier` to identify glomeruli and tubulointerstitium samples from the study GSE32591 (Berthier et al., 2012). I selected three samples representing healthy normal glomeruli and tubulointerstitium as reference, respectively. I tested the performance of my tool on the remaining 87 samples, which contained in addition to healthy normal also samples representing lupus nephritis from glomeruli and tubulointerstitium. `sampleClassifier` identified all glomeruli and tubulointerstitium samples with a mean similarity score of 0.86 or 0.81, respectively, independent of their disease status (healthy normal or lupus nephritis) (Figure 4.4). The healthy normal and lupus nephritis glomeruli samples were identified with a mean similarity score of 0.9 or 0.85, respectively. The healthy normal and lupus nephritis tubulointerstitium samples were identified with a mean similarity score of 0.86 or 0.8, respectively.

4.3.2 Classification using random samples

I report the performance of my tool on the curated microarray data using three randomly selected samples for each disease type as reference. The classification accuracies obtained over the 100 performed runs are illustrated using boxplots in Figure 4.5. The performance of `sampleClassifier` differs depending on the disease types and on which samples were considered as reference. The overall mean classification accuracy for glomeruli samples over the 100 runs is 62%, whereas the lowest and highest classification accuracy is 38% or 76%, respectively. For tubulointerstitium samples, the mean accuracy over the 100 runs is 63.4%, whereas the lowest and highest classification accuracy is 43.4% or 81%, respectively. Figure 4.6 shows the classification results of the 87 glomeruli test samples obtained in the run with the median classification accuracy for each disease type. A total of 54 samples (62%) were classified correctly with a mean similarity score of 0.5. Figure 4.7 shows the classification results of the 83 tubulointerstitium test samples obtained in the run with the median classification accuracy. A total of 54 samples (65%) were classified correctly with a mean similarity score of 0.61.

4.3.3 Classification using pseudo samples

The classification accuracies obtained over the 100 performed runs are illustrated using boxplots in Figure 4.8. The overall mean classification accuracy obtained for the glomeruli test samples over the 100 runs is 71% (range: 64.4 - 83%). For tubulointerstitium samples, the overall mean classification accuracy over the 100 runs is 70% (range: 58 - 83%). Figure 4.9 shows the classification results of the 87 glomeruli test samples obtained in the run with the median classification accuracy. A total of 64 samples (74%) were classified correctly with a mean similarity score of 0.61. Figure 4.10 shows the classification results of the 83 tubulointerstitium test samples obtained in the run with the median classification accuracy. A total of 58 samples (70%) were classified correctly with a mean similarity score of 0.61.

4.3.4 Robust marker genes

Based on the 100 classification runs performed using pseudo samples, I identified lists of robust marker genes (i.e. predicted as markers in different runs) associated with each of

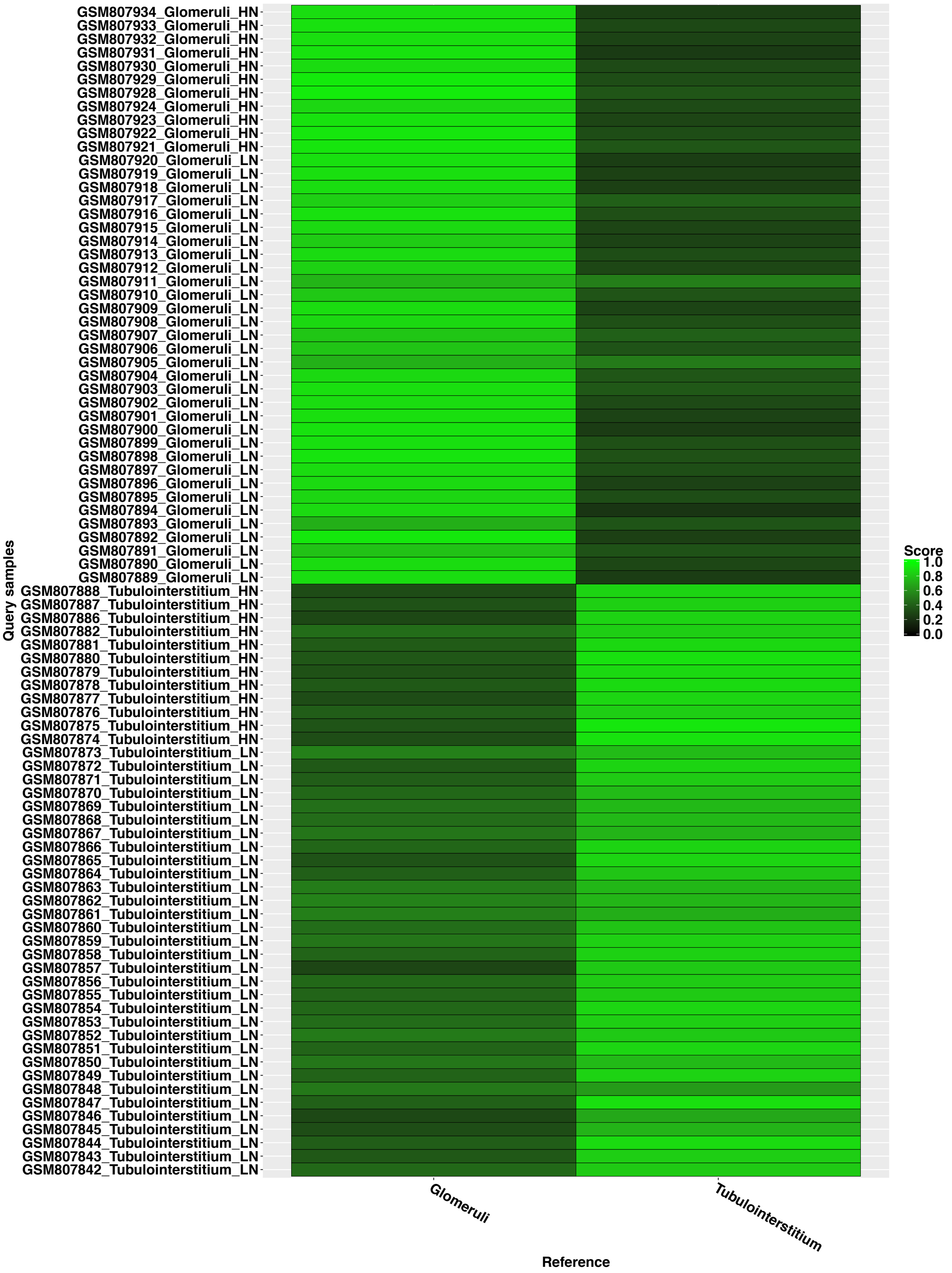


Figure 4.4: Classification heatmap of 87 samples from the study GSE32591 using `sampleClassifier`. The healthy normal and lupus nephritis samples are labeled HN or LN, respectively.

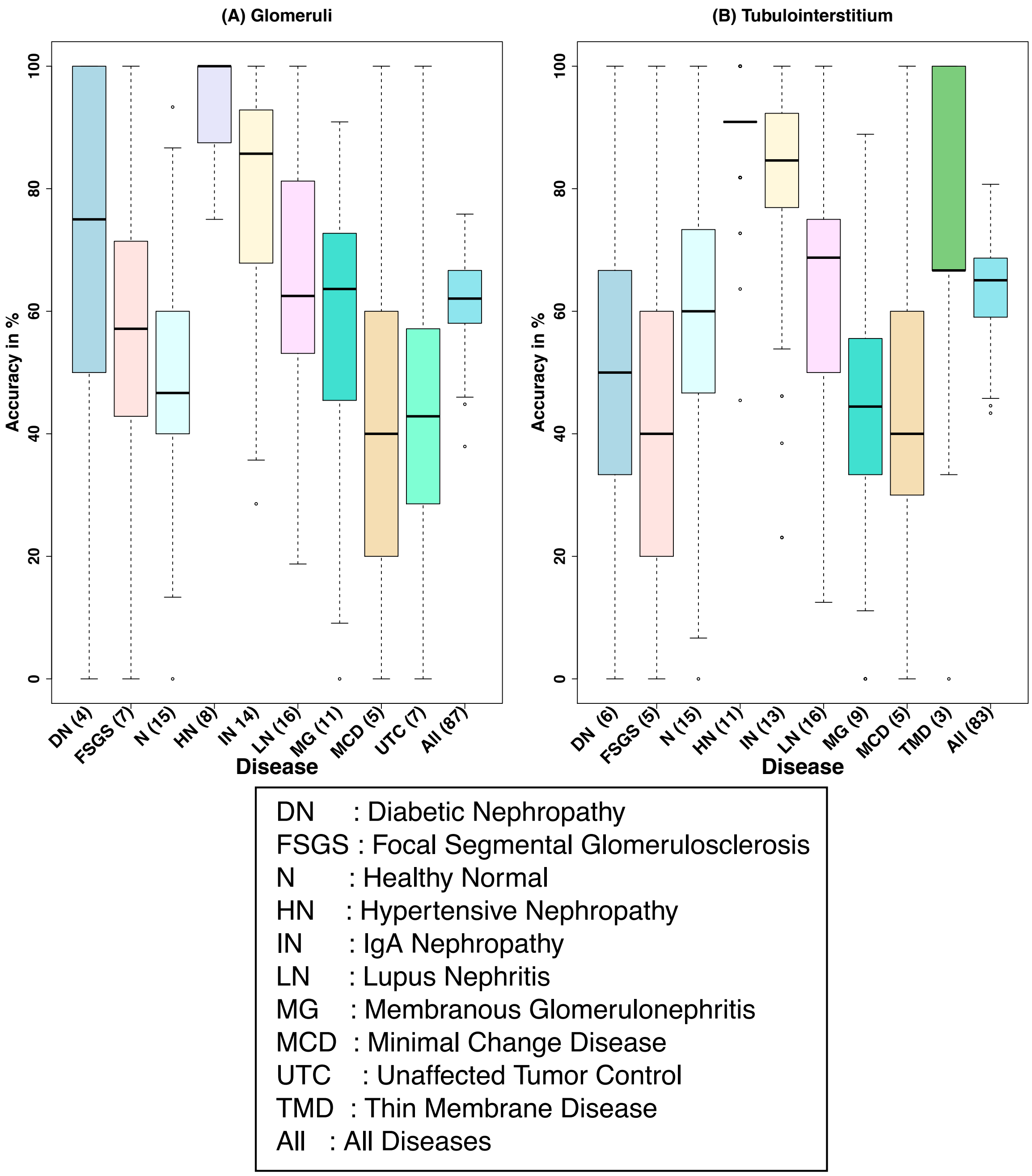
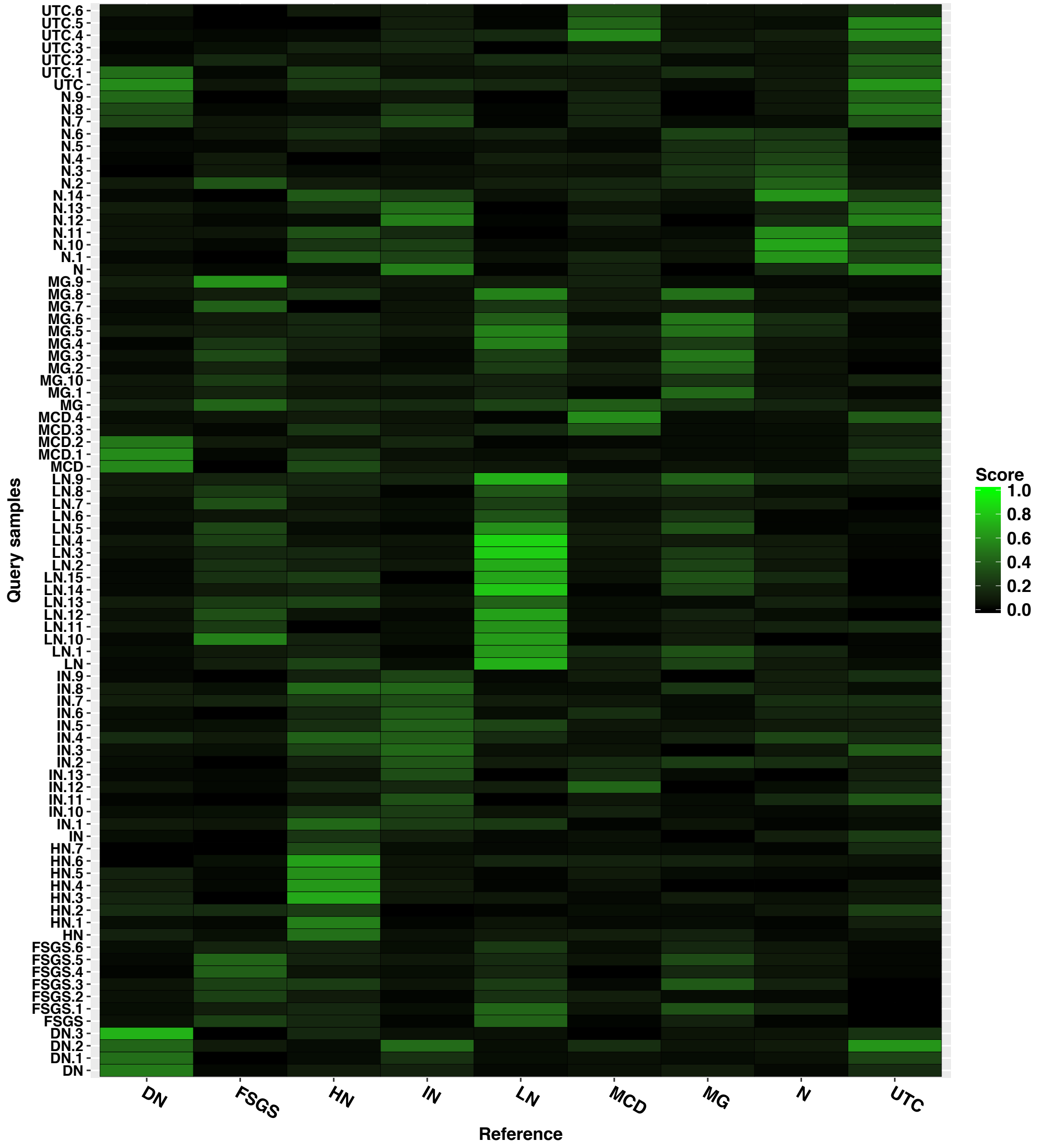
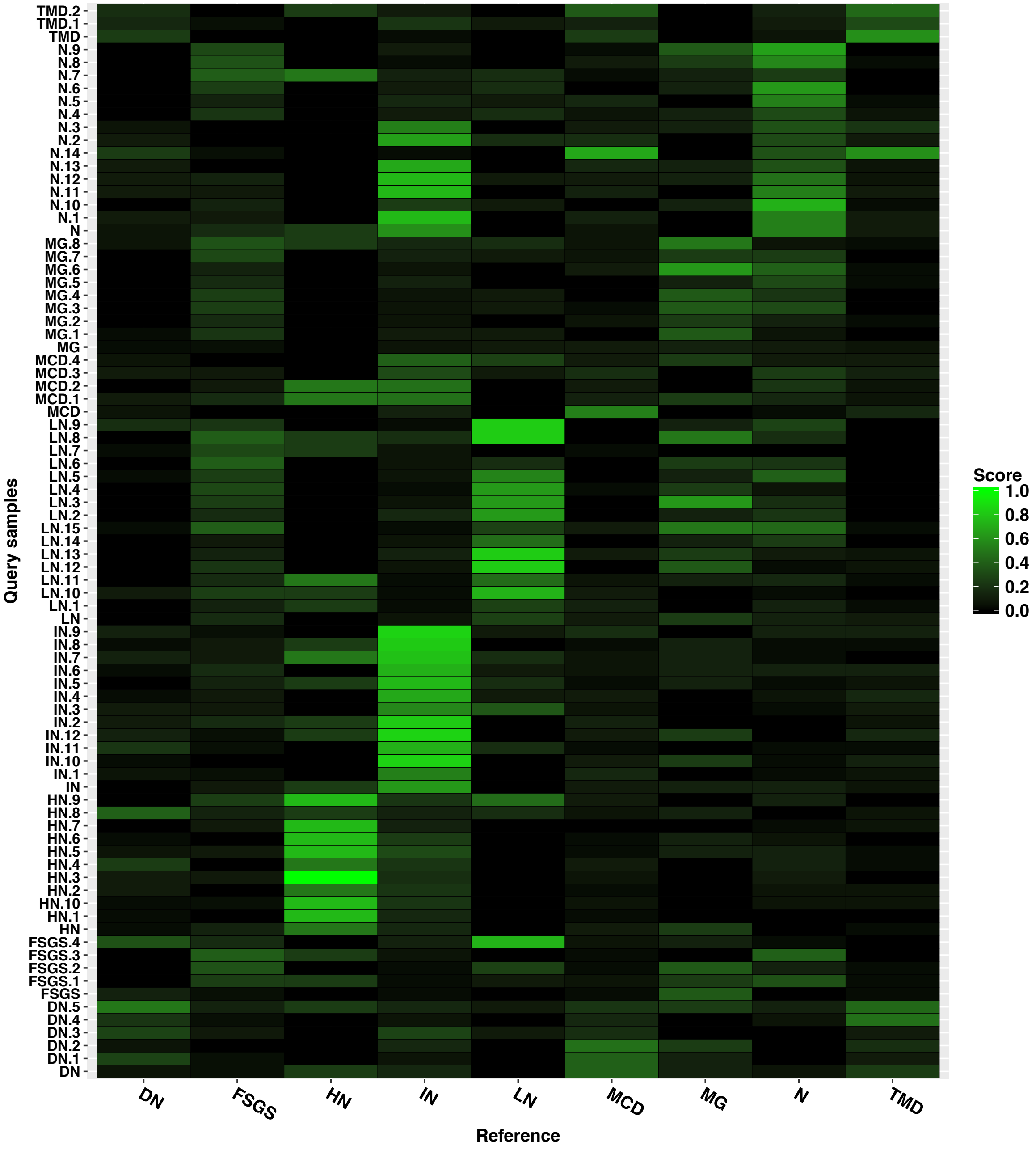


Figure 4.5: Classification accuracy obtained by `sampleClassifier` using microarray gene expression profiles of diseased kidney samples for (A) Glomeruli and (B) Tubulointerstitium, using three random samples for reference. The number of test samples for each disease is shown in brackets.



DN : Diabetic Nephropathy
FSGS : Focal Segmental Glomerulosclerosis
HN : Hypertensive Nephropathy
IN : IgA Nephropathy
LN : Lupus Nephritis
MCD : Minimal Change Disease
MG : Membranous Glomerulonephritis
N : Healthy Normal
UTC : Unaffected Tumor Control

Figure 4.6: Classifcation heatmap of 87 glomeruli samples obtained in the run with the median classification accuracy, using three random samples for reference.



DN : Diabetic Nephropathy
FSGS : Focal Segmental Glomerulosclerosis
HN : Hypertensive Nephropathy
IN : IgA Nephropathy
LN : Lupus Nephritis
MCD : Minimal Change Disease
MG : Membranous Glomerulonephritis
N : Healthy Normal
TMD : Thin Membrane Disease

Figure 4.7: Classifcation heatmap of 83 tubulointerstitium samples obtained in the run with the median classification accuracy, using three random samples for reference.

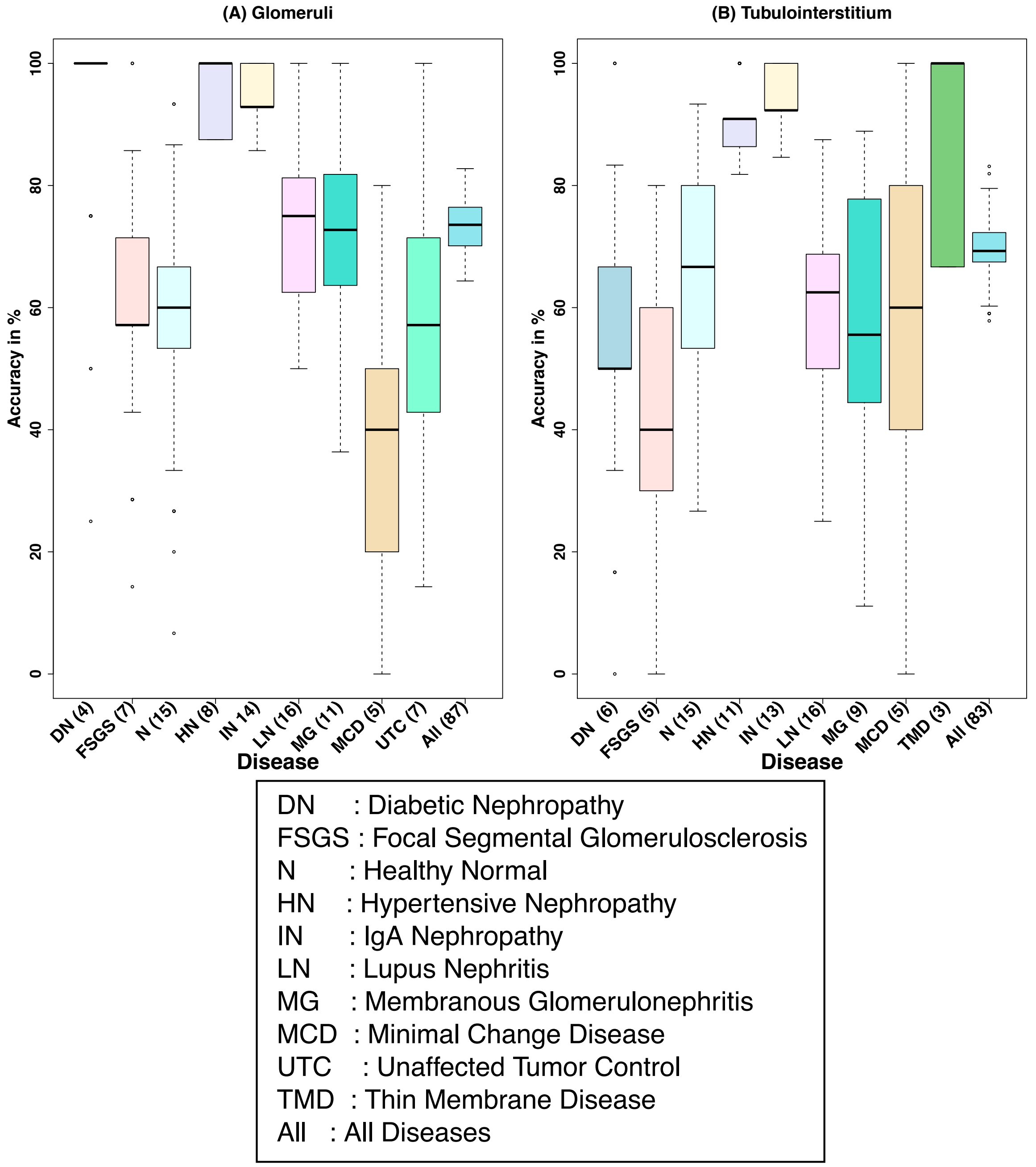
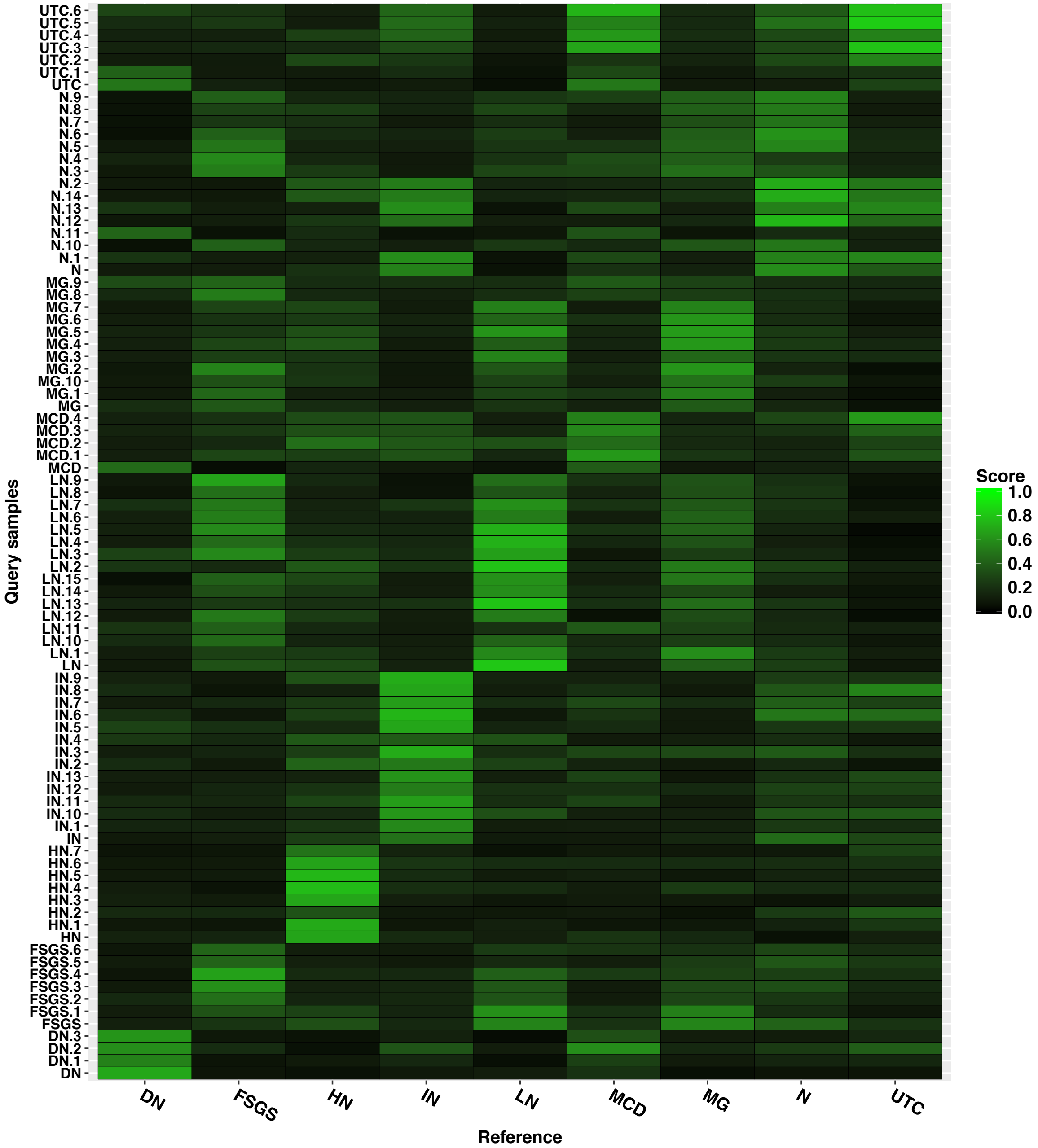
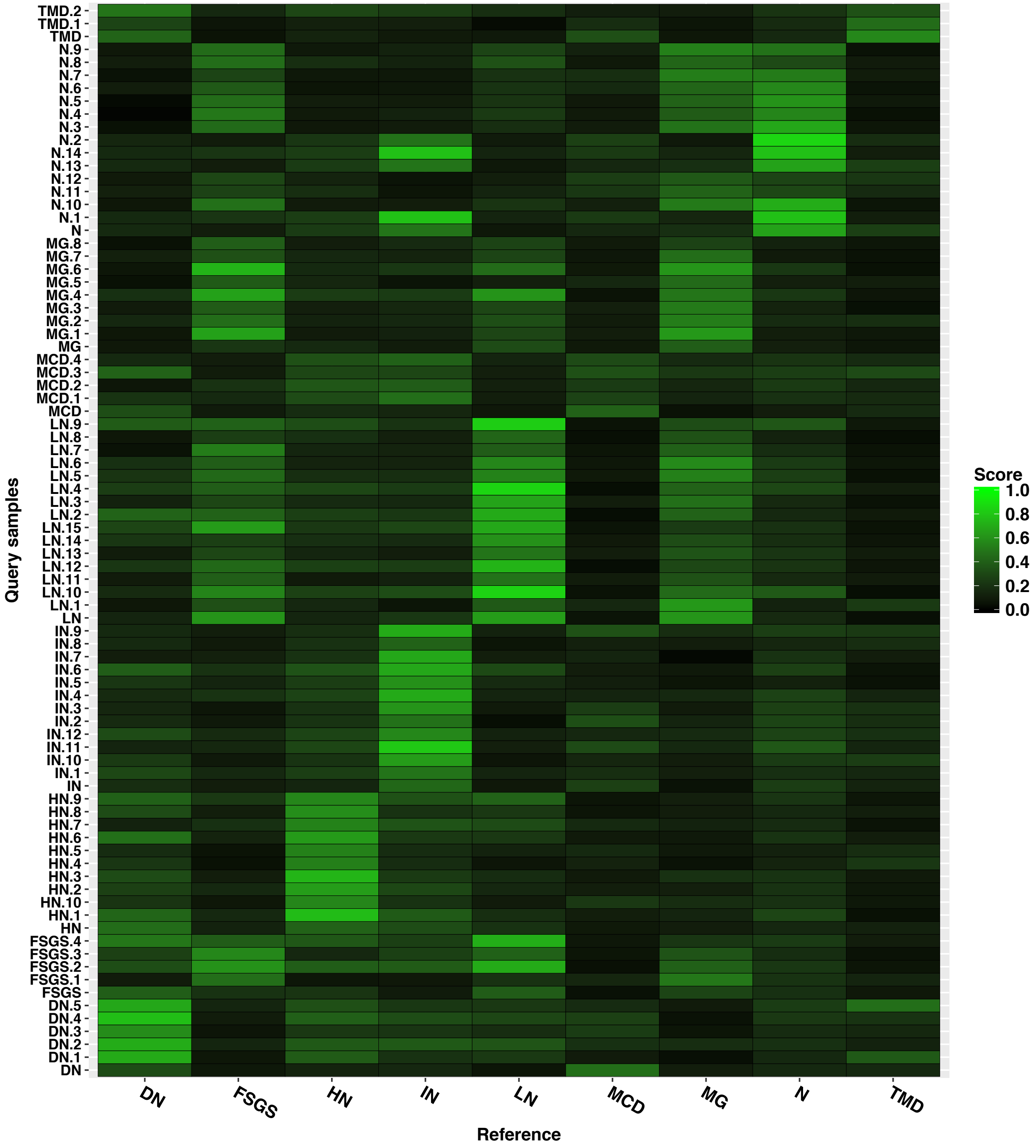


Figure 4.8: Classification accuracy obtained by `sampleClassifier` using microarray gene expression profiles of diseased kidney samples for (A) Glomeruli and (B) Tubulointerstitium, using pseudo samples for reference. The number of test samples for each disease is shown in brackets.



DN : Diabetic Nephropathy
FSGS : Focal Segmental Glomerulosclerosis
HN : Hypertensive Nephropathy
IN : IgA Nephropathy
LN : Lupus Nephritis
MCD : Minimal Change Disease
MG : Membranous Glomerulonephritis
N : Healthy Normal
UTC : Unaffected Tumor Control

Figure 4.9: Classifcation heatmap of 87 glomeruli samples obtained in the run with the median classification accuracy, using pseudo samples for reference.



DN : Diabetic Nephropathy
FSGS : Focal Segmental Glomerulosclerosis
HN : Hypertensive Nephropathy
IN : IgA Nephropathy
LN : Lupus Nephritis
MCD : Minimal Change Disease
MG : Membranous Glomerulonephritis
N : Healthy Normal
TMD : Thin Membrane Disease

Figure 4.10: Classifcation heatmap of 83 tubulointerstitium samples obtained in the run with the median classification accuracy, using pseudo samples for reference.

the kidney diseases and healthy normal for both glomeruli and tubulointerstitium. As a cutoff for the specificity score of all selected marker genes, I used 0.7. The predicted robust marker genes were ranked using a score ranging from 0 to 100, which is the number of times a gene was selected as a marker. The corresponding marker genes are available on github at https://github.com/khadija-a/Marker-genes/blob/master/Glomeruli_robust_marker_genes.xlsx and https://github.com/khadija-a/Marker-genes/blob/master/Tubulointerstitium_robust_marker_genes.xlsx for glomeruli and tubulointerstitium, respectively. The expression of the top 10 robust marker genes that were also predicted as markers for glomeruli in the run with the median classification accuracy, are shown in a heatmap in Figure 4.11. In addition, I provide the marker genes obtained in the run with the best classification accuracy for both glomeruli and tubulointerstitium. These genes might guide other studies to identify novel markers for the examined kidney diseases. These markers are available on github at https://github.com/khadija-a/Marker-genes/blob/master/Marker_Genes_Diseased_Glomeruli.xlsx or https://github.com/khadija-a/Marker-genes/blob/master/Marker_Genes_Diseased_Tubulointerstitium.xlsx for glomeruli and tubulointerstitium, respectively.

4.4 Discussion

Disease classification and prediction of novel genes associated with diseases are fundamental problems in diagnostics, and biomedical research. In this work, I curated publicly available biopsy-based microarray data from eight diverse kidney diseases from GEO (diabetic nephropathy, focal and segmental glomerulosclerosis, hypertensive nephropathy, IgA nephropathy, lupus nephritis, membranous glomerulonephritis, minimal change disease, and thin membrane disease). The gene expression profiles were from human glomeruli and tubulointerstitium obtained using Affymetrix HG-U133A arrays (GPL96). I applied the classification tool `sampleClassifier` to classify the different kidney disease types based on their gene expression profiles. First, I tested if my tool can discriminate between glomeruli and tubulointerstitium samples from one study. `sampleClassifier` identified all 87 test samples correctly. Next, I separated the glomeruli and tubulointerstitium samples, and challenged `sampleClassifier` to identify the different kidney disease types based on the gene expression profiles. `sampleClassifier` is a supervised method and uses samples from each class as a reference. In order to evaluate how its performance changes depending on the used reference sample set, I performed 100 runs, varying the reference set in each run. I partitioned the samples of each disease type randomly into an initial reference set consisting of half of the samples and a test set containing the second half. I tested two strategies for the selection of the final three reference samples for each disease type. In the first strategy,

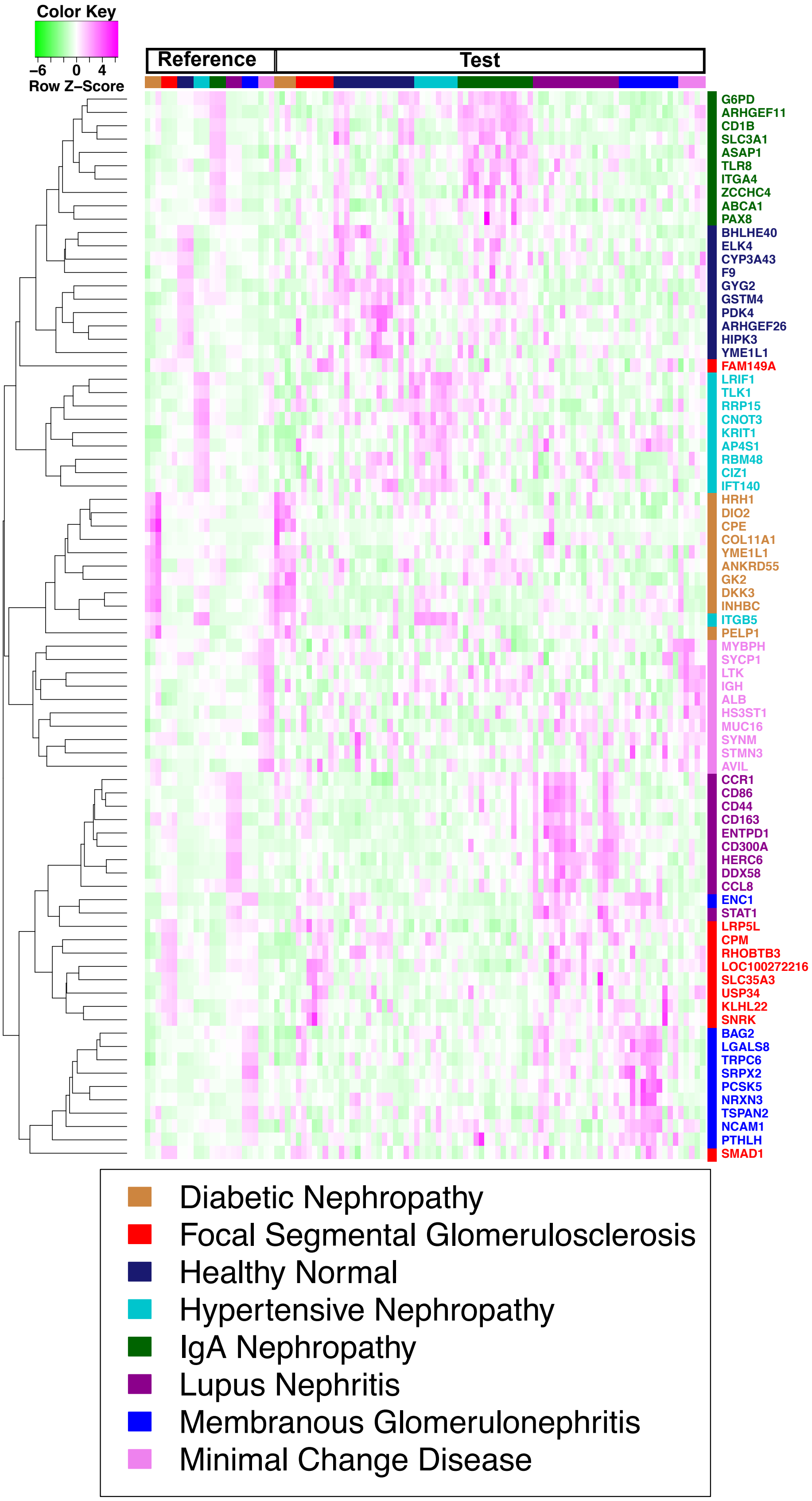


Figure 4.11: Expression of the top 10 robust glomeruli marker genes that were also predicted as markers in the run with the median classification accuracy in each of the examined kidney diseases. The scaled expression of each gene, denoted as the row Z-score, is plotted in green–magenta colour scale with magenta indicating high expression and green indicating low expression.

I selected three random samples from the initial reference set of each disease type. In the second strategy, I constructed pseudo samples based on the samples in the initial reference set (for more details see section 4.2.2). In both cases I used the same set of samples for testing. I showed that using pseudo samples with information from combined samples achieved better classification results than using fixed three random samples as reference. For example, using pseudo samples improved the mean classification accuracy for glomeruli samples by 9% compared to random samples (from 62% to 71%). This is due to the fact that using pseudo samples, more reference samples per gene are used in contrast to the use of three random samples.

The classification accuracy of tubulointerstitium samples was similar to that of glomerular samples, despite the fact that the examined diseases are glomerular diseases, which primarily affect glomeruli. This confirms that these diseases cause changes in the tubulointerstitium as well (Ong and Fine, 1994).

Next, I applied the marker tool `MGFM` to identify marker genes associated with each of the kidney disease types in each of the 100 performed runs. I combined marker genes from the performed runs, identified robust markers, and ranked them using a score ranging from 0 to 100, which is the number of times a gene was selected as a marker. Since the classification accuracy was higher using the pseudo samples, I provide the marker genes obtained based on this strategy. The set of robust markers contained three genes, which have been previously found to show high expression levels in the corresponding diseases. For example, the gene *HAVCR1* (hepatitis A virus cellular receptor 1, also known as kidney injury molecule 1 or *KIM-1*) (Peters et al., 2011) in IgA nephropathy, *IFIH1* (interferon induced with helicase C domain 1, also known as *MDA5*) (Imaizumi et al., 2012) in lupus nephritis, and *TRPC6* (transient receptor potential cation channel subfamily C member 6) (Möller et al., 2007) in membranous glomerulonephritis. The predicted robust disease marker genes for the eight kidney diseases are available on github at https://github.com/khadija-a/Marker-genes/blob/master/Glomeruli_robust_marker_genes.xlsx and https://github.com/khadija-a/Marker-genes/blob/master/Tubulointerstitium_robust_marker_genes.xlsx for glomeruli and tubulointerstitium, respectively. In addition, I provide marker genes obtained in the run with the best classification accuracy based on pseudo samples. Since these marker genes enabled the discrimination between the kidney disease types in 83% of the test samples, I suggest them for further investigation in future studies. These markers are available on github at https://github.com/khadija-a/Marker-genes/blob/master/Marker_Genes_Diseased_Glomeruli.xlsx or https://github.com/khadija-a/Marker-genes/blob/master/Marker_Genes_Diseased_Tubulointerstitium.xlsx for glomeruli and tubulointerstitium, respectively.

While my approach is effective, it has several limitations that future studies can address. I tested my tool on datasets from the same platform. It would be interesting to test its performance on samples from different platforms. In my classification procedure, I considered all eight kidney disease types together; another possible approach would be to test the diseases pairwise. For example, disease vs. healthy normal or two disease types against each other. Microarrays have been extensively used by the scientific community compared to RNA-seq. Consequently, over the years, more microarray data are available in public repositories. Once more RNA-seq data from kidney diseases are available, applying my tools to RNA-seq data and comparing the results to that from microarrays might lead to more accurate and reliable results.

Finally, it would be interesting to test if the suggested marker genes or their products can be detected in cells shed from the kidney into the urine of patients with the corresponding kidney disease. The predicted markers from the present study might be useful to develop a non-invasive screening test using the biomolecules in urine to uniquely identify the kidney disease types based on single marker genes or combinations of them.

Chapter 5

Summary and outlook

The use of high-throughput technologies such as microarrays or RNA-seq for gene expression profiling has revolutionized genetic and biomedical research. Consequently, vast amounts of data of gene expression have accumulated in many public repositories, such as GEO and ArrayExpress. Hence, computational methods to make use of these data are in high demand. This thesis describes bioinformatics tools for marker gene detection and sample classification using gene expression data.

In Chapter 2, I introduced `MGFR`, a bioinformatics tool for marker gene detection from RNA-seq data. This tool is an extension of the original marker tool `MGFM`, which I have developed for marker gene detection from microarray data. Both tools are available as R packages from the Bioconductor website. Using publicly available microarray and RNA-seq datasets from 16 human tissues, I applied both tools to detect marker genes for each of the examined tissues. I assessed the performance of the tools using tissue-specific genes taken from the TiGER database. Further, I compared the overlap of marker genes obtained using each tool, identified robust marker genes, found by both tools, for each of the examined tissues, and suggested novel candidate marker genes that were not previously associated with the examined tissues, for further investigation. Further, I confirmed the tissue-specific expression of top marker genes, predicted by `MGFM` as markers for a set of five human tissues, by RT-PCR.

In contrast to very comprehensive but static databases of tissue-specific genes such as TiGER, or PaGenBase, my tools enable users to easily modify and adapt the sample types in the reference to their set of interest.

In Chapter 3 I described the tool `sampleClassifier` for the classification of samples based on gene expression profiles. The tool supports the classification of microarray and RNA-seq gene expression data, and requires a training and a test dataset. `sampleClassifier` uses a simple algorithm called "Shared Marker Genes" (SMG). As the name implies, the number of shared marker genes between a reference and a query sample is used as a similarity measure. Marker genes are detected using the tool MGFM for microarray data and MGFR for RNA-seq data, which are available as Bioconductor R packages. Using publicly available microarray and RNA-seq data, I demonstrated the utility and effectiveness of my approach. Furthermore, I compared my tool to SVMs, and showed that my tool performed better or comparable to SVMs. Nevertheless, my tool has the following advantages compared to SVMs: i) sampleClassifier compares the similarity of each query to all sample types in the reference and calculates a similarity score; ii) the similarity score provides information about the marker genes shared between the query and a reference profile; iii) the algorithm used for classification (SMG) is easy to understand and use. Further, sampleClassifier can be applied: i) to evaluate the similarity of experimentally derived cells with their desired target cell type; ii) to compare *in vitro* derived organoids (e.g. kidney organoids) to their *in vivo* counterparts; iii) to classify different types of diseases.

In the last part of this thesis, in Chapter 4, I applied the previously described tools (MGFM and sampleClassifier) to publicly available biopsy-based microarray data from 8 diverse kidney diseases. I identified marker genes, and classified the different kidney disease types. Since sampleClassifier uses reference samples for each disease type, and due to the heterogeneity across different data sources, I randomly selected the reference samples and repeated the procedure 100 times. Based on the performed 100 runs, I identified robust marker genes for each disease type, and scored the markers based on how many times they were predicted as markers for the corresponding disease type. Finally, I suggested these marker genes for further investigation by future studies.

5.1 Outlook

The introduced tools could be further enhanced and improved in many ways.

First, the marker tools MGFM and MGFR could be modified to return negative marker genes (segregating samples from one tissue at low expression) as well. Second, the classification tool sampleClassifier could be modified to consider the number of negative marker genes (in addition to the currently used positive marker genes) in the calculation of the similarity score, which may increase the accuracy. Providing information about the number of both positive and negative markers, is of great importance for the evaluation of the similarity of experimentally derived cells with their desired target cell type. Third, integrating the developed tools into the CellFinder platform (http://cellfinder.org) and connecting them with its molecular database, which contains preprocessed gene expression data derived from different tissues and cell types, would disseminate the tools to a wide group of users and increase its application. Until now, only the marker tool MGFM is integrated into CellFinder, and can only be used with predefined datasets. I aim to connect the presented tools with CellFinder's database to serve as a data source, in order to enable the identification of marker genes associated with a set of samples of interest, and classification of samples, in a convenient, interactive, and fast way.

Although the focus in this thesis was on microarray and RNA-seq data, the presented tools can be easily adapted to other data types such as for example proteomics data.

Recent advances in single cell RNA-seq will allow a detailed analysis of the different cell types of different organs and tissues. Finally, update of the presented tools to work with single cell RNA-seq data will further broaden their application range, further increase their usability, and may contribute to their improvement.

Appendix A

Supplementary data to Chapter 2

Table A.1: The corresponding IDs to the samples used in the microarray reference dataset

Tissue	Sample IDs
adrenal gland cortex	GSM80605, GSM80606, GSM80608
bone marrow	GSM80576, GSM80577, GSM80603
colon cecum	GSM80624, GSM80625, GSM80632
endometrium	GSM80672, GSM80673, GSM80685
esophagus	GSM80695, GSM80696, GSM80697
heart atrium	GSM80655, GSM80656, GSM80698
kidney cortex	GSM80686, GSM80687, GSM80689
liver	GSM80729, GSM80730, GSM80739
lung	GSM80707, GSM80710, GSM80712
lymph node	GSM80735, GSM80737, GSM80738
midbrain	GSM80700, GSM80701, GSM80702
prostate	GSM80805, GSM80806, GSM80824
salivary gland	GSM80821, GSM80822, GSM80823
spleen	GSM80808, GSM80825, GSM80826
testis	GSM80853, GSM80868, GSM80869
thyroid gland	GSM80864, GSM80865, GSM80867

Table A.2: The corresponding IDs to the samples used in the RNA-seq reference dataset

Tissue	Sample IDs
adipose tissue	*fat_a*, *fat_e*, *fat_x1*
adrenal gland	*adrenal_4a*, *adrenal_4c*, *adrenal_4d*
bone marrow	*bonemarrow_5a*, *bonemarrow_6a*, *bonemarrow_6c*
brain	*brain_a*, *brain_3b*, *brain_3c*
colon	*colon_a*, *colon_b*, *colon_c*
endometrium	*endometrium_4a*, *endometrium_4b*, *endometrium_5a*
Esophagus	*esophagus_5a*, *esophagus_5b*, *esophagus_5c*
heart	*heart_5b*, *heart_6a*, *heart_6b*
kidney	*kidney_b*, *kidney_c*, *kidney_d*
liver	*liver_a*, *liver_c*, *liver_d*
lung	*lung_4a*, *lung_4d*, *lung_3e*
lymph nodes	*lymphnode_4a*, *lymphnode_4b*, *lymphnode_5a*
prostate	*prostate_4a*, *prostate_4b*, *prostate_4c*
salivary gland	*salivarygland_6a*, *salivarygland_6b*, *salivarygland_6c*
spleen	*spleen_3a*, *spleen_3c*, *spleen_3d*
thyroid gland	*thyroid_5a*, *thyroid_5b*, *thyroid_5c*

Table A.3: Primer Sequences for PCR amplification

Primers used for liver

Gene name	forward primer (5'-3')	Reverse primer (3'-5')	Product size
AKR1D1	TCTCAGTGCTGCAAGTCACC	CTCCCCAACTTCGTGTTCAT	193
FGG	GAATTTTGGCTGGGAAATGA	ATCATCGCCAAAATCAAAGC	222
APOA2	GAGCTTTGGTTCGGAGACAG	TGTGTTCCAAGTTCCACGAA	235
CYP2C8	GCAGGAAAAGGACAACCAAA	GTGTAAGGCATGTGGCTCCT	228
GC	GAAACACCAGCCACAGGAAT	GGTACAGCAGGACCCTACCA	196
CPS1	ATTCCTTGGTGTGGCTGAAC	ATGGAAGAGAGGCTGGGATT	150
CYP2E1	ACCCGAGACACCATTTTCAG	TCCAGCACACACTCGTTTTC	201
APOC3	CTCCCTTCTCAGCTTCATGC	GTCTGACCTCAGGGTCCAAA	200
SERPINC1	TTTACTTCAAGGGCCTGTGG	CTTTGAAGGGCAACTCAAGC	171
AHSG	CAGAACAACGGCTCCAATTT	CTGTGTTTGGAACACCATGC	240
AMBP	AGTGGTACAACCTGGCCATC	AAGCTCCAGACGTCTCCTCA	165

Primers used for lung

Gene name	forward primer (5'-3')	Reverse primer (3'-5')	Product size
CLDN18	GGTATCCATCTTTGCCCTGA	GGTCTGAACAGTCTGCACCA	217
NKX2-1	ACAAGAAAGTGGGCATGGAG	GTTCCTCATGGTGTCCTGGT	248
SCGB1A1	GTCACACTGGCTCTCTGCTG	TGATGCTTTCTCTGGGCTTT	205
SFTPB	CAAACGGCATCTGTATGCAC	CGGAGAGATCCTGTGTGTGA	195
CYP4B1	CCTGGACAAAGTGGTGTCCT	CCAATCCACTGGAGGAAGAA	171
CD52	GCGCTTCCTCTTCCTCCTAC	GAGGTGGATTATGGCATTGG	166
LAMP3	ACTTCAACATCGACCCCAAC	CACTCACGCACTTGAAGGAA	246
AGER	GCTGTCAGCATCAGCATCAT	ATTCAGTTCTGCACGCTCCT	225
LYZ	GCCAAATGGGAGAGTGGTTA	ATCACGGACAACCCTCTTTG	213
SFTPD	AAGTGGGCTTCCAGATGTTG	CTGTGCCTCCGTAAATGGTT	181
SFTPC	ACACTGCCACCTTCTCCATC	CTGGCCCAGCTTAGACGTAG	221
SLC34A2	TCGCCACTGTCATCAAGAAG	TGGATAAGCCCTCTCAATGG	185

Primers used for heart

Gene name	forward primer (5'-3')	Reverse primer (3'-5')	Product size
MYOZ2	CCGGAGCTTTTAGAGGCTTT	CCAGAAAGGGGATTGACAAA	206
TNNI3	TGCAGATTGCAAAGCAAGAG	CAGATCTGCAATCTCCGTGA	224
SYNPO2L	CGACCTGGTCCTCATCTCAT	GCCGCCGTTTCTTAAACATA	232
MYH6	CTTCAACCACCACATGTTCG	GGCTTCTGGAAATTGTTGGA	234
CSRP3	CCTTGGCACAAGACCTGTTT	TTGTGTAAGGCCTCCAAACC	150
CKM	TGAAAACCTCAAGGGTGGAG	TCCTTCTCCGTCATGCTCTT	213
PLN	GAGAAAGTCCAATACCTCACTCG	GAGAAGCATCACGATGATACAGA	153
MB	GAGATGAAGGCGTCTGAGGA	TCTGCAGAACCTGGATGATG	190
TTN	AGGCTGGAAACGGTGTAATG	TAGGGCATCCTGCTCTCACT	233
MYL7	GTCTTCCTCACGCTCTTTGG	CCACCTCAGCTGGAGAGAAC	166
MYH7	TGTGTCACCGTCAACCCTTA	TGGCTGCAATAACAGCAAAG	238
TPM1	AGTCGAGCCCAAAAAGATGA	TTCTTCCAGCTGTCGGACTT	192

Primers used for brain

Gene name	forward primer (5'-3')	Reverse primer (3'-5')	Product size
GAP43	GGGAGGCTTGAGGAAAAATC	TCAGCAGCTTGGACATCATC	240
GFAP	AGGAAGATTGAGTCGCTGGA	ATACTGCGTGCGGATCTCTT	165
TMEFF1	CCCTGTTTTGTTCGAGAAGC	CACATTTTCCATGGATGCAG	229
FUT9	CAAAAGAGTGGCATTGAGCA	AATGCTTGCCCGTAGGTATG	236
SYT1	GTGAGCGAGAGTCACCATGA	ACGGTGGCAATGGAATTTTA	172
SNAP25	ATGCCCGAGAAAATGAAATG	AGCATCTTTGTTGCACGTTG	190
MBP	GGGAGGACAACACCTTCAAA	ACTTGCTGTGGCCAGGTACT	155
GRIA2	GTGGCTAGAGTGCGGAAGTC	ACGCCTTGCTCACTGAGTTT	206
KIF5C	TGCAGGATGCTGAAGAAATG	GCTTCTCCTGTTCCAGTTGC	168
STMN2	AGCTGTCCATGCTGTCACTG	CAGCCTCCAGTTTCTTCTGG	231
NEFM	TCAACGTCAAGATGGCTCTG	CTTCCACCTTGGGTTTCTGA	165
GABBR2	CGCCTGTTCTAGCCGATAAG	GGTACAGGGATCGTTGGAGA	236

Primers used for kidney

Gene name	forward primer (5'-3')	Reverse primer (3'-5')	Product size
SLC12A1	TTTGGAGCTGTTTTGTGCTG	ATGGGTCCCCCTGTTAAGAC	245
SLC3A1	GGGAACAGCGTGTATGAGGT	GGAGTTCCAGGGAGTGTGAA	167
UMOD	AAGAGTCTGGGCTTCGACAA	GCTGTAAGTGGCATGGGTTT	162
AOC1	AGGCATGCAGACCAAGTACC	GGCATTTCAAAGAGGCAGAG	171
CD24	ACCCACGCAGATTTATTCCA	ACCACGAAGAGACTGGCTGT	153
HSD11B2	GACCTGACCAAACCAGGAGA	GCCAAAGAAATTCACCTCCA	174
CA12	AGTACAAAGGCCAGGAAGCA	GCCGGAAGTTGTTGATCATT	246
PDZK1IP1	TTGCAGTCAACCACTTCTGG	GCATTCTCATGCTCACTGGA	151
FXYD2	CGTGGACCCGTTCTACTATGA	GGCTCATCTTCATTGATTTGC	153
CDH16	GGGAAACCTCTACGTGACCA	AGCTCAGGGATGCTGACTGT	189
SLC22A8	CTGAGCACCGTCATCTTGAA	GACCAACCAGCGTATGGACT	219
CLDN8	GGGGACAATGAGAAGGTGAA	GAGCTCCTCCAACAATCAGC	217

Table A.4: Description of brain marker genes predicted by MGFM and verified by RT-PCR

Gene symbol	Gene name	Description
FUT9	fucosyltransferase 9 (alpha (1,3) fucosyltransferase)	*FUT9* encodes a member of the glycosyltransferase family, and has been reported to synthesize the Lewis X carbohydrate structure in the brain (Nishihara et al., 2003).
GABBR2	gamma-aminobutyric acid type B receptor subunit 2	*GABBR2* encodes a membrane protein from the G-protein coupled receptor 3 family and Gamma-aminobutyric acid type B (*GABA-B*) receptor subfamily. The *GABA-B* receptors inhibit neuronal activity through G protein-coupled second-messenger systems, which regulate the release of neurotransmitters and the activity of ion channels and adenylyl cyclase (Kaupmann et al., 1998; Kerr and Ong, 1995). *GABBR2* have been associated with autism (Fatemi et al., 2009). Additionally, significant reductions in *GABBR1* and *GABBR2* expression in lateral cerebellum of subjects with schizophrenia, bipolar disorder, and major depression have been reported (Fatemi et al., 2011).
GAP43	growth associated protein 43	The protein encoded by this gene has been termed a 'growth' or 'plasticity' protein because it is expressed at high levels in neuronal growth cones during development and axonal regeneration. This protein is considered a crucial component of an effective regenerative response in the nervous system [1].
GFAP	glial fibrillary acidic protein	*GFAP* encodes one of a family of intermediate filament proteins. *GFAP* is known as astrocytes marker. Mutations in this gene have been associated with Alexander disease (Brenner et al., 2001), a rare disorder of astrocytes in the central nervous system.
GRIA2	glutamate receptor, ionotropic, AMPA 2	*GRIA2* encodes the *GLUR2* subunit of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors, which is mainly expressed in the brain, and play important roles in normal brain function (Isaac et al., 2007). In addition, Hackmann et al. (Hackmann et al., 2013) hypothesized that *GRIA2* is involved in intellectual disability.
KIF5C	kinesin family member 5C	*KIF5C* gene encodes a member of the kinesin superfamily of proteins that are motor proteins involved in various processes in the brain, such as neuronal functioning, development, and survival (Aizawa et al., 1992; Willemsen et al., 2014). In a recent study, Willemsen et al. (Willemsen et al., 2014) suggested that mutations in *KIF4A* and *KIF5C* cause intellectual disability by tipping the balance between excitatory and inhibitory synaptic excitability.

[1] https://www.ncbi.nlm.nih.gov/gene/2596

MBP	myelin basic protein	The protein encoded by the classic MBP gene is a major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system. However, MBP-related transcripts are also present in the bone marrow and the immune system [2].
NEFM	neurofilament, medium polypeptide	*NEFM* encodes the medium neurofilament protein, one of the 3 subunits forming the neurofilaments, that localizes to neuronal axons and dendrites (Myers et al., 1987). Various human brain diseases have been associated with neurofilaments proteins.
SNAP25	synaptosomal-associated protein, 25kDa	*SNAP25* encodes a presynaptic plasma membrane protein involved in the regulation of neurotransmitter release. *SNAP25* is expressed by neurons in the hippocampus, is suggested to play major role in long-term memory formation and has been associated with cognitive ability (Gosso et al., 2006; Spellmann et al., 2008).
STMN2	stathmin 2	*STMN2*, also known as *SCG10*, encodes a member of the stathmin family of phosphoproteins, which are involved in microtubule dynamics and signal transduction. Reductions in the expression of this gene have been associated with Down's syndrome (Bahn et al., 2002) and Alzheimer's disease (Zhang et al., 2005)
SYT1	synaptotagmin I	*SYT1* encodes an integral membrane protein of synaptic vesicles, which is thought to serve as Ca(2+) sensors in the process of vesicular trafficking and exocytosis. Calcium binding to synaptotagmin-1 participates in triggering neurotransmitter release at the synapse (Fernández-Chacón et al., 2001). The corresponding protein was suggested to be important for synaptic function and may be related to cognitive impairments in Alzheimer's disease (Reddy et al., 2005).
TMEFF1	transmembrane protein with EGF-like and two follistatin-like domains 1	*TMEFF1* encodes a transmembrane protein containing two follistatin-like modules and an epidermal growth factor-like domain. Gery et al. (Gery et al., 2003) investigated *TMEFF1* expression in normal brain and brain cancer cells. According to this study, *TMEFF1* was highly expressed in the normal brain and at lower levels in brain cancer. These results suggest that *TMEFF1* may function as a tumor suppressor gene in brain cancers.

[2] https://www.ncbi.nlm.nih.gov/gene/4155

Table A.5: Description of heart marker genes predicted by `MGFM` and verified by RT-PCR

Gene symbol	**Gene name**	**Description**
CKM	creatine kinase, muscle	*CKM* is a cytoplasmic enzyme involved in energy homeostasis and is an important serum marker for myocardial infarction.
CSRP3	cysteine and glycine-rich protein 3 (cardiac LIM protein)	*CSRP3* plays an important role in the organization of cytosolic structures in cardiomyocytes. Mutations in this gene have been associated with hypertrophic cardiomyopathy and dilated cardiomyopathy (Geier et al., 2003).
MB	myoglobin	*MB* functions as an intracellular oxygen carrier in the skeletal and heart muscles of most vertebrates (Wittenberg, 2003).
MYH6	myosin heavy chain 6, cardiac muscle, alpha	*MYH6* encodes the alpha heavy chain subunit of cardiac myosin.
MYH7	myosin heavy chain 7, cardiac muscle, beta	*MYH7* encodes the beta heavy chain subunit of cardiac myosin. Mutations in this gene are associated with familial hypertrophic cardiomyopathy, myosin storage myopathy, and dilated cardiomyopathy.
MYL7	myosin light chain 7	*MYL7* binds calcium and has been shown to be a useful molecular marker for cardiac chamber specification.
MYOZ2	myozenin 2	*MYOZ2* belongs to a family of sarcomeric proteins that bind to calcineurin. Mutations in this gene cause cardiomyopathy familial hypertrophic type 16, a hereditary heart disorder.
PLN	phospholamban	*PLN* has been postulated to regulate the activity of the calcium pump of cardiac sarcoplasmic reticulum.
SYNPO2L	synaptopodin 2-like	*SYNPO2L* encodes a cytoskeletal protein. Beqqali et al. (Beqqali et al., 2010) recently reported the corresponding protein as a novel protein that interacts and colocalizes with α-actinin at the Z-disc of the sarcomere.
TNNI3	troponin I type 3 (cardiac)	*TNNI3* is one of 3 subunits that form the troponin complex of the thin filaments of striated muscle. It serves as a calcium-sensitive switch that regulates striated muscle contraction (Bhavsar et al., 1996). Mutations in this gene have been associated with familial hypertrophic cardiomyopathy type 7 and familial restrictive cardiomyopathy.
TPM1	tropomyosin 1	*TPM1* is a member of the tropomyosin family of highly conserved actin binding proteins, which are involved in the control of thin filament function in striated muscle contraction. Mutations in this gene are associated with type 3 familial hypertrophic cardiomyopathy.

TTN	titin	*TTN* encodes a large abundant protein of striated muscle. Mutations in this gene have been identified in patients with peripartum cardiomyopathy and dilated cardiomyopathy (van Spaendonck-Zwarts et al., 2014).

Table A.6: Description of kidney marker genes predicted by MGFM and verified by RT-PCR

Gene symbol	Gene name	Description
AOC1	amine oxidase, copper containing 1	*AOC1* (formerly known as amiloride-binding protein 1) is a homodimeric glycoprotein, which deaminates putrescine and histamine. *AOC1* was described to be highly expressed in the kidney (Schwelberger and Bodner, 1998), placenta (Morel et al., 1992) and intestine (Biegański et al., 1983).
CA12	carbonic anhydrase XII	Carbonic anhydrases (CAs) are a large family of zinc metalloenzymes that catalyze the reversible hydration of carbon dioxide. They participate in a variety of biological processes, including respiration, calcification, acid-base balance, bone resorption, and the formation of aqueous humor, cerebrospinal fluid, saliva, and gastric acid [3].
CD24	CD24 molecule	*CD24* encodes a sialoglycoprotein that is expressed on mature granulocytes and B cells and modulates growth and differentiation signals to these cells. Sagrinati et al. identified a population of parietal epithelial cells, isolated from the Bowman's capsule of human adult kidneys, which are $CD24+$ (Sagrinati et al., 2006).
CDH16	cadherin 16	*CDH16* encodes Kidney-specific (Ksp)-cadherin, a member of the cadherin superfamily of calcium-dependent cell adhesion molecules. It has been identified as a specific marker for terminal differentiation of the basolateral membrane of renal tubular epithelial cells (Thomson and Aronson, 1999; Thomson et al., 1995). Additionally, Ksp-cadherin has been identified as a highly sensitive marker for chromophobe renal cell carcinoma and oncocytoma (Shen et al., 2005).
CLDN8	claudin 8	*CLDN8* encodes a member of the claudin family. Claudins are integral membrane proteins of the tight junction that are involved in the permeation of solutes across epithelia via the paracellular pathway. Yu et al. (Yu et al., 2003) suggested that *CLDN8* plays an important role in the paracellular cation barrier of the distal renal tubule.
FXYD2	FXYD domain containing ion transport regulator 2	*FXYD2* encodes a member of the *FXYD* family of transmembrane proteins, which regulates the function of the Na,K-ATPase in mammalian kidney epithelial cells (Arystarkhova et al., 1999; Béguin et al., 1997).

[3] https://www.ncbi.nlm.nih.gov/gene/771

HSD11B2	hydroxysteroid (11-beta) dehydrogenase 2	*HSD11B2* encodes the type II isoform of 11-beta-hydroxysteroid dehydrogenase, a microsomal enzyme complex responsible for the interconversion of biologically active cortisol and inactive cortisone. *HSD11B2* is highly expressed in kidney and placenta (Ferrari, 2010).
PDZK1IP1	PDZK1 interacting protein 1	*PDZK1IP1*, also known as *MAP17*, encodes a membrane associated protein. *PDZK1IP1* was earlier detected in normal renal proximal tubules (Kocher et al., 1995). In addition, it has been reported to be upregulated in carcinomas arising from kidney, colon, lung, and breast (Kocher et al., 1996).
SLC12A1	solute carrier family 12 (sodium/potassium/chloride transporter), member 1	*SLC12A1* encodes a kidney-specific sodium-potassium-chloride cotransporter and accounts for renal salt reabsorption.
SLC22A8	solute carrier family 22 (organic anion transporter), member 8	*SLC22A8* is a member of the organic anion transporter *SLC22* gene family. Cha et al. (Cha et al., 2001) reported that *SLC22A8* is exclusively expressed in the kidney, and plays important roles in the basolateral uptake of organic anions in proximal tubular cells.
SLC3A1	solute carrier family 3 (amino acid transporter heavy chain), member 1	*SLC3A1* encodes a type II membrane glycoprotein which is one of the components of the renal amino acid transporter which transports neutral and basic amino acids in the renal tubule and intestinal tract. Mutations and deletions in this gene are associated with cystinuria [4].
UMOD	uromodulin	*UMOD* encodes a glycoprotein, also known as Tamm-Horsfall protein, which is produced by renal cells of ascending limb of loop of Henle and is largely excreted in urine (Serafini-Cessi et al., 1993).

[4] https://www.ncbi.nlm.nih.gov/gene/6519

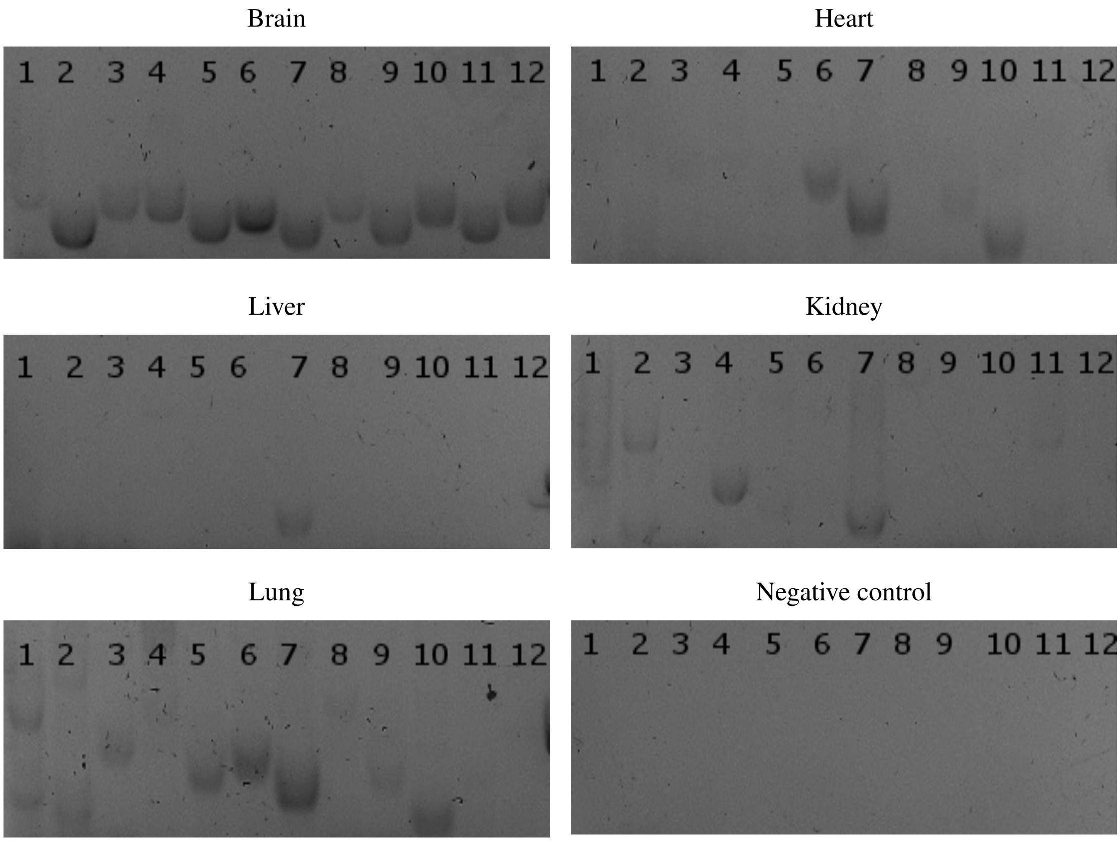
Table A.7: Description of liver marker genes predicted by MGFM and verified by RT-PCR

Gene symbol	**Gene name**	**Description**
AKR1D1	aldo-keto reductase family 1, member D1	*AKR1D1* encodes an enzyme, which is responsible for the catalysis of the 5-beta-reduction of bile acid intermediates and steroid hormones carrying a delta(4)-3-one structure.
AHSG	alpha-2-HS-glycoprotein	*AHSG* is synthesized by hepatocytes and secreted into serum. It is involved in insulin resistance and fat accumulation in the liver (Stefan, 2006).
AMBP	alpha-1-microglobulin/bikunin precursor	*AMBP* gene encodes the two plasma glycoproteins alpha-1-Microglobulin (A1M) and bikunin. *A1M* belongs to the superfamily of lipocalin transport proteins and may play a role in the regulation of inflammatory processes, whereas bikunin is an urinary trypsin inhibitor.
APOA2	apolipoprotein A-II	*APOA2* encodes apolipoprotein (apo-) A-II, which is implicated in triglyceride, fatty acid and glucose metabolism.
APOC3	apolipoprotein C-III	*APOC3* is a very low density lipoprotein (VLDL) protein. *APOC3* inhibits lipoprotein lipase and lipoprotein remnant uptake by the liver (Virgil Brown and Baginsky, 1972; Windler and Havel, 1985).
CPS1	carbamoyl-phosphate synthase 1, mitochondrial	*CPS1* is the rate-limiting enzyme in the first step of the urea cycle and an indispensable enzyme in the metabolism of human liver.
CYP2C8, *CYP2E1*	cytochrome P450, family 2, subfamily C, -E, polypeptide 8, -2	*CYP2C8* and *CYP2E1* encode members of the cytochrome P450 superfamily of enzymes. Cytochrome P450 epoxygenases are predominantly expressed in the liver (Zanger and Schwab, 2013) and they catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.
FGG	fibrinogen gamma chain	Fibrinogen, a coagulation factor responsible for normal blood clotting, is synthesized in the liver by the hepatocytes.
GC	group-specific component (vitamin D binding protein)	*GC* encodes a protein that belongs to the albumin gene family. It functions predominantly as a transporter protein for vitamin D and its metabolites (Cooke and David, 1985).
SERPINC1	serpin peptidase inhibitor, clade C (antithrombin), member 1	*SERPINC1* is a plasma protease inhibitor, synthesized in the liver and a member of the serpin superfamily. It is the principal plasma serpin of blood coagulation proteases and functions as inhibitor of thrombin and other factors by the formation of covalently linked complexes.

Table A.8: Description of lung marker genes predicted by `MGFM` and verified by RT-PCR

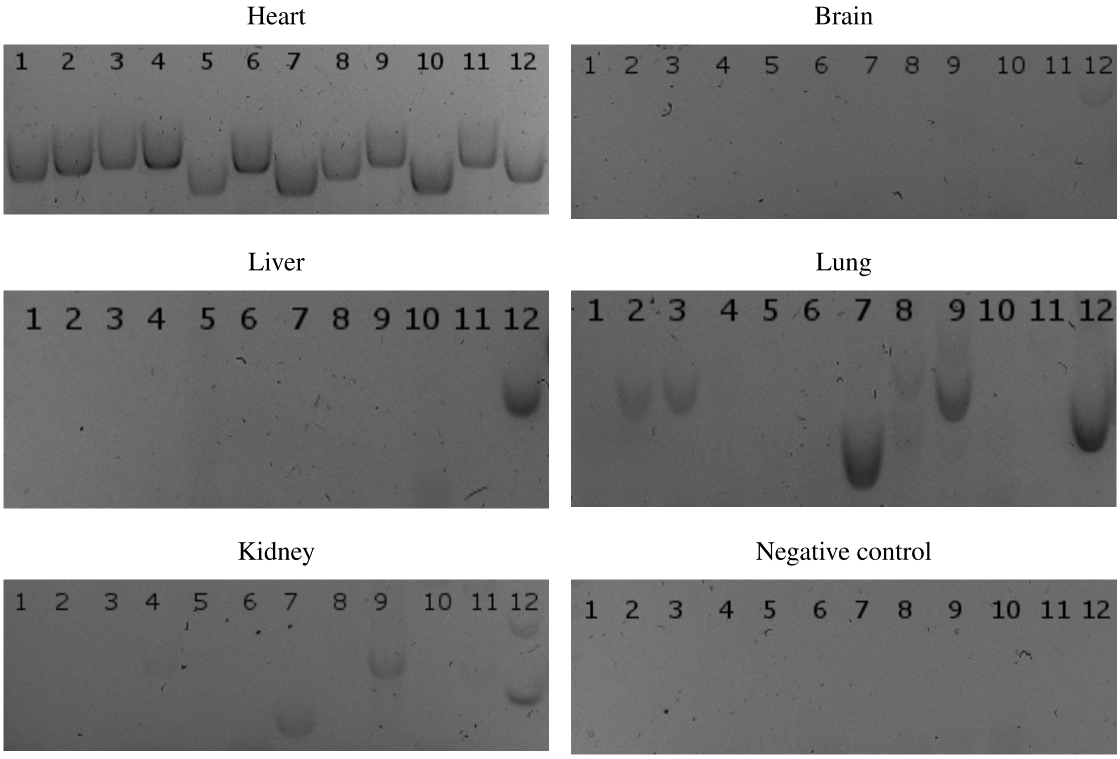
Gene symbol	Gene name	Description
AGER	advanced glycosylation end product-specific receptor	*AGER*, also known as *RAGE*, is a cell surface receptor. *AGER* is highly expressed in the lung, in particular alveolar epithelial cells (Fehrenbach et al., 1998). *AGER* expression is significantly decreased in human lung carcinomas (Jing et al., 2010).
CD52	CD52 molecule	CD52 encodes a glycoprotein expressed on normal as well as leukemic immune cells (Vojdeman et al., 2017).
CLDN18	claudin 18	CLDNs are components of cellular tight junctions regulating the permeability of cellular layers between different tissue compartments (Soini, 2005). Mutations in *CLDN18* are associated with adenocarcinomas (Merikallio et al., 2011).
CYP4B1	cytochrome P450, family 4, subfamily B, polypeptide 1	This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. Several cytochrome P450 enzymes are selectively expressed in the lung, including CYP4B1 (Poch et al., 2005).
LAMP3	lysosomal-associated membrane protein 3	*LAMP3* is the third member of the lysosome-associated membrane glycoprotein (*LAMP*) family, and it has been reported to play important roles in the occurrence and metastasis of lung cancer (Wang et al., 2013).
LYZ	lysozyme	*LYZ* is an essential component of innate defense in lung epithelia.
NKX2-1	NK2 homeobox 1	*NKX2-1* is a homeodomain-containing transcription factor, encoding thyroid transcription factor-1 (*TTF-1*). *NKX2-1* plays a crucial role in normal lung function and morphogenesis (Minoo et al., 1999), and has been suggested as a lineage marker for the terminal respiratory unit in lung carcinogenesis (Yatabe et al., 2002).
SCGB1A1	secretoglobin, family 1A, member 1	*SCGB1A1* is a member of the secretoglobin family of small secreted proteins, implicated in numerous functions including anti-inflammation. Defects in *SCGB1A1* are associated with a susceptibility to asthma.
SFTPB, *SFTPC*, *SFTPD*	surfactant protein B, -C, -D	*SFTPB*, *SFTPC*, and *SFTPD* are surfactant proteins which are involved in defense against microbial invasion. Moreover, they act by lowering the surface tension and are crucial for gaseous exchange between air and blood.

SLC34A2	solute carrier family 34 (type II sodium/phosphate co-transporter), member 2	*SLC34A2* is a phosphate transport protein. Mutations in the corresponding gene has been associated with pulmonary alveolar microlithiasis (Corut et al., 2006).



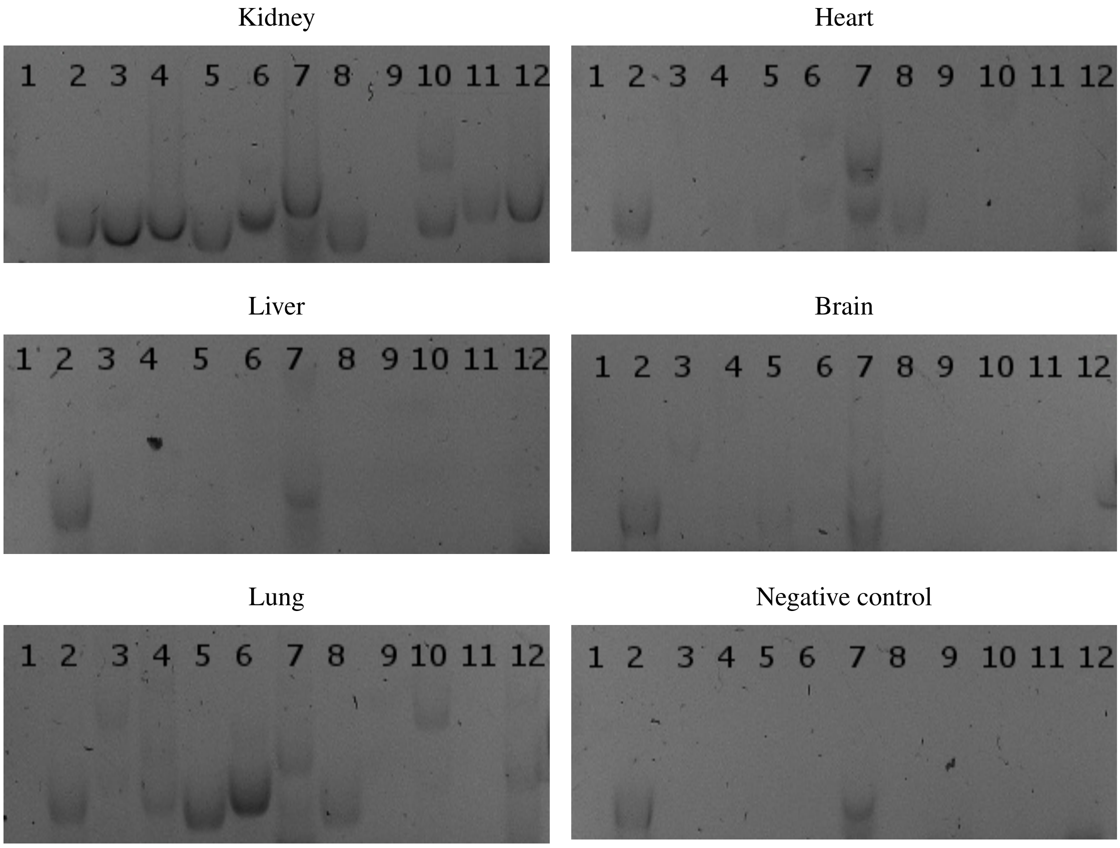
1: *GAP43*, 2: *GFAP*, 3: *TMEFF1*, 4: *FUT9*, 5: *SYT1*, 6: *SNAP25*, 7: *MBP*, 8: *GRIA2*, 9: *KIFSC*, 10: *STMN2*, 11: *NEFM*, 12: *GABBR2*

Figure A.1: Expression of human gene transcripts predicted by `MGFM` as markers for brain, detected by RT-PCR and agarose gel electrophoresis in brain, liver, lung, heart, kidney and the negative control (no template).



1: *MYOZ2*, 2: *TNNI3*, 3: *SYNPO2L*, 4: *MYH6*, 5: *CSRP3*, 6: *CKM*, 7: *PLN*, 8: *MB*, 9: *TTN*, 10: *MYL7*, 11: *MYH7*, 12: *TPM1*

Figure A.2: Expression of human gene transcripts predicted by `MGFM` as markers for heart, detected by RT-PCR and agarose gel electrophoresis in heart, liver, kidney, brain, lung and the negative control (no template).



Kidney

Heart

Liver

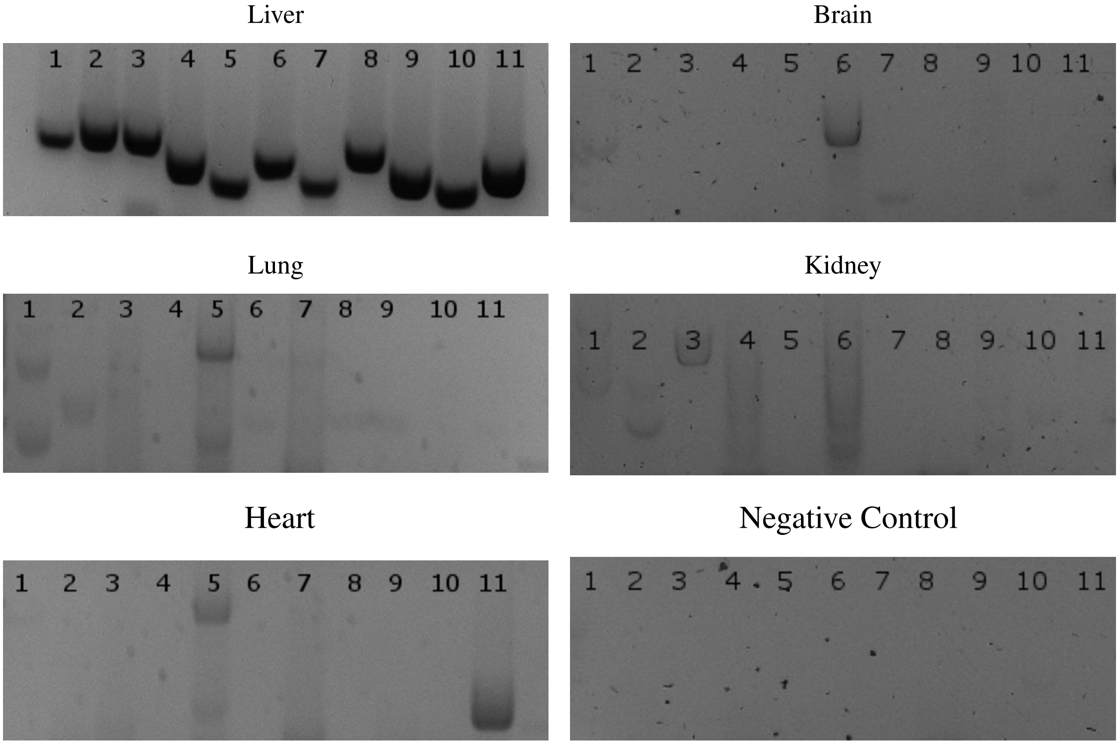
Brain

Lung

Negative control

1: *SLC12A1*, 2: *SLC3A1*, 3: *UMOD*, 4: *AOC1*, 5: *CD24*, 6: *HSD11B2*, 7: *CA12*, 8: *PDZK1IP1*, 9: *FXYD2*, 10: *CDH16*, 11: *SLC22A8*, 12: *CLDN8*

Figure A.3: Expression of human gene transcripts predicted by `MGFM` as markers for kidney, detected by RT-PCR and agarose gel electrophoresis in kidney, liver, lung, heart, brain and the negative control (no template).



Liver

Brain

Lung

Kidney

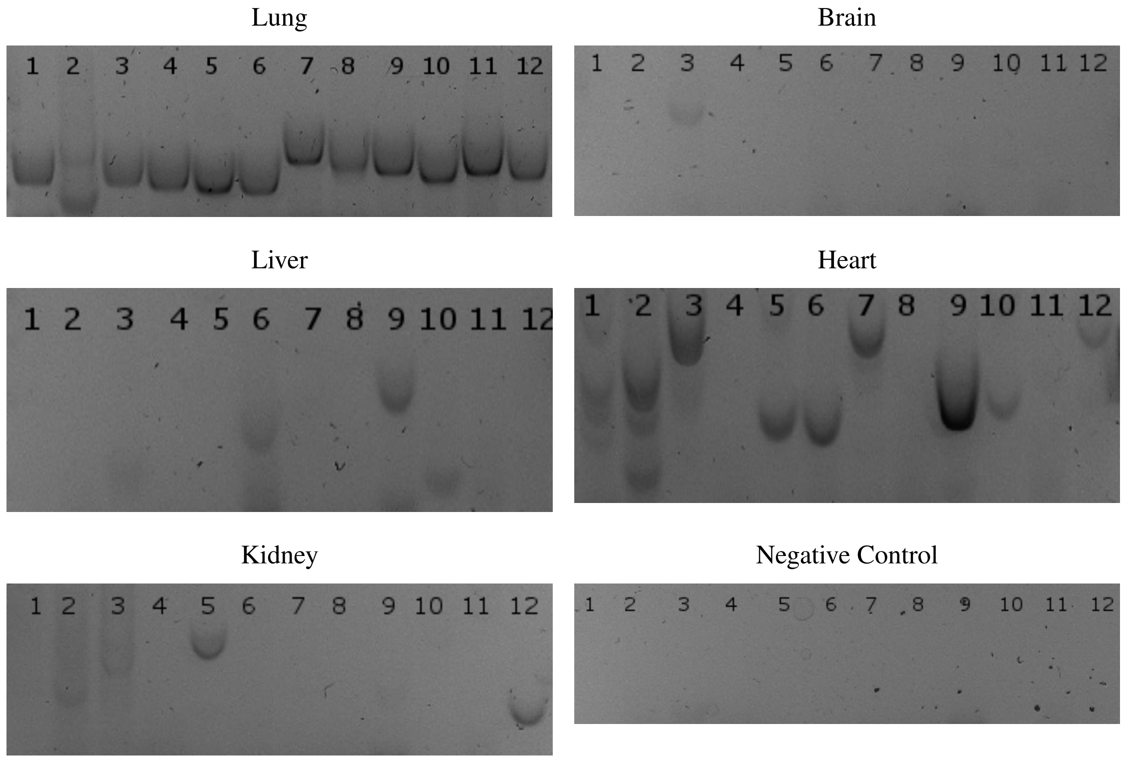
Heart

Negative Control

1: *AKR1D1*, 2: *APOA2*, 3: *CYP2C8*, 4: *GC*, 5: *CPS1*, 6: *CYP2E1*, 7: *SERPINC1*, 8: *AHSG*, 9: *APOC3*, 10: *AMBP*, 11: *FGG*

1: *AKR1D1*, 2: *FGG*, 3: *APOA2* 4: *CYP2C8*, 5: *GC*, 6: *CPS1*, 7: *CYP2E1*, 8: *APOC3*, 9: *SERPINC1*, 10: *AHSG*, 11: *AMBP*

Figure A.4: Expression of human gene transcripts predicted by `MGFM` as markers for liver, detected by RT-PCR and agarose gel electrophoresis in liver, lung, heart, brain, kidney and the negative control (no template).



1: *CLDN18*, 2: *NKX2-1*, 3: *SCGB1A1*, 4: *SFTPB*, 5: *CYP4B1*, 6: *CD52*, 7: *LAMP3*, 8: *AGER*, 9: *LYZ*, 10: *SFTPD*, 11: *SFTPC*, 12: *SLC34A2*

Figure A.5: Expression of human gene transcripts predicted by `MGFM` as markers for lung, detected by RT-PCR and agarose gel electrophoresis in lung, liver, kidney, brain, heart and the negative control (no template).

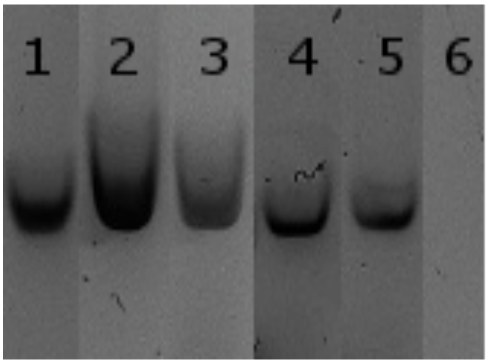


Figure A.6: Expression of the human gene transcript β-actin, detected by RT-PCR and agarose gel electrophoresis in liver (1), lung (2), heart (3), brain (4), kidney (5) and the negative control (no template) (6).

Appendix B

Supplementary data to Chapter 3

Table B.1: The corresponding IDs to the samples used in the microarray reference dataset

Tissue	Sample IDs
adipose tissue omental	GSM80561, GSM80562, GSM80564
bone marrow	GSM80602, GSM80603, GSM80604
cerebellum	GSM80637, GSM80638, GSM80639
dorsal root ganglia	GSM80612, GSM80613, GSM80614
heart atrium	GSM80655, GSM80656, GSM80698
kidney cortex	GSM80686, GSM80687, GSM80689
liver	GSM80728, GSM80730, GSM80739
lung	GSM80707, GSM80710, GSM80712
lymph nodes	GSM80735, GSM80737, GSM80738
oral mucosa	GSM80776, GSM80777, GSM80778
ovary	GSM80758, GSM80759, GSM80780
parietal lobe	GSM80744, GSM80745, GSM80746
pituitary gland	GSM80802, GSM80803, GSM80804
prostate gland	GSM80805, GSM80806, GSM80824
putamen	GSM80581, GSM80595, GSM80596
saphenous vein	GSM80788, GSM80789, GSM80793
skeletal muscle	GSM80790, GSM80791, GSM80792
spinal cord	GSM80784, GSM80786, GSM80787
spleen	GSM80808, GSM80825, GSM80826
testes	GSM80853, GSM80868, GSM80869
thalamus	GSM80838, GSM80841, GSM80863
thyroid gland	GSM80864, GSM80865, GSM80867
trigeminal ganglia	GSM80875, GSM80877, GSM80878
urethra	GSM80911, GSM80912, GSM80913
vestibular nuclei superior	GSM80879, GSM80880, GSM80881
vulva	GSM80898, GSM80899, GSM80900

Table B.2: The corresponding IDs to the samples used in the RNA-seq reference dataset

Tissue	Sample IDs
adipose tissue	*fat_a, fat_e, fat_x1*
adrenal gland	*adrenal_4a, adrenal_4c, adrenal_4d*
Appendix	*appendix_4a, appendix_4b, appendix_4c*
bone marrow	*bonemarrow_5a, bonemarrow_6a, bonemarrow_6c*
brain	*brain_a, brain_3b, brain_3c*
colon	*colon_a, colon_b, colon_c*
endometrium	*endometrium_4a, endometrium_4b, endometrium_5a*
Esophagus	*esophagus_5a, esophagus_5b, esophagus_5c*
gallbladder	*gallbladder_5a, gallbladder_5b, gallbladder_5c*
heart	*heart_5b, heart_6a, heart_6b*
kidney	*kidney_b, kidney_c, kidney_d*
liver	*liver_a, liver_c, liver_d*
lung	*lung_4a, lung_4d, lung_3e*
lymph nodes	*lymphnode_4a, lymphnode_4b, lymphnode_5a*
ovary	*ovary_6a, ovary_6b*
placenta	*placenta_6a, placenta_6b, placenta_6c*
prostate	*prostate_4a, prostate_4b, prostate_4c*
salivary gland	*salivarygland_6a, salivarygland_6b, salivarygland_6c*
skin	*skin_5e, skin_5f, skin_6a*
small intestine	*smallintestine_4a, smallintestine_4b, smallintestine_4c*
spleen	*spleen_3a, spleen_3c, spleen_3d*
stomach	*stomach_a, stomach_3a, stomach_3b*
testis	*testis_7a, testis_7b, testis_7c*
thyroid gland	*thyroid_5a, thyroid_5b, thyroid_5c*

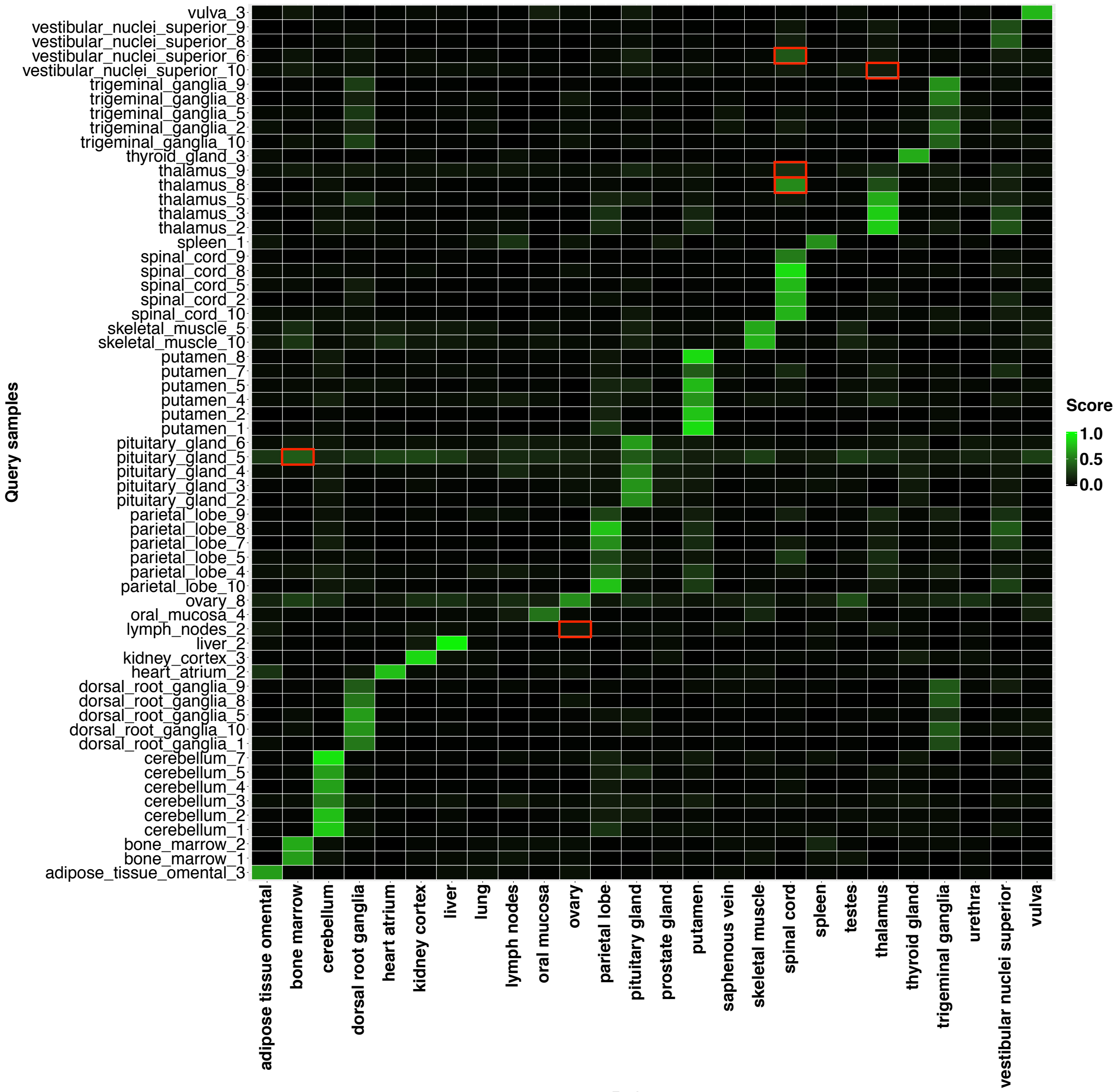


Figure B.1: Classification heatmap of 61 samples from the study GSE3526 using `sampleClassifier`, and `YuGene` for normalization. The misclassified samples are marked in red.

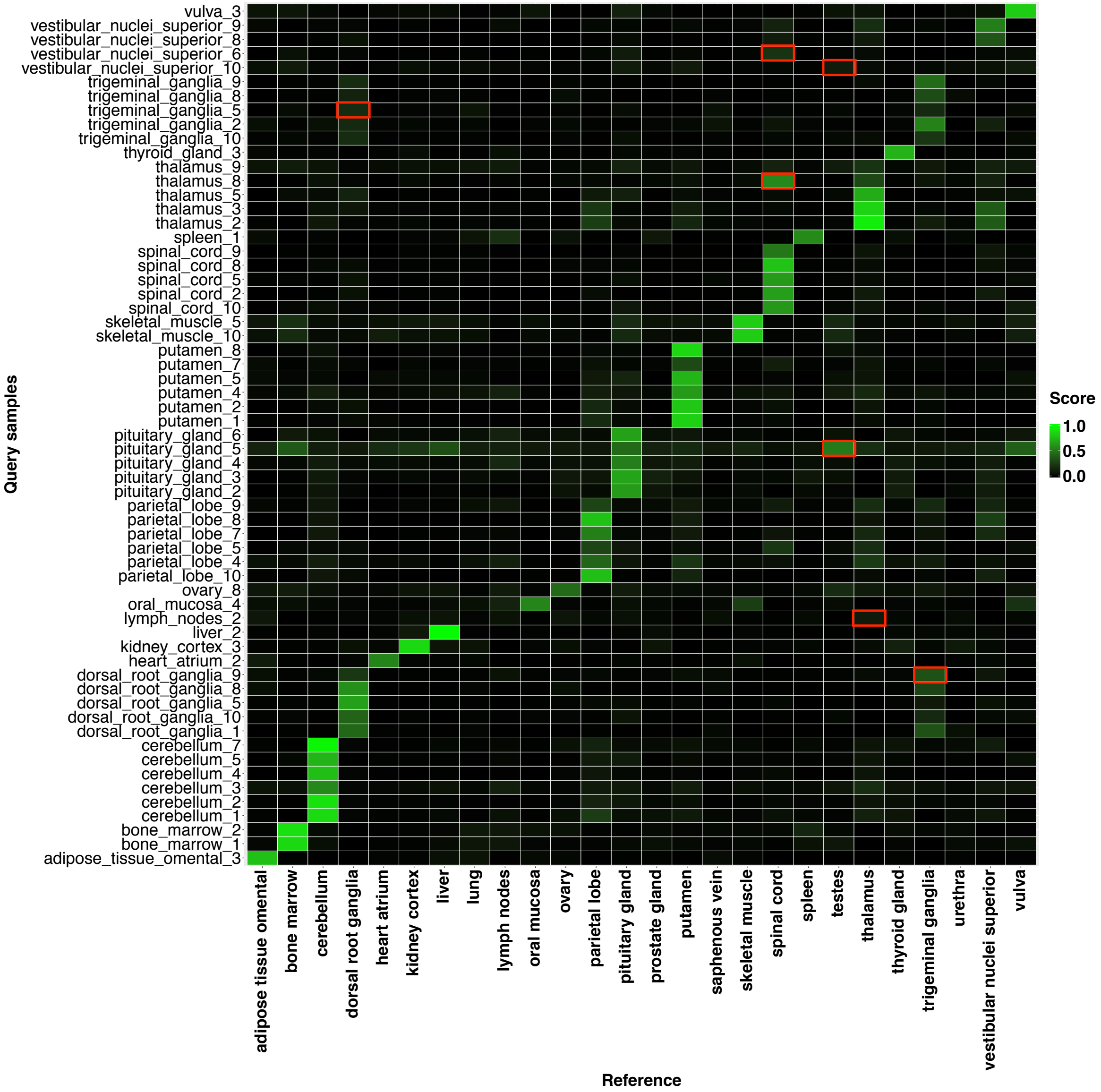


Figure B.2: Classification heatmap of 61 samples from the study GSE3526 using `sampleClassifier`, and `RMA` for normalization. The misclassified samples are marked in red.

Table B.3: Misclassified samples from the study GSE3526 by SVMs using `RMA` or `YuGene` for normalization

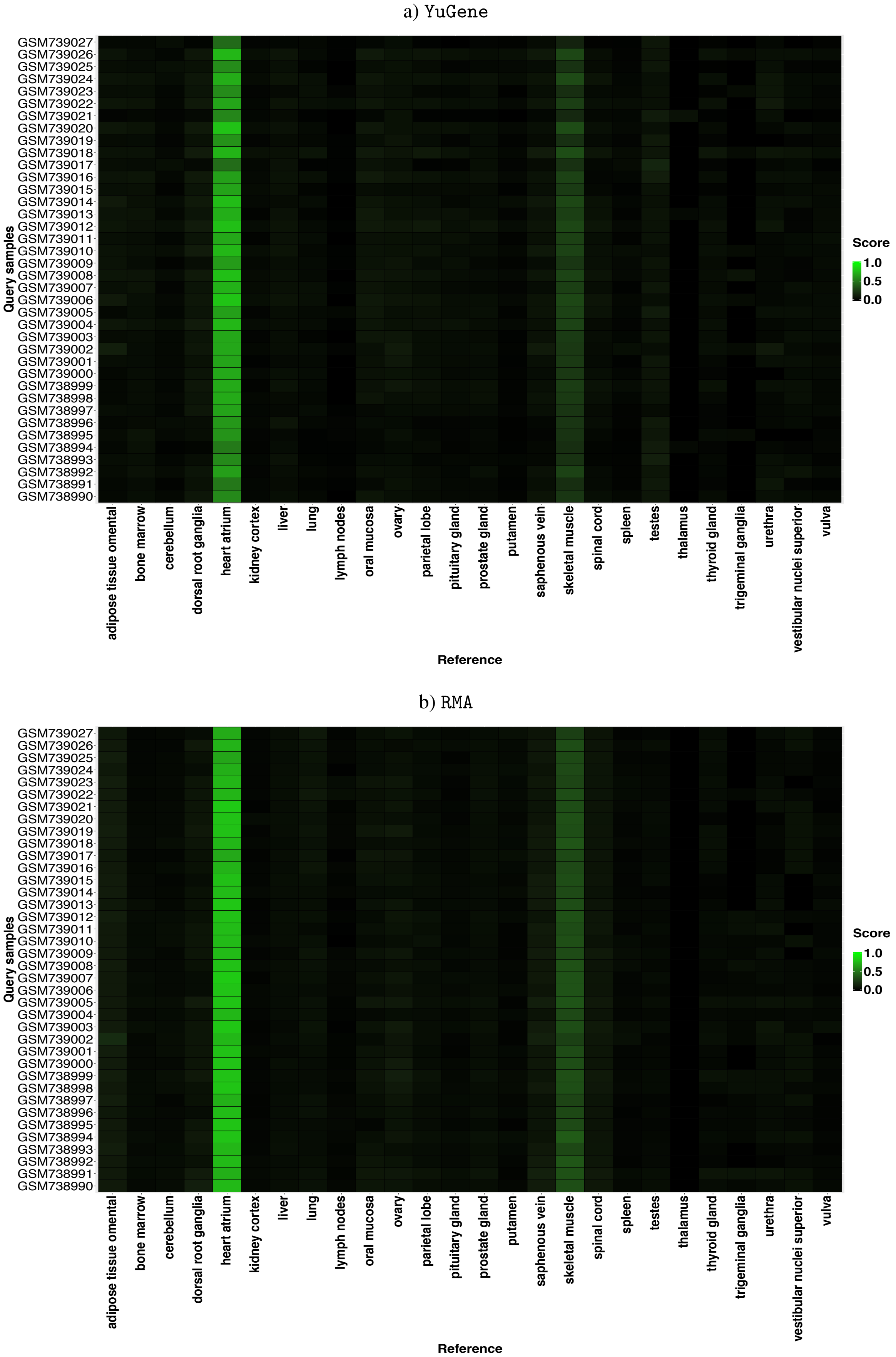
Query name	**Predicted class (RMA)**	**Predicted class (YuGene)**
*dorsal_root_ganglia*_9	trigeminal ganglia	trigeminal ganglia
*lymph_nodes*_2	adipose tissue omental	adipose tissue omental
*ovary*_8	ovary	bone marrow
*pituitary_gland*_5	bone marrow	bone marrow
*trigeminal_ganglia*_5	dorsal root ganglia	dorsal root ganglia
*vestibular_nuclei_superior*_6	thalamus	spinal cord
*vestibular_nuclei_superior*_10	thalamus	thalamus
*parietal_lobe*_9	thalamus	parietal lobe
*thalamus*_8	spinal cord	thalamus
*trigeminal_ganglia*_10	dorsal root ganglia	trigeminal ganglia

Table B.4: Misclassified samples from the study GSE2361 by SVMs using `RMA` or `YuGene` for normalization

Query name	**Predicted class (RMA)**	**Predicted class (YuGene)**
GSM44671 : Heart	pituitary gland	heart atrium
GSM44673 : Spleen	bone marrow	adipose tissue omental
GSM44674 : Ovary	pituitary gland	ovary
GSM44675 : Kidney	pituitary gland	kidney cortex
GSM44678 : Prostate	pituitary gland	urethra
GSM44689 : Cerebellum	pituitary gland	cerebellum
GSM44698 : Thalamus	pituitary gland	thalamus
GSM44700 : Spinal Cord	pituitary gland	spinal cord
GSM44701 : Testis	pituitary gland	testes
GSM44705 : Fetal Lung	pituitary gland	lung
GSM44706 : Fetal Liver	bone marrow	bone marrow

Table B.5: Misclassified samples (from E-MTAB-1733 or E-MTAB-513) by SVMs

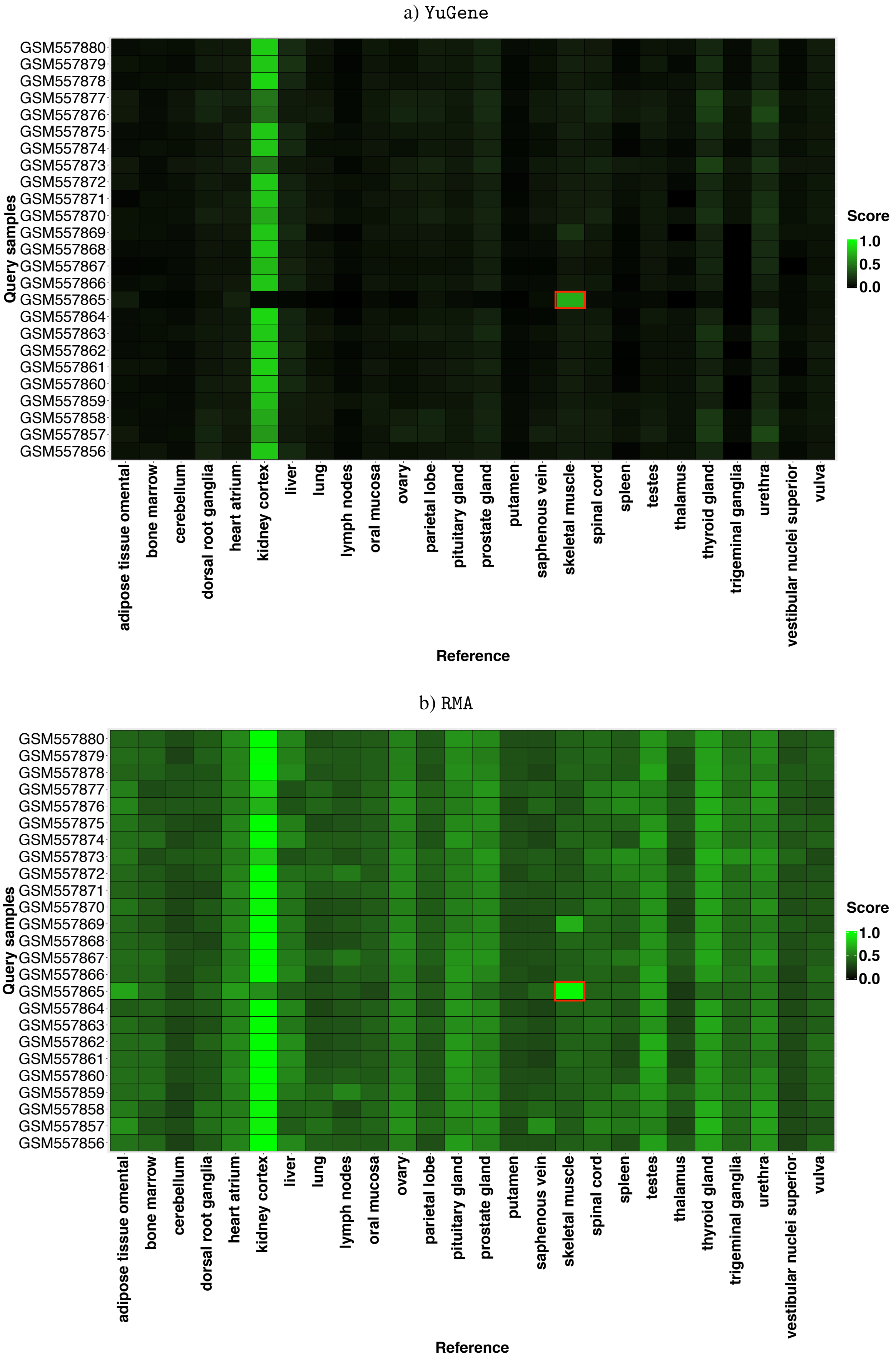
Study ID	**Query name**	**Predicted class**
E-MTAB-1733	*prostate_a*	endometrium
E-MTAB-513	adrenal	appendix
	colon	endometrium
	lymph node	appendix
	ovary	endometrium
	prostate	endometrium



a) YuGene

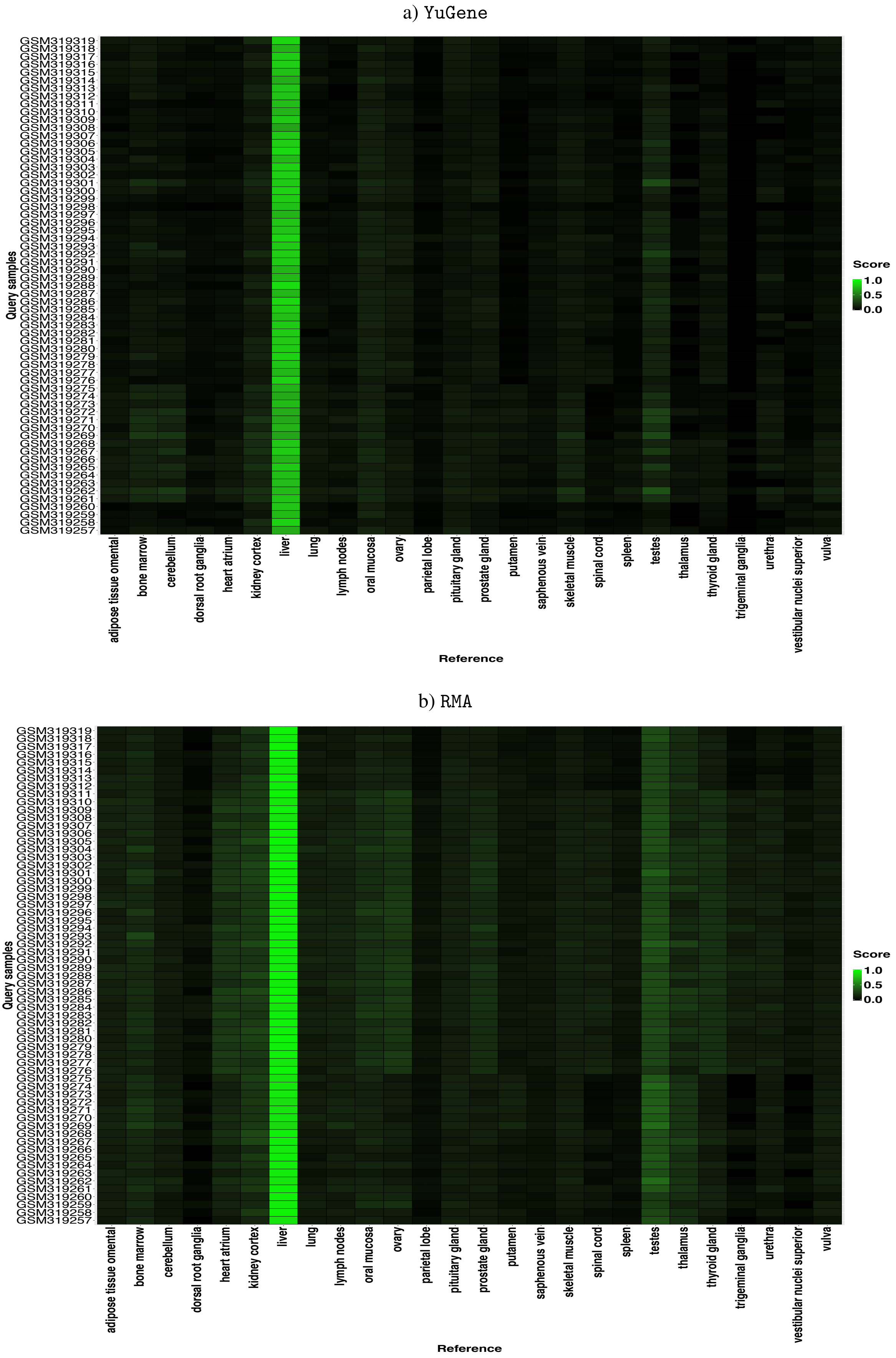
b) RMA

Figure B.3: Classification heatmaps of 38 heart samples using `sampleClassifier`, and a) `YuGene` or b) `RMA` for normalization. The mean similarity scores for the `YuGene` or `RMA` normalized samples were 0.66 or 0.76, respectively.


a) YuGene

b) RMA

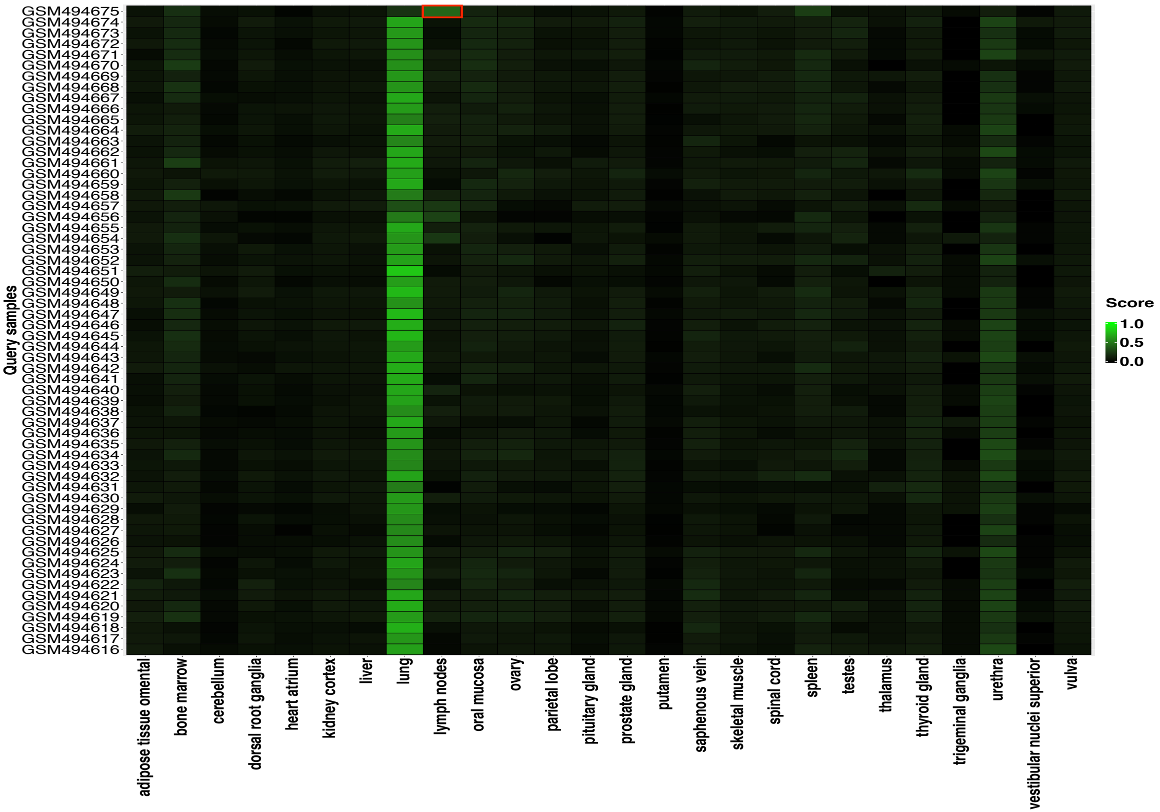
Figure B.4: Classification heatmaps of 25 kidney samples using `sampleClassifier`, and a) `YuGene` or b) `RMA` for normalization. The mean similarity scores for the `YuGene` or `RMA` normalized samples were 0.74 or 0.96, respectively.


a) YuGene

b) RMA

Figure B.5: Classification heatmaps of 63 liver samples using `sampleClassifier`, and a) `YuGene` or b) `RMA` for normalization. The mean similarity scores for the `YuGene` or `RMA` normalized samples were 0.78 or 0.94, respectively.

a) YuGene



b) RMA

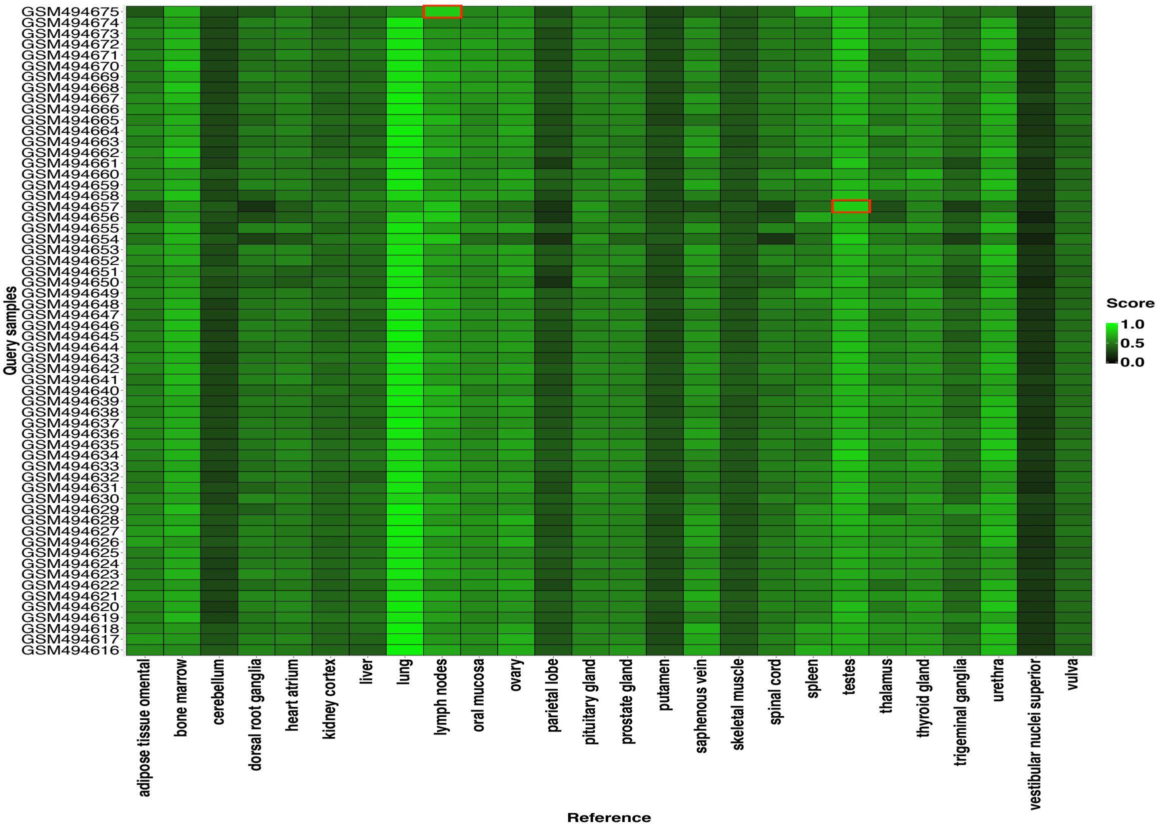


Figure B.6: Classification heatmaps of 60 lung samples using `sampleClassifier`, and a) `YuGene` or b) `RMA` for normalization. The mean similarity scores for the `YuGene` or `RMA` normalized samples were 0.63 or 0.9, respectively.

Appendix C

Zusammenfassung

Die Identifizierung von gewebe- oder zelltypspezifischen Markergenen, sowie die Unterscheidung zwischen verschiedenen Klassen von Proben, wie z. B. verschiedenen Zelltypen oder Geweben, basierend auf ihren Genexpressionsprofilen sind wichtige Aspekte innerhalb der Zellforschung. Herauszufinden, inwiefern sich die Genexpressionsprofile verschiedener Probenarten unterscheiden bzw. ähneln, ist von großer Bedeutung für das Verständnis der Zelldifferenzierung, Entwicklung und Erkrankungen.

In dieser Doktorarbeit stelle ich neue bioinformatische Ansätze vor, um Markergene zu detektieren und Proben anhand von Genexpressionsprofilen zu klassifizieren. Die Beiträge der Arbeit können in drei Teilprojekte unterteilt werden:

Erstens habe ich das Marker-Tool `MGFM` (Marker Gene Finder in Microarray gene expression data) optimiert und erweitert um die Vorhersage von Markergenen aus RNA-seq-Daten zu unterstützen. Zu diesem Zweck habe ich ein R-Paket namens `MGFR` (Marker Gene Finder in RNA-seq data) implementiert. Darüber hinaus präsentiere ich eine Vergleichsstudie zwischen Microarrays und RNA-seq. Ich identifiziere robuste Markergene (vorhergesagt durch `MGFM` und `MGFR`) für 16 humane Gewebe, und schlage neue Kandidatenmarkergene für jedes der untersuchten Gewebe vor. Als nächstes vergleiche ich die vorhergesagten Markergene mit einer Gold-Standard Liste von Markergenen, die aus der TiGER (Tissue-specific Gene Expression and Regulation) Datenbank extrahiert wurden. Darüber hinaus habe ich die expression von Top-Markergenen durch reverse Transkriptase-Polymerase-Kettenreaktion (RT-PCR) für fünf Gewebe validiert.

Zweitens habe ich `sampleClassifier` entwickelt, ein neuartiges bioinformatisches Tool, das einen einfachen Algorithmus namens "Shared Marker Genes" (SMG) verwendet, um Proben basierend auf ihrem Genexpressionsprofil zu klassifizieren. Wie der Name schon sagt, wird die Anzahl der gemeinsamen Markergene zwischen einer Referenz und einer Abfrageprobe als Ähnlichkeitsmaß verwendet. Ich zeige den Nutzen und die Wirksamkeit des vorgeschlagenen Ansatzes durch die Klassifizierung verschiedener Gewebe unter Verwendung von öffentlichen Microarray- und RNA-seq-Datensätzen. Darüber hinaus habe ich mein Tool mit Support Vector Machines (SVMs) verglichen. Die Genauigkeit meines Tools ist besser oder vergleichbar mit der der SVMs. Der SMG Algorithmus ist als R-Paket implementiert, das auf der Bioconductor Website (http://www.bioconductor.org) verfügbar ist.

Zum Schluss wende ich `MGFM` und `sampleClassifier` auf der Grundlage von öffentlich zugänglichen Biopsie-basierten Microarraydaten von acht verschiedenen Nierenerkrankungen an. Ich identifiziere Markergene für jede der untersuchten Krankheiten, und demonstriere die Performance des Klassifizierungstools bei der Unterscheidung zwischen Genexpressionsprofilen von normalem und erkranktem Gewebe, sowie zwischen verschiedenen Arten von Nierenerkrankungen.

Bibliography

Aizawa, H., Sekine, Y., Takemura, R., Zhang, Z., Nangaku, M., and Hirokawa, N. (1992). Kinesin family in murine central nervous system. *The Journal of cell biology*, 119(5):1287–96.

Arystarkhova, E., Wetzel, R. K., Asinovski, N. K., and Sweadner, K. J. (1999). The gamma subunit modulates Na(+) and K(+) affinity of the renal Na,K-ATPase. *The Journal of biological chemistry*, 274(47):33183–5.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25:25–29.

Bahn, S., Mimmack, M., Ryan, M., Caldwell, M. A., Jauniaux, E., Starkey, M., Svendsen, C. N., and Emson, P. (2002). Neuronal target genes of the neuron-restrictive silencer factor in neurospheres derived from fetuses with Down's syndrome: a gene expression study. *Lancet*, 359(9303):310–5.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*, 41(D1).

Béguin, P., Wang, X., Firsov, D., Puoti, A., Claeys, D., Horisberger, J. D., and Geering, K. (1997). The gamma subunit is a specific component of the Na,K-ATPase and modulates its transport function. *The EMBO journal*, 16(14):4250–60.

Beqqali, A., Monshouwer-Kloots, J., Monteiro, R., Welling, M., Bakkers, J., Ehler, E., Verkleij, A., Mummery, C., and Passier, R. (2010). CHAP is a newly identified Z-disc protein essential for heart and skeletal muscle function. *Journal of cell science*, 123(Pt 7):1141–50.

Berthier, C. C., Bethunaickan, R., Gonzalez-Rivera, T., Nair, V., Ramanujam, M., Zhang, W., Bottinger, E. P., Segerer, S., Lindenmeyer, M., Cohen, C. D., Davidson, A., and Kretzler, M. (2012). Cross-Species Transcriptional Network Analysis Defines Shared Inflammatory Responses in Murine and Human Lupus Nephritis. *The Journal of Immunology*, 189(2).

Bhavsar, P. K., Brand, N. J., Yacoub, M. H., and Barton, P. J. (1996). Isolation and characterization of the human cardiac troponin I gene (TNNI3). *Genomics*, 35(1):11–23.

Biegański, T., Kusche, J., Lorenz, W., Hesterberg, R., Stahlknecht, C. D., and Feussner, K. D. (1983). Distribution and properties of human intestinal diamine oxidase and its relevance for the histamine catabolism. *Biochimica et biophysica acta*, 756(2):196–203.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2):185–93.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brenner, M., Johnson, A. B., Boespflug-Tanguy, O., Rodriguez, D., Goldman, J. E., and Messing, A. (2001). Mutations in GFAP, encoding glial fibrillary acidic protein, are associated with Alexander disease. *Nature genetics*, 27(1):117–20.

Bumgarner, R. (2013). Overview of dna microarrays: Types, applications, and their future. *Current Protocols in Molecular Biology*, Chapter 22(SUPPL.101):Unit 22.1.

Cahan, P., Li, H., Morris, S., Lummertz da Rocha, E., Daley, G., Collins, J., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V., Ren, B., and et Al. (2014). CellNet: Network Biology Applied to Stem Cell Engineering. *Cell*, 158(4):903–915.

Cha, S. H., Sekine, T., Fukushima, J. I., Kanai, Y., Kobayashi, Y., Goya, T., and Endou, H. (2001). Identification and characterization of human organic anion transporter 3 expressing predominantly in the kidney. *Molecular pharmacology*, 59(5):1277–86.

Chaudhry, A. S., Thirumaran, R. K., Yasuda, K., Yang, X., Fan, Y., Strom, S. C., and Schuetz, E. G. (2013). Genetic Variation in Aldo-Keto Reductase 1D1 (AKR1D1) Affects the Expression and Activity of Multiple Cytochrome P450s. *Drug Metabolism and Disposition*, 41(8):1538–1547.

Cooke, N. E. and David, E. V. (1985). Serum vitamin D-binding protein is a third member of the albumin and alpha fetoprotein gene family. *The Journal of clinical investigation*, 76(6):2420–4.

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

Corut, A., Senyigit, A., Ugur, S. A., Altin, S., Ozcelik, U., Calisir, H., Yildirim, Z., Gocmen, A., and Tolun, A. (2006). Mutations in SLC34A2 cause pulmonary alveolar microlithiasis and are possibly associated with testicular microlithiasis. *American journal of human genetics*, 79(4):650–6.

Corvol, H., Blackman, S. M., Boëlle, P.-Y., Gallins, P. J., Pace, R. G., Stonebraker, J. R., Accurso, F. J., Clement, A., Collaco, J. M., Dang, H., Dang, A. T., Franca, A., Gong, J., Guillot, L., Keenan, K., Li, W., Lin, F., Patrone, M. V., Raraigh, K. S., Sun, L., Zhou, Y.-H., O'Neal, W. K., Sontag, M. K., Levy, H., Durie, P. R., Rommens, J. M., Drumm, M. L., Wright, F. A., Strug, L. J., Cutting, G. R., and Knowles, M. R. (2015). Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nature Communications*, 6:8382.

Crick Mc, F. (1970). Central Dogma of Molecular Biology. *NATURE*, 227(8).

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 edition.

D'Agati, V. D., Kaskel, F. J., and Falk, R. J. (2011). Focal Segmental Glomerulosclerosis. *New England Journal of Medicine*, 365(25):2398–2411.

Ding, Q., Kang, J., Dai, J., Tang, M., Wang, Q., Zhang, H., Guo, W., Sun, R., and Yu, H. (2016). AGXT2L1 is down-regulated in heptocellular carcinoma and associated with abnormal lipogenesis. *Journal of Clinical Pathology*, 69(3):215–220.

Du Puy, L., Beqqali, A., Monshouwer-Kloots, J., Haagsman, H. P., Roelen, B. A., and Passier, R. (2009). CAZIP, a novel protein expressed in the developing heart and nervous system. *Developmental Dynamics*, 238(11):2903–2911.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440.

Eckardt, K.-U., Coresh, J., Devuyst, O., Johnson, R. J., Köttgen, A., Levey, A. S., and Levin, A. (2013). Evolving importance of kidney disease: from subspecialty to global health burden. *The Lancet*, 382(9887):158–169.

El Amrani, K., Stachelscheid, H., Lekschas, F., Kurtz, A., and Andrade-Navarro, M. A. (2015). MGFM: a novel tool for detection of tissue and cell specific marker genes from microarray gene expression data. *BMC genomics*, 16(1):645.

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C. A.-K., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., Nilsson, P., Schwenk, J. M., Lindskog, C., Danielsson, F., Mardinoglu, A., Sivertsson, A., von Feilitzen, K., Forsberg, M., Zwahlen, M., Olsson, I., Navani, S., Huss, M., Nielsen, J., Ponten, F., and Uhlén, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & cellular proteomics : MCP*, 13(2):397–406.

Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics (Oxford, England)*, 23(2):257–8.

Fatemi, S. H., Folsom, T. D., Reutiman, T. J., and Thuras, P. D. (2009). Expression of GABA(B) receptors is altered in brains of subjects with autism. *Cerebellum (London, England)*, 8(1):64–9.

Fatemi, S. H., Folsom, T. D., and Thuras, P. D. (2011). Deficits in GABA(B) receptor system in schizophrenia and mood disorders: a postmortem study. *Schizophrenia research*, 128(1-3):37–43.

Fehrenbach, H., Kasper, M., Tschernig, T., Shearman, M. S., Schuh, D., and Müller, M. (1998). Receptor for advanced glycation endproducts (RAGE) exhibits highly differential cellular and subcellular localisation in rat and human lung. *Cellular and molecular biology (Noisy-le-Grand, France)*, 44(7):1147–57.

Fernández-Chacón, R., Königstorfer, A., Gerber, S. H., García, J., Matos, M. F., Stevens, C. F., Brose, N., Rizo, J., Rosenmund, C., and Südhof, T. C. (2001). Synaptotagmin I functions as a calcium regulator of release probability. *Nature*, 410(6824):41–9.

Ferrari, P. (2010). The role of 11β-hydroxysteroid dehydrogenase type 2 in human hypertension. *Biochimica et biophysica acta*, 1802(12):1178–87.

Fogo, A. B. (2015). Causes and pathogenesis of focal segmental glomerulosclerosis. *Nature reviews. Nephrology*, 11(2):76–87.

Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S. M., and Aburatani, H. (2005). Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, 86(2):127–41.

Geier, C., Perrot, A., Ozcelik, C., Binner, P., Counsell, D., Hoffmann, K., Pilz, B., Martiniak, Y., Gehmlich, K., van der Ven, P. F. M., Fürst, D. O., Vornwald, A., von Hodenberg, E., Nürnberg, P., Scheffold, T., Dietz, R., and Osterziel, K. J. (2003). Mutations in the human muscle LIM protein gene in families with hypertrophic cardiomyopathy. *Circulation*, 107(10):1390–5.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome research*, 17(6):669–81.

Gery, S., Yin, D., Xie, D., Black, K. L., and Koeffler, H. P. (2003). TMEFF1 and brain tumors. *Oncogene*, 22(18):2723–7.

Gosso, M. F., de Geus, E. J. C., van Belzen, M. J., Polderman, T. J. C., Heutink, P., Boomsma, D. I., and Posthuma, D. (2006). The SNAP-25 gene is associated with cognitive ability: evidence from a family-based study in two independent Dutch cohorts. *Molecular psychiatry*, 11(9):878–86.

Hackmann, K., Matko, S., Gerlach, E.-M., von der Hagen, M., Klink, B., Schrock, E., Rump, A., and Di Donato, N. (2013). Partial deletion of GLRB and GRIA2 in a patient with intellectual disability. *European journal of human genetics : EJHG*, 21(1):112–4.

Haroun, M. K., Jaar, B. G., Hoffman, S. C., Comstock, G. W., Klag, M. J., and Coresh, J. (2003). Risk factors for chronic kidney disease: a prospective study of 23,534 men and women in Washington County, Maryland. *Journal of the American Society of Nephrology : JASN*, 14(11):2934–41.

Hennessey, J. A., Marcou, C. A., Wang, C., Wei, E. Q., Wang, C., Tester, D. J., Torchio, M., Dagradi, F., Crotti, L., Schwartz, P. J., Ackerman, M. J., and Pitt, G. S. (2013). FGF12 is a candidate Brugada syndrome locus. *Heart rhythm*, 10(12):1886–94.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121.

Igarashi, P., Vanden Heuvel, G. B., Payne, J. A., and Forbush, B. (1995). Cloning, embryonic expression, and alternative splicing of a murine kidney-specific Na-K-Cl cotransporter. *The American journal of physiology*, 269(3 Pt 2):F405–18.

Imaizumi, T., Aizawa-Yashiro, T., Tsuruga, K., Tanaka, H., Matsumiya, T., Yoshida, H., Tatsuta, T., Xing, F., Hayakari, R., and Satoh, K. (2012). MDA5 Regulates PolyIC-Induced CXCL10 in Mesangial Cells Melanoma Differentiation-Associated Gene 5 Regulates the Expression of a Chemokine CXCL10 in Human Mesangial Cells: Implications for Chronic Inflammatory Renal Diseases. *Tohoku J. Exp. Med. Tohoku J. Exp. Med*, 228(2281):17–26.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4):e15.

Isaac, J. T. R., Ashby, M. C., and McBain, C. J. (2007). The role of the GluR2 subunit in AMPA receptor function and synaptic plasticity. *Neuron*, 54(6):859–71.

James, M. T., Hemmelgarn, B. R., and Tonelli, M. (2010). Early recognition and prevention of chronic kidney disease. *The Lancet*, 375(9722):1296–1309.

Jiang, L., Kwong, D. L.-W., Li, Y., Liu, M., Yuan, Y.-F., Li, Y., Fu, L., and Guan, X.-Y. (2015). HBP21, a chaperone of heat shock protein 70, functions as a tumor suppressor in hepatocellular carcinoma. *Carcinogenesis*, 36(10):1111–1120.

Jing, R., Cui, M., Wang, J., and Wang, H. (2010). Receptor for advanced glycation end products (RAGE) soluble form (sRAGE): a new biomarker for lung cancer. *Neoplasma*, 57(1):55–61.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, 8(1):118–27.

Ju, W., Greene, C. S., Eichinger, F., Nair, V., Hodgin, J. B., Bitzer, M., Lee, Y.-S., Zhu, Q., Kehata, M., Li, M., Jiang, S., Rastaldi, M. P., Cohen, C. D., Troyanskaya, O. G., and Kretzler, M. (2013). Defining cell-type specificity at the transcriptional level in human disease. *Genome research*, 23(11):1862–73.

Kaupmann, K., Malitschek, B., Schuler, V., Heid, J., Froestl, W., Beck, P., Mosbacher, J., Bischoff, S., Kulik, A., Shigemoto, R., Karschin, A., and Bettler, B. (1998). GABA(B)-receptor subtypes assemble into functional heteromeric complexes. *Nature*, 396(6712):683–7.

Kerr, D. I. and Ong, J. (1995). GABAB receptors. *Pharmacology & therapeutics*, 67(2):187–246.

Kocher, O., Cheresh, P., Brown, L., and Lee, S. (1995). Identification of a novel gene, selectively up-regulated in human carcinomas, using the differential display technique. *Clin. Cancer Res.*, 1(10):1209–1215.

Kocher, O., Cheresh, P., and Lee, S. W. (1996). Identification and partial characterization of a novel membrane-associated protein (MAP17) up-regulated in human carcinomas and modulating cell replication and tumor growth. *The American journal of pathology*, 149(2):493–500.

Kramer, H., Luke, A., Bidani, A., Cao, G., Cooper, R., and McGee, D. (2005). Obesity and Prevalent and Incident CKD: The Hypertension Detection and Follow-Up Program. *American Journal of Kidney Diseases*, 46(4):587–594.

Krupp, M., Marquardt, J. U., Sahin, U., Galle, P. R., Castle, J., and Teufel, A. (2012). RNA-Seq Atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics (Oxford, England)*, 28(8):1184–5.

Kubiak, M., Januszko-Giergielewicz, B., Moczulska, B., and Gromadziński, L. (2014). Hypertensive nephropathy – A yet unsolved problem. *Polish Annals of Medicine*, 21(2):147–151.

Kurts, C., Panzer, U., Anders, H.-J., and Rees, A. J. (2013). The immune system and kidney disease: basic concepts and clinical implications. *Nature Publishing Group*, 13.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13057–62.

Lê Cao, K.-A., Rohart, F., Mchugh, L., Korn, O., and Wells, C. A. (2014). YuGene: A simple approach to scale gene expression data derived from different platforms for integrated analyses. *Genomics*, 103:239–251.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2016). *sva: Surrogate Variable Analysis. R package version 3.22.0*.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):1724–1735.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595.

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–8.

Li, L., Darden, T. A., Weinberg, C. R., Levine, A. J., and Pedersen, L. G. (2001). Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinatorial chemistry & high throughput screening*, 4(8):727–39.

Li, Q., Birkbak, N. J., Gyorffy, B., Szallasi, Z., and Eklund, A. C. (2011). Jetset: selecting the optimal microarray probe set to represent a gene. *BMC bioinformatics*, 12(1):474.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714.

Lindenmeyer, M. T., Eichinger, F., Sen, K., Anders, H.-J., Edenhofer, I., Mattinzoli, D., Kretzler, M., Rastaldi, M. P., and Cohen, C. D. (2010). Systematic Analysis of a Novel Human Renal Glomerulus-Enriched Gene Expression Dataset. *PLoS ONE*, 5(7):e11545.

Liu, X., Yu, X., Zack, D. J., Zhu, H., and Qian, J. (2008). TiGER: a database for tissue-specific gene expression and regulation. *BMC bioinformatics*, 9(1):271.

Markovich, D. (2014). Na+-sulfate cotransporter SLC13A1. *Pflugers Archiv : European journal of physiology*, 466(1):131–7.

Martini, S., Nair, V., Keller, B. J., Eichinger, F., Hawkins, J. J., Randolph, A., Böger, C. A., Gadegbeku, C. A., Fox, C. S., Cohen, C. D., Kretzler, M., cDNA European Renal cDNA Bank, t. E. R., C-PROBE Cohort, C.-P., and CKDGen Consortium, C. (2014). Integrative biology identifies shared transcriptional networks in CKD. *Journal of the American Society of Nephrology : JASN*, 25(11):2559–72.

Matovinović, M. S. (2009). Pathophysiology and classification of kidney. *eJIFCC*, 20(1):1–10.

McMahon, A. P. (2016). Development of the Mammalian Kidney. *Current topics in developmental biology*, 117:31–64.

Merikallio, H., Pääkkö, P., Harju, T., and Soini, Y. (2011). Claudins 10 and 18 are predominantly expressed in lung adenocarcinomas and in tumors of nonsmokers. *International journal of clinical and experimental pathology*, 4(7):667–73.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2017). Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

Minoo, P., Su, G., Drum, H., Bringas, P., and Kimura, S. (1999). Defects in tracheoesophageal and lung morphogenesis in Nkx2.1(-/-) mouse embryos. *Developmental biology*, 209(1):60–71.

Möller, C. C., Wei, C., Altintas, M. M., Li, J., Greka, A., Ohse, T., Pippin, J. W., Rastaldi, M. P., Wawersik, S., Schiavi, S., Henger, A., Kretzler, M., Shankland, S. J., and Reiser, J. (2007). Induction of TRPC6 Channel in Acquired Forms of Proteinuric Kidney Disease. *J Am Soc Nephrol*, 18:29–36.

Morel, F., Surla, A., and Vignais, P. V. (1992). Purification of human placenta diamine oxidase. *Biochemical and biophysical research communications*, 187(1):178–86.

Müller, F.-J., Schuldt, B. M., Williams, R., Mason, D., Altun, G., Papapetrou, E. P., Danner, S., Goldmann, J. E., Herbst, A., Schmidt, N. O., Aldenhoff, J. B., Laurent, L. C., and Loring, J. F. (2011). A bioinformatic assay for pluripotency in human cells. *Nature methods*, 8(4):315–7.

Myers, M. W., Lazzarini, R. A., Lee, V. M., Schlaepfer, W. W., and Nelson, D. L. (1987). The human mid-size neurofilament subunit: a repeated protein sequence and the relationship of its gene to the intermediate filament gene family. *The EMBO journal*, 6(6):1617–26.

Narkiewicz, K. (2006). Obesity and hypertension–the issue is more complex than we thought. *Nephrology Dialysis Transplantation*, 21(2):264–267.

Nielsen, K. L., Høgh, A. L., Emmersen, J., M.J., J., M.O., B., H.B., K., P., H., N., B., A.M., S., and S.M., S. (2006). DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Research*, 34(19):e133–e133.

Nishihara, S., Iwasaki, H., Nakajima, K., Togayachi, A., Ikehara, Y., Kudo, T., Kushi, Y., Furuya, A., Shitara, K., and Narimatsu, H. (2003). Alpha1,3-fucosyltransferase IX (Fut9) determines Lewis X expression in brain. *Glycobiology*, 13(6):445–55.

Ong, A. C. and Fine, L. G. (1994). Loss of glomerular function and tubulointerstitial fibrosis:Cause or effect? *Kidney International*, 45(2):345–351.

Pan, J. B., Hu, S. C., Shi, D., Cai, M. C., Li, Y. B., Zou, Q., and Ji, Z. L. (2013). PaGenBase: A pattern gene database for the global and dynamic understanding of gene function. *PLoS ONE*, 8(12):e80747.

Pang, A., Hu, Y., Zhou, P., Long, G., Tian, X., Men, L., Shen, Y., Liu, Y., and Cui, Y. (2015). Corin is down-regulated and exerts cardioprotective action via activating pro-atrial natriuretic peptide pathway in diabetic cardiomyopathy. *Cardiovascular diabetology*, 14(1):134.

Peters, H. P. E., Waanders, F., Meijer, E., van den Brand, J., Steenbergen, E. J., van Goor, H., and Wetzels, J. F. M. (2011). High urinary excretion of kidney injury molecule-1 is an independent predictor of end-stage renal disease in patients with IgA nephropathy. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association*, 26(11):3581–8.

Poch, M. T., Cutler, N. S., Yost, G. S., and Hines, R. N. (2005). MOLECULAR MECHANISMS REGULATING HUMAN CYP4B1 LUNG-SELECTIVE EXPRESSION. *Drug Metabolism and Disposition*, 33(8).

R Development Core Team (2016). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria*, 0:{ISBN} 3–900051–07–0.

Ramond, F., Janin, A., Di Filippo, S., Chanavat, V., Chalabreysse, L., Roux-Buisson, N., Sanlaville, D., Touraine, R., and Millat, G. (2017). Homozygous PKP2 deletion associated with neonatal left ventricle noncompaction. *Clinical genetics*, 91(1):126–130.

Reddy, P. H., Mani, G., Park, B. S., Jacques, J., Murdoch, G., Whetsell, W., Kaye, J., and Manczak, M. (2005). Differential loss of synaptic proteins in Alzheimer's disease: implications for synaptic dysfunction. *Journal of Alzheimer's disease : JAD*, 7(2):103–17; discussion 173–80.

Reich, H. N., Tritchler, D., Cattran, D. C., Herzenberg, A. M., Eichinger, F., Boucherot, A., Henger, A., Berthier, C. C., Nair, V., Cohen, C. D., Scholey, J. W., and Kretzler, M. (2010). A Molecular Signature of Proteinuria in Glomerulonephritis. *PLoS ONE*, 5(10):e13451.

Reverter, A., Ingham, A., and Dalrymple, B. P. (2008). Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. *BioData mining*, 1:8.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40.

Roost, M. S., Van Iperen, L., Ariyurek, Y., Buermans, H. P., Arindrarto, W., Devalla, H. D., Passier, R., Mummery, C. L., Carlotti, F., De Koning, E. J. P., Van Zwet, E. W., Goeman, J. J., and Chuva De Sousa Lopes, S. M. (2015). KeyGenes, a Tool to Probe Tissue Differentiation Using a Human Fetal Transcriptional Atlas. *Stem Cell Reports*, 4(6):1112–1124.

Roth, R., Hevezi, P., Lee, J., Willhite, D., Lechner, S., Foster, A., and Zlotnik, A. (2006a). Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, 7(2):67–80.

Roth, R. B., Hevezi, P., Lee, J., Willhite, D., Lechner, S. M., Foster, A. C., and Zlotnik, A. (2006b). Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, 7(2):67–80.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., Kurbatova, N., Malone, J., Mani, R., Mupo, A., Pedro Pereira, R., Pilicheva, E., Rung, J., Sharma, A., Tang, Y. A., Ternent, T., Tikhonov, A., Welter, D., Williams, E., Brazma, A., Parkinson, H., and Sarkans, U. (2013). ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Research*, 41(D1):D987–D990.

Sagrinati, C., Netti, G. S., Mazzinghi, B., Lazzeri, E., Liotta, F., Frosali, F., Ronconi, E., Meini, C., Gacci, M., Squecco, R., Carini, M., Gesualdo, L., Francini, F., Maggi, E., Annunziato, F., Lasagni, L., Serio, M., Romagnani, S., and Romagnani, P. (2006). Isolation and characterization of multipotent progenitor cells from the Bowman's capsule of adult human kidneys. *Journal of the American Society of Nephrology : JASN*, 17(9):2443–56.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235).

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels : support vector machines, regularization, optimization, and beyond.* Adaptive computation and machine learning. MIT Press.

Schulze, A. and Downward, J. (2001). Navigating gene expression using microarrays – a technology review. *Nature Cell Biology*, 3(8):E190–E195.

Schwelberger, H. G. and Bodner, E. (1998). Identity of the diamine oxidase proteins in porcine kidney and intestine. *Inflammation research : official journal of the European Histamine Research Society ... [et al.]*, 47 Suppl 1:S58–9.

Serafini-Cessi, F., Malagolini, N., Hoops, T. C., and Rindler, M. J. (1993). Biosynthesis and oligosaccharide processing of human Tamm-Horsfall glycoprotein permanently expressed in HeLa cells. *Biochemical and biophysical research communications*, 194(2):784–90.

Shen, S. S., Krishna, B., Chirala, R., Amato, R. J., and Truong, L. D. (2005). Kidney-specific cadherin, a specific marker for the distal portion of the nephron and related renal neoplasms. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 18(7):933–40.

Shimobaba, S., Taga, S., Akizuki, R., Hichino, A., Endo, S., Matsunaga, T., Watanabe, R., Yamaguchi, M., Yamazaki, Y., Sugatani, J., and Ikari, A. (2016). Claudin-18 inhibits cell proliferation and motility mediated by inhibition of phosphorylation of PDK1 and Akt in human lung adenocarcinoma A549 cells. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1863(6):1170–1178.

Smith, A. D., Xuan, Z., and Zhang, M. Q. (2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, 9(1):128.

Soini, Y. (2005). Expression of claudins 1, 2, 3, 4, 5 and 7 in various types of tumours. *Histopathology*, 46(5):551–60.

Spellmann, I., Müller, N., Musil, R., Zill, P., Douhet, A., Dehning, S., Cerovecki, A., Bondy, B., Möller, H.-J., and Riedel, M. (2008). Associations of SNAP-25 polymorphisms with cognitive dysfunctions in Caucasian patients with schizophrenia during a brief trail of treatment with atypical antipsychotics. *European archives of psychiatry and clinical neuroscience*, 258(6):335–44.

Stachelscheid, H., Seltmann, S., Lekschas, F., Fontaine, J.-F., Mah, N., Neves, M., Andrade-Navarro, M. A., Leser, U., and Kurtz, A. (2014). CellFinder: a cell data repository. *Nucleic acids research*, 42(Database issue):D950–8.

Steele, M. P., Luna, L. G., Coldren, C. D., , Murphy, E., , Hennessy, C. E., , Heinz, D., , Evans, C. M., , Groshong, S., , Cool, C., , Cosgrove, G. P., , Brown, K. K., , Fingerlin, T. E., , Schwarz, M. I., , Schwartz, D. A., , and Yang, I. V. (2015a). Relationship between gene expression and lung function in Idiopathic Interstitial Pneumonias. *BMC Genomics*, 16(1):869.

Steele, M. P., Luna, L. G., Coldren, C. D., Murphy, E., Hennessy, C. E., Heinz, D., Evans, C. M., Groshong, S., Cool, C., Cosgrove, G. P., Brown, K. K., Fingerlin, T. E., Schwarz, M. I., Schwartz, D. A., and Yang, I. V. (2015b). Relationship between gene expression and lung function in Idiopathic Interstitial Pneumonias. *BMC genomics*, 16(1):869.

Stefan, N. (2006). 2-Heremans-Schmid Glycoprotein/ Fetuin-A Is Associated With Insulin Resistance and Fat Accumulation in the Liver in Humans. *Diabetes Care*, 29(4):853–857.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101:6062–6067.

Takasato, M., Er, P. X., Chiu, H. S., Maier, B., Baillie, G. J., Ferguson, C., Parton, R. G., Wolvetang, E. J., Roost, M. S., Chuva de Sousa Lopes, S. M., and Little, M. H. (2015). Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. *Nature*, 526(7574):564–568.

Thomson, R. B. and Aronson, P. S. (1999). Immunolocalization of Ksp-cadherin in the adult and developing rabbit kidney. *The American journal of physiology*, 277(1 Pt 2):F146–56.

Thomson, R. B., Igarashi, P., Biemesderfer, D., Kim, R., Abu-Alfa, A., Soleimani, M., and Aronson, P. S. (1995). Isolation and cDNA cloning of Ksp-cadherin, a novel kidney-specific member of the cadherin multigene family. *The Journal of biological chemistry*, 270(29):17594–601.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–78.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., Feilitzen, K. V., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., Heijne, G. V., Nielsen, J., and Pontén, F. (2015). Tissue-based map of the human proteome. *Science (New York, N.Y.)*, 347(6220):1260419–1260419.

van Spaendonck-Zwarts, K. Y., Posafalvi, A., van den Berg, M. P., Hilfiker-Kleiner, D., Bollen, I. A. E., Sliwa, K., Alders, M., Almomani, R., van Langen, I. M., van der Meer, P., Sinke, R. J., van der Velden, J., Van Veldhuisen, D. J., van Tintelen, J. P., and Jongbloed, J. D. H. (2014). Titin gene mutations are common in families with both peripartum cardiomyopathy and dilated cardiomyopathy. *European heart journal*, 35(32):2165–73.

Virgil Brown, W. and Baginsky, M. (1972). Inhibition of lipoprotein lipase by an apoprotein of human very low density lipoprotein. *Biochemical and Biophysical Research Communications*, 46(2):375–382.

Vojdeman, F. J., Herman, S. E. M., Kirkby, N., Wiestner, A., van t' Veer, M. B., Tjønnfjord, G. E., Itälä-Remes, M. A., Kimby, E., Farooqui, M. Z., Polliack, A., Wu, K. L., Doorduijn, J. K., Alemayehu, W. G., Wittebol, S., Kozak, T., Walewski, J., Abrahamse-Testroote, M. C. J., van Oers, M. H. J., Geisler, C. H., and Niemann, C. U. (2017). Soluble CD52 is an indicator of disease activity in chronic lymphocytic leukemia. *Leukemia & Lymphoma*, 58(10):2356–2362.

Waldron, L., Steimle, J. D., Greco, T. M., Gomez, N. C., Dorr, K. M., Kweon, J., Temple, B., Yang, X. H., Wilczewski, C. M., Davis, I. J., Cristea, I. M., Moskowitz, I. P., and Conlon, F. L. (2016). The Cardiac TBX5 Interactome Reveals a Chromatin Remodeling Network Essential for Cardiac Septation. *Developmental cell*, 36(3):262–75.

Wang, X.-M., Li, J., Yan, M.-X., Liu, L., Jia, D.-S., Geng, Q., Lin, H.-C., He, X.-H., Li, J.-J., and Yao, M. (2013). Integrative analyses identify osteopontin, LAMB3 and ITGB1 as critical pro-metastatic genes for lung cancer. *PloS one*, 8(2):e55714.

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

Watson, J. and Crick, F. C. (1953). MOLECULAR STRUCTURE OF NUCLEIC ACIDS: A Structure for Deoxyribose Nucleic Acid.

Wei, R.-R., Zhang, M.-Y., Rao, H.-L., Pu, H.-Y., Zhang, H.-Z., and Wang, H.-Y. (2012). Identification of ADH4 as a novel and potential prognostic marker in hepatocellular carcinoma. *Medical Oncology*, 29(4):2737–2743.

Wheelan, S. J., Martínez Murillo, F., and Boeke, J. D. (2008). The incredible shrinking world of DNA microarrays. *Molecular bioSystems*, 4(7):726–32.

Willemsen, M. H., Ba, W., Wissink-Lindhout, W. M., de Brouwer, A. P. M., Haas, S. A., Bienek, M., Hu, H., Vissers, L. E. L. M., van Bokhoven, H., Kalscheuer, V., Nadif Kasri, N., and Kleefstra, T. (2014). Involvement of the kinesin family members KIF4A and KIF5C in intellectual disability and synaptic function. *Journal of medical genetics*, 51(7):487–94.

Windler, E. and Havel, R. (1985). Inhibitory effects of C apolipoproteins from rats and humans on the uptake of triglyceride-rich lipoproteins and their remnants by the perfused rat liver. *J. Lipid Res.*, 26(5):556–565.

Wittenberg, J. B. (2003). Myoglobin function reassessed. *Journal of Experimental Biology*, 206(12):2011–2020.

Wrzesiński, T., Szelag, M., Cieślikowski, W. a., Ida, A., Giles, R., Zodro, E., Szumska, J., Poźniak, J., Kwias, Z., Bluyssen, H. a. R., and Wesoly, J. (2015). Expression of pre-selected TMEMs with predicted ER localization as potential classifiers of ccRCC tumors. *BMC cancer*, 15(2015):518.

Yatabe, Y. M., Mitsudomi, T. M., and Takahashi, T. M. (2002). TTF-1 expression in pulmonary adenocarcinomas. *The American journal of surgical pathology*, 26(6):767–73.

Yu, A. S. L., Enck, A. H., Lencer, W. I., and Schneeberger, E. E. (2003). Claudin-8 expression in Madin-Darby canine kidney cells augments the paracellular barrier to cation permeation. *The Journal of biological chemistry*, 278(19):17350–9.

Zanger, U. M. and Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & therapeutics*, 138(1):103–41.

Zhang, X., Zhang, Q.-Y., Liu, D., Su, T., Weng, Y., Ling, G., Chen, Y., Gu, J., Schilling, B., and Ding, X. (2005). Expression of cytochrome p450 and other biotransformation genes in fetal and adult human nasal mucosa. *Drug metabolism and disposition: the biological fate of chemicals*, 33(10):1423–8.

Zhang, Z., Liang, X., Gao, L., Ma, H., Liu, X., Pan, Y., Yan, W., Shan, H., Wang, Z., Chen, Y. H., and Ma, C. (2015). TIPE1 induces apoptosis by negatively regulating Rac1 activation in hepatocellular carcinoma cells. *Oncogene*, 34(20):2566–2574.

Zheng, H., Yang, S., Yang, Y., Yuan, S.-X., Wu, F.-Q., Wang, L.-L., Yan, H.-L., Sun, S.-H., and Zhou, W.-P. (2015). Epigenetically silenced long noncoding-SRHC promotes proliferation of hepatocellular carcinoma. *Journal of Cancer Research and Clinical Oncology*, 141(7):1195–1203.

Zhou, X., Popescu, N. C., Klein, G., and Imreh, S. (2007). The interferon-α responsive gene TMEM7 suppresses cell proliferation and is downregulated in human hepatocellular carcinoma. *Cancer Genetics and Cytogenetics*, 177(1):6–15.

Lebenslauf

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten