Aalto University
School of Science
Master's Degree Programme in Computational and Systems Biology

Karolis Uziela

Making microarray and RNA-seq gene expression data comparable

Master's Thesis
Espoo, June 30, 2012

Supervisors: Professor Juho Rousu, Aalto University
Professor Erik Aurell, KTH Royal Institute of Technology
Instructor: Antti Honkela D.Sc. (Tech.)

Aalto University
School of Science
Master's Degree Programme in Computational and Systems Biology

ABSTRACT OF MASTER'S THESIS

Author:	Karolis Uziela		
Title:	Making microarray and RNA-seq gene expression data comparable		
Date:	June 30, 2012	**Pages:**	vi + 61
Professorship:	Information and Computer Science	**Code:**	T-61
Supervisors:	Professor Juho Rousu Professor Erik Aurell		
Instructor:	Antti Honkela D.Sc. (Tech.)		

Measuring gene expression levels in the cell is an important tool in biomedical sciences. It can be used in new drug development, disease diagnostics and many other areas. Currently, two most popular platforms for measuring gene expression are microarrays and RNA-sequencing (RNA-seq). Making the gene expression results more comparable between these two platforms is an important topic which has not yet been investigated enough.

In this thesis, we present a novel method, called PREBS, that addresses this issue. Our method adjusts RNA-seq data computational processing in a way that makes the resulting gene expression measures more similar to microarray-based gene expression measures. We compare our method against two other RNA-seq processing methods, RPKM and MMSEQ, and evaluate each method's agreement with microarrays by calculating correlations between the platforms. We show that our method reaches the highest level of agreement among all of the methods in absolute expression scale and has a similar level of agreement as the other methods in differential expression scale.

Additionally, this thesis provides some background on gene expression, its measurement and computational analysis of gene expression data. Moreover, it gives a brief literature review on the past microarray–RNA-seq comparisons.

Keywords:	Microarray, RNA-seq, gene expression, bioinformatics
Language:	English

Acknowledgements

I would like to thank all of my friends and family for supporting me during the difficult time of writing the thesis. Moreover, I would like to thank my supervisors Prof. Juho Rousu and Prof. Erik Aurell, and especially my instructor Dr Antti Honkela for useful advice and comments regarding the thesis. Last but not least, I would like to thank my friends Andrew Keating and Maia Malonzo for keeping me a company during the lunch and coffee breaks.

Espoo, June 30, 2012

Karolis Uziela

Abbreviations and Acronyms

CDF	Chip Description File
DCPM	Average depth of coverage per million reads
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase of exon per Million fragments mapped
HIV	Human Immunodeficiency Virus
LIMMA	Linear Models for Microarray Data
Log2FC	Log_2 Fold Change
miRNA	MicroRNA
MM probe	Mismatch probe
MPSS	Massively Parallel Signature Sequencing
mRNA	Messenger RNA
PCR	Polymerase chain reaction
PM probe	Perfect match probe
PREBS	Probe Region Expression Based on Sequencing
qPCR (qRT-PCR)	Quantitative real time PCR
RMA	Robust Multichip Average
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
RPKM	Reads Per Kilobase of exon per Million mapped reads
rRNA	Ribosomal RNA
SAGE	Serial Analysis of Gene Expression
siRNA	Small interfering RNA
SNP	Single-nucleotide polymorphism
TAR	Transcriptionally Active Region
tRNA	Transfer RNA

Contents

Chapter 1

Introduction

1.1 Problem setting

Gene expression is a fundamental process in the cell during which the DNA is transcribed to the corresponding RNA and the RNA is translated to the corresponding protein. Gene expression levels can differ between different cells, tissues or points in time. Measuring gene expression has proven to be a very important tool in biomedical sciences, because it can be used for disease diagnostics, search for new drug targets and for analysis of diseases, such as cancer, Alzheimer's disease, schizophrenia and HIV infection [1]. Therefore, gene expression measurement has been of great interest to scientists and many gene expression measurement methods have been developed.

Nowadays, two most popular gene expression measurement platforms are RNA-seq (RNA-sequencing) and microarrays. RNA-seq is a newer and more accurate technology [2, 3], but microarrays are still popular because they are cheaper and have a well established infrastructure [4]. On the other hand, it is predicted that in the future RNA-seq might fully replace microarrays [5]. This is also supported by Table 1.1 which shows that the number of RNA-seq experiments is rapidly increasing while the number of microarray experiments is slowly decreasing. That shows that the scientists are increasingly turning towards the newer gene expression measurement technology. However, even if the RNA-seq technology fully replaces microarrays we will still want to use the existing microarray data as a reference. There is a huge number of microarray experiments available in gene expression databases, such as ArrayExpress [6] or GEO [7], and therefore it is important to be able to compare the newly conducted RNA-seq experiments with existing microarray data in a meaningful way.

In the past, there have been many experimental RNA-seq–microarray

	2010	**2011**	**2012**
Microarray	5796	5196	4724
RNA-seq	147	256	564

Table 1.1: Number of microarray and RNA-seq experiments in ArrayExpress [6] database in the last three years. Data for the year 2012 is extrapolated based on the date when the query was made (June 25, 2012).

comparisons (they will be reviewed in Chapter 3). However, none of these comparisons tried to make RNA-seq and microarray data more similar, or, in other words, more comparable, by adjusting the computational processing of the data. In this thesis, we describe a method called PREBS that processes RNA-seq data in a way that the results become more similar to the microarray gene expression results. We evaluate our method by calculating gene expression correlations with microarrays and compare it against two other RNA-seq processing methods—RPKM [8] and MMSEQ [9]. We show that our method reaches the highest level of agreement with microarrays in absolute expression scale and a similar level of agreement as the other methods in differential expression scale.

1.2 Structure of the thesis

The thesis is organized as follows. In Chapter 2, we provide the background on gene expression, its measurement tools and analysis of the data. Two gene expression measurement tools, microarray and RNA-seq, are described in more detail.

In Chapter 3, we review past studies which were comparing or trying to combine/visualize inter-platform gene expression data. This review helps to get an understanding of the related work that has already been done.

In Chapter 4, we introduce our method and explain how it works. Additionally, we list all of the tools that were used for the implementation.

In Chapter 5, we compare our method against two other RNA-seq processing methods: RPKM and MMSEQ. We demonstrate that our method agrees better with microarrays than the other two methods.

Finally, Chapter 6 concludes our work. Section 6.1 summarizes the work and Section 6.2 discusses the results and gives suggestions about possible future work.

Chapter 2

Gene expression and its measurement

In this chapter, we will review the necessary background that will be required to understand the rest of the thesis. In Section 2.1, we will explain biological background of gene expression and, in Section 2.2, we will give a brief overview of its measurement methods and data analysis. Additionally, in Sections 2.3 and 2.4, we will look at the two currently most popular gene expression measurement tools, microarrays and RNA-seq, in more detail.

2.1 Biology of gene expression

All of the genetic information of the cell is encoded into DNA, a long double-stranded helical molecule composed of nucleotides. These nucleotides differ among themselves because of their side chains which are also called *bases*. In DNA there are four different bases: adenine (A), guanine (G), cytosine (C) and thymine (T). The sequence of these four bases determines the genetic information encoded into DNA.

In order to make use of the genetic information, DNA first has to be transcribed to RNA. RNA chemical structure is very similar to DNA, except that instead of thymine (T) base it has uracil (U) base. RNA can sometimes be the final *gene product*, as it is in case of tRNA, rRNA, miRNA or siRNA. However, most common type of RNA is mRNA (messenger RNA) that is translated to proteins, molecules which are responsible for the most of the functions of the cell. The whole process of transcribing the DNA to the RNA and translating the RNA to the protein is called *gene expression* [10] (see Figure 2.1).

A single DNA strand can be viewed as a long string of letters A, C, G,

Figure 2.1: Basic scheme of gene expression. The DNA is first transcribed to RNA and then the RNA is translated into a protein

T, each of which stand for the base of a particular nucleotide. Not all of the DNA molecule is transcribed to RNA in a single transcription process, but only a small part of it. The substring of DNA which codes for a single RNA molecule is called *a gene*. During the process of transcription an enzyme called *RNA polymerase* reads the gene portion of the DNA and transcribes it to an RNA with *complementary* sequence. That is, adenine in the DNA is replaced with uracil in an RNA, cytosine is replaced with guanine, guanine is replaced with cytosine and thymine is replaced with adenine.

Before the mRNA can be translated into the protein, it undergoes *post-transcriptional modifications*. During this phase, 5' cap and 3' poly-A tail are added to the mRNA which protect the mRNA from degradation. Moreover, during *splicing* process the non-coding parts of mRNA, *introns*, are removed and coding parts, *exons*, are joined together. The exons can be often joined in a number of alternative ways, giving rise to different versions of the same gene—*gene isoforms*. RNA splicing and other post-transcriptional modifications occur only in eukaryotes, but not prokaryotes.

Next, during the translation phase the mRNA is translated into a protein. A protein is also a long macro-molecule like DNA and RNA, except that its subunits are not nucleotides, but amino acids. In most of organisms proteins are composed of 20 types of amino acids. Some organisms might, however, include two additional types of amino acids: selenocysteine and pyrrolysine [11].

Every three nucleotides in the mRNA code for one amino acid in a protein. During the translation phase, every three nucleotides in the mRNA are read and corresponding amino acid molecule is added to the protein chain which is being synthesized. After the whole protein is synthesized, it undergoes some more post-processing steps and folds into the correct shape. Then, it is transported to the appropriate place in the cell and can perform its function.

Gene expression does not happen for all of the genes at the same time. If a gene is expressed at a particular point of time it is said to be *active*, otherwise, it is said to be *passive*. The rate of gene expression can also be different for different genes. If a lot of gene product is produced for a gene, the gene is said to have a *high expression level*, if only a little product is produced, the gene is said to have a *low expression level*.

2.2 Gene expression measurement and data analysis

In this section, we will review the available tools for gene expression measurement. Moreover, we will provide an outline of the basic steps in gene expression analysis.

2.2.1 Gene expression measurement methods

Gene expression is defined as the conversion of genetic information into the actual protein [10]. Therefore, intuitively, *gene expression level* can be thought of as the amount of the corresponding protein present in the cell. There are some tools, such as Western Blot [12], which measure protein expression levels in the cell. However, mRNA quantification is technically an easier task and it can be used as an approximation of the final gene product—protein [13]. So for the purposes of this thesis we will define gene expression level as the mRNA level and use these two terms interchangeably as it is also done by other scientists [13].

There is a variety of technologies available for quantifying mRNA levels in a cell. These technologies can be divided into two broad categories: hybridization-based and sequencing-based. The difference between the two is that hybridization-based technologies (Northern blot [14], qPCR [15], microarrays [16]) use hybridization probes—short sequences which are complementary to some part of expressed gene sequence. Designing these probes requires prior knowledge about the transcriptome which is being analyzed. On the other hand, sequencing-based technologies (RNA-seq [17], SAGE [18], MPSS [19]) do not use probes for gene expression analysis. There, the principle is to sequence the whole transcriptome and to determine the amount of reads which originate from each of the gene regions and, in this way, to evaluate gene expression levels.

Gene expression measurement technologies can also be divided into low-throughput and high-throughput categories. Low-throughput technologies can be used to analyze only from one up to several genes in the same experiment. On the other hand, high-throughput technologies allow us to analyze whole transcriptome—thousands of genes in the same experiment. All of the mentioned sequencing-based technologies and one hybridization-based technology (microarrays) fall into the high-throughput category, while the rest of the hybridization-based technologies (Northern blot, qPCR) fall into the low-throughput category.

High-throughput technologies are particularly interesting because of the

huge amount of genes that they can interrogate in one experiment. Microarrays used to be a dominant high-throughput gene expression measurement platform, but nowadays RNA-seq is taking over its place [5]. Other sequencing-based high-throughput technologies (SAGE, MPSS) can be considered as older variants of RNA-seq and are rarely used any more.

2.2.2 Analysis of gene expression measurement data

Gene expression data are often visualized as a *gene expression matrix*, a table where rows represent different genes, columns represent different sample conditions and each value represents the gene expression of a specific gene under a specific condition. Sample conditions can correspond to a number of different things, for example, different tissues, different disease states or samples taken at different points of time. Genes in the table can represent either all or some subset of the genes from the organism being analyzed [13, 20].

Gene expression measures inside the matrix can be either in an absolute or a relative scale. In case of an absolute scale, the values in gene expression matrix represent an absolute gene expression measurement in some abstract units, while, in case of a relative scale, the values are gene expression ratios between two conditions. The ratios of gene expression are often more interesting to the scientists, because they show how much expression values differ between two conditions of an experiment. Genes which have statistically significant changes in expression are called *differentially expressed genes* [20].

A natural way to analyze gene expression matrix is either to compare rows or columns of the table. However, for a comparison we need to decide what similarity measure to use. Most commonly used similarity measures are Euclidean distance, Pearson correlation, Spearman correlation or mutual information. It is difficult to tell which similarity measure is the best to use and it can often depend on the type of experiment being conducted [13].

After choosing the similarity measure, there are two major ways to analyze the data: *supervised learning* and *unsupervised learning*. In case of supervised learning, the data rows or columns have to be associated with some known features. It could be, for example, gene functions for rows or disease states for columns. Then some sort of a classifier is built which is trained to predict these features. Some of the popular methods used in gene expression analysis for classification are linear discriminants, decision trees and Support Vector Machines. After building a classifier, it can be used to predict features for new data. Moreover, if a classifier is built with some relatively simple classification rules, it can also be used to infer the underlying biological mechanisms of the system [13].

	Affymetrix	**Agilent**	**Illumina**	**Nimblegen**	**Other**
Count	13832	3091	1715	800	7456

Table 2.1: Number of experiments in ArrayExpress database for each microarray platform as of June 20, 2012

In unsupervised learning the aim is to group objects (genes or samples) with similar properties. Some of the popular clustering methods that are used are K-means, Hierarchical clustering and Self-organizing maps. These methods can be, for example, used to cluster the genes with similar expression patterns in order to identify common transcription-control mechanisms. Clustering genes can also help to infer function for an unknown gene, because genes with similar expression patterns tend to share a similar function [13]. Furthermore, clustering of the samples, can, for example, be used to infer new sub-classes of tumors as it was done by Alizadeh *et al.* [21].

2.3 Microarrays

Microarrays have evolved from Southern Blotting, a technique which is used to identify a specific DNA sequence in DNA samples [22]. The first studies that involved microarrays were published in 1980s, but the real microarrays take-off began with a publication by Fodor *et al.* [23] from Affymax, a company which later changed its name to Affymetrix and became the market leader for microarray technology. Fodor *et al.* described protein and nucleotide microarrays, their uses and design principles. Since then microarray technology became more and more popular and, in the beginning of 21st century, one could hardly find a modern biology or genetics journal which does not mention microarrays [16].

Nowadays, besides Affymetrix, the other popular platforms include Agilent, Illumina and Nimblegen. In order to see what are the approximate market shares of the commercial microarray platforms in academic use, we queried each platform name on ArrayExpress [6] database. The results are displayed in Table 2.1. As the results show, Affymetrix is by far the most popular platform and it takes up more than 50% of the whole market. Therefore, in this thesis most emphasis will be put on the Affymetrix microarray technology description and data analysis.

In this section we will shortly describe the basic principles of microarrays and review different types of technology. In addition, we will give a brief introduction to computational analysis of microarray data.

2.3.1 Technical principles of microarray technology

Typically, a microarray consists of a large number microscopic DNA spots attached to a solid surface. Each of the spots contains many short DNA sequences called *probes*. All of the probes inside one spot have the same sequence. Also, all of the probe sequences are complementary to a part of the gene from an organism which is being analyzed. Different spots on the microarray correspond to different genes, so the more spots a microarray has, the more genes can be interrogated at the same time.

In order to conduct a microarray experiment, one has to extract RNA from the cell and *reverse transcribe* it to cDNA (complementary DNA) (see Figure 2.2). In this step, the *reverse transcriptase* enzyme synthesizes a DNA molecule which has complementary sequence to the given RNA molecule. The reason of doing this is that an RNA molecule is less stable than a DNA molecule, so the molecule conversion helps to improve the stability without losing the genetic information.

In the next step, the cDNA sample has to be fluorescently labeled (other labeling methods are also possible). The labeled sample is poured on top of the microarray and the cDNA molecules hybridize onto the microarray probes based on their complementarity. The spots where cDNA molecules have attached can be identified because of the fluorescent dye. The stronger the signal (fluorescence intensity) is for a spot, the higher is the gene expression level for the gene to which that spot corresponds.

The above described principles hold only for what is called *gene expression microarrays* [1]. The purpose of these microarrays is to measure gene expression (mRNA) levels inside a cell. There are many other kinds of microarrays that can be used for different purposes. For example, *antibody microarrays* [24] can be used to measure protein expression levels, *SNP microarrays* [25] can be used to detect Single Nucleotide Polymorphisms, *tiling microarrays* [26] can be used for *ChIP-on-chip* [27] studies.

Tiling microarrays can also be used for gene expression profiling like gene expression microarrays, but there are some important differences between the technologies. The main difference is that the probe sequences for tiling arrays are not selected from the genes being investigated, but instead they are taken at regular intervals from the whole genome. In this way, there is a possibility for new genes to be identified in the genome places which were previously thought to be non-transcribed. However, tiling microarrays are more expensive and are not as commonly used for gene expression profiling as gene expression microarrays. In this thesis, we will concentrate on analyzing only gene expression microarrays.

Gene expression microarrays can be further divided into two categories:

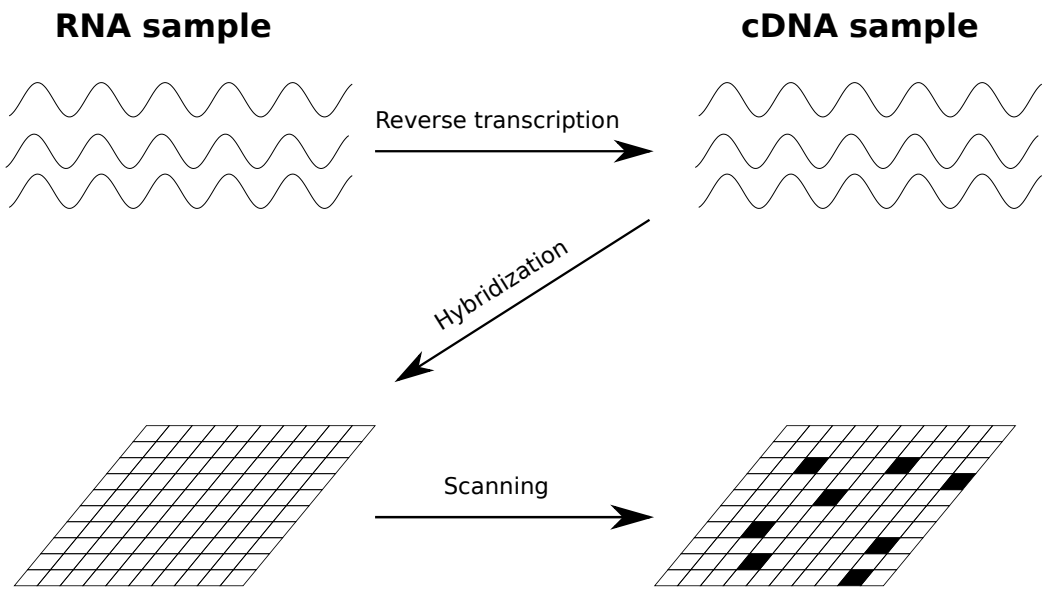


Figure 2.2: Microarray workflow scheme. First, RNA sample is reverse transcribed to cDNA. Then cDNA sample is poured on top of the microarray to allow probes with complementary sequence to hybridize. Finally, the microarray is analyzed by a special machine to determine fluorescent dye intensities (scanning)

spotted microarrays and *oligonucleotide microarrays* [1]. The main difference between the two is that in spotted microarrays the probes usually consist of long cDNA molecules (several hundreds of base pairs) which are prepared beforehand and are printed on the microarray by a robotic arm. In oligonucleotide microarrays, the probes are short oligonucleotides (typically, 25-100 bp long) which are synthesized directly on the microarray plate.

Another important difference lies in the target preparation. In oligonucleotide microarrays, the target is labeled by one fluorescent dye and, in order to identify the differentially expressed genes, the signal intensity is compared between two microarrays. However, the probe preparation procedure for spotted microarrays is not as accurate as for oligonucleotide microarrays and comparison between two separate microarrays would not be possible. Therefore, for spotted microarrays two target samples are labeled by two different fluorescent dyes (typically Cy3 and Cy5) and the differential gene expression is identified by the ratio of two different signals on a single microarray [1].

Spotted microarrays are often termed *in-house* microarrays, because they are prepared in individual labs [28]. For this type of microarrays, the researchers conducting an experiment can decide which genes they want to

interrogate and design a microarray suiting their needs. On the other hand, oligonucleotide microarrays are usually called *commercial microarrays*, because they are mass-produced by industrial companies. These microarrays usually aim to interrogate all known genes in the target organism transcriptome. Spotted microarrays were more popular in the past, because they were cheaper and more customizable. Nowadays, however, spotted microarrays are rarely used any more, because oligonucleotide microarrays have become more affordable and provide more accurate results [16]. Also, the design of oligonucleotide microarrays became customizable via availability of custom oligonucleotide arrays [29]. In this thesis we will analyze only oligonucleotide microarrays.

2.3.2 Computational analysis of microarray data

Computational analysis of microarray data largely depends on the type of microarray platform being used. Since, as we mentioned before, Affymetrix is, at the moment, the dominant microarray platform, we will discuss the data analysis only for this platform.

Affymetrix microarrays consist of many 25 nucleotides long probes attached to a solid surface [30]. There are of two types of probes: perfect match (PM) and mismatch (MM). Perfect match probes are perfectly complementary to some part of a target gene which they are interrogating. Mismatch probes, on the other hand, have the same sequence as perfect match probes, but the middle nucleotide (13th) is changed. Some microarray methods use mismatch probes to account for non-specific probe binding while other methods simply ignore the intensity values of the mismatch probe hybridization.

The probes are grouped into probe sets by the manufacturer. Typically, a probe set contains 15–20 PM/MM probe pairs and it interrogates one target gene. The information about the probe sets is available in so called *Chip Description Files* (CDFs). However, transcriptome annotations change over time and some of the probes might appear to be mapping not to one, but to several genes or do not correspond to any genes at all. The manufacturer's CDFs are rarely updated, therefore, many people prefer to use custom CDFs [31] which account for the latest changes in the transcriptome annotation. In these files, probes that do not map uniquely to a single gene are filtered out and the rest of them are regrouped to new probe sets each of which corresponds to a single gene in the latest transcriptome annotation.

In general, probes belonging to one probe set can give different hybridization intensity values in an experiment. Furthermore, the microarray experiments are often done with replicates—the same experiment is done several times to reduce variability. The first step in microarray computational anal-

ysis is to summarize probe intensities from all probe sets and all microarray replicates to get a single gene expression measure for each gene. Many algorithms have been developed for this procedure, most popular of them being RMA [30] and MAS5 [32]. RMA is open source software, while MAS5 is Affymetrix proprietary method.

Both of these algorithms use some background-correction, normalization and summarization procedures. During the background-correction phase the probe intensity values are adjusted to account for technical noise, then, during normalization phase, probe intensity values are normalized to be comparable across different microarrays and, finally, during summarization phase, probe intensities are converted to gene expression measures.

The main difference between RMA and MAS5 is that RMA uses only values from PM probes while MAS5 uses values from both PM and MM probes. In MAS5 algorithm, MM values are subtracted from PM values to account for non-specific binding while RMA algorithm simply ignores MM values. Another important difference is that MAS5 normalizes every microarray independently while RMA normalizes all microarrays at the same time. Nowadays, RMA algorithm is more often used than MAS5, therefore, for the rest of the thesis we will restrict ourselves only on analyzing results with RMA.

Further analysis of microarray data often include identification of differentially expressed genes. There are many tools for this task, but one of the most popular of such tools is LIMMA [33]. Among other results, LIMMA outputs log_2 *fold change* values. These values show how much gene expression has changed from one condition to another. They are determined, by calculating the ratio of absolute gene expression measures in the two samples and then taking base 2 logarithm of the ratio. In addition to that, LIMMA uses moderated t-test to calculate p-values. A p-value is a probability, assuming that the null hypothesis is true, to obtain a statistic value at least as inconsistent with the null hypothesis as the one observed [34]. In our case, the null hypothesis says that the gene is not differentially expressed, therefore, p-value measures statistical significance of differentially expressed genes. LIMMA also outputs *adjusted p-values* which are more commonly known as *False Discovery Rate* (FDR) values. False Discovery Rate is a technique which is used to account for multiple testing problem and it denotes the percentage of false positives among the significant hypotheses [35].

2.4 RNA-sequencing

For over a decade, microarrays were the dominant platform in the high-throughput analysis of gene expression [36]. Sequencing-based methods, such as SAGE or MPSS, used to be the major alternative methods. One of the advantages of these methods is that they provided precise digital gene expression measures instead of analog expression measures provided by microarrays[1]. These methods, however, were based on a conventional Sanger sequencing, and they were not as efficient as later developed methods based on *next-generation sequencing* [17].

There are several different technologies which correspond to the next-generation sequencing, but all of them have one thing in common—sequencing is done via massive parallelization. At the moment, three most popular next-generation sequencing technologies are Roche/454, Illumina and AB SOLiD [36]. Their ability to sequence transcriptome cost-effectively and in a high depth gave birth to a new technology for gene expression measurement—RNA-sequencing (RNA-seq) [17].

In this section, we review the basic principles of the next-generation sequencing technologies and the steps needed to take in order to conduct an RNA-seq experiment. Also, we give a brief overview of the computational methods involved in RNA-seq data analysis.

2.4.1 Technical principles of RNA-seq technology

The first key step in the next-generation sequencing is sample preparation. The procedure varies from technology to technology, but the basic principles remain the same: coding RNA (mRNA) has to be separated from the rest of the sample, reverse transcribed, fragmented and amplified [17]. For separation purposes, poly-A tail of the mRNA is often targeted by poly-T oligonucleotides attached to a given substrate. Next, the mRNA is reverse-transcribed to cDNA and fragmented into sizes required by the specific protocol. The amplification can be carried out in a few different ways: 454 and SOLiD use *emulsion PCR* [39] while Illumina uses *bridge amplification* [40]. The end result for any sample preparation is the same: a number of short single-stranded cDNA molecules separated into clusters or microscopic wells on a plate and ready to be sequenced [41].

The next-generation sequencing technology which was developed the first

[1]Some of the studies claim that another advantage of SAGE is the ability to detect novel transcripts [37, 38]. However, one of the studies [8] explicitly state that SAGE is not able to detect novel transcripts which is a little bit confusing.

is Roche/454 [41, 42]. This sequencing method is based on *sequencing by synthesis* methodology. Sequencing is done by synthesizing a complementary DNA strand for each of the oligonucleotides on a plate. During the sequencing process, all four types of nucleotides (A, G, C, T) are added to the sequencing reaction sequentially, one at a time. If some particular DNA strand can be extended by the added nucleotide based on the complementarity principle, a DNA polymerase adds that nucleotide to the DNA strand being synthesized. This causes a special reaction where an inorganic phosphate ion is released and a flash of light is observed. In case of repetitive sequence, several nucleotides are added during one cycle and the intensity of the light being released becomes proportionally stronger. Light intensities and positions on a plate are captured by a monitor. Unincorporated nucleotides are washed away and the sequencing reaction becomes ready for the next cycle. This reaction is repeated many times, until all of the oligonucleotide sequences are determined.

Illumina uses a similar approach as 454 which is also based on sequencing by synthesis [41, 40]. The main difference is that following Illumina technology, all four types of nucleotides are added at the same time. Each of them, however, are labeled by a different fluorescent dye and have a terminating group which prevents the chain extension by more than one nucleotide. After each chain on a plate is extended by one nucleotide, unincorporated nucleotides are removed and the types of incorporated nucleotides are determined by color imaging. Next, fluorescent dyes and terminating groups are removed and the sequencing reaction is ready for the next cycle. As it was in case of 454, the cycles are repeated until all of the sequences are determined.

SOLiD system is based on a different methodology which is called sequencing by ligation [41, 43]. First, a universal primer and 8-base long fluorescently labeled oligonucleotide probes are added to the reaction. Of these 8 bases only the first two are meaningful, the rest of them are *degenerate*, meaning that they can pair with any other base. The oligonucleotide probe binds to the DNA strand being sequenced and a *DNA ligase* enzyme links the oligonucleotide to the growing strand. The unlinked oligonucleotides are washed away and the fluorescent label is read by a scanner. Next, three trailing degenerate nucleotides are cleaved off and new oligonucleotides are added. This process continues until the new strand is fully synthesized. After this, the whole new strand is denaturated and the whole process is repeated with the only difference that a new primer is one nucleotide shorter than the previous one. As a result, new bases are read during the process. The whole reaction is repeated five times, to ensure that each nucleotide on the strand is interrogated twice. In this system, only 4 fluorescent colors are used to label 16 types of oligonucleotide probes, but this is sufficient, because the sequence

can be later inferred based on a set of logical rules, known as 4-color coding scheme [41].

After the sequencing is completed, the data has to be computationally analyzed. The exact type of analysis depends on what kind of experiment we want to conduct. In addition to gene expression profiling, RNA-seq data can be used for non-coding RNA discovery and detection, transcript rearrangement discovery or single-nucleotide variation profiling [44]. We, however, will focus only on RNA-seq applications for the gene expression profiling.

2.4.2 Computational analysis of RNA-seq data

In the computational analysis description we will assume that the genome of the species being investigated is known. In that case, the first step is to map the sequencing reads to the reference genome in order to know where they have originated from. This process is not complicated for individual reads, but the problem arises because of the huge amount of reads that need to be mapped. Conventional alignment programs such as BLAST or BLAT would simply be too slow for this task [45]. Hence, new alignment tools have been developed which are aimed at aligning a large amount of short read data. One of the most popular tools used for this task is Bowtie [46], a program which aligns the short reads in a very fast and memory efficient way.

Another problem for read mapping might be caused because of repetitive regions in a genome. Reads originating from these regions usually cannot be mapped unambiguously. In higher eukaryotic organisms, these regions constitute almost 50% of the genome [45], so discarding all of those reads would result in a substantial loss of the data. Therefore, many studies use *pair-ended* reads. The idea is that a DNA fragment is sequenced from both of its ends giving rise to two reads with approximately known gap length between them. In this way, aligning one of the paired reads could help to align the other one unambiguously. Nowadays, pair-ended reads are supported by most of the sequencing platforms and alignment tools [45].

Yet another problem is caused by reads originating from the locations of splice junctions. These reads cannot be straightforwardly mapped to the original genome, because the read sequence is split into two parts and separated by an intron sequence in the original genome. Some of the alignment programs take into consideration the existing transcriptome annotations or even try to find novel splice junctions in order to map such reads. One of the most popular program among these is TopHat [47]. Original Bowtie software was not able to deal with such reads, but the problem is addressed in Bowtie 2 [48], a new version of the tool which is currently under the development.

After the reads are mapped, the gene expression levels can be inferred simply by counting how many mapped reads fall into the regions of known genes [8]. In the fragmentation step, longer genes get more fragments for sequencing, therefore, the counts for each gene have to be normalized by gene lengths. Moreover, these counts have to be normalized by the total number of mapped reads for a sample, because some samples might have more mapped reads than the others which would result in a sample bias. A popular gene expression measure which follows these principles is called RPKM (Reads Per Kilobase of exon model per Million mapped reads) [8]. It normalizes read counts by gene lengths in kilobases and by millions of mapped reads for a sample.

More advanced methods, such as Cufflinks [49], MMSEQ [9] or BitSeq [50], measure gene expression levels not on the gene level, but on the isoform level. Since gene isoform sequences are very similar, many reads can often be mapped to several gene isoforms. If we discarded all of those reads, it would be difficult to estimate the expression levels for separate isoforms. Therefore, a statistical model has to be created which aims to tell how many of these reads originate from each of the isoform. Parameters of the model are usually adjusted by looking at the reads which map to the distinct parts of the isoforms. Having isoform level expressions, gene expressions can be derived by summing up all of the isoform expressions belonging to a single gene.

As it was in the case of microarrays, the downstream analysis often include identification of differentially expressed genes. One of the most popular tools used for this task are DESeq [51] and edgeR [52]. Similarly to microarray differential expression analysis tool LIMMA, DESeq and edgeR calculate $\log_2$ fold change values, p-values and FDR values. However, unlike LIMMA, for calculating p-values these tools use a statistical test that is based on a negative binomial distribution.

Chapter 3

Inter-platform gene expression data comparisons

RNA-seq technology offers many advantages over conventional microarrays, such as a low background signal, a possibility to detect novel transcripts and an increased dynamic range of measurements [2, 17]. Therefore, RNA-seq has already become a popular alternative to microarrays and it is likely that in the future it will fully replace microarrays [5]. However, there are some important practical considerations which favor microarrays: lower cost, known biases and well-established experimental pipelines [4]. Thus, microarrays still remain a popular technology for dealing with gene expression data.

Naturally, there is a need to compare microarray experiments against RNA-seq experiments. Previously, such comparisons usually aimed at evaluating the reliability of RNA-seq technology [2, 3], but in the future such comparisons might more often be used in order to retrieve similar microarray–RNA-seq experiment pairs and get more insight on a particular study. These comparisons will still be relevant even if microarrays are fully replaced by RNA-seq, because there are big databases of microarray experiments available [6, 7] and they could be used as a reference.

In order to make these comparisons more effective, we have to find a way, how to convert the measurements of microarrays and RNA-seq or how to change the processing of their raw data in order to make these measurments more similar or more "comparable". However, to the best of our knowledge, there are no past studies which would had been aiming at doing that. On the other hand, there were previous studies that are 1) comparing the platforms in order to estimate their accuracy, 2) performing the same experiment on two platforms in order to validate results, 3) visualizing data from the two platforms 4) combining the data from two platforms in order to extract some new information about the experiment. In this chapter, we will first present

the studies which were comparing the two platforms (1 and 2) and then we will give a short overview of the studies which were combining or visualizing data from the two platforms (3 and 4).

3.1 Microarray–RNA-seq comparisons

One of the earliest and most prominent microarray–RNA-seq comparisons was done by Marioni *et al.* [2]. This study evaluated RNA-seq technical reproducibility and compared the RNA-seq and microarray gene expression measurements. The RNA-seq platform used in the study was Illumina Genome Analyzer and the microarray platform was Affymetrix U133 Plus 2. RNA-seq and microarray experiments were done using the same samples taken from human kidney and liver. Sequencing was done in 2 runs, each of the runs was using 7 lanes where some of the lanes contained kidney samples while the other lanes contained liver samples. Microarray experiment was done with 3 technical replicates for each kidney and liver sample.

RNA-seq data was shown to be highly reproducible and the technical variance across the lanes was very small. In addition, RNA-seq gene expression measures were found to agree with microarray gene expression measures rather well both on an absolute and differential gene expression scale. The Spearman correlation was 0.75 for kidney, 0.73 for liver and 0.73 for $\log_2$ fold changes between the two conditions (Fig. 3.1). 11493 genes were found differentially expressed at FDR of 0.1% according to RNA-seq and 8113 genes according to microarray technology. Among these, the majority (6534) were found to be differentially expressed according to both platforms. In order to find out which of the platforms gives more accurate results, they were compared against the third technology—qRT-PCR. The results of qRT-PCR were found to agree better with RNA-seq giving an indication that RNA-seq is a more sensitive method at detecting differentially expressed genes.

Another prominent study by Fu *et al.* [3] compared microarrays and RNA-seq with the third dataset coming from an independent source—proteomics. Sequencing platform used was Illumina's Solexa Sequencer, microarray platform was Affymetrix Human Exon 1.0 ST and protein expression levels were measured by 2D LC-MS/MS system. Absolute gene expression measurements of a human brain sample were calculated for both platforms. The Pearson correlation between RNA-seq and microarray measurements was reasonably good: $r = 0.67$ for 8441 genes. For a subset of 520 genes RNA-seq and microarrays were compared against proteomics. The Pearson correlation was of a moderate level, but it was better for RNA-seq than microarrays ($r = 0.36$ for RNA-seq and $r = 0.24$ for microarrays). Hence, RNA-seq was

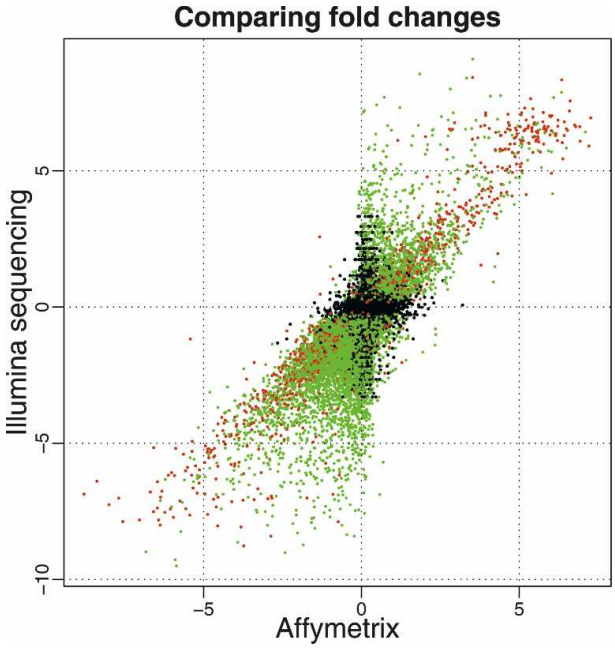


Figure 3.1: A comparison of $\log_2$ fold changes between Affymetrix and Illumina platforms by Marioni *et al.* [2]. Differentially expressed genes having more than 250 reads are colored red and genes having less than 250 reads are colored green. Reads that are not differentially expressed at FDR of 0.1% are colored black.

concluded to be a more precise method for an absolute gene expression level estimation.

A particularly interesting inter-platform comparison was done by Agarwal *et al.* [53] comparing tiling arrays against Solexa/Illumina sequencing technology. Tiling arrays are the only type of microarrays which is able to detect novel transcripts. Therefore, RNA-seq and tiling arrays were compared not only by their gene expression measurements, but also by their ability to detect known transcriptionally active regions (TARs). The detected TARs were compared with a gold standard set—known transcriptome annotations. RNA-seq was found to be in a better agreement with known transcriptome annotations, hence, it was suggested that RNA-seq could be used to calibrate tiling arrays for detecting unknown transcripts.

An another interesting study was done by Bradford *et al.* [54]. Expression measurements of human breast cell lines were compared between SOLiD sequencing platform and Affymetrix Human Exon 1.0 ST microarrays. The difference between this study and the previously mentioned studies is that in

this study, expression measures were compared not on a gene level, but on an exon level. The Pearson correlations were $r = 0.55$ for MCF-10a cell line and $r = 0.53$ for MCF-7 cell line. The $\log_2$ fold change correlation for the exons expressed in both conditions was $r = 0.59$.

All of the above mentioned and some of the other microarray–RNA-seq comparisons are listed in Table 3.1. The list is not exhaustive, but includes most of the published comparisons up to this date. As we can see, most of the comparisons were done on human samples, but there were some comparisons on mouse, yeast and other organisms. Most of the studies compared Illumina sequencing platform against Affymetrix microarray platform, but there were some comparisons between other types of platforms, too. For many of the studies, the inter-platform comparison and evaluation of RNA-seq platform reliability was the major purpose, however, some of the studies were aimed at biology-related questions and the comparisons were done just as an additional validation of experimental results [55, 56].

Most of the studies calculated either absolute gene expression measurement or $\log_2$ fold change correlation between the two platforms. Both absolute gene expression and $\log_2$ fold change correlations varied between studies. For example, in Beane *et al.* [55] $\log_2$ fold change correlation between Illumina sequencing and Affymetrix HGU133A 2.0 microarray was only 0.16, while in Marioni *et al.* [2] it was 0.73. The correlation levels rarely depend on which type, Pearson or Spearman, correlation measurement is used. However, they depend a lot on the way experiment samples are prepared. If the exact same experiment samples are being used for both RNA-seq and microarray studies and the sample preparation protocol is very similar, the correlation between the two platforms is significantly better. For $\log_2$ fold change correlations there is an another important factor: the similarity of the two sample conditions. If the conditions are very similar, $\log_2$ fold change values are very small and the two platforms are not able to measure them precisely because of the noise. Therefore, in this case, the two platforms do not agree as well as in the case of where two sample conditions are very different. Finally, one should remember that the correlation significance largely depends on the number of points being correlated [57]. Therefore, in general, one could expect a lower correlation coefficient if the correlation is calculated for a large number of genes and a higher correlation coefficient if it is calculated for a small number of genes.

In the past, there were also many studies which compared gene expression measurements across different microarray platforms [28, 69, 70, 37]. In some of these studies, the correlation between different microarray platforms is even worse than the correlations between RNA-seq and microarray in the above mentioned studies. For example in Liu *et al.* [37], the absolute gene

Article	**Organism**	**Sequencing pl.**	**Microarray pl.**
Marioni *et al.* [2]	Human	Illumina	Affymetrix
Fu *et al.* [3]	Human	Illumina	Affymetrix
Beane *et al.* [55]	Human	Illumina	Affymetrix
Cheung *et al.* [56]	Human	Illumina	Affymetrix
Sultan *et al.* [58]	Human	Illumina	Illumina
Bradford *et al.* [54]	Human	SOLiD	Affymetrix
Labaj *et al.* [59]	Human	SOLiD	Affymetrix
Cloonan *et al.* [60]	Mouse	SOLiD	Illumina
Mortazavi *et al.* [8]	Mouse	Illumina	Affymetrix
Tang *et al.* [61]	Mouse	SOLiD	Affymetrix
Rathi1 *et al.* [62]	Mouse	Illumina	Affymetrix
Bottomly *et al.* [63]	Mouse	Illumina	Affymetrix
Su *et al.* [64]	Rat	Illumina	Affymetrix
Wang *et al.* [17]	Yeast	Illumina	Affymetrix
Wilhelm *et al.* [65]	Yeast	Illumina	Affymetrix
Arino *et al.* [66]	Yeast	454	2-color microarray
Bloom *et al.* [67]	Yeast	Illumina	2-color microarray
Agarwal *et al.* [53]	Roundworm	Illumina	Affymetrix
Malone *et al.* [4]	Fruit fly	Illumina	Nimblegen
Liu *et al.* [68]	Chimpanzee	Illumina	HJAY

Table 3.1: Publications that compare RNA-seq – microarray gene expression data

expression measurement correlations between different microarray platforms ranged only from 0.42 to 0.60. This is an another indication that we cannot expect a perfect correlation between different platforms, especially given different sample preparation protocols. However, we can still look for the ways to reduce the gene expression measurement differences by computational analysis and aim at increasing the inter-platform correlation as we will be doing in this thesis.

3.2 Methods that combine or visualize inter-platform data

There have been a few studies which developed tools for combined microarray–RNA-seq gene expression data processing, visualization and statistical analysis [71–73]. The common feature of these tools is that they all support several different RNA-seq and microarray platforms, employ traditional gene expression data processing algorithms and have a few different ways of visualizing the data.

Among these tools, probably the most universal and well-known tool is Mayday SeaSight [71] which is an extension of an older version of the tool Mayday [74, 75]. The main new feature of the extension is the ability to handle sequencing data, in addition to microarray data. As sequencing data input SeaSight takes a mapped reads file in SAM or BAM format while as microarray data input it takes raw output files from GenePix, Affymetrix, Agilent or ImaGene platforms. Alternatively, it can take generic tabular format files as sequencing or microarray data input. SeaSight has a number of different possible data processing methods both for microarray (background correction, two-channel array normalization, inter-array normalization and summarization) and sequencing (RPKM, DCPM and locus-dependent functions). Also, the original Mayday software offers many plugins which can be used for data clustering, filtering, classification and for finding significantly differentially expressed genes. Finally, the data can be visualized in a number of different ways: scatter plots, box plots, profile plots or enhanced heat maps.

Some of the past studies created methods that combine gene expression measurements from several microarray platforms in order to get some additional information about the experiment [76, 77]. Warnat *et al.* [76] has used gene expression data from several different microarray platforms to classify the samples between three different diseases: prostate cancer, breast cancer and acute myeloid leukemia. The gene expression measurements were trans-

formed using either Median Rank Score or Quantile discretization algorithm in order to make them more comparable and then they were used as an input for Support-Vector-Machine-based classifier. In another study, Wang *et al.* [77] used factor analysis to unify gene expression data from several microarray platforms. The factor analysis was performed either on gene expression level or probe expression level. The method accuracy was evaluated by comparing it to an independent dataset—gene expression measurements from SAGE platform.

To the best of our knowledge, so far there have not been any attempts to combine gene expression data from microarray and sequencing platforms or from two different sequencing platforms. There was one study which combined results from two RNA-seq experiments, but the study was concerned not with the gene expression measurements, but with the transcriptome annotation [77]. The two experiment results were combined in order to identify more novel transcripts and splice sites. On the other hand, combination of RNA-seq and microarray gene expression results is possible in principle and might be done in the future. This could be another motivation to make RNA-seq and microarray gene expression measurements more comparable, as we are aiming in this thesis.

Chapter 4

Methods

In this chapter, we present our method that aims to make microarray and RNA-seq gene expression data more comparable. Section 4.1 explains the main idea of our method, Section 4.2 gives the details about the statistical model which we used to infer expression levels and Section 4.3 provides the information about the tools used for implementation.

4.1 Basic idea of the method

One of the fundamental differences between microarrays and RNA-seq is that RNA-seq measures gene expression based on the whole gene sequence while microarrays rely only on the sub-portion of the gene where the microarray probe sequences are located. The idea of our method is to eliminate this difference, by calculating gene expression for RNA-seq using only the gene regions where probe sequences are located.

The basic idea of our method is illustrated in the Figure 4.1. Usually, in RNA-seq technology, gene expression is calculated based on the number of mapped reads overlapping with gene regions. We, on the other hand, counted the number of reads overlapping with probe regions. Probe region locations were retrieved from Custom CDF probe annotation files [31].

Based on the read counts we estimated probe region expressions using a probabilistic approach (more details on this step will be provided in Section 4.2). As a result, we got probe region expressions using the sequencing data. Each probe region expression has a corresponding probe expression in the microarray. Therefore, from this point we treated probe region expressions the same way as probe expressions are treated in microarray data analysis. That is, we applied one of the most popular algorithms for microarray data analysis—RMA.

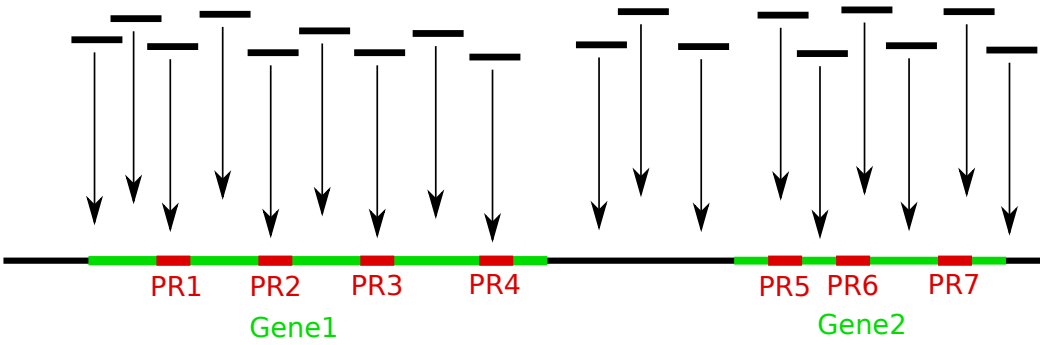


Figure 4.1: Counting reads overlapping with probe regions. Gene1 and Gene2 denotes gene regions, while PR1-PR7 denotes probe regions inside the gene regions.

The RMA algorithm (which was also discussed in Section 2.3.2 of Chapter 2) consist of three steps: background-correction, normalization and summarization. During background-correction step, technical noise is removed, during normalization step, the expression values from different microarray replicates are normalized and, finally, during summarization step, probe-level expressions are converted to gene-level expressions. Since sequencing data contains very little technical noise [17], we skipped background-correction step and applied only the two latter steps of RMA.

We call our method PREBS—Probe Region Expression Based on Sequencing. The basic pipeline for the whole method is depicted in Figure 4.2. In short, we count read overlaps with probe regions to get probe region expressions and then we apply RMA to get gene expressions. This type of sequencing data processing resembles microarray data processing and, as we will see later, the gene expression results which we get this way are more similar to microarray gene expression results.

Figure 4.2: Basic pipeline to get gene expression measurements using PREBS method

4.2 Expression level inference

Counting the number of reads that overlap with regions of interest (probe regions or gene regions) is a stochastic process. As a result, there is a lot of uncertainty about the observed counts for these regions, especially, when the counts are small. In order to account for the uncertainty, we decided to use a probabilistic approach.

We use Bayesian inference [78] which is based on Bayes rule:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \propto P(E|H) \cdot P(H). \tag{4.1}$$

Bayes rule says that the *posterior* $P(H|E)$ is equal to the *likelihood* $P(E|H)$ times *prior* $P(H)$ divided by the *evidence* $P(E)$. We can often choose to ignore the normalizing factor (evidence) and then we get that the posterior is proportional to the likelihood times the prior. In our case, the evidence E will be the number of mapped reads that overlap with the region of interest in a genome and the hypothesis H will be the distribution of those reads.

Each sampled read has two possibilities: either it overlaps with the region of interest or it does not. Therefore, read sampling can be modeled as a Bernoulli process and the read distribution converges to a Poisson distribution as the number of reads approaches infinity [79]. The number of reads k mapped to the region of interest depends on the Poisson distribution with a rate parameter λ and it can be modeled by

$$p(k|\lambda) = \text{Poisson}(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{4.2}$$

where λ parameter can be viewed as the gene expression level of the region of interest or, in other words, the rate by which reads are sampled from that region.

Equation (4.2) will be our likelihood $P(E|H)$. In order to select the prior $P(H)$, we have to select a distribution for λ parameter. In this case, the most convenient distribution for the prior is the Gamma distribution

$$p(\lambda) = \text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \tag{4.3}$$

because it is conjugate to the Poisson distribution. That means, if the likelihood is a Poisson distribution and the prior is a Gamma distribution then the posterior $P(H|E)$ will also be a Gamma distribution [80]. If we have a

single measurement of k, it can be shown that posterior will be

$$\begin{aligned} p(\lambda|k) &= \frac{1}{Z} p(k|\lambda)p(\lambda) \\ &= \frac{1}{Z} \text{Poisson}(k;\lambda)\text{Gamma}(\lambda;\alpha,\beta) \\ &= \text{Gamma}(\lambda;\alpha+k,\beta+1), \end{aligned} \tag{4.4}$$

where $Z = p(k) = \int p(k|\lambda)p(\lambda)\mathrm{d}\lambda$.

The prior distribution can be viewed as the expression level distribution before we see any evidence (number of counts for a region of interest). The posterior distribution can be viewed as the expression level distribution, after we observe the evidence. Therefore, in order to estimate expression level for a region of interest we will calculate an expected value of the posterior distribution. The expected value of Gamma distribution is equal to the ratio of its scale parameter and rate parameter, therefore, in our case, it is

$$\mathbf{E}[p(\lambda|k)] = \frac{\alpha+k}{\beta+1}. \tag{4.5}$$

We have estimated α and β parameters, by comparing cumulative distribution functions for read counts and Gamma (Figure 4.3). We found out that the best values for α and β parameters were (0.2, 0.03) and (0.17, 0.01) for Marioni *et al.* and Cheung *et al.* data sets respectively.

4.3 Tools used for implementation

Affymetrix microarray data were processed using Affy package [81] from R/Bioconductor [82] and custom CDF files [31]. Microarray expression values were summarized using RMA algorithm [30].

Sequencing data was converted from .sra format to .fastq format using SRA Toolkit version 2.1.9 [83]. Next, sequencing data was processed using three different methods: MMSEQ [9], RPKM [8] and PREBS.

To get MMSEQ gene expression measures, MMSEQ software (version 0.9.18) was used. Bowtie software (version 0.12.7) [46] was used to map the reads to the transcriptome, as recommended by MMSEQ manual. MMSEQ options were set to default and Bowtie options were set as recommended by MMSEQ (`-a --best --strata -S -m 100 -X 400`). Homo Sapiens transcriptome version GRCh37.65 from Ensembl database was used. SAMtools (version 0.1.18) [84] was used for alignment format conversion.

For the RPKM method, reads were mapped by TopHat software (version 1.4.1) [47]. Option `--transcriptome-only` was used for TopHat to get

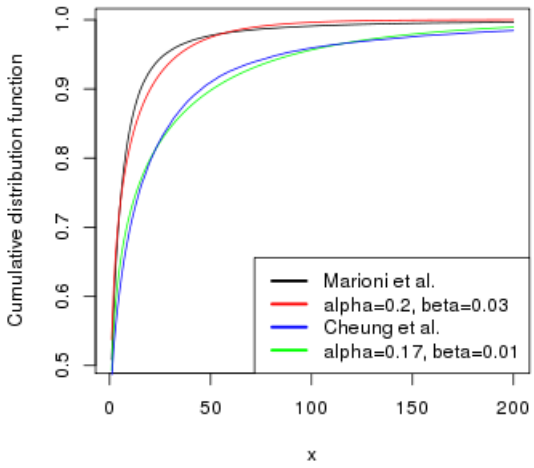


Figure 4.3: Comparison of read count and Gamma cumulative distribution function

read alignments only to known transcriptome annotations. Ensembl GTF annotation file (version GRCh37.65) was used along with the genome file of the same version. Next, Ensembl genomic annotations were downloaded using GenomicFeatures package and read overlaps with gene regions were calculated using GenomicRanges package [85]. Then, RPKM values were calculated and base 2 logarithm values were taken.

PREBS values were calculated using the same tools as for RPKM values, but the read overlaps were calculated not for genes, but for probe regions taken from custom CDF file annotations. RMA algorithm was applied using Affy package [81] slightly modified to accept custom probe expression measures.

All necessary data processing tasks were done using R, Perl and Python scripts. Script running was coordinated by Bash scripts. Additionally, some Unix shell utilities, such as grep, awk and sed, were used.

Chapter 5

Results

In this chapter, we will introduce the results of the PREBS method and compare it against two other methods: MMSEQ and RPKM. In order to evaluate the methods, we will compare each of the method results against microarray results and calculate correlations. The methods will be tested in three different categories: absolute expression (Section 5.2), differential expression (Section 5.3) and cross-platform differential expression (Section 5.4). In addition, we will have a look at some more technical results: we will examine our method performance using manufacturer's CDF files instead of custom CDF files (Section 5.5) and we will calculate the differential expression for microarray replicates (Section 5.6).

5.1 Data sets

In our study we used two data sets: Marioni *et al.* [2] and Cheung *et al* [56]. These data sets were selected, because they were easily accessible, used popular sequencing and microarray platforms (Illumina and Affymetrix respectively). An additional advantage was that most of the samples in both datasets had microarray technical replicates which makes the data more reliable.

The Marioni *et al.* [2] data set consisted of two samples: human kidney and liver. The sequencing was done using Illumina Genome Analyzer sequencer for two runs where each of the runs was using 7 lanes. Some of the lanes contained 3 pM concentration while the others contained 1.5 pM concentration samples from human kidney or liver. We used only data from the lanes that contained 3 pM concentration. Altogether, there were 5 such lanes for the kidney sample and 5 such lanes for the liver sample. Each of these lanes gave 12.9-14.7 million 32 nucleotides long reads. The microarray

experiment was done using Affymetrix U133 Plus 2 platform. 3 microarray technical replicates were used for kidney and 3 for liver sample. Microarray and sequencing experiments were done on exactly the same samples using as similar sample preparation protocols as possible. The microarray data set is available in the GEO database [7] under accession number GSE11045 and sequencing data is available in NCBI short read archive under accession number SRA000299.

The Cheung *et al.* [56] data set consisted of 41 samples of B-cells taken from CEPH HapMap individuals. Illumina Genome Analyzer sequencer was used giving around 40 million 50 nucleotides long reads for each of the sample. Microarray experiments were done using Affymetrix Human HG-Focus Target Array. 25 out of 41 samples had two microarray technical replicates while the rest of the samples did not have any microarray technical replicates. Sequencing and microarray samples were taken from the same individuals, hence, they were biological replicates. The data is available in the GEO database [7] under accession numbers GSE16921 and GSE16778 for sequencing and microarray data, respectively.

We tried to find more coupled microarray–RNA-seq data sets, but it proved to be a difficult task. Most of the ones we found had some problems, for example, the microarray and RNA-seq samples were prepared in very different ways, old type of sequencing platforms were used or the data was not easily accessible. Therefore, in this thesis, we will only use the two aforementioned data sets.

5.2 Absolute expression comparison

We have processed the RNA-seq data using 3 different methods: MMSEQ, RPKM and PREBS. Then, we processed microarray data for the same samples using a single method—RMA. In order to find out which one of the 3 methods agrees best with microarray gene expression measurements, we have calculated Pearson and Spearman correlations and made scatter plots of microarray–RNA-seq data. We used 2 data sets: Marioni *et al.* [2] and Cheung *et al.* [56]. The scatter plots for all of the samples in each of the data sets looked similar, therefore, here we provide the scatter plots only for first sample in each data set: kidney sample in the Marioni *et al.* and GM06985 sample in the Cheung *et al.* data set (see Figure 5.1). On the other hand, the performance tables include the average correlations among all of the samples (see Tables 5.1, 5.2, 5.3 and 5.4).

For evaluation we used only those genes that were present on all of the RNA-seq processing methods and microarrays. Moreover, we filtered points

that had FPKM < 0.3 according to the MMSEQ method. The reason we did this is because MMSEQ measurements for low expressed genes are not very reliable and they decrease the correlation with microarrays quite a lot (see Figure 5.2). We also tried to filter low expression genes in RPKM and PREBS methods, but the correlation stayed virtually unaffected (data not shown), so we decided to include unfiltered results for these methods. In order to make the comparison fair, the three methods were evaluated on the exactly same set of genes.

From the scatter plots and the tables it is evident that PREBS agrees best with microarray data both in terms of Pearson and Spearman correlations. For the kidney sample in the Marioni *et al.* data set the Pearson correlations were 0.74, 0.68 and 0.83 for MMSEQ, RPKM and PREBS respectively (see Figures 5.1a, 5.1c and 5.1e). Similarly, for a GM06985 sample in the Cheung *et al.* data set the Pearson correlations were 0.76, 0.71 and 0.81. Thus, PREBS performs best on both samples. The improvements in the correlation are just as evident when we look at Tables 5.1 and 5.2 where the correlations are averaged over all samples in the data sets (2 samples for the Marioni *et al.* data set and 41 samples for the Cheung *et al.* data set). Pearson correlations for PREBS are 0.84 and 0.8, while for MMSEQ they are 0.7 and 0.76, for the Marioni *et al.* and the Cheung *et al.* data sets respectively. RPKM correlations are the lowest among all of the methods for both data sets.

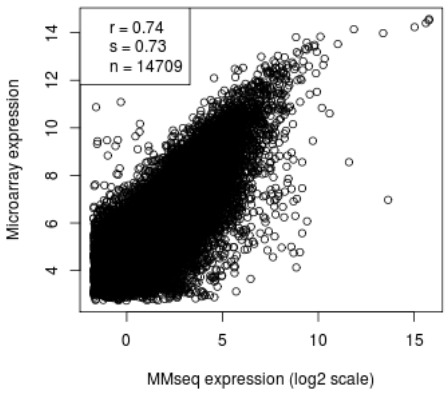
Tables 5.3 and 5.4 show the correlation improvements of PREBS compared to other methods. The calculation formula for improvements was $\frac{(\mathrm{PREBS}-\mathrm{OTHER})}{(1-\mathrm{OTHER})} \cdot 100\%$, where PREBS is the correlation of PREBS vs microarray and OTHER is the correlation of the other method vs microarray. The normalizing factor $(1 - \mathrm{OTHER})$ was chosen, because it corresponds to a total amount of how much correlation can be improved up until the perfect correlation.

Both for the Marioni *et al.* and the Cheung *et al.* data sets, PREBS reached reasonable amounts of improvement over the other two RNA-seq processing methods, but the improvement was bigger for the Marioni *et al.* data set. The reason for that might have been, because RNA-seq and microarray experiments for the Marioni *et al.* data set were conducted on exactly the same samples using very similar protocols, while for the Cheung *et al.* data set the RNA-seq and microarray samples were biological replicates which might have caused more differences in the expression levels. Therefore, the Marioni *et al.* data set is better suited for RNA-seq vs microarray comparisons and it is easier to reach some improvement in correlation using a novel method.

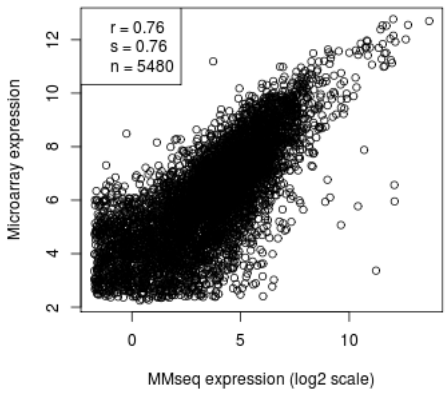
In Figures 5.1a and 5.1e, there are some low MMSEQ and PREBS values

that have high microarray values (in the upper left corner). This effect is more evident for the Marioni *et al.* data set than the Cheung *et al.* data set. It is hard to tell what is the cause and it would require some further investigation. Interestingly, RPKM seems to be the only method for which this effect is not evident.

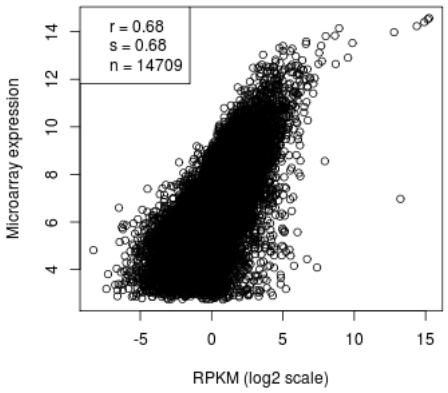
Moreover, in the plots for the Marioni *et al.* data set (Figures 5.1a, 5.1c and 5.1e) we see that expression values based on sequencing can go up even after microarray values reach the saturation level. This suggests that RNA-seq has a larger dynamic range than microarrays. However, we cannot see such behaviour in the scatter plots for the Cheung *et al.* data set (Figures 5.1b, 5.1d and 5.1f).



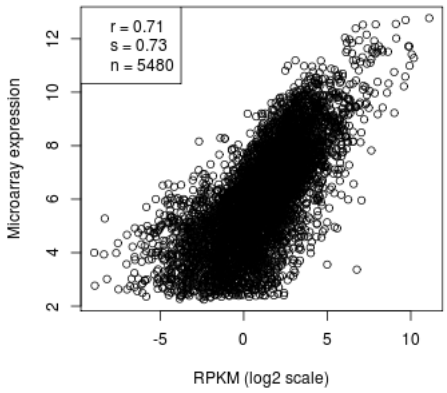
(a) MMSEQ, Marioni *et al.* dataset



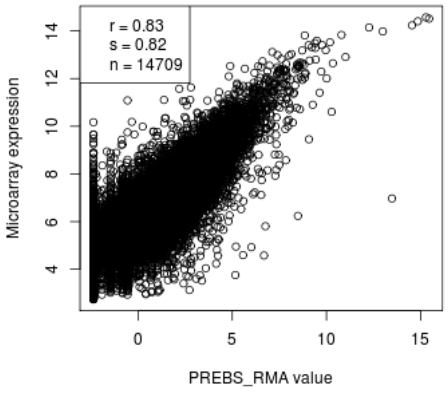
(b) MMSEQ, Cheung *et al.* dataset



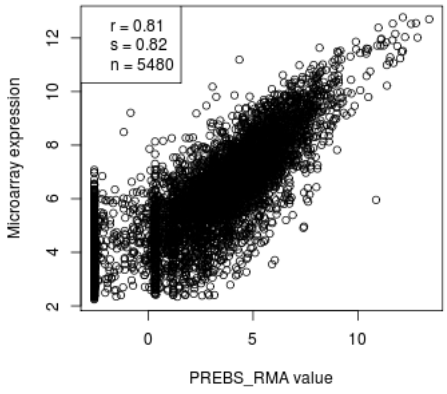
(c) RPKM, Marioni *et al.* dataset



(d) RPKM, Cheung *et al.* dataset

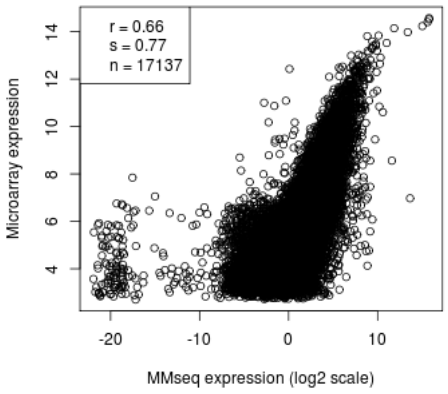


(e) PREBS, Marioni *et al.* dataset

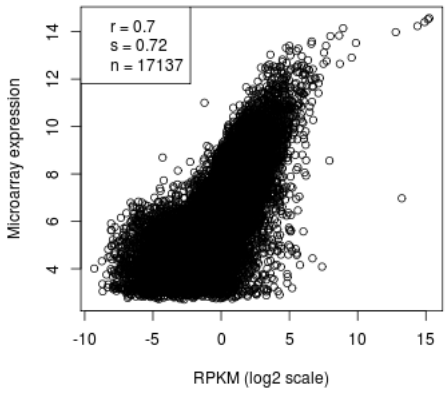


(f) PREBS, Cheung *et al.* dataset

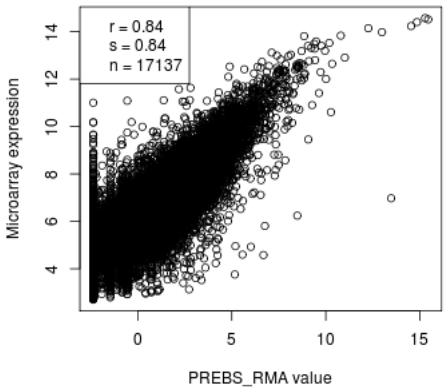
Figure 5.1: RNA-seq expressions plotted against microarray expressions. Kidney sample was used for the Marioni *et al.* dataset and GM06985 sample was used for the Cheung *et al.* dataset. Points with FPKM < 0.3 were filtered. The legend contains Pearson correlation (r), Spearman correlation (s) and the number of genes (n).



(a) MMSEQ, Marioni *et al.* dataset



(b) RPKM, Marioni *et al.* dataset



(c) PREBS, Marioni *et al.* dataset

Figure 5.2: Same as Figure 5.1, but no points are filtered.

	MMSEQ	**RPKM**	**PREBS**
Pearson	0.70	0.66	0.84
Spearman	0.68	0.64	0.83

Table 5.1: Absolute expression correlations for Marioni *et al.* dataset.

	MMSEQ	**RPKM**	**PREBS**
Pearson	0.76	0.71	0.80
Spearman	0.76	0.72	0.81

Table 5.2: Absolute expression correlations for Cheung *et al.* dataset.

	PREBS–MMSEQ	**PREBS–RPKM**
Pearson	44.93%	51.73%
Spearman	45.08%	51.08%

Table 5.3: Absolute expression correlation improvements comparing PREBS against MMSEQ and RPKM on Marioni *et al.* dataset.

	PREBS–MMSEQ	**PREBS–RPKM**
Pearson	16.49%	29.48%
Spearman	18.35%	30.01%

Table 5.4: Absolute expression correlation improvements comparing PREBS against MMSEQ and RPKM on Cheung *et al.* dataset.

5.3 Differential expression comparison

In this section, we will compare differential expression measures between RNA-seq and microarray platforms. More precisely, we will compare differences in $\log_2$ fold changes between the two platforms. We will not compare p-values or False Discovery Rate values, because these values in each platform are calculated in different ways and therefore they are not as comparable as $\log_2$ fold change values.

For the Marioni *et al.* data set $\log_2$ fold changes were calculated between liver and kidney samples. For the Cheung *et al.* data set we calculated $\log_2$ fold changes for several pairs of samples, but they were all quite similar, so we will include the results only for the pair of first two samples in the data set (GM06985 and GM06993).

MMSEQ and RPKM vs microarray scatter plots (Figures 5.3a and 5.3c) look rather similar to the original scatter plot from the Marioni *et al.* publication (Figure 3.1 on page 18). The Spearman correlations (0.74 and 0.75) are also similar to the one reported by the authors (0.75). MMSEQ and RPKM methods perform similarly both with respect to Pearson and Spearman correlations. Unfortunately, the PREBS method (Figure 5.3e) did not significantly improve the differential expression correlations with microarray. Compared to RPKM, there is a small improvement for Pearson correlation (from 0.74 to 0.75), but there is no improvement for Spearman correlation.

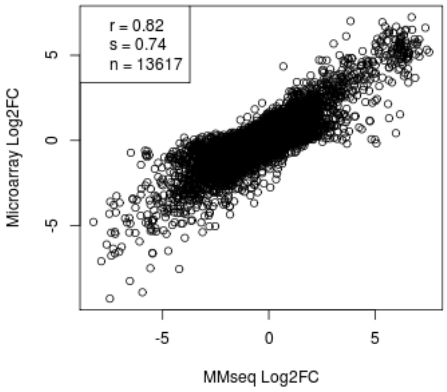
Scatter plots for the Cheung *et al.* data set (Figures 5.3b, 5.3d and 5.3f) do not look as diagonal as the Marioni *et al.* scatter plots and the correlations between RNA-seq and microarray platforms are much smaller. One reason for this might be because the expression changes between the conditions are much smaller in the Cheung *et al.* data set. As we can see in Table 5.9, the 95% quantiles of absolute $\log_2$ fold change values are smaller in the Cheung *et al.* data set than in the Marioni *et al.* data set both for microarrays and RNA-seq. Smaller changes in gene expression are harder to detect by both of the platforms, and therefore there is less agreement between the two platforms.

Tables 5.5, 5.6, 5.7 and 5.8 include differential expression correlations and correlation improvements for the same samples which were used for the scatter plots. From these tables we can see that PREBS does not reach any significant improvement on the Marioni *et al.* dataset and performs slightly worse than the other two methods on Cheung *et al.* data set. It is hard to tell why PREBS does not improve differential expression results even though it significantly improves absolute expression results. One of the reasons might be that differential expression measures have more "degrees of freedom" than absolute expression measures, that is, differential expression of a single gene

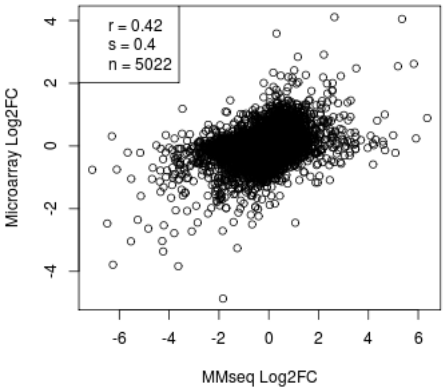
depends on expressions of that gene in two samples instead of one. Therefore, there is more room for errors and it is harder to achieve as accurate results. Furthermore, the bad PREBS performance on the Cheung *et al.* data set can be attributed to that the low expression changes in the Cheung *et al.* data set. When the expression changes are low, the noise level compared to the signal is high and therefore it is hard to create a computational method that interprets the data precisely.

Finally, we present Venn diagrams of differentially expressed genes according to PREBS, MMSEQ and microarrays (see Figures 5.4 and 5.5). RPKM method was not included here, because, as we already saw, it is inferior to MMSEQ method in terms of agreement with microarrays. The genes were regarded as differentially expressed if their $\log_2$ fold change value was bigger than $\log_2(1.5)$ or smaller than $-\log_2(1.5)$ (a similar criteria for identifying differentially expressed genes was used by Beane *et al.* [55]).

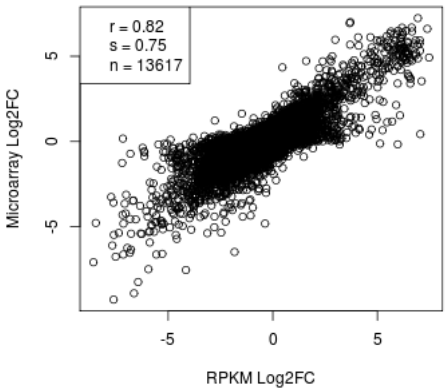
In Figure 5.4 we can see that the number of differentially expressed genes that are found only by MMSEQ is bigger than the number of differentially expressed genes that are found only by PREBS (3076 vs 1174). This is probably because PREBS uses only a part of the gene regions for counting overlapped reads, and therefore it is less sensitive in detecting differentially expressed genes. On the other hand, we want to create an RNA-seq processing method which is similar to the less sensitive microarray method, so the loss of sensitivity is probably inevitable. Moreover, even at lower sensitivity, the number of overlapping differentially expressed genes between PREBS and microarrays is bigger than the number of overlapping differentially expressed genes between MMSEQ and microarrays (622 vs 574, excluding differentially expressed genes that are found by all of the 3 methods). The situation for the Cheung *et al.* data set is rather similar to what we discussed for the Marioni *et al.* data set. However, the numbers of differentially expressed genes for the Cheung *et al.* data set are much smaller according to all three methods, once again confirming the fact that gene expression levels do not differ much between the samples in the Cheung *et al.* data set.



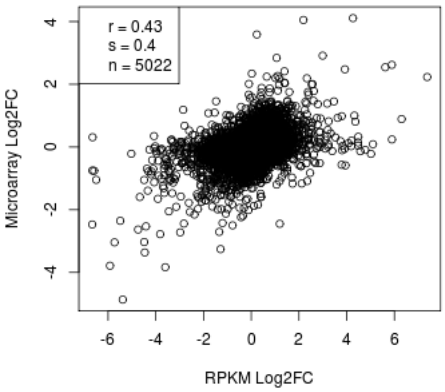
(a) MMSEQ, Marioni *et al.* dataset



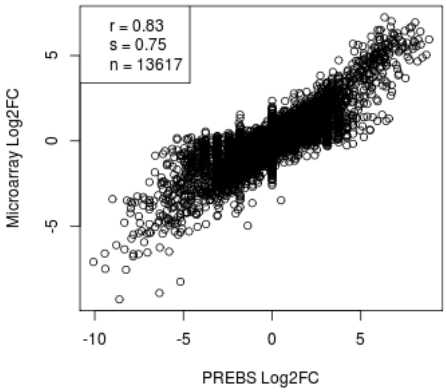
(b) MMSEQ, Cheung *et al.* dataset



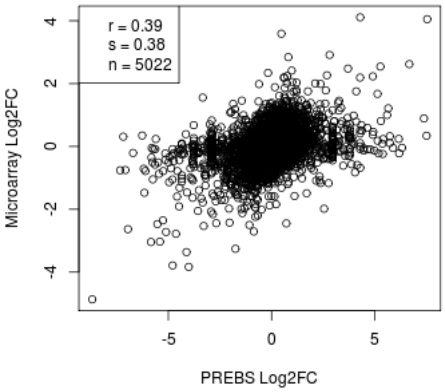
(c) RPKM, Marioni *et al.* dataset



(d) RPKM, Cheung *et al.* dataset



(e) PREBS, Marioni *et al.* dataset



(f) PREBS, Cheung *et al.* dataset

Figure 5.3: RNA-seq Log2FC plotted against microarray Log2FC. Kidney and liver samples were used for Marioni *et al.* dataset, GM06985 and GM06993 were used for Cheung *et al.* dataset. Points with FPKM < 0.3 were filtered. The legend contains Pearson correlation (r), Spearman correlation (s) and the number of genes (n).

	MMSEQ	**RPKM**	**PREBS**
Pearson	0.82	0.82	0.83
Spearman	0.74	0.75	0.75

Table 5.5: Differential expression correlations for Marioni *et al.* dataset.

	MMSEQ	**RPKM**	**PREBS**
Pearson	0.42	0.43	0.39
Spearman	0.40	0.40	0.38

Table 5.6: Differential expression correlations for Cheung *et al.* dataset.

	PREBS–MMSEQ	**PREBS–RPKM**
Pearson	2.91%	6.11%
Spearman	1.91%	-1.25%

Table 5.7: Differential expression correlation improvements comparing PREBS against MMSEQ and RPKM on Marioni *et al.* dataset.

	PREBS–MMSEQ	**PREBS–RPKM**
Pearson	-5.06%	-7.22%
Spearman	-2.72%	-2.86%

Table 5.8: Differential expression correlation improvements comparing PREBS against MMSEQ and RPKM on Cheung *et al.* dataset.

	Microarray	**MMSEQ**	**RPKM**	**PREBS**
Marioni *et al.*	2.24	3.03	3.40	3.75
Cheung *et al.*	1.09	2.33	2.18	2.75

Table 5.9: 95% quantiles for absolute $\log_2$ fold change values in Marioni *et al.* and Cheung *et al.* datasets.

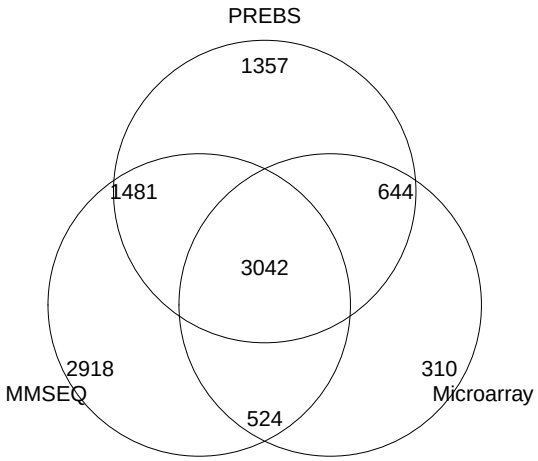


Figure 5.4: Venn diagram of differentially expressed genes for Marioni *et al.* dataset

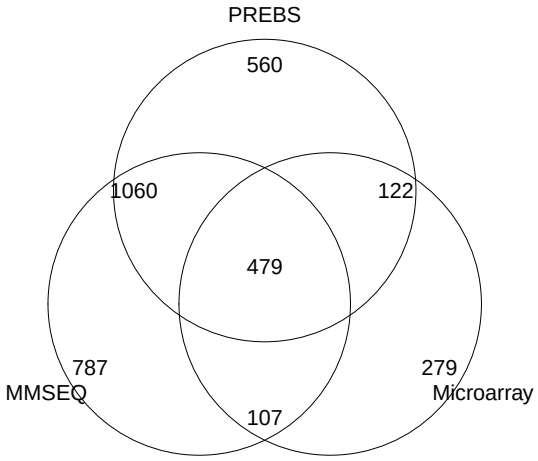


Figure 5.5: Venn diagram of differentially expressed genes for Cheung *et al.* dataset

5.4 Cross-platform differential expression

As the third category of RNA-seq methods evaluation, we compared cross-platform differential expressions. That is, instead of calculating differential expression between two microarray or two RNA-seq samples, we calculated differential expression between one microarray and one RNA-seq sample. Such differential expression calculation can be useful in some cases and it can be seen as one further step in microarray and RNA-seq data integration. To evaluate which of the three methods performs best in this category, we compared cross-platform $\log_2$ fold changes with microarray–microarray $\log_2$ fold changes and calculated the correlations as before (Figure 5.7).

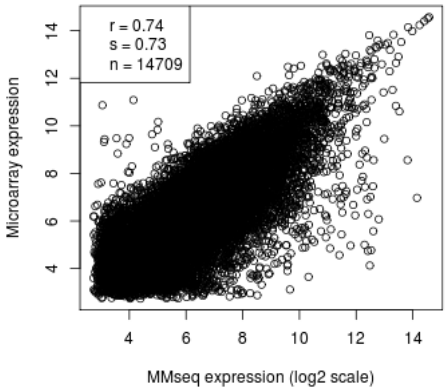
The problem with calculation of microarray–RNA-seq $\log_2$ fold change values is that dynamic ranges of the two platforms are very different and therefore such values are hard to interpret. To overcome this problem, we assigned ranks for microarray and RNA-seq gene expression measurements from the highest expressed to the lowest expressed. To avoid equal ranks we used random tie breaking method for genes with equal expression levels. Then, we replaced the RNA-seq gene expression values by microarray gene expression values from the second sample with corresponding ranks. That way, we achieved that RNA-seq dynamic range is exactly the same as microarray dynamic range and the plots became easier to interpret. This rank replacement method was applied for making cross-differential expression plots (see Figure 5.7). Additionally, to get an idea how absolute expression plots look like after rank replacement, we can have a look at the Figure 5.6.

In the cross-platform differential expression scatter plots for the Marioni *et al.* data set (Figures 5.7a, 5.7c and 5.7e) we can see that there are two clusters of genes: those that agree well between the platforms and those that have very low microarray–microarray $\log_2$ fold changes, but rather high microarray–RNA-seq $\log_2$ fold changes (this effect seem to be stronger for the MMSEQ and RPKM methods than PREBS). Such behavior can probably be explained by the fact that RNA-seq is a more sensitive technology and it can detect expression of the genes that are below microarray noise floor. Since we calculated $\log_2$ fold changes as a ratio of RNA-seq expression divided by microarray expression, those genes, whose expression levels are detected only by RNA-seq, will have a high $\log_2$ fold change according to cross-platform evaluation, but small $\log_2$ fold change according to microarray–microarray evaluation. Another question is why the two clusters are so well-separated and there is no continuity of the points joining the two clusters. This question requires some extra analysis which is out of scope of this thesis.

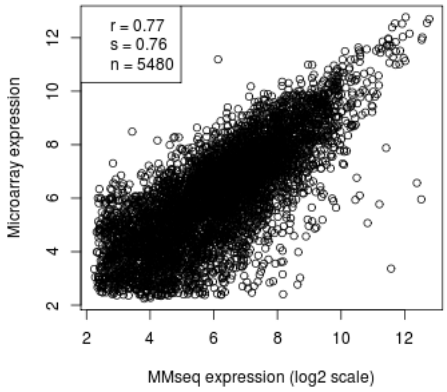
In the correlation tables for the Marioni *et al.* data set (Tables 5.10

and 5.12) we can see that PREBS clearly outperforms MMSEQ and RPKM methods. On the other hand, PREBS performs the worst among the three methods on the Cheung *et al.* data set (Tables 5.11 and 5.13). However, we already saw that the Cheung *et al.* data set is not as well suited for our analysis as Marioni *et al.* data set.

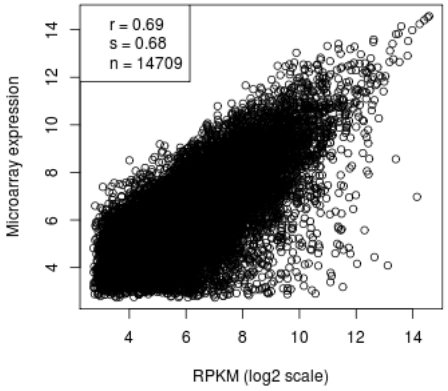
Overall, the correlations of microarray–microarray vs microarray–RNA-seq $\log_2$ fold change values (Section 5.4) were worse than the correlation between microarray–microarray vs RNA-seq—RNA-seq $\log_2$ fold change values (Section 5.3). For example, MMSEQ method had Pearson correlation equal to 0.82 for regular differential expression, but only 0.56 for cross-platform differential expression in the Marioni *et al.* dataset (see Tables 5.5 and 5.10). That suggest us that $\log_2$ fold change values calculated within platform (RNA-seq–RNA-seq) are more accurate than the $\log_2$ fold change values calculated between different platforms (microarray–RNA-seq). However, it is still good to know that we are able to reach reasonable cross-platform vs within-platform $\log_2$ fold change correlations which suggests that cross-platform $\log_2$ fold change correlations can be meaningful.



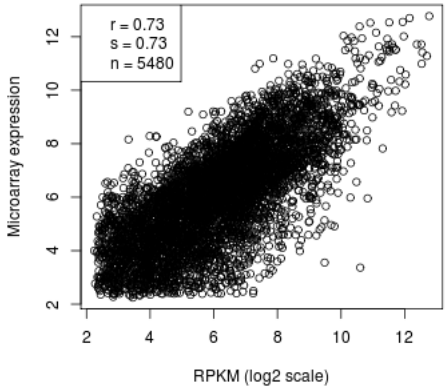
(a) MMSEQ, Marioni *et al.* dataset



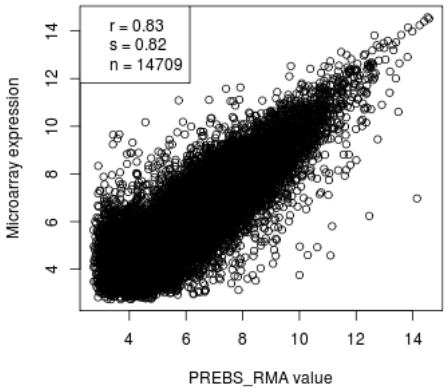
(b) MMSEQ, Cheung *et al.* dataset



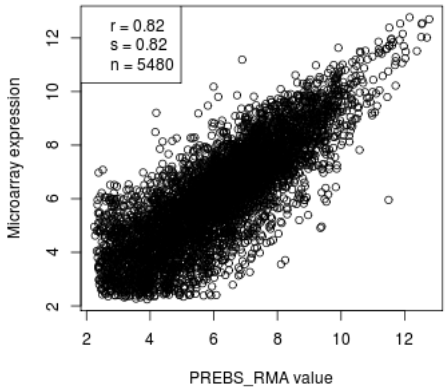
(c) RPKM, Marioni *et al.* dataset



(d) RPKM, Cheung *et al.* dataset

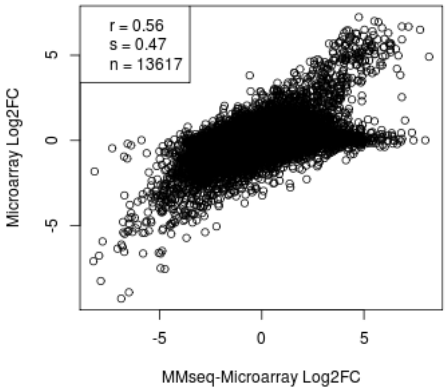


(e) PREBS, Marioni *et al.* dataset

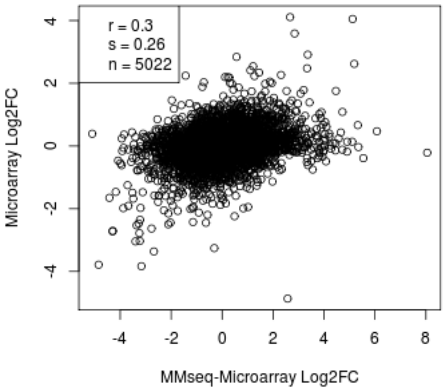


(f) PREBS, Cheung *et al.* dataset

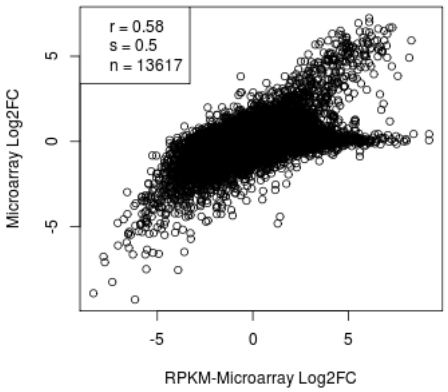
Figure 5.6: Same as Figure 5.1, but RNA-seq expression values are replaced by microarray expression values with corresponding ranks.



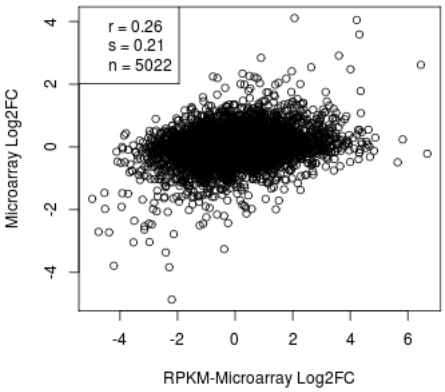
(a) MMSEQ, Marioni *et al.* dataset



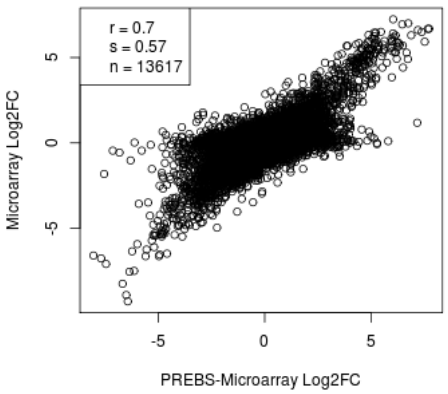
(b) MMSEQ, Cheung *et al.* dataset



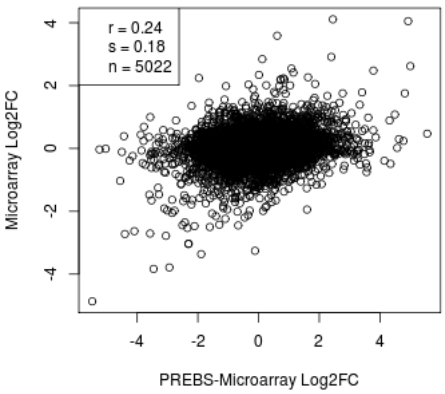
(c) RPKM, Marioni *et al.* dataset



(d) RPKM, Cheung *et al.* dataset



(e) PREBS, Marioni *et al.* dataset



(f) PREBS, Cheung *et al.* dataset

Figure 5.7: RNA-seq–microarray Log2FC plotted against microarray–microarray Log2FC. Kidney and liver samples were used for Marioni *et al.* dataset, GM06985 and GM06993 were used for Cheung *et al.* dataset. Points with FPKM < 0.3 were filtered. The legend contains Pearson correlation (r), Spearman correlation (s) and the number of genes (n).

	MMSEQ	**RPKM**	**PREBS**
Pearson	0.56	0.58	0.70
Spearman	0.47	0.50	0.57

Table 5.10: Cross-platform differential expression correlations for Marioni *et al.* dataset.

	MMSEQ	**RPKM**	**PREBS**
Pearson	0.30	0.26	0.24
Spearman	0.26	0.21	0.18

Table 5.11: Cross-platform differential expression correlations for Cheung *et al.* dataset.

	PREBS–MMSEQ	**PREBS–RPKM**
Pearson	31.82%	28.57%
Spearman	18.87%	14.00%

Table 5.12: Cross-platform differential expression correlation improvements comparing PREBS against MMSEQ and RPKM on Marioni *et al.* dataset.

	PREBS–MMSEQ	**PREBS–RPKM**
Pearson	-8.57%	-2.70%
Spearman	-10.81%	-3.8%

Table 5.13: Cross-platform differential expression correlation improvements comparing PREBS against MMSEQ and RPKM on Cheung *et al.* dataset.

5.5 Manufacturer's CDF

So far, all of the microarray data in the results we provided were processed using custom CDF files [31]. Even though, as we mentioned in Section 2.3.2, custom CDF files provide more accurate results, some of the people might still want to use manufacturer's CDF files. For this reason, we tested how PREBS compares to microarray data processed with manufacturer's CDF.

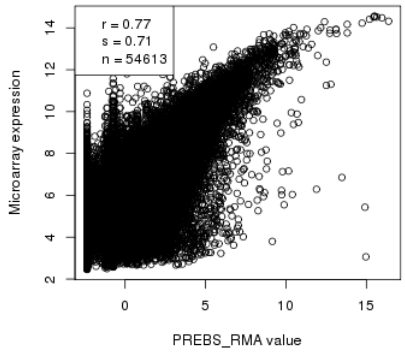
When custom CDF files are used to process microarray data, expression estimates grouped by a chosen type of gene identifiers are obtained as an output (for example, we used Ensembl gene identifiers). On the other hand, when manufacturer's CDF files are used, expression estimates are obtained for probe set identifiers defined by the manufacturer. In order to get gene expression measurements using manufacturer's CDF files, probe set identifiers have to be further mapped to genes identifiers. However, we decided to stick with probe set identifiers when comparing PREBS and microarrays processed using manufacturer's CDF. Since PREBS uses microarray processing tools, it was easy to get PREBS values for microarray probe sets just by changing custom CDF files to manufacturer's CDF files in PREBS processing pipeline. However, such values would be harder to get using MMSEQ or RPKM methods, so we decided to exclude these methods in this step of analysis.

The scatter plots of PREBS vs microarray gene expression values using manufacturer's CDF are provided in Figure 5.8. The correlation for kidney sample in the Marioni *et al.* data set using manufacturer's CDF (Figure 5.8a) is worse than the correlation for the same sample using custom CDF (Figure 5.1e). The reason for this is probably because Figure 5.8a (54613) includes many more points than Figure 5.1e (14709) and as we discussed in Section 3.1, the correlation depends on the number of points being examined. The number of points in the two figures is different, because the number of probe set identifiers in manufacturer's CDF largely exceeds the number of gene identifiers in custom CDF. On the other hand, for GM06985 sample in the Cheung *et al.* data set, the correlation is better when manufacturer's CDF files are used (Figure 5.8c and 5.1f). The reason of this is not clear. However, the the number of points in this case does not differ so much, because another microarray platform is used (8746 points in Figure 5.8c and 5480 points in Figure 5.1f).

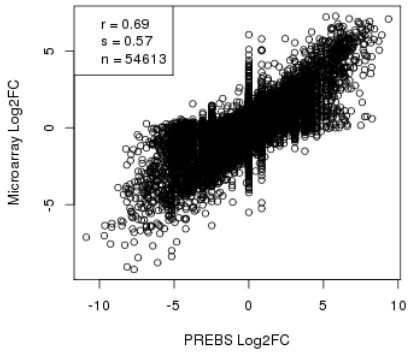
Differential expression correlations using manufacturer's CDF files are also slightly worse than differential expression correlations using custom CDF both for the Marioni *et al.* data set and the Cheung *et al.* data set (compare Figures 5.8b and 5.8d against Figures 5.3e and 5.3f). The differences

in differential expression correlations can again be attributed to the larger amount of points in manufacturer's CDF scatter plots.

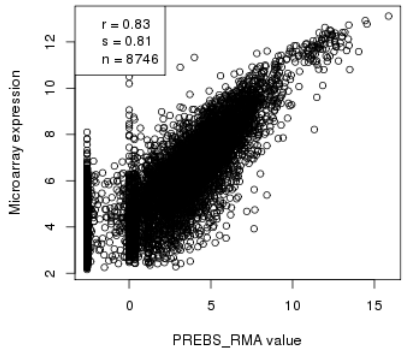
The shapes of output scatter plots using manufacturer's CDF files (Figure 5.8) is similar to the ones using custom CDF files (Figures 5.1 and 5.3), except that the former one has more points. Similar scatter plot shapes and the fact that correlation levels are comparable suggest that PREBS can be successfully used with manufacturer's CDF files in addition to custom CDF files.



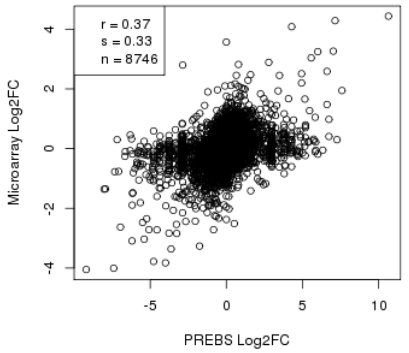
(a) Absolute expression, Marioni *et al.* dataset



(b) Differential expression, Marioni *et al.* dataset



(c) Absolute expression, Cheung *et al.* dataset



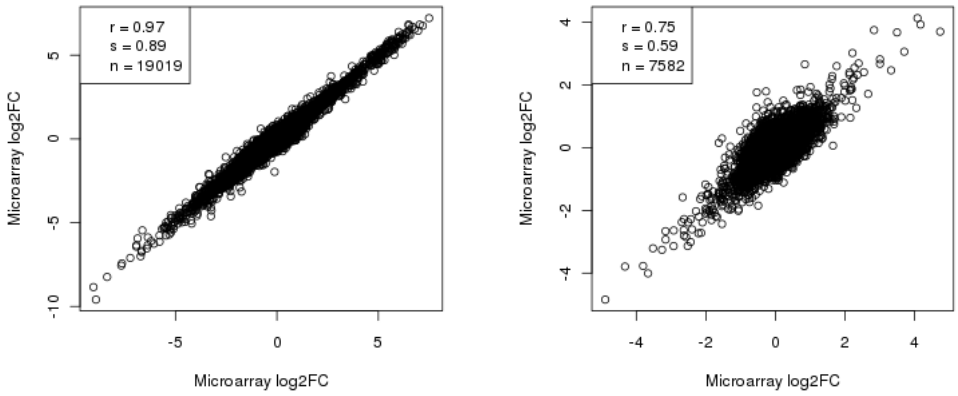
(d) Differential expression, Cheung *et al.* dataset

Figure 5.8: Absolute and differential expression plots for PREBS vs microarray using manufacturer's CDF files.

5.6 Differential expression in microarray technical replicates

Since PREBS correlations for differential expression estimates were not as good as for absolute expression estimates, we wanted to find out what is the best possible correlation for differential expression estimates that any method could reach. For that we analyzed differential expression in microarray replicates from both of the data sets. We calculated differential expression between two samples for two pairs of microarray replicates and then plotted $\log_2$ fold changes according to the first pair of replicates against the second pair of replicates (Figure 5.9). Obviously, the correlation of $\log_2$ fold changes between microarray replicates should be better than the correlation of $\log_2$ fold changes between microarray and RNA-seq platforms. Therefore, microarray replicate $\log_2$ fold change can be seen as the ceiling of the best microarray–RNA-seq $\log_2$ fold change correlation we could achieve.

We found out that the correlation between technical replicates was much better for the Marioni *et al.* data set than the Cheung *et al.* data set. This again confirms the fact that differential expression calculation for the Cheung *et al.* data set is more difficult. Spearman correlation for the Cheung *et al.* data set is particularly low—0.59. Knowing that this within-platform correlation is basically the ceiling of the inter-platform correlation gives some justification of low inter-platform correlations which we saw earlier (Figure 5.3 and 5.7). As we mentioned before, the low differential expression correlations on the Cheung *et al.* data set are probably the result of the fact that gene expression changes in this data set are very small and the measurements are strongly affected by the noise.



(a) Marioni *et al.* data set

(b) Cheung *et al.* data set

Figure 5.9: Microarray replicate differential expression for Marioni *et al.* and Cheung *et al.* datasets. The legend contains Pearson correlation (r), Spearman correlation (s) and the number of genes (n).

Chapter 6

Conclusion

This is the concluding chapter for this thesis. In this chapter, the thesis contents are summarized and obtained results are discussed. Also, some ideas about the future work are provided.

6.1 Summary

This thesis addressed the issue of gene expression comparison for two platforms: microarrays and RNA-seq. A novel computational method, called PREBS, was developed that adjusts RNA-seq data processing in a way that the results are more comparable to microarray results. PREBS results were compared to two other RNA-seq processing methods, RPKM and MMSEQ. PREBS was shown to have the best agreement with microarrays in absolute expression comparison, and have a similar level of agreement with the other methods in differential expression and cross-platform differential expression comparisons.

Additionally, this thesis provided a brief background on gene expression and its measurement. Several different gene expression measurement tools were presented and two of them, RNA-seq and microarrays, are explained in more detail. This thesis also gave an overview of the past microarray–RNA-seq comparisons and studies which were combining/visualizing the inter-platform data.

6.2 Discussion and future work

Measuring gene expression is an important tool for biomedical sciences. Gene expression measurements can be applied to disease diagnostics, new drug development and other areas [1]. Two most popular platforms for measuring

gene expression, RNA-seq and microarrays, give results that are not completely consistent [2, 3]. Therefore, it is important to compare these two platforms and understand their differences.

In the past, there have been studies that made such comparisons experimentally, however, none of them addressed the issue of developing computational methods for making RNA-seq and microarray data more comparable. We have developed and presented a novel method, called PREBS, that addresses this issue. We compared our method against two other RNA-seq processing methods, MMSEQ and RPKM, and we showed that our method has the best correlation with microarrays in absolute expression scale and has similar levels of correlation in differential and cross-platform differential expression scale.

It is difficult to tell why our method cannot reach any correlation improvement in differential expression, even though it reaches a big improvement in absolute expression. One reason could be that differential expression calculation is more complicated because it includes measurements from two samples instead of one, so there is more room for errors and it hard to achieve accurate results. However, it is clear that the future work for our method should concentrate on increasing differential expression correlations, as it is the weakest point of our method.

Other future work for our method could include testing it on more datasets. However, as we mentioned before, good datasets with paired microarray and RNA-seq experiments are hard to find. Moreover, implementing a user friendly version of our method is quite important. The best way for this would probably be to make it available from Bioconductor [82]. Finally, one more thing we could do for our method is to make the dynamic range of PREBS more comparable with microarray dynamic range. At the moment, even though the correlations of PREBS and microarrays are good, the absolute values are still not directly comparable. If we could find a way to make the dynamic ranges more similar, it would extend the number of possible applications of our method.

One of the possible applications of our method could be retrieval of similar microarray–RNA-seq experiments. Microarrays were the dominant gene expression measurement platform for more than a decade, so we have large microarray experiment databases available. One way to reuse that data is to develop a method which for a query microarray experiment retrieves microarray experiments with similar gene expression patterns as was done by Caldas *et al.* [86]. Using PREBS we could extend such method by allowing queries or results to be RNA-seq experiments in addition to microarray experiments.

Another application could be for machine learning related tasks. Let us assume someone is developing a machine learning model which aims to

predict drug response. He wants to train the model on microarray data, because there is more such data available at the moment, but he wants to make predictions also on RNA-seq data in addition to microarray data. In this case, he can use PREBS to process RNA-seq data in order to make input RNA-seq values for machine-learning-based predictor more similar to microarray values. However, in this case, our method could work better, if we could make dynamic ranges of the two platforms more similar.

Potentially, there could be many other types of applications for our method. Together with increasing popularity of RNA-seq technology the need for microarray–RNA-seq experiment comparisons is likely to increase, too. Therefore, in the future, researchers might find new ways of integrating RNA-seq–microarray data and our method might play an important role in their research.

Bibliography

[1] A. Schulze and J. Downward. Navigating gene expression using microarrays–a technology review. *Nat Cell Biol*, 3(8):E190–E195, Aug 2001.

[2] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–1517, Sep 2008.

[3] X. Fu, N. Fu, S. Guo, Z. Yan, Y. Xu, H. Hu, C. Menzel, W. Chen, Y. Li, R. Zeng, and P. Khaitovich. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, 10:161, 2009.

[4] J. H. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol*, 9:34, 2011.

[5] J. Shendure. The beginning of the end for microarrays? *Nat Methods*, 5(7):585–587, Jul 2008.

[6] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra, and S.-A. Sansone. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 31(1):68–71, Jan 2003.

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

[8] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7):621–628, Jul 2008.

[9] E. Turro, S. Su, Â. Gonçalves, L. Coin, S. Richardson, and A. Lewin. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol*, 12(2):R13, 2011.

[10] K. Nill. Glossary of biotechnology terms. 2002.

[11] Y. Zhang and V. N. Gladyshev. High content of proteins containing 21st and 22nd amino acids, selenocysteine and pyrrolysine, in a symbiotic deltaproteobacterium of gutless worm Olavius algarvensis. *Nucleic Acids Res*, 35(15):4952–4963, 2007.

[12] H. Towbin, T. Staehelin, and J. Gordon. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc Natl Acad Sci U S A*, 76(9):4350–4354, Sep 1979.

[13] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS Lett*, 480(1):17–24, Aug 2000.

[14] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. New York: Garland, 5 edition, 2008.

[15] H. D. VanGuilder, K. E. Vrana, and W. M. Freeman. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques*, 44(5):619–626, Apr 2008.

[16] S. J. Wheelan, F. M. Murillo, and J. D. Boeke. The incredible shrinking world of DNA microarrays. *Mol Biosyst*, 4(7):726–732, Jul 2008.

[17] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[18] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, Oct 1995.

[19] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 18(6):630–634, Jun 2000.

[20] M. Babu. Introduction to microarray data analysis. *Computational Genomics: Theory and Application*, pages 225–249, 2004.

[21] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, Feb 2000.

[22] E. M. Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol*, 98(3):503–517, Nov 1975.

[23] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–773, Feb 1991.

[24] P. Angenendt. Progress in protein and antibody microarray technology. *Drug Discov Today*, 10(7):503–511, Apr 2005.

[25] B. Sobrino, M. Brión, and A. Carracedo. SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Sci Int*, 154(2-3):181–194, Nov 2005.

[26] T. C. Mockler, S. Chan, A. Sundaresan, H. Chen, S. E. Jacobsen, and J. R. Ecker. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, 85(1):1–15, Jan 2005.

[27] M. J. Buck and J. D. Lieb. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, Mar 2004.

[28] T. Bammler, R. P. Beyer, S. Bhattacharya, G. A. Boorman, A. Boyles, B. U. Bradford, R. E. Bumgarner, P. R. Bushel, K. Chaturvedi, D. Choi, M. L. Cunningham, S. Deng, H. K. Dressman, R. D. Fannin, F. M. Farin, J. H. Freedman, R. C. Fry, A. Harper, M. C. Humble, P. Hurban, T. J. Kavanagh, W. K. Kaufmann, K. F. Kerr, L. Jing, J. A. Lapidus, M. R. Lasarev, J. Li, Y.-J. Li, E. K. Lobenhofer, X. Lu, R. L. Malek, S. Milton, S. R. Nagalla, J. P. O'malley, V. S. Palmer, P. Pattee, R. S. Paules, C. M. Perou, K. Phillips, L.-X. Qin, Y. Qiu, S. D. Quigley, M. Rodland, I. Rusyn, L. D. Samson, D. A. Schwartz, Y. Shi, J.-L. Shin, S. O. Sieber, S. Slifer, M. C. Speer, P. S. Spencer, D. I. Sproles, J. A. Swenberg, W. A. Suk, R. C. Sullivan, R. Tian, R. W. Tennant, S. A. Todd,

C. J. Tucker, B. V. Houten, B. K. Weis, S. Xuan, H. Zarbl, and Members of the Toxicogenomics Research Consortium. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*, 2(5):351–356, 2005.

[29] Affymetrix. MyGeneChip Custom Array Program. `http://www.affymetrix.com/browse/staticHtmlContentTemplate.jsp?staticHtmlMediaId=m891202&isHtmlStatic=true&navMode=35810&aId=productsNav`, 2012. [Online; accessed 27-April-2012].

[30] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003.

[31] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, 33(20):e175, 2005.

[32] Affymetrix. Statistical algorithms description document. `http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf`, 2002. [Online; accessed 20-June-2012].

[33] G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[34] R. Peck and J. Devore. *Statistics: the exploration and analysis of data*. Duxbury Press, 2011.

[35] G. McLachlan, K. Do, and C. Ambroise. *Analyzing microarray gene expression data*. Wiley series in probability and statistics. Hoboken, N. J: Wiley-Interscience, 2004.

[36] S. Marguerat and J. Bähler. RNA-seq: from technology to biology. *Cell Mol Life Sci*, 67(4):569–579, Feb 2010.

[37] F. Liu, T.-K. Jenssen, J. Trimarchi, C. Punzo, C. L. Cepko, L. Ohno-Machado, E. Hovig, and W. P. Kuo. Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates. *BMC Genomics*, 8:153, 2007.

[38] S. Lee, J. Bao, G. Zhou, J. Shapiro, J. Xu, R. Z. Shi, X. Lu, T. Clark, D. Johnson, Y. C. Kim, C. Wing, C. Tseng, M. Sun, W. Lin, J. Wang, H. Yang, J. Wang, W. Du, C.-I. Wu, X. Zhang, and S. M. Wang. Detecting novel low-abundant transcripts in drosophila. *RNA*, 11(6):939–946, Jun 2005.

[39] D. S. Tawfik and A. D. Griffiths. Man-made cell-like compartments for molecular evolution. *Nat Biotechnol*, 16(7):652–656, Jul 1998.

[40] Illumina. Illumina Sequencing Technology. `http://www.illumina.com/technology/sequencing_technology.ilmn`, 2012. [Online; accessed 8-May-2012].

[41] S. Moorthie, C. Mattocks, and C. Wright. Review of massively parallel DNA sequencing technologies. *The HUGO Journal*, 5:1–12, 2011.

[42] Roche/454. Roche/454 Sequencing Technology. `http://454.com/products/technology.asp`, 2012. [Online; accessed 8-May-2012].

[43] SOLiD. SOLiD Sequencing Technology. `http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html`, 2012. [Online; accessed 8-May-2012].

[44] O. Morozova, M. Hirst, and M. A. Marra. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet*, 10:135–151, 2009.

[45] B. T. Wilhelm and J.-R. Landry. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3):249–257, Jul 2009.

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

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

[48] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, Apr 2012.

[49] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and

isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

[50] P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, May 2012.

[51] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

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

[53] A. Agarwal, D. Koppstein, J. Rozowsky, A. Sboner, L. Habegger, L. W. Hillier, R. Sasidharan, V. Reinke, R. H. Waterston, and M. Gerstein. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*, 11:383, 2010.

[54] J. R. Bradford, Y. Hey, T. Yates, Y. Li, S. D. Pepper, and C. J. Miller. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics*, 11:282, 2010.

[55] J. Beane, J. Vick, F. Schembri, C. Anderlind, A. Gower, J. Campbell, L. Luo, X. H. Zhang, J. Xiao, Y. O. Alekseyev, S. Wang, S. Levy, P. P. Massion, M. Lenburg, and A. Spira. Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. *Cancer Prev Res (Phila)*, 4(6):803–817, Jun 2011.

[56] V. G. Cheung, R. R. Nayak, I. X. Wang, S. Elwyn, S. M. Cousins, M. Morley, and R. S. Spielman. Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol*, 8(9), 2010.

[57] R. Lowry. Concepts and applications of inferential statistics. `http://faculty.vassar.edu/lowry/webtext.html`. Chapter 4. A First Glance at the Question of Statistical Significance.

[58] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, Aug 2008.

[59] P. Łabaj, G. Leparc, B. Linggi, L. Markillie, H. Wiley, and D. Kreil. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, 27(13):i383–i391, Jul 2011.

[60] N. Cloonan, A. R. R. Forrest, G. Kolle, B. B. A. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*, 5(7):613–619, Jul 2008.

[61] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*, 6(5):377–382, May 2009.

[62] R. D. Thiagarajan, N. Cloonan, B. B. Gardiner, T. R. Mercer, G. Kolle, E. Nourbakhsh, S. Wani, D. Tang, K. Krishnan, K. M. Georgas, B. A. Rumballe, H. S. Chiu, J. A. Steen, J. S. Mattick, M. H. Little, and S. M. Grimmond. Refining transcriptional programs in kidney development by integration of deep RNA-sequencing and array-based spatial profiling. *BMC Genomics*, 12:441, 2011.

[63] D. Bottomly, N. A. R. Walter, J. E. Hunter, P. Darakjian, S. Kawane, K. J. Buck, R. P. Searles, M. Mooney, S. K. McWeeney, and R. Hitzemann. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One*, 6(3):e17820, 2011.

[64] Z. Su, Z. Li, T. Chen, Q.-Z. Li, H. Fang, D. Ding, W. Ge, B. Ning, H. Hong, R. G. Perkins, W. Tong, and L. Shi. Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chem Res Toxicol*, 24(9):1486–1493, Sep 2011.

[65] B. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. Penkett, J. Rogers, and J. Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, Jun 2008.

[66] J. Ariño, A. Casamayor, J. Pérez, L. Pedrola, M. Alvarez-Tejado, M. Marba, J. Santoyo, and J. Dopazo. Assessing Differential Expression

Measurements by Highly Parallel Pyrosequencing and DNA Microarrays: A Comparative Study. *OMICS*, Sep 2011.

[67] J. S. Bloom, Z. Khan, L. Kruglyak, M. Singh, and A. A. Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, 10:221, 2009.

[68] S. Liu, L. Lin, P. Jiang, D. Wang, and Y. Xing. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic Acids Res*, 39(2):578–588, Jan 2011.

[69] L. Shi, L. Reid, W. Jones, R. Shippy, J. Warrington, S. Baker, P. Collins, F. De Longueville, E. Kawasaki, K. Lee, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–1161, Sep 2006.

[70] W. P. Kuo, F. Liu, J. Trimarchi, C. Punzo, M. Lombardi, J. Sarang, M. E. Whipple, M. Maysuria, K. Serikawa, S. Y. Lee, D. McCrann, J. Kang, J. R. Shearstone, J. Burke, D. J. Park, X. Wang, T. L. Rector, P. Ricciardi-Castagnoli, S. Perrin, S. Choi, R. Bumgarner, J. H. Kim, G. F. Short, M. W. Freeman, B. Seed, R. Jensen, G. M. Church, E. Hovig, C. L. Cepko, P. Park, L. Ohno-Machado, and T.-K. Jenssen. A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotechnol*, 24(7):832–840, Jul 2006.

[71] F. Battke and K. Nieselt. Mayday SeaSight: combined analysis of deep sequencing and microarray data. *PLoS One*, 6(1):e16345, 2011.

[72] B. A. Friedman and T. Maniatis. ExpressionPlot: a web-based framework for analysis of RNA-Seq and microarray gene expression data. *Genome Biol*, 12(7):R69, 2011.

[73] J. Kim, K. Patel, H. Jung, W. P. Kuo, and L. Ohno-Machado. AnyExpress: integrated toolkit for analysis of cross-platform gene expression data using a fast interval matching algorithm. *BMC Bioinformatics*, 12:75, 2011.

[74] J. Dietzsch, N. Gehlenborg, and K. Nieselt. Mayday–a microarray data analysis workbench. *Bioinformatics*, 22(8):1010–1012, Apr 2006.

[75] F. Battke, S. Symons, and K. Nieselt. Mayday–integrative analytics for expression data. *BMC Bioinformatics*, 11:121, 2010.

[76] P. Warnat, R. Eils, and B. Brors. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6:265, 2005.

[77] X. V. Wang, R. G. W. Verhaak, E. Purdom, P. T. Spellman, and T. P. Speed. Unifying gene expression measures from multiple platforms using factor analysis. *PLoS One*, 6(3):e17691, 2011.

[78] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman & Hall texts in statistical science series. London: Chapman & Hall, 1995.

[79] S. Dowdy, S. Wearden, and D. Chilko. *Statistics for research*, volume 512 of *Wiley series in probability and statistics; 1345*. Hoboken, NJ: Wiley-Interscience, 3 edition, 2011.

[80] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume I. Upper Saddle River (NJ): Prentice Hall, 2 edition, 2001.

[81] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, Feb 2004.

[82] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[83] S. Sherry. NCBI SRA Toolkit Technology for Next Generation Sequence Data. In *Plant and Animal Genome XX Conference (January 14-18, 2012)*. Plant and Animal Genome.

[84] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[85] P. Aboyoun, H. Pages, and M. Lawrence. *GenomicRanges: Representation and manipulation of genomic intervals.* R package version 1.8.3.

[86] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25(12):i145–i153, 2009.