Turun yliopisto
University of Turku

# DATA ANALYSIS TOOLS AND METHODS FOR DNA MICROARRAY AND HIGH-THROUGHPUT SEQUENCING DATA

Asta Laiho

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. AI osa - tom. 529 | Astronomica - Chemica - Physica - Mathematica | Turku 2015

## University of Turku

Faculty of Mathematics and Natural Sciences
Department of Information Technology
Computer Science
The Doctoral Programme in Mathematics and Computer Sciences (MATTI)
Turku Centre for Biotechnology of University of Turku and Åbo Akademi University


## Supervised by


Docent Attila Gyenesei
Department of Information Technology
University of Turku
Finland

Docent Laura Elo
Department of Mathematics
University of Turku
Finland


## Reviewed by


Professor GarryWong
Faculty of Health Sciences
University of Macau
China

Professor Mauno Vihinen
Department of Experimental Medical Science
Lund University
Sweden


## Opponent


Professor Arndt von Haeseler
Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories (MFPL)
University of Vienna and Medical University Vienna
Austria


The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.


ISBN 978-951-29-6354-6 (PRINT)
ISBN 978-951-29-6355-3 (PDF)
ISSN 0082-7002
Painosalama Oy - Turku, Finland 2015

## ABSTRACT

**Asta Laiho**

**Data analysis tools and methods for DNA microarray and high-throughput sequencing data**

The recent rapid development of biotechnological approaches has enabled the production of large whole genome level biological data sets. In order to handle these data sets, reliable and efficient automated tools and methods for data processing and result interpretation are required. Bioinformatics, as the field of studying and processing biological data, tries to answer this need by combining methods and approaches across computer science, statistics, mathematics and engineering to study and process biological data. The need is also increasing for tools that can be used by the biological researchers themselves who may not have a strong statistical or computational background, which requires creating tools and pipelines with intuitive user interfaces, robust analysis workflows and strong emphasis on result reporting and visualization.

Within this thesis, several data analysis tools and methods have been developed for analyzing high-throughput biological data sets. These approaches, covering several aspects of high-throughput data analysis, are specifically aimed for gene expression and genotyping data although in principle they are suitable for analyzing other data types as well. Coherent handling of the data across the various data analysis steps is highly important in order to ensure robust and reliable results. Thus, robust data analysis workflows are also described, putting the developed tools and methods into a wider context. The choice of the correct analysis method may also depend on the properties of the specific data set and therefore guidelines for choosing an optimal method are given.

The data analysis tools, methods and workflows developed within this thesis have been applied to several research studies, of which two representative examples are included in the thesis. The first study focuses on spermatogenesis in murine testis and the second one examines cell lineage specification in mouse embryonic stem cells.

**Keywords:** next-generation sequencing, DNA microarrays, data analysis, gene expression, statistical testing, genotyping, functional analysis, gene-gene interaction analysis, biclustering

# YHTEENVETO (FINNISH SUMMARY)

**Asta Laiho**

## Analyysityökaluja ja -menetelmiä DNA-mikrosiru- ja syväsekvensointimittausaineistoille

Viime vuosina nopeasti kehittyneet bioteknologian tekniikat ovat mahdollistaneet laajojen koko genomin tason mittausaineistojen tuottamisen. Jotta nämä mittausaineistot saataisiin käsiteltyä ja tulkittua tarvitaan luotettavia ja tehokkaita automatisoituja menetelmiä ja työkaluohjelmia. Bioinformatiikka on biologisten mittausaineistojen käsittelyyn ja tulkintaan keskittyvä ala, jossa yhdistetään lähestymistapoja ja menetelmiä tietotekniikan, tilastotieteen, matematiikan ja insinööritieteiden aloilta. Yhä useammat bioalojen tutkijat, joilla ei usein ole kovinkaan vahvaa tilastotieteen tai tietotekniikan osaamista, tarvitsevat helposti käytettäviä ja tehokkaita työkaluja biotekniikan mittausaineistojen analysointiin. Jotta mahdollisimman monet tutkijat pystyisivät hyödyntämään näitä työkaluja, on tärkeää varustaa ne graafisella käyttöliittymäll ja luotettavilla valmiiksi suunnitelluilla analyysiskenaarioilla sekä mahdollistaa havainnollisten tulosraporttien ja -kuvien tuottaminen.

Tässä väitöskirjassa on kehitetty ohjelmistotyökaluja, menetelmiä ja työkulkuja laajojen biotekniikan mittausaineistojen analysointiin, erityisesti geeniekspressio- ja genotyypitysaineistoille. Kehitetyt lähestymistavat helpottavat eri analyysivaiheita, ja koska on tärkeää valita kussakin vaiheessa kokonaisanalyysiin soveltuvia menetelmiä, väitöskirjassa käsitellään myös analyysityökulkuja eri tyyppisille mittausaineistoille. Optimaalisen analyysimenetelmän valinta on usein hyvä suorittaa tarkastelemalla käsiteltävän aineiston ominaispiireitä. Työssä onkin vertailtu eri menetelmiä, minkä perusteella voidaan antaa suosituksia analyysimenetelmän valintaan.

Väitöskirjassa kehitettyjä ohjelmistotyökaluja, menetelmiä ja analyysityökulkuja on käytetty useiden tutkimusaineistojen analysointiin. Kaksi edustavaa esimerkkitutkimusta on sisällytetty tähän väitöskirjaan: ensimmäinen keskittyy hiiren kiveskudoksen spermantuotannon tutkimukseen ja toinen hiiren sikiön kantasolujen solulinjan määräytymisen tutkimukseen.

**Asiasanat:** syväsekvensointi, DNA-mikrosiru, data-analyysi, geenieksprressio, tilastollinen testaus, genotyypitys, funktionaalinen analyysi, geenien interaktioanalyysi, biklusterointi

## CONTENTS

## LIST OF INCLUDED ARTICLES

PI Laiho A, Kiraly A, Gyenesei A. GeneFuncster: A Web Tool for Gene Functional Enrichment Analysis and Visualisation. *CMSB 2012, LNCS 7605*, 2012.

PII Kiraly A, Laiho A, Gyenesei A. Biclustering of high-throughput gene expression data with BiclusterMiner. *IEEE 12th International Conference on Data Mining Workshops*, 2012.

PIII Gyenesei A, Moody J, Laiho A, Semple CA, Haley CS, Wei WH. BiForce Toolbox: powerful high-throughput computational analysis of gene-gene interactions in genome-wide association studies. *Nucleic Acids Res.*, 2012.

PIV Laiho A, Elo L. A Note on an exon-based strategy to identify differentially expressed genes in RNA-seq experiments. *PLoS One*, 2014.

PV Seyednasrollah F, Laiho A, Elo L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, 2015.

PVI Laiho A, Kotaja N, Gyenesei A, Sironen A. Transcriptome profiling of the murine testis during the first wave of spermatogenesis. *PLoS One*, 2013.

PVII Koh KP, Yabuuchi A, Rao S, Huang Y, Cunniff K, Nardone J, Laiho A, Tahiliani M, Sommer CA, Mostoslavsky G, Lahesmaa R, Orkin SH, Rodig SJ, Daley GQ, Rao A. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell*, 2011.

# 1 INTRODUCTION

## 1.1 Background and motivation

Most living organisms comprise of cells containing DNA (deoxyribonucleic acid) which carries the hereditary information across generations. Functionally the genome is divided into genes which are sequences of DNA that can encode RNA (ribonucleic acid) molecules that are used as templates for producing proteins. Proteins are needed for performing the functions within living organisms, determining their phenotypes. In principle, the cells of an organism contain an almost exact copy of the DNA although they can have very distinct appearances and functions and respond differently to their environment. This is possible due to the differences in the produced proteins and their abundances which depend on the present cell states and current external signals. The protein levels are governed by the complex regulatory system encoded in the DNA sequence and structure [49].

Since the discovery of DNA structure by James Watson and Francis Crick in 1953, the field of molecular biology has advanced enormously. The appearance of novel biotechnological approaches such as DNA microarrays at the end of the 1990's and more recently the high-throughput (often referred to as next-generation) sequencing (HTS), has enabled the measurement of biological signals at a whole genome level in an efficient and accurate manner [60]. These novel technologies produce vast and ever growing data sets that require reliable and efficient methods and tools for data processing and analysis on various levels. This requirement has given a rise to the field of *bioinformatics*, which combines methods and approaches across computer science, statistics, mathematics and engineering to study and process biological data.

Microarray and high-throughput sequencing based technologies in general allow studying various different aspects of DNA and RNA at a whole genome level. Possible applications include for example DNA sequence variation analysis, transcriptome analysis (i.e. study of the transcribed elements of the genome such as genes) and epigenome analysis (i.e. study of traits that are heritable but not caused by changes in DNA sequence) to study various regulatory mecha-

nisms. In this thesis, the focus is on the analysis and interpretation of microarray and HTS data, mainly concerning variant and especially gene expression analysis, although some of the approaches used are more broadly applicable across a much wider range of different applications.

When proteins are produced, DNA is first transcribed into messenger RNA molecules (mRNAs) which will then be used as templates for proteins (see Fig. 1). Thus by studying the differences in the abundances of mRNA molecules it is in principle possible to compare the differences in the levels of proteins produced between given biological samples. However, due to the complex nature of gene expression process and limitations of the measurement technologies, mRNA levels do not perfectly correlate to the observed protein levels [59]. Therefore, efficient and robust data analysis techniques and data driven method choices are needed in order to make correct biological interpretations from this kind of noisy measurement data. In addition, strong experience in data analysis and good knowledge of various kinds of analysis tools and methods are required from a data analyst. Typically the main goal of a gene expression study is to identify genes whose expression levels significantly differ between two or more sample groups. For example, to understand the effect of a treatment such as exposure to varying conditions (e.g. extreme heat or coldness), we may ask which genes are up-regulated (increased in expression) or down-regulated (decreased in expression) between treatment and control groups. In addition, mapping the altered genes to biological processes and other higher level biological categorizations is highly important in order to interpret the results.

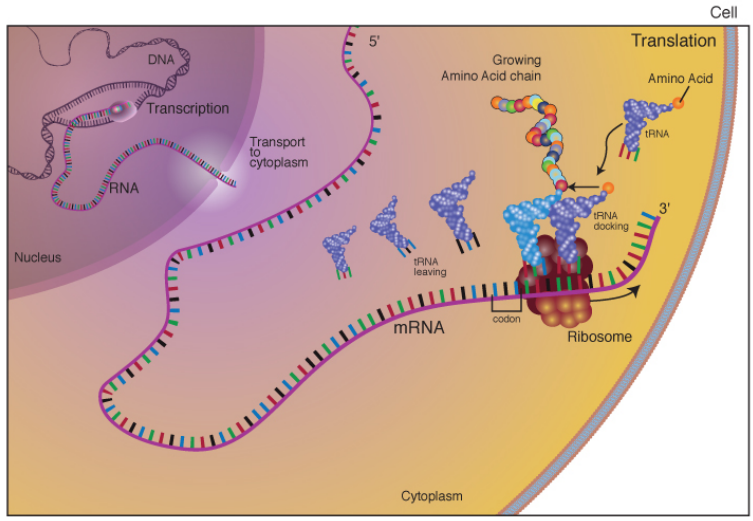



FIGURE 1 Central dogma of molecular biology. DNA is first transcribed into messenger RNA (mRNA) and then translated into amino acid chain forming a protein. This process is called gene expression. Figure from [69].

Numerous different factors and processes determine the gene expression levels in any given cell at any given time, including the small differences in DNA sequence between individuals or the different cells. These variations; single nucleotide polymorphisms (SNPs) and small insertions and deletions are dispersed throughout the 3-billion-base human genome every 100 to 300 bases. They can occur in both coding and non-coding regions of the genome and many of them have no effect on cell function while others strongly affect gene expression and may predispose people to disease or influence their response to a drug or some other factor. The general aim of the variant analysis typically is to detect differences that can be linked to certain physical traits or help explain how diseases are developed or how humans respond to pathogens, chemicals, drugs, vaccines, and other agents. In genetic epidemiology, the term *genome-wide association study* (GWAS) is commonly used when the aim is to detect variants associated with traits or diseases. In addition to studying individual SNPs, approaches for studying the complex interactions between the variant loci have recently started to emerge [31].

## 1.2 Objectives of the thesis

In practice, a number of different data analysis steps is required in order to produce a comprehensive view of the underlying biology within any data set, broadly divided in preprocessing and downstream data analysis parts. Currently, many bioinformatics method developers are focusing on a specific step or a couple of steps and consider these in almost complete isolation of the rest of the data processing flow. However, in the worst case for example the wrong choice of a preprocessing method can lead to biased or even erroneous results. Thus there is a strong need for more comprehensive approaches where the different parts of the data analysis are considered together in order to ensure coherent handling of the data throughout the analysis. As the number of data sets produced is growing fast, there is also an increasing need for tools that can be used by the biological researchers themselves who may not have a strong statistical or computational background. This requires creating pipelines with intuitive user interfaces, robust analysis workflows and strong emphasis on result reporting and visualization. In this thesis, these important aspects have been carefully taken into consideration in the development of data analysis methods and tools and in designing data analysis workflows for studying complex biological settings.

Specifically, the objectives of the thesis are as follows:

1. To develop efficient methods, user-friendly tools and robust workflows for analyzing high-throughput biological data.

2. To apply these tools and methods for biological research questions.

Publications I-IV of the thesis introduce the details of the novel data analysis tools and methods specifically aimed at gene expression and genotyping data analysis, while Publication V presents a comparison of the statistical testing tools for RNA-sequencing data. These methods, tools and approaches have been used in many research studies with results published in international peer-reviewed journals, for example [28, 73, 33, 40, 65, 37, 94, 93, 44, 42]. Two representative examples where the high-throughput data and its analysis plays an important role have been included in this thesis: 1) a study on spermatogenesis on murine testis (Publication VI) and 2) a study on cell lineage specification in mouse embryonic stem cells (Publication VII). Both of these studies illustrate the importance of careful consideration of all the various data processing steps from raw data to result interpretation and visualization in order to generate deeper biological insight.

## 1.3 Outline of the thesis

This thesis consists of four parts: the first part (Chapters 1-3) gives a general introduction to the thesis and its research goals and introduces the measurement technologies used for generating the types of data targeted within the thesis and gives an overview on the basic data analysis steps typically applied for these data. The second part (Chapter 4) describes the methods used within the thesis combined with the necessary background information, putting the methods also in context with the broader data analysis workflow. Third part (Chapters 5-6) summarizes the results of the publications included in the thesis and gives the general conclusion of this work. The last part (Appendix: included articles) contains the original publications included in the thesis.

# 2 HIGH-THROUGHPUT MEASUREMENT TECHNOLOGIES FOR MOLECULAR BIOLOGY: DNA MICROARRAYS AND HIGH-THROUGHPUT SEQUENCING

## 2.1 DNA microarrays

DNA microarrays have become a widely used standard tool in molecular biology during the last decade and they can be used for a number of purposes including gene expression profiling and alternative splicing analysis, comparative genomic hybridization analysis (CGH) to discover genetic amplifications and deletions, chromatin immunopercipitation on chip (ChIP) to detect binding sites of DNA binding proteins or genotyping by single nucleotide polymorphism (SNP) detection and fusion gene analysis [98]. In this thesis the main focus is on gene expression and genotyping (SNP) analysis.

DNA microarray technology can be used for measuring the relative abundance of the biological sequences of interest in a given sample. The technology is based on a use of fluorescent labeled and slide attached interrogation probe sequences [98, 89]. The technology takes advantage of the ability of the complementary single-stranded sequences of nucleic acids to form double stranded hybrids. After removing the unattached sequences by washing, a laser is used to excite the attached fluorescent dyes to produce light which is then detected by a confocal scanner. The scanner generates a digital image from the excited microarray and the digital image is further processed by specialized software to transform the image of each spot to a numerical reading. These numeric values are considered as the relative target sequence concentrations which can then be compared between different samples. The DNA microarray principle in the context of gene expression analysis is further illustrated in Figure 2. In the early days of the microarrays they were often manufactured by custom spotting complete probe sequences onto glass slides using a pin-spotting device [89]. However, nowadays there are many vendors providing commercial high quality catalogue or custom

microarrays based on advanced technologies such as base by base construction of probe sequences directly on the slide by photolithography [89]. Microarrays provided by different vendors, even for the same purpose, vary for example in their fabrication principle, probe length and design, accuracy, efficiency and cost.

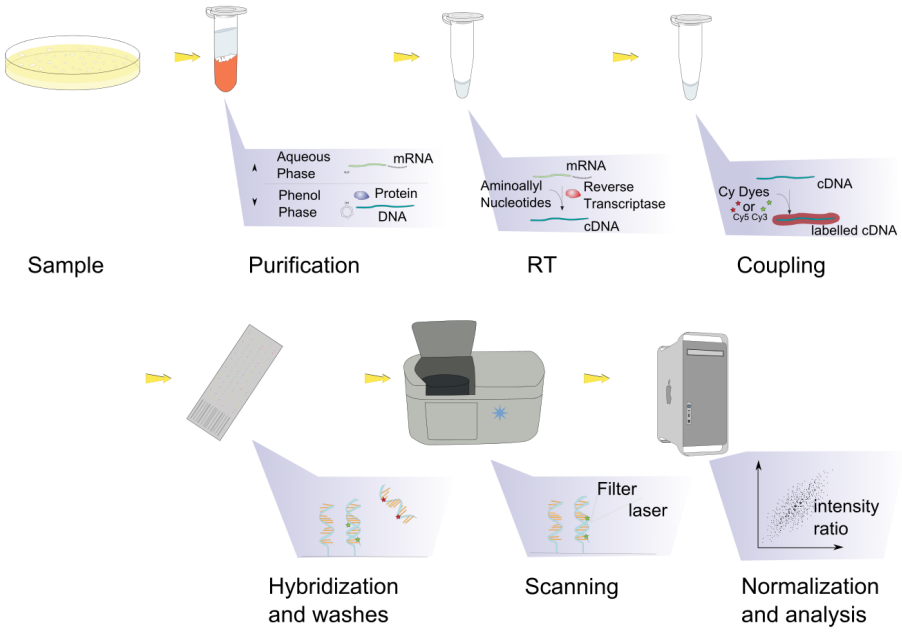



FIGURE 2 Microarray principle. mRNA is first extracted from sample and then reverse transcribed to complementary DNA (cDNA). After this the sample is labelled. Next the labelled sample is applied to a microarray where sequences matching an interrogation probe sequence will be hybridized to the array. Finally signals are scanned and quantitated with a scanner software after which the data is ready for analysis. Figure adapted from [1].

### 2.1.1 Gene expression microarrays

Gene expression microarrays are the most widely used type of microarrays and they are typically used for measuring relative messenger RNA (mRNA) expression abundances between samples across known genes. The largest manufacturers currently are Affymetrix, Agilent and Illumina [89]. In Affymetrix technology, 25-mer probes are printed to the array base by base in a process that employs a combination of chemistry and photolitography. Each gene is represented by a set of probes distributed across the full length of the gene or near the 3′ end of the gene depending on the array type. The probe values are typically summa-

rized for each gene during the preprocessing of the data. Agilent applies 60-mer probes that are deposited onto specially prepared glass slides, base by base, using an inkjet printing process. Most genes are represented by a single probe, some by a couple of different probes. In Illumina technology, the 50-mer probes are bound to magnetic beads randomly distributed across the microarray. Specific index sequences are then used to decode the identity of each bead which allows mapping to genes. Most genes are represented by a single probe sequence, some by a couple of probes for different isoforms of the gene. Each probe appears in 15 to 30 copies within each microarray and these technically replicated measurements are then combined during the automatic data preprocessing. From all vendors, custom microarrays are also available that can be designed according to specific needs of the study in question.

### 2.1.2 Genotyping microarrays

Genotyping or SNP arrays can be used for detecting known single nucleotide polymorphisms. There are around 150 million SNPs that have been identified in the human genome [3]. The basic principles of SNP arrays are the same as for the gene expression microarrays: DNA hybridization, fluorescence microscopy, and solid surface DNA capture. For simplicity, manufacturers often arbitrarily label the two alleles of a SNP as A and B. Therefore, since each individual usually inherits one copy of each SNP position from each parent, the individual's genotype at a SNP site is typically either AA (homozygous reference), AB (heterozygous) or BB (homozygous alternate). Genotyping microarrays are then based on interrogation probes designed against these three alternative genotype clusters for the two alleles. Using the same arrays it is also typically possible to detect small insertions and deletions or loss of heterozygosity (LOH). LOH occurs when one allele of a gene is mutated in a deleterious way and the normally-functioning allele is lost, a phenomena often observed in oncogenesis.

## 2.2 High-throughput sequencing

During the recent years, high-throughput (or next-generation) sequencing technologies have rapidly gained popularity by parallelizing the sequencing process and producing concurrently up to hundreds of millions of sequences. These technologies are generally used for the same purposes as the DNA microarrays and for many applications they already provide a cost effective competitive alternative. Popular technology platforms for high-throughput sequencing during the recent years have been Thermo Fisher Scientific's (earlier Life Technologies) SOLiD and Ion Torrent, Roche's 454 and PacBio's RS in addition to Illumina, whose MiSeq and HiSeq platforms currently by far remain the most popular platforms in use [100]. As most of the current deep sequencing data is produced using Illumina platform, this technology is here used as an example in introducing the

modern sequencing technologies.

llumina sequencing technology is based on sequencing by synthesis where the bases of the fragmented sample material are sequentially identified from signals emitted as each fragment is re-synthesized from a DNA template strand [100]. The Illumina sequencing principle is further illustrated in Figure 3. The technology is currently able to produce up to 300 base pair sequence reads. In contrast to microarray technology, HTS technology can easily be tuned to provide a variable resolution depending on the needs of each project. This can be done by adjusting the coverage generated for a particular type of experiment, coverage generally referring to the average number of sequencing reads that align to each base within the sample. For example, a whole genome sequenced at 30x coverage means that on average each base in the genome is covered by 30 sequencing reads. Increased coverage thus improves the resolution of the analysis by increasing the sensitivity of the detection of gene expression (especially for lowly expressed genes) and genetic variants. While microarrays measure continuous signal intensities, HTS thus quantifies discrete, digital sequencing read counts.

In addition to studying mRNA expression, RNA-sequencing (RNA-seq) can simultaneously be potentially used for analysing novel transcripts and isoforms, alternative splice sites, gene fusion and SNPs in a single experiment [64]. While relatively short reads (50-75bp) and single-end sequencing approach are typically sufficient for basic gene level analysis, longer read length (>75 bp), higher coverage and paired-end sequencing is usually required for more in-depth analyses such as the detection of alternative splicing events. With paired-end sequencing both ends of the sample fragments are sequenced in order to improve the precision of the read alignment to the reference genome and enhance the sensitivity of the downstream data analysis. Also the number of reads required per sample varies depending on the goals of the study. For gene expression level analysis, it is typically sufficient to use mRNA as starting material as the analysis concentrates on the protein coding genes. In order to detect also non-coding transcripts total RNA can be used as input. In this case, however, the number of reads required per sample is much higher. Table 1 summarizes Illumina's latest recommendations regarding the sequencing specifications for RNA-sequencing. The effect of sequencing specifications on downstream analysis of RNA-seq has also been investigated in recent studies [85, 108].

DNA resequencing can be performed for whole genomes (typically in low resolution) but it is more common to run targeted resequencing of exomes (the transcribed part of the genome) or other specified regions to produce higher sequencing read coverage cost effectively on the most interesting regions [10, 32]. Exonic protein coding regions, although representing less than two percent of the human genome, yet contain the majority of known disease causing mutations. Resequencing can be used for revealing single-nucleotide variants (SNVs), small insertions and deletions and large structural variants (such as inversions or translocations) and copy number variants (CNVs). According to a recent study by Meynert et al. [63] exome-seq achieves 95% SNP detection sensitivity at a mean on-target depth of 40 reads, whereas WGS only requires a mean of 14 reads.

| Applications | Read type | Read depth/sample (mRNA/total RNA) |
|---|---|---|
| Gene profiling (gene-level counts) | 1 x 50 bp | >5 M / >10 M |
| Discovery (alternative transcripts, gene fusions, etc.) | 2 x 50 - 75 bp | $\geq$ 50 M / >100 M |
| Complete transcriptome annotation | 2 x 75 - 100 bp | $\geq$ 100 M / $\geq$ 200 M |

TABLE 1 Sequencing platform manufacturer Illumina's recommendations for sequencing specifications for different RNA-sequencing experiment types [39]. Read type is given as a combination of single end sequencing (1 x) or paired-end sequencing (2 x) and read length in base pairs (bp). Read depth is given per million reads (M).

High-throughput sequencing provides several benefits over microarrays and is rapidly gaining popularity over them [111, 68]. For example, HTS results can be updated any time new reference sequence information is obtained while microarrays are limited to the reference information available during the design of the array. In principle HTS also enables the discovery of novel sequences, although typically the data are analyzed against known genomic features. Hybridization issues seen with microarrays, such as cross-hybridization or non-ideal hybridization kinetics are also eliminated in sequencing experiments. HTS can also potentially provide larger dynamic measurement range as the sensitivity of the detection can be increased by sequencing deeper (i.e. producing more reads for a sample). In general, the number of reads required for the analysis largely varies depending on the application, the goals of the analysis and the sample material itself. As a downside, handling of HTS data is more complicated and the analysis pipelines not as well established yet, as for microarrays.

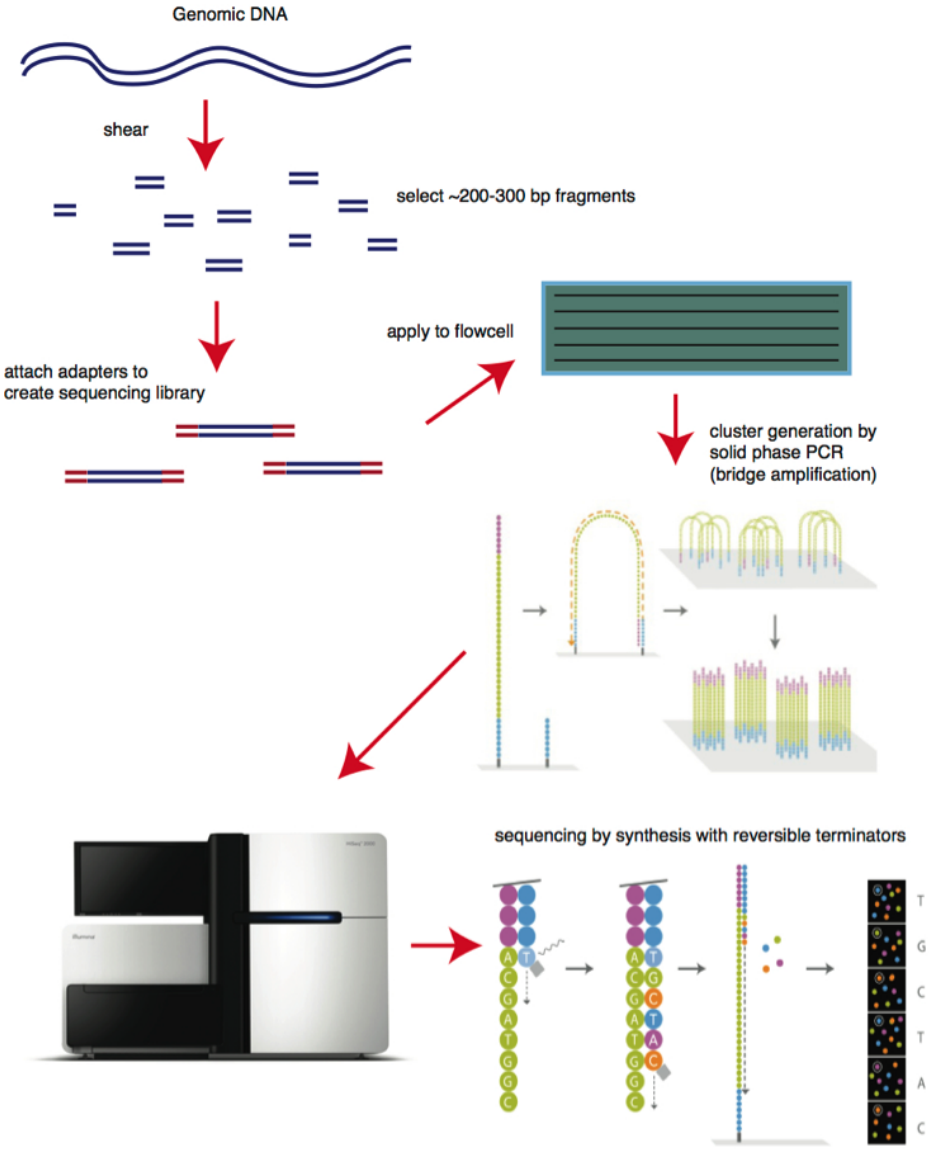



FIGURE 3 Illumina sequencing principle. The genomic DNA is first sheared to fragments of around 200-300 nucleotides. Sequencing adapters are next ligated to the sequence fragments which are then applied to a sequencing flowcell. Next the sequences are amplified in order to produce clusters with a large number of identical copies of the sequences to be analyzed. During the sequencing run the sequences of each cluster are read base by base using fluorescent labeled terminator molecules which can be detected by a camera. [15].

# 3 OVERVIEW OF TYPICAL DATA ANALYSIS PIPELINES

In order to interpret the results of the high-throughput biological experiments, several consecutive data analysis steps need to be applied for the given data set. These can be broadly divided into preprocessing and down-stream analysis parts. Preprocessing is typically needed to evaluate the quality of the data and to correct for potential technical biases in order to ensure data comparability. Although some of the approaches used are application specific, many of them can be applied more broadly to different data types. The data also need to be processed to a format feasible for statistical analysis typically performed to produce results whose reliability is indicated by a significance score (typically p-value). Depending on the experimental setup and data type, various analysis methods can be applied in order to interpret the biological relevance of the measurements. In this chapter, the different steps required in a typical high-throughput data analysis pipeline for gene expression and variant data are introduced. Preprocessing is described in detail, separately for microarrays and next-generation sequencing data. For the convenience of the reader, the pipelines are summarized visually: gene expression data analysis pipeline in Figure 4 and NGS variant analysis pipeline in Figure 5.

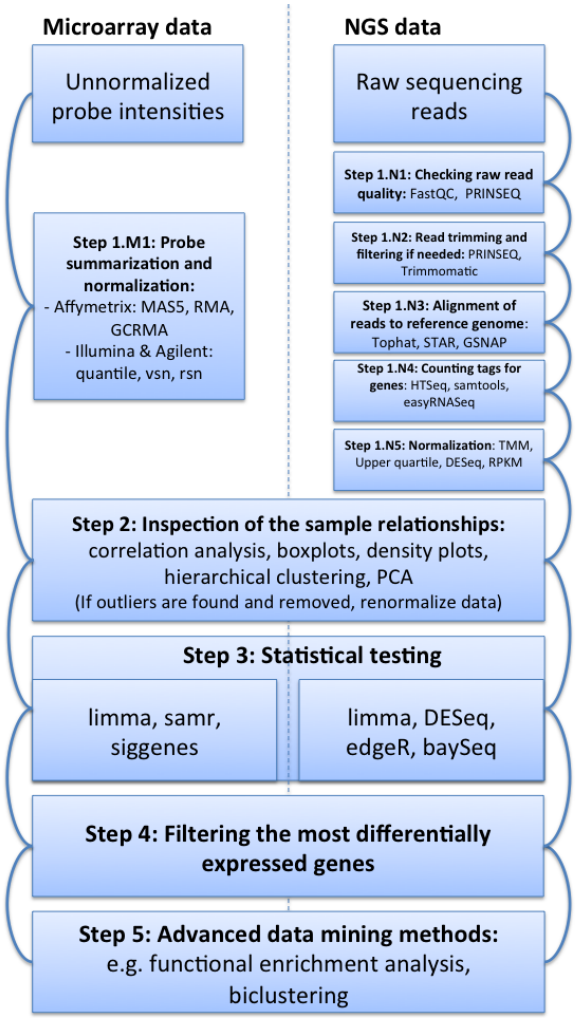



FIGURE 4 Typical data analysis pipeline for microarray and high-throughput sequencing gene expression data with common tools and methods.

## 3.1 Gene expression microarray data preprocessing

Preprocessing (Figure 4, steps 1-2) can be divided into platform specific and general parts which fundamentally differ between microarray and high-throughput sequencing data. For microarrays the scanner software automatically quantitates the scanner images into numeric intensity values corresponding to the relative abundance of the target sequences in the sample [23]. Quantitation includes various signal processing steps for example for reducing background and handling outlier measurements. The scanner output file format depends on the platform; Affymetrix provides binary CEL files while Agilent and Illumina produce delimited text files. Data analysis software tools have specific methods for handling output from the different platforms.

Many free and commercial software packages are available for data analysis, R/Bioconductor [92, 30] being the most popular free tool among them. In R there are several packages available for preprocessing data from each microarray platform (Figure 4, step 1.M1). Preprocessing methods for Affymetrix gene expression data typically contain steps for summarizing probe level signals to gene level values while for Illumina and Agilent the probe values are not summarized due to the genes being represented with only one or a couple of different probe sequences. Normalization is also applied in order to remove non-biological variation and to make the measurement values comparable across the sample set. With Affymetrix the summarization step is coupled with the normalization, MAS5 (Microarray suite 5), RMA (Robust multiarray average) and GCRMA (GC - Robust multiarray average) being the most popular methods [78]. For Agilent and Illumina the most popular normalization approaches are quantile, vsn (variance stabilization normalization) and rsn (robust spline normalization) [82]. Normalization for gene expression data is typically based on the general assumption that only a small proportion of the assayed genes are differentially expressed between the samples and that roughly an equal proportion of the genes are up-regulated and down-regulated. Thus the signal distribution across samples is expected to roughly follow the same pattern. Expression intensity values are also log transformed in order to make the value distribution better suitable for statistical testing methods.

Careful inspection of the data quality is essential for ensuring the reliability of the analysis results. This covers reports of the sample material quality analysis performed by the microarray facility that has performed the experiments as well as platform specific hybridization reports which help in evaluating the technical quality of the samples and the hybridization process performance. Various data mining and related visualization techniques can be used for exploring the data set [23] (Figure 4, step 2). Measurement value distributions for the samples can be inspected using for example boxplots and density plots. Correlation analysis and various clustering based approaches such as hierarchical clustering and principal component analysis are typically used for exploring the relationships between samples. This way it is possible to see how the sample groups and the

replicates within them relate to each other in a general level and to detect possible outliers. If outliers are observed, it needs to be considered whether they should be removed prior to downstream analysis.

## 3.2 Genotyping microarray data preprocessing and analysis

For Affymetrix genotyping microarrays, Affymetrix Genotyping Console (GTC) Software [4] is commonly used to preprocess the data and to perform the quality analysis and SNP calling. Illumina data can be analyzed using the array manufacturer's GenomeStudio software [38] while Agilent Genomic Workbench [5] is available for Agilent data. In addition to the array manufacturers' tools there are also alternative software packages available for genotyping data analysis (e.g. crimm [16], RLMM [77] and beadarraySNP [71]). Typically the genotyping analysis consists of the quality control and genotype calling (e.g. using Birdseed algorithm [4]) steps. Depending on the array type, it may additionally be possible to perform copy number or loss of heterozygosity analysis from the same data.

## 3.3 High-throughput gene expression sequencing data preprocessing

For high-throughput sequencing data the initial output from the sequencing instrument are the raw read files that also contain the base call quality values. It is important to check the sequencing read quality information (Figure 4, step 1.N1) using tools like FastQC [9] or PRINSEQ [83] which can provide information for example on sequence length, base and GC content, quality scores, sequence duplication levels and overrepresented sequences. Read trimming and filtering (Figure 4, step 1.N2) may be needed to account for example for reads with low general quality or dropping quality towards the ends of the reads or reads containing sequencing adapters, and can be performed using for example PRINSEQ or Trimmomatic [13].

The next step in processing HTS data typically is to align the reads to a genome reference (Figure 4, step 1.N3). Human, mouse and many model organisms have a good quality reference genome and gene annotation available which usually allows a good alignment rate even with stringent alignment scheme. Various different alignment tools are available for this purpose such as Tophat [96], STAR [22] or GSNAP [109] for transcriptome data and Bowtie [47] and BWA [50] for variant data. Read alignment is computationally a very intensive process and it is typically performed using a powerful computer cluster. It is also possible to analyze HTS data when reference genome is not complete or it is completely missing although this is currently considered a highly error prone approach. In this situation full or reference based assembly approaches can be used to construct

the reference based on the sequencing data prior to the read alignment. Various different tools are available for transcriptome and genome read assembly [26]. Read alignment reports give useful information on the experiment quality - the more uniquely aligned reads the better as only these are typically used for the downstream analysis. Low number of aligned reads or an elevated number of duplicated reads may suggest technical problems in the sample processing.

In order to carry out gene-level expression analysis, the next step after the read alignment is to count the reads associated with the annotated genes (Figure 4, step 1.N4) (using e.g. HTSeq [7], samtools [51] or easyRNASeq [19]). In this step the number of reads overlapping the exonic gene regions are counted to produce a total read count for each gene. If the data is analyzed for alternative splicing, the counts can be summarized on transcript or exon level as well. Alignment and count calculation are both complex tasks involving several adjustable parameters. However, according to our investigation (Publication V), the default parameters may work reasonably robustly in many settings in the context of transcriptome analysis. Methods also exist for reference based novel transcript discovery and abundance estimation (e.g. [97]) although this is still generally considered a very challenging task.

The choice of the normalization method for RNA-seq data is commonly coupled with the statistical analysis method used (Figure 4, step 1.N5). Count data normalization methods primarily aim at dealing with the variable sequencing depths across samples making the read counts generally higher in some samples compared to others. Various scaling based methods are typically used (see [21] for a comprehensive review). Count values are also often transformed to *RPKM (reads per kilobase (i.e. thousand bases) per million mapped reads)* normalized read counts which may be helpful in getting a better overview of the expression levels across genes as they also normalize against the variable gene length. However, these values are generally not recommended to be used in the context of statistical testing methods as they depend on the mean expressed transcript length [102].

## 3.4 Statistical testing for high-throughput gene expression data

The primary goal of a gene expression study typically is to identify genes whose expression levels differ between two or more sample condition groups. For example, to understand the effect of a treatment such as exposure to varying conditions (e.g. extreme heat or coldness), we may ask which genes are up-regulated (increased in expression) or down-regulated (decreased in expression) between treatment and control groups. Typically the analysis is carried out against known genes found in RefSeq [75] or Ensembl [29] databases. Methods also exist for assembling the gene and transcript models from the sequencing data and calculating abundance estimates based on these models and analysing differential expression of isoforms based on exon-level expression signals but these still remain challenging tasks [8].

Initially, comparative experiments were done even with few, if any replicates, and statistical criteria were not used for identifying differentially expressed (DE) genes. Instead, simple criteria were used such as fold-change, with 2-fold being a popular cut-off. Nowadays the requirement for having replicated measurements for conducting gene expression studies is highly recognized although the high experimental cost still in many cases limits the number of replicates measured. Thus, studies with two or three biological replicates per sample group are still common, which poses challenges for the analysis methods used. Experimental design and the number of required biological replication has been recently discussed for example in [108]. Liu et al. [53] also indicated that increasing the number of biological replicates leads to better sensitivity in detecting differentially expressed genes compared to sequencing fewer samples with increased sequencing coverage. Nowadays, technical replicates are rarely measured as the technical variation within the modern commercial systems is typically insignificant compared to the between-sample biological variation.

Today, a plethora of different statistical tools and methods are available for the statistical analysis of both microarray and HTS gene expression profiling data, many of the most popular packages being based on R/Bioconductor software for statistical computing [77, 30]. The largest difference between microarray and next-generation sequencing data is that microarrays produce continuous values while NGS generates discrete count values and thus different approaches are needed for handling the two different types of data. Popular packages for microarray analysis include for example limma based on linear modeling [87], samr [99] based on the non-parametric SAM algorithm taking advantage of randomized permutations and siggenes [84] based on the SAM and EBAM [25] methods. Commonly used RNA-seq data analysis packages include for example DESeq [6], edgeR [80] and baySeq [34] based on negative binomial models and limma [79] based on transformations of read counts for linear modeling.

Differentially expressed genes are typically filtered using cutoffs for both statistical significance score (typically corrected p-value such as false-discovery rate [11]) and expression fold-change. P-value correction is needed in order to account for the multiple testing of thousands of genes, easily resulting in false positive findings if not controlled for. Although thresholds such as 0.05 for the p-value and two for the fold-change are often applied by default, there is not a single correct way or method to determine the thresholds. In fact, it is advisable to base the choice of the filtering thresholds on the individual characteristics of the data set in question and the purpose of the filtered gene list [6]. For example, when the gene list is directly included in a publication it is important to minimize the false positive findings and thus apply very strict filtering thresholds. On the other hand, the filtering criteria can be markedly relaxed when the produced gene list will be used as input for functional enrichment analysis where the focus is in general trends rather than individual genes. In this case a longer gene list may help in detecting more subtle trends as the statistical power is increased, while random false positive genes are not very likely to have a strong influence on the result of the test. Different visualizations such as hierarchical

clustering dendrograms, PCA plots, volcano plots and MA-plots are also helpful for choosing the filtering cutoffs [72]. Varying by the study and the purpose of the filtering, the length of the list of the DE genes can be anything between a handful or hundreds of genes; typically the aim in any case is to list a reasonable number of the strongest influenced genes.

## 3.5 Variant sequencing data preprocessing and analysis

The typical analysis pipeline for resequencing data is presented in Figure 5. Quality control and read alignment steps are similar to those for the gene expression data. After the initial read alignment of the resequencing data, sequence variants are detected using a variant calling method. Methods also exist for detecting larger structural rearrangements, such as deletions or duplications of whole chromosomal arms, but in this thesis the focus is on the detection of small insertions, deletions and single nucleotide polymorphisms. For resequencing data it is possible to detect known as well as novel sequence variants (see [67] for a comprehensive review on variant analysis). Typical steps in the variant calling programs include include realignment of reads, removal of duplicated reads and recalibration of quality scores in order to increase the sensitivity and reduce false positive call rate. Read filtering is also important as only the highest quality reads and alignments are considered for variant analysis to avoid problems caused by single base errors easily occuring with the current sequencing technologies. Popular tools include for example GATK [61] and samtools [51]. Depending on the experimental design and the goals of the study the variant calling is then applied on the samples individually or using a multi-sample procedure. In the latter case a shallow sequencing of a large number of samples has typically been produced and the interest is in the general population level differences rather than single individuals. In this case also other prior information such as allele frequencies and patterns of linkage disequilibrium are commonly used to enhance the analysis. In order to interpret the variant calling results, it is important to connect them with information on the previously known variants and their known effects, nearest genes and the predicted consequences of the previously unknown variants. Many free and commercial tools exist for SNP prioritization, such as ANNOVAR [103] HaploReg [104] and RegulomeDB [14]) and for variant annotation, e.g. SPOT [81] and SNPranker [62].

## 3.6 Summary

In this chapter the basic steps included in a typical analysis pipeline of high-throughput gene expression and genotyping data have been introduced, many of which can also be applied to other types of high-throughput data sets as well.

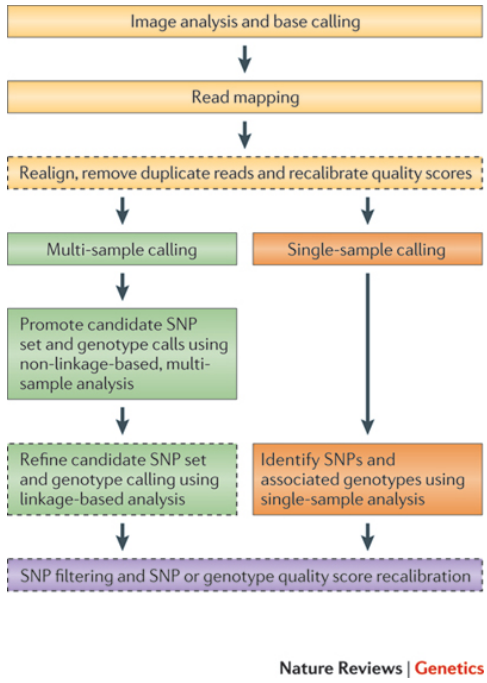



FIGURE 5 Variant data analysis pipeline (Figure from [67], reprinted by permission from Macmillan Publishers Ltd: [Nature Reviews Genetics, copyright 2011). Pre-processing steps are shown in yellow, multi-sample calling in green, single-sample calling in orange and post-processing in purple. Optional steps are shown in dashed lines.

For example normalization, quality analysis and statistical testing are needed for most high-throughput data sets and are a prerequisite for conducting more advanced analysis. When choosing the method for each step, it is important to consider the prerequisites for different tools and their compatibility in order to perform a coherent analysis. Thus it is important to plan the general data analysis flow keeping in mind the limitations and requirements of the chosen methods at each step. An experienced data analyst is typically needed for conducting a fully streamlined analysis and thus seizing the full potential of a given data set. Within this thesis, a carefully considered data processing workflow, covering many of the above represented analysis steps, was developed for the gene expression analysis of high-throughput data and applied for two real experimental setups (Publications VI and VII).

# 4 METHODS AND TOOLS FOR HIGH-THROUGHPUT DATA ANALYSIS

The methods and tools developed in this thesis are presented in this chapter with further background information on the related topics. Section 4.1 concentrates on differential gene expression detection for high-throughput sequencing. To date, many algorithms and software packages have been published for this purpose but as there is currently no consensus on which methods should be applied under different conditions we conducted a comparison of the existing approaches. Thus, our aim was to generate useful information to help researchers in selecting a robust and usable method under various circumstances. Independent of the method selected, in order to carry out gene-level differential expression analysis the data are typically summarized across exons prior conducting the statistical testing. Within this thesis our aim was also to investigate whether the detection of differential gene expression could be improved by an alternative strategy of conducting the testing on exon level prior to summarizing the results at gene level.

Many biotechnical high-troughput measurement technologies generate lists of genes as the primary result, the length of which can range from a handful to many hundreds or even thousands. Therefore functional enrichment analysis, described in Section 4.2, has become an important tool aiding the interpretation of such gene lists. In general, functional enrichment analysis refers to analyses taking the gene functional annotations into account and focusing on co-operational gene modules rather than individual genes. In order to facilitate an efficient functional enrichment analysis, a new web based tool, GeneFuncster, was developed within this thesis.

The topic of Section 4.3 is biclustering which is a data mining approach that can be applied to identify groups of genes following the same expression pattern across a set of samples (and thus likely belonging to the same functional module). Biclustering can also be used for sample classification by grouping together the samples with the most similar expression patterns across genes. Novel BiclusterMiner tool is presented as an efficient solution to detecting co-operational gene modules within gene expression data.

Section 4.4 introduces the gene-gene interaction analysis (also called epistasis analysis) for genome wide association studies based on high-throughput variant/genotyping data. BiForce Toolbox is also presented, a novel tool developed within this thesis for epistasis detection.

## 4.1 Differential gene expression analysis of high-throughput sequencing data

Methods for detecting differential expression in high-throughput sequencing data are currently under rapid development: new packages are frequently published and the existing algorithms are often revised. Thus it can be challenging for a researcher to choose a method for differential expression analysis of their data set. To help this choice, comparison studies have been recently published, discussing the different methods' performance, strengths and weaknesses and applicability to different kinds of data sets [101, 45, 88]. However, these comparison studies are based on only simulated data sets or they include very few biological replicates. Thus, in Publication V, we investigated the performance of the different methods on real data sets with large numbers of replicates. Our aim was specifically to investigate 1) the number of detections at different numbers of replicates 2) the consistency of the detections within and between pipelines 3) the estimated proportion of false discoveries and 4) the runtimes. We included eight popular software packages for our comparison: DESeq [6], edgeR [80] and baySeq [34] based on negative binomial models, SAMseq [52] and NOISeq [91] based on non-parametric approaches, limma [79] based on transformations of read counts for linear modeling and Cuffdiff [95] and EBSeq [48] which are transcript-based methods also enabling gene level analysis.

Our study showed that the choice of the analysis method can markedly affect the outcome of the data analysis and that no single tool is likely to be optimal under all circumstances. We also discovered that the number of replicates and the heterogeneity of the samples should be taken into account when selecting the pipeline. General usability and the quality of documentation also varied across methods. As many users in practice prefer a method that is user friendly and works fast and is robust under a wide range of conditions, we recommend limma for these users, based on our comparisons.

In order to detect differential gene expression based on RNA-seq data, the read counts are typically summarized at the gene level prior to carrying out the statistical testing. In Publication IV, we investigated an alternative strategy in which statistical testing at the exon level is performed prior to the summary of the results at the gene level and specifically the effect on the sensitivity and specificity of the detections. This was motivated by the earlier reported observations with Affymetrix gene expression microarrays indicating that statistical testing of probe-level expression signals, rather than gene-level summary values, can markedly improve the detection of differential gene expression [27].

The presented exon-based strategy can in principle be used in the context of any existing or future statistical testing approach. For our examination we chose to use two popular packages, limma [79] and edgeR [80]. With our approach, the statistical testing is first performed separately (with limma or edgeR) for each exon. The gene-level scores are then calculated as the medians of the exon-level significance p-values while taking the directions of the changes into account. More formally, the gene-level score is defined as the median over the signed log-transformed $p$-values $y_i = -sgn(x_i)logp_i, i = 1, ..., n$, where $x_i$ is the estimated log2 fold change of an exon $i$, $p_i$ the corresponding $p$-value obtained from the statistical testing, $n$ the number of exons in the particular gene and $sgn$ the sign function. The calculated $p$-values are then corrected using the Benjamini-Hochberg multiple testing adjustment method [11].

As described in Publication IV, using two publicly available data sets we were able to show that the suggested exon-based strategy improved the statistical testing results over the conventional gene-based strategy by increasing sensitivity and specificity of the detections. The improvement was especially pronounced for genes with moderate but systemic gene expression changes that were missed by the gene-based strategy relying on single gene-level summary counts only. Our results also showed how the gene-based approaches are prone to effects of single exons, while the exon-based strategy is robust against them.

## 4.2 Functional enrichment analysis

Genes act in co-operational groups, modules, to carry out coordinated tasks. Taking this information into account during data analysis may help interpreting the results from any experiments producing candidate lists of genes. Therefore, it is a common step during the high-throughput data analysis to apply some technique for functional analysis, also called *gene set enrichment analysis* or *pathway analysis*, to detect enrichment in gene lists towards known gene modules, thus suggesting altered performance of the related functions (see [56] and [36] for recent reviews). These known gene modules or gene sets can be derived from various different sources, Gene Ontology (GO) [35] and Kyoto Encyclopaedia of Genes and Genomes (KEGG) [41] being the most popular of them and also freely available.

To take advantage of GO, genes are organized into a hierarchy where gene products with similar functions are placed together under the same GO term. In this hierarchy, a gene belonging to a category is automatically part of all its parent classifications as well. Thus the number of genes placed in the nodes decreases as traversing down the tree, gradually leading to more specific terms. GO is divided into three main hierarchies, namely *biological processes*, *molecular functions* and *cellular components*. Genes are mapped to the nodes under each hierarchy at different confidence levels ranging from manually curated to computationally predicted.

KEGG database provides manually curated searchable pathways related to molecular interaction and reaction networks for metabolism, various cellular processes and human diseases. In KEGG, the relationship between the genes belonging to each pathway is defined and it can thus be visualized as a pathway map (example shown in Figure 6).

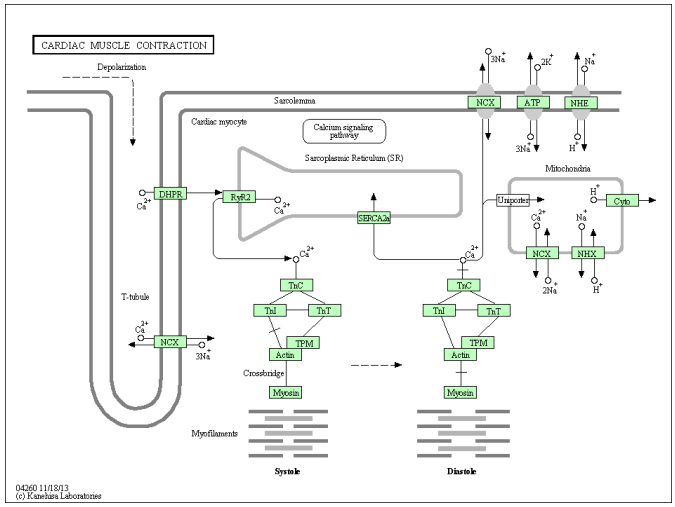



FIGURE 6 Example of a KEGG pathway map for cardiac muscle contraction [2].

In general, there are two distinct commonly used approaches for functional enrichment analysis; detection of functional enrichment in short filtered gene list (such as list of differentially expressed genes) and detecting functional enrichment towards the top of a ranked gene list. For example in the context of differential gene expression analysis the ranked list can be the whole set of microarray genes sorted according to the decreasing likelihood of the gene being differentially expressed between the sample groups. Figure 7 summarizes the differences between the two approaches.

In the filtered gene list enrichment analysis the defined gene sets are tested for over-representation within the input gene list (e.g. the list of differentially expressed genes) and a statistical significance score (typically p-value) is calculated for each set. Testing is conducted against a background list of genes which is often the list of all known genes. Typically sets with too few (<15) or too many (>200) genes are excluded from the analysis as the results are not likely to be meaningful or reliable for them. When the gene set is too small, the test can easily become significant due to a single or a couple of genes only and for the very large sets its difficult to interpret the results. Statistical testing is commonly based on a hypergeometric or binomial model [56] and the results are represented as a table of the most enriched gene sets in a decreasing order according to the test significance score. Popular tools for this kind of analysis include David [20] and GOrilla [24].

While analysing the functional enrichment among the filtered genes is very

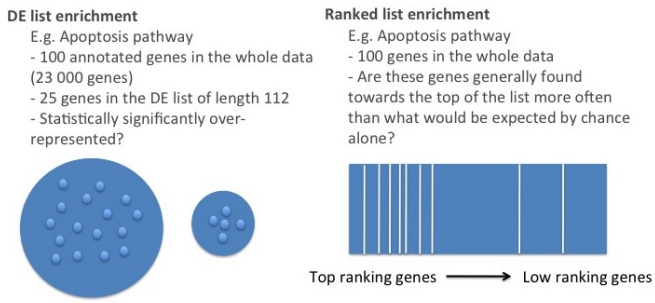
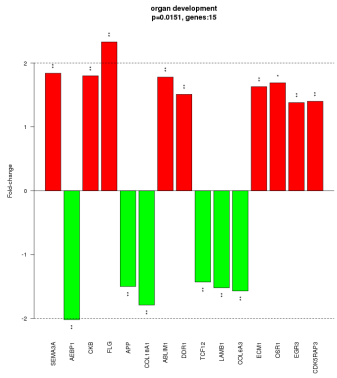


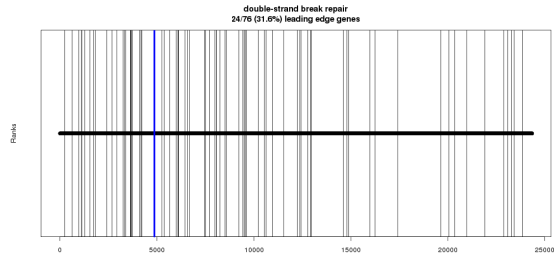

FIGURE 7 Difference between the filtered list (on left) and ranked list enrichment analysis (on the right).

useful, the choice of filtering thresholds can have a significant effect on the analysis outcome. It is also possible that although many genes associated with a particular functional module are showing consistent yet subtle expression changes only a few or none of them are affected strongly enough to be included in the filtered list of genes. Thus this functional module may be missed in the enrichment analysis based on the short filtered gene list. As a solution, threshold free, ranking based enrichment analysis approaches can be used to detect these more subtle changes. Popular tools are typically based on non-parametric tests such as Kolmogorov-Smirnov test, e.g. GSEA [90] and GAGE [55]. In general, the two enrichment analysis approaches efficiently complement each other and are thus ideally applied in parallel to gain a complete view of all the affected functional gene modules.

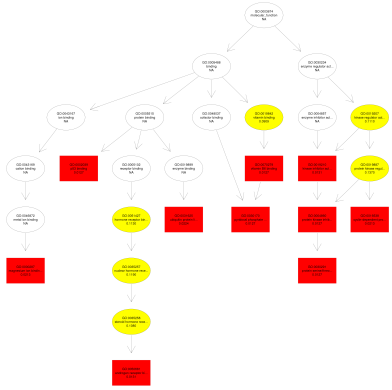
The different tools available for short or long ranked gene list enrichment analysis in general vary for example based on the organisms supported, the selection of different statistical test approaches, the databases available and also regarding the way the results are reported and visualized. Most freely available tools only support one or the other enrichment analysis approach and the result visualization is typically also very limited. Thus in Publication I we developed a new web tool, GeneFuncster, which combines both analysis approaches, based on popular algorithms, with versatile result visualization abilities including gene plots, GO graphs, colored KEGG pathway maps and ranking plots (examples in Figure 8).



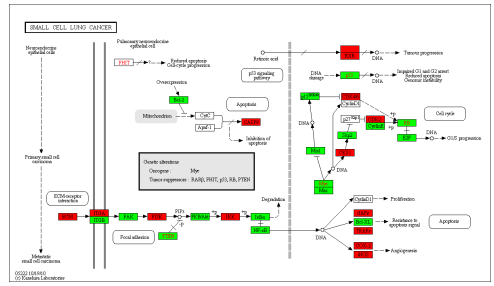

(a)




(b)



(c)




(d)

FIGURE 8 Various visualization plot types provided by GeneFuncster. **a)** An example of a gene plot with genes associated with an enriched database term, fold-change values as the height of the bars and p-value levels marked by asterisks on top of each gene. **b)** Ranking plot for unfiltered list enrichment analysis where each gene associated with the database query term in question has been marked with a vertical bar along the ranked list (from left to right). Enrichment can be observed as a clear bias of vertical lines towards the high ranking end (left side) of the graph. **c)** GO graph describing the relations between the most enriched GO terms with stronger colors (scale from white to yellow, orange and red) signifying stronger enrichment. **d)** Enriched KEGG pathway map with upregulated genes marked red and downregulated genes green.

## 4.3 Biclustering

In situations where there is a need to interpret large high-throughput data sets with a large number of genes (or other similar biological measurement features) and condition groups or potential sub-groupings within the sample groups, various clustering techniques become highly beneficial. The data may then be analysed to detect general patterns, potentially giving further insight on the underlying biological processes.

In general, clustering based techniques are used to analyze data matrices with the aim of detecting patterns across rows or columns or both simultaneously. High-throughput gene expression data is typically represented in matrix format with genes on rows and samples (or sample condition groups) on columns. Clustering techniques can then be used for arranging the genes and samples based on their general expression value similarity. With traditional clustering techniques each gene is included in the clustering result exactly once as items can not belong to multiple clusters or be excluded completely from the clustering result [12], although in practice, genes act in modules to carry out coordinated tasks and each gene may participate in multiple processes. Traditional clustering also includes all columns (samples or sample condition groups) in all patterns despite that gene sets are typically co-expressed only under a subset of samples or sample condition groups. A natural solution to these problems is provided by a technique called *biclustering* (also called *co-clustering* or *two-mode clustering*) that is able to cluster rows and columns simultaneously and does not set *a priori* constrains on the organization of the resulting clusters. In other words, it allows any gene to belong to multiple or none of the biclusters and detects also gene groups with similar expression patterns over only a subset of the samples or sample condition groups [57].

Despite the clear theoretical benefits of biclustering, this approach has not so far gained very wide popularity among the gene expression research community. As we speculate in Publication II, possible reasons include for example the unrealization of the various complementary ways in which biclustering can be applied to high-throughput gene expression data and the lack of reliable and fast algorithms. To tackle these problems, in Publication II we have developed an efficient novel biclustering method, BiclusterMiner, and illustrated various complementary scenarios for applying biclustering on gene expression data. Depending on the scenario, normalized expression data, filtered differentially expressed genes with fold-change values or functional enrichment analysis results across group comparisons can be used as input for biclustering as depicted in Figure 9. Review through the biclustering literature revealed that the steps for preprocessing gene expression data prior to biclustering are typically poorly described which may also have hindered the adoption of biclustering tools for downstream data analysis. Thus we illustrate in our publication how biclustering in general fits to the overall gene expression analysis workflow to aid its wider application (also described in Figure 9.). Hence, our work provides a good reference for the biclus-

tering method developers aiming to work with gene expression data.

Some biclustering methods are able to work with real-valued data but generally this is considered a demanding task. The size of the input data also needs to be decreased by applying heavy prefiltering of data which in turn may lead to a loss of important information. Typically the number of resulting biclusters also has to be defined before the biclustering which may lead to unmeaningful results when a nonoptimal number of biclusters is selected. Thus in practice the most popular methods rely on data discretization, often binarization [57]. Different data discretization schemes and their performance in the context of various clustering techniques has recently been discussed in [58]. Discretization is often performed simply by applying certain cutoff values: expression values higher than the cutoff are marked with 1 and lower with 0. In addition to redocung the computational demand and the need for heavy prefiltering there are also other benefits to data discretization as we discuss in Publication II. For example, when performing meta analysis of data sets derived from different sources, discretization may help in making the data better comparable across the different data sets as the normalization issues with the real values are effectively avoided. In addition, the differential expression status of a gene (up-regulated/down-regulated/not affected) or a pathway (affected/not affected) can be naturally represented with a few discrete value categories. Yet, most of the previously published algorithms work on binarized data which makes them unable to distinguish between the direction of the gene regulation, although this is biologically very important. Ignorance of the direction of the regulation leads to the inclusion of erroneous genes and results in invalid biclusters as gene profiles with uncorrelated patterns are grouped together.

Various biclustering approaches developed include iterative row and column clustering combination, divide and conquer, greedy iterative search, exhaustive bicluster enumeration and distribution parameter identification [57]. Some of the widely used approaches apply a greedy search scheme and are thus unable to discover all maximal biclusters (i.e. biclusters not entirely contained in any other bicluster) and important patterns may thus be missed, as described in Publication II. Different methods and approaches have been recently compared in [70]. The algorithm of BiclusterMiner is a generalization of the original problem, presented in [74], where only binary data was used.The algorithm discovers the biclusters recursively as explained in detail in our publication. To our knowledge, BiclusterMiner is the first published method able to work on three discretized value categories and yet discover all maximal biclusters. The tool is also simple and fast to use.

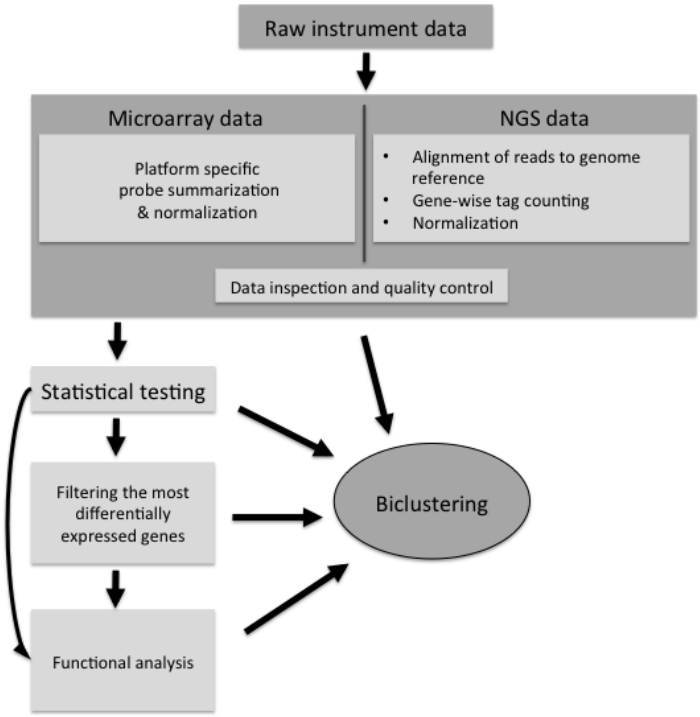



FIGURE 9 Data processing pipeline for biclustering of gene expression data. Biclustering can be applied for normalized expression data, filtered differentially expressed genes with fold-change values or functional enrichment analysis results across group comparisons.

## 4.4 Gene-gene interaction analysis

Genome-wide association studies (GWAS) focus on the detection of common genetic variants associated with traits like major diseases. Genetic variants impact gene function mainly by influencing promoter activity directly controlling the gene expression, stability of messenger RNA molecules and subcellular localization of these molecules or the protein produced [86]. The effect and severeness of a variation depends on whether it causes an amino acid change in the protein (nonsynonymous change) or not (synonymous change) or whether it occurs in the noncoding regions controlling the regulation of the gene, as discussed in [31].

Traditionally the focus of GWAS has been in the search of genes that increase (or decrease) disease susceptibility through the examination of individual single-nucleotide polymorphisms (SNPs). In practice, however, it has turned out to be difficult to confirm the results across several studies based on these SNP level results [31]. This is speculated largely to be explained by the overwhelming number of genetic markers that are typically analyzed in a very limited number of subjects, making the statistical inference very challenging. In addition, the general complexity of the mapping between genotype and phenotype, arising for instance from the nonlinear interactions with the genetic and environmental factors are likely to complicate the analysis. Especially gene-gene interaction analysis (i.e. detection of genes with effects that are dependent on other genes) has lately been considered a promising approach in improving GWAS and has already been successfully applied for a wide variety of different common human diseases and clinical endpoints including e.g. bladder cancer, amytrophic lateral sclerosis, and eczema [31]. For this, several tools for detecting significant pairwise SNP combinations have been developed [17] ranging from simple exhaustive search (implemented e.g. in PLINK Tool Set [76]) to various data-mining and machine learning approaches (e.g. random forests [110]) and bayesian model selection approaches (e.g. WinBUGS [54]). It has also been demonstrated how pathway-based approaches may narrow the search space and enhance power of GWAS studies [106].

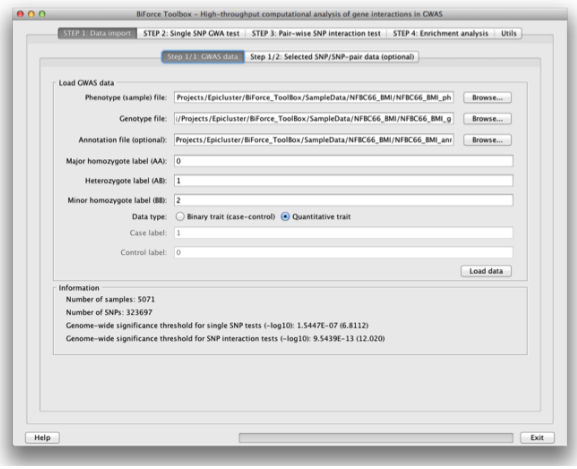
The analysis of billions of SNP combinations is computationally a highly challenging task and it has limited the efficient study of gene-gene interaction in GWAS. For this reason, many of the previously developed methods are restricted to the analysis of binary disease or quantitative traits and are often specifically designed for computers equipped with particular graphical processing units, even further limiting wider application of these methods. Tools have also been missing for the meta-analysis of multiple GWAS simultaneously which could even further enhance the power of gene-gene interaction detection. To address these limitations, BiForce Toolbox was developed in Publication III.

BiForce Toolbox provides fast screening of pairwise interactions in GWAS of complex disease and quantitative traits, relying on enhanced computational power derived from the bitwise computing and multi-threaded parallelization. The toolbox is implemented as a stand-alone software package in Java to enable

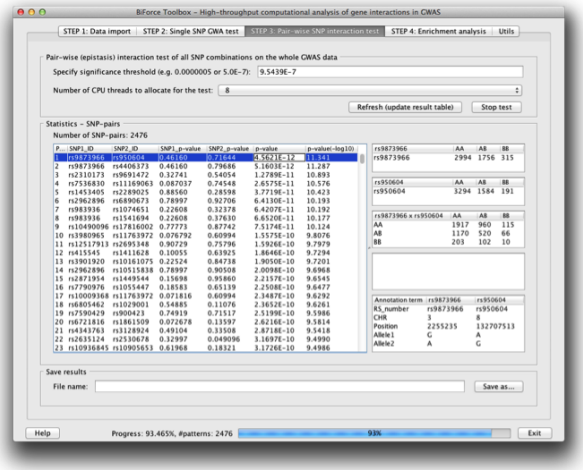
its use on all the most commonly used computer systems. This makes the software suitable also for local secure analysis that is important when handling sensitive data.

The software includes two consecutive genome scans: single SNP-based genome-wide association tests and pairwise interaction tests of all SNP combinations. Additionally, marginal-SNPs (SNPs that do not lead to any marginal correlation between genotype and phenotype when each locus is considered individually) are identified in the first scan and then separately tested for interaction. Association tests are based on linear regression models, where the genotypes of each SNP (i.e. homozygote of the minor (i.e. least common) allele, homozygote of the major (i.e. most common) allele and heterozygote) are fitted as fixed factors. Pairwise SNP interactions are assessed using contingency tables which makes BiForce Toolbox applicable to both quantitative and binary disease traits. The details of the algorithms have been earlier published in [46] and [105].

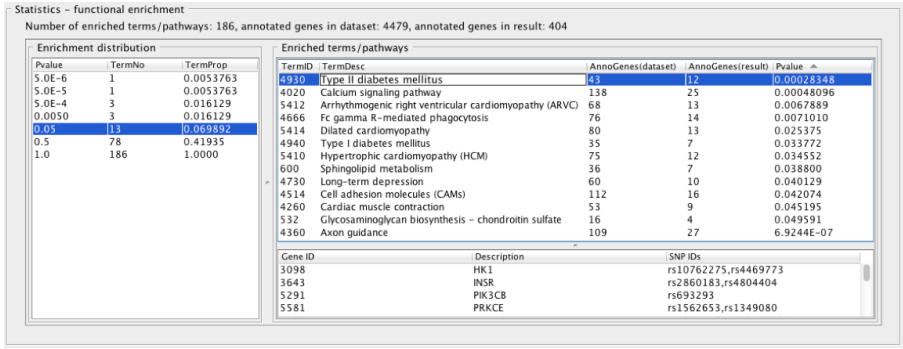
The toolbox can also perform an analysis of pathways enriched within groups of genes showing interaction signals, potentially giving insight to the biology underlying these signals. The enrichment analysis towards Gene Ontology categories and KEGG pathways is performed based on the mapping of the SNPs to nearest genes and then performing a typical gene over-representation analysis. Special attention has been paid to the general ease and usability of the tool and informative result report formats and visualizations. The BiForce Toolbox workflow with snapshots of the tool are illustrated in Figure 10.




(a)




(b)




(c)

FIGURE 10 BiForce Toolbox workflow. **a)** Snapshot of the BiForce Toolbox GUI Step 1: data loading. **b)** Snapshot of the BiForce Toolbox GUI Step 3: running the pairwise genome scan for the example data. **c)** Snapshot of BiForce Toolbox GUI Step 4: running pathway enrichment analysis after the pairwise genome scan using interacting genes as input.

# 5 SUMMARY OF THE PUBLICATIONS

Within this thesis new tools and methods for differential gene expression analysis, functional enrichment analysis, biclustering and gene-gene interaction analysis are introduced with special focus on performance, usability and result visualization. Comprehensive and robust data analysis workflows and analysis schemes for various data types are also illustrated. These tools and approaches have been used in research studies concentrating on both mouse and human, cell culture and tissue samples.

## I GeneFuncster: A Web Tool for Gene Functional Enrichment Analysis and Visualisation

In Publication I, we developed a tool for gene functional enrichment analysis and visualization. The project was motivated by the need to have a tool to analyze both short filtered and long unfiltered gene lists for enrichments towards functional categorizations and pathways available in public databases and to visualize the results in a comprehensive manner. While there were many free tools available for these types of analyses, the most useful functionality was scattered across various tools and especially visualization was poor even with the most popular software. In order to perform a comprehensive functional enrichment analysis, both short filtered and long unfiltered gene lists need to be analyzed in order to provide a complete view of the underlying biology. Although the analysis of filtered lists is efficient in detecting functions with strongly affected genes, some important functions with consistent but subtle changes maybe missed due to the application of filtering cutoff values. Thus the two approaches complement each other and are best applied in parallel, for example for high-throughput gene expression data. Many of the available tools also simply report the results as a table ranked according to the statistical enrichment significance and maybe difficult to interpret, especially when many related categories appear on the results mainly due to the shared genes between the categories. This is especially true to

Gene Ontology categories that are hierarchical in nature and contain thousands of terms with genes annotated to them with varying confidence levels. In this case it is of utmost importance to inspect the enrichment results by visualizing the affected categories and their relationships. For KEGG pathways the problem on the other hand is the relatively low number of genes annotated to each of the couple of hundred pathways, which causes even the most enriched pathways to typically contain only a few affected genes. Visual inspection of the affected genes on the pathway map context may thus significantly aid the interpretation of the results. In addition, the ability to easily check the significance and fold-change of the associated genes further enhances the conclusion making based on the enrichment results. All these features are present in our novel tool, GeneFuncster, that is available as an online web tool (http://bioinfo.utu.fi/GeneFuncster) providing an intuitive user interface and fast analysis. The tool is able to analyze data from human and mouse and various other organisms and can be used for high-throughput gene expression data or any other biological data type generating gene lists with potentially enriched functions.

## II Biclustering of high-throughput gene expression data with BiclusterMiner

In Publication II, a novel tool for biclustering of high-throughput gene expression data was developed. In general, biclustering holds the promise to avoid the major problems of the traditional clustering based approaches, namely the inclusion of each gene to resulting clusters exactly once and the consideration of patterns across all of the conditions without exceptions. As genes are known to act in modules and each gene may participate in more than one function, biclustering is an attractive approach for handling gene expression data. It also allows the simultaneous detection of patterns only across a subset of the samples or sample condition groups which makes it especially useful when working with large data sets consisting of a high number of samples and sample condition groups. Despite the clear theoretical benefits, biclustering has not gained wide popularity within the gene expression research community which may be due to several different reasons, the lack of usable and efficient methods being one of them. Another possible hindrance is the lack of good guidelines for preprocessing the data for biclustering which we also aimed at correcting by providing a comprehensive summary in our publication on the important issues to take into account. Moreover, we discuss the various different schemes in which biclustering can be applied to gene expression, a topic that also has not been comprehensively handled in the previous literature. Most of the popular biclustering approaches for gene expression data work on discretized data as handling real valued data is computationally very demanding and result interpretation is difficult without sufficient data dimensionality reduction. In practice typically the data is binarized to two value categories. This however leads to the ignorance of the direction of the gene

regulation, although this is known to be biologically important. On the other hand, some methods apply greedy search due to which they may miss important gene modules and patterns. To our knowledge, our new method, BiclusterMiner, is the first discrete biclustering method able to work on three value categories (i.e. take the direction of the gene regulation into account) and yet discover all maximal biclusters (i.e. those not fully contained by any other bicluster). Finally, comparison to popular approaches shows the superiority of our method in regard to the gene expression patterns detected and the running time.

## III BiForce Toolbox: powerful high-throughput computational analysis of gene-gene interactions in genome-wide association studies

In Publication III, we developed a toolbox for gene-gene interaction analysis of genome-wide association studies (GWAS). GWAS are generally applied to detect genomic loci, typically single nucleotide polymorphisms (SNPs), related to quantitative traits or disease. Traditionally in GWAS, single SNPs have been analyzed independently but more recently more attention has been paid to the potential interactions between SNPs. However, the lack of usable and powerful software has highly limited the study of gene-gene interaction analysis. Interaction analysis is computationally highly demanding as it requires handling of billions of pairwise SNP combinations. Earlier methods are also confined to either binary disease or quantitative traits and many of them are designed specifically for computers equipped with particular graphical processing units. To address these limitations we developed BiForce Toolbox, a stand-alone Java program that allows efficient gene-gene interaction analysis in quantitative and disease traits across all commonly used computer systems. The implementation is taking advantage of the efficient bitwise computing technologies and multi-threaded parallelization and the software allows full pairwise genome scans via a graphical user interface or the command line. The combined search algorithm implemented in BiForce Toolbox includes two consecutive genome scans: single SNP-based genome-wide association tests and pairwise epistatic interaction tests of all SNP combinations. Association tests are based on linear regression models and pairwise SNP interactions are assessed using contingency tables. As it has been indicated that pathway-based approaches may narrow the search space and also enhance power in GWAS, enrichment analysis towards Gene Ontology and KEGG databases was also included in the Toolbox. The tool is available for download at http://bioinfo.utu.fi/biforcetoolbox.

## IV A Note on an Exon-Based Strategy to Identify Differentially Expressed Genes in RNA-Seq Experiments

In Publication IV, we investigated the effect of the exon-based strategy compared to the traditional gene-based strategy for the differential gene expression analysis of RNA sequencing data. To detect the differentially expressed genes between sample groups, read counts are typically summarized on the gene-level prior to statistical testing. In our publication we present an alternative approach where the statistical testing is first carried out on the exon level and the results are only then summarized at the gene level. The work was motivated by the earlier reported observations with Affymetrix gene expression microarrays indicating that statistical testing of probe-level expression signals, rather than gene-level summary values, can markedly improve the detection of differential gene expression. The proposed exon-based strategy can be applied in the context of any statistical testing package developed for RNA-seq data. Using publicly available data sets we demonstrate how the proposed strategy can markedly improve the sensitivity and specificity of the detections especially for genes with moderate but systemic changes.

## V Comparison of software packages for detecting differential expression in RNA-seq studies

Publication V presents a comparison study of eight widely used software packages for detecting differential expression in RNA-seq studies. Earlier comparison studies were based on only simulated data sets or they included very few biological replicates and thus the aim of our study was to investigate the performance of the methods on real data sets with a relatively large number of replicates. In particular, we investigated the number of detections at different numbers of replicates, their consistency within and between pipelines, the estimated proportion of false discoveries and the runtimes. Our study shows that the choice of the analysis method can markedly affect the outcome of the data analysis and that no single tool is likely to be optimal under all circumstances. We also discovered that the number of replicates and the heterogeneity of the samples should be taken into account when selecting the pipeline and thus we also provide general guidelines for choosing a robust pipeline.

## VI Transcriptome profiling of the murine testis during the first wave of spermatogenesis

In Publication VI, we designed experiments and analyzed the produced RNA sequencing data to study gene expression during sperm development in mouse to elucidate the stage specificity and complexity of testicular transcription machinery. During the project we developed gene expression data analysis pipelines and functional enrichment analysis tools, which were applied for the mouse data and later included in Publications I-III. In the study, testis samples were examined in two replicates at five different time points (post natal days 7, 14, 17, 21 and 28) during the first wave of spermatogenesis in the context of germ cell differentiation. The samples were sequenced with the Life Technologies' SOLiD 4 deep sequencing platform at 50 bp read length. The first wave of spermatogenesis in the mouse provides an invaluable tool for the characterization of gene expression and cellular events during spermatogenesis, which were previously largely unknown. The RNA-seq data were analyzed for differential expression in the gene level as well as in the isoform level. Across four chronological comparisons between the consecutive time points, altogether 2494 genes were detected as differentially expressed, which represents approximately 9% of all annotated genes detected expressed in the mouse testis. Many of the detections were found to be specific to a certain comparison. Extensive functional enrichment analysis revealed that the differentially expressed genes were highly enriched in many biological processes important for correct sperm development and related to spermatogenesis, reproduction, meiosis and fertilization. In isoform level analysis over 160 000 forms were identified of which nearly 40 000 did not have any overlap to previously known genes and 57% of all expressed genes were found to have at least two isoforms. Large differences in the promoter and transcription start site usage were detected even between the first two time points that did not show many differences in the gene level, illustrating the complexity of the transcriptional regulation in the testis. The differentially expressed isoforms were found to be mostly involved in protein domain specific binding, nucleotide binding and telomeric DNA binding. The data were also analyzed for differential expression of the long non-coding RNAs, the expression of which was found to be highly specific to each time point.

## VII Tet1 and Tet2 regulate 5-hydroxymethylcytocine production and cell lineage specification in mouse embryonic stem cells

Publication VII presents the results of a study to clarify the functional roles of Tet proteins during mammalian development. This newly-discovered family of DNA-modifying enzymes was studied in mouse embryonic stem cells (ESC) and induced pluripotent stem cells (iPSC). These cells can be maintained in the prolif-

erative, undifferentiated state in culture, possessing features of self-renewal and ability to differentiate, characteristic of a pluripotent state, known to require a high degree of epigenetic plasticity. Tet1 and Tet2 are able to alter the methylation status of DNA, known to influence many biological processes during mammalian development and being highly aberrant in cancer. Dynamic changes in DNA methylation occur during early embryogenesis and both methylation and Tet expression/activity are tightly regulated during ESC differentiation. For this study, the whole genome expression analyses of Tet1 knock down and control knock down ES clones from ES or 4-week trophoblast stem (TS) cell culture were performed using the Illumina mouse WG-6 v2.0 expression beadchip array at the Finnish Microarray and Sequencing Centre following the array manufacturer's standard protocols. At least 3 independent clones were analysed in each group. This study shows that Tet1 and Tet2 are the key enzymes responsible for the presence of 5-hydroxyme-tylcytocine (5hmC) methylation in mouse ESCs and iPSCs and that their expression is regulated by Oct4. The data suggest a complex relation between Tet proteins and DNA methylation and highlight a strong correlation between Tet1 and Tet2 expression and the pluripotent state. Thus Tet proteins are identified as key regulators of early embryonic differentiation. The data also indicate that these enzymes do not act alone but, rather, operate in coordination with developmental signals to regulate lineage determination at decision points that are critical for early lineage commitment. Taken together, these data suggest that dysregulation of DNA methylation via TET proteins and hmC may have a role in ESC pluripotency, oncogenic transformation and neuronal function. The analysis of the microarray data involved applying and optimizing various methods and approaches for preprocessing, statistical testing, functional analysis and result visualization, later to be included in publications I-III.

# 6 DISCUSSION

The completion of the sequencing of the human genome in the beginning of the last decade and the rapid development of several high-throughput measurement technologies during the last couple of decades has resulted in an exponential growth in the amount of biological data produced per year. This has made the efficient and reliable data analysis and result interpretation a serious bottle neck in the field – a situation which is well condensed in the nowadays often cited statement: *researchers are drowning in data but starving for knowledge*. The measurement technologies also continue to develop very fast which is predisposing considerable challenges for the method and software developers. While DNA microarray technology and the related data analysis methods have started to mature and stabilize, novel deep sequencing technologies and upgrades to currently available technologies are frequently being introduced. Thus the data analysis tools also need to be constantly updated to match the technological developments and format changes.

Consequently, the field of bioinformatics is also evolving fast and a huge number of novel methods and data analysis tools are published each year. However, applying these tools for analyzing data sets in real research projects is not always that straight-forward for many practical reasons. Firstly, many methods are still published without a working or openly available implementation. Secondly, the implementation can be based on a specific programming language or computing environment that the user has to be familiar with in order to use the method. Some of the methods are also based on commercial software packages, like Matlab, which further restricts their availability to all interested researchers. Installation of many methods and tools also requires advanced IT skills beyond those possessed by most biological researchers. Thirdly, as many freely available tools are developed by a single researcher or a research group, their general usability and the level and quality of documentation greatly varies and many tools also lack proper user support and maintenance schemes and thus easily become obsolete. Thus it is important not just to develop new methods but to provide them for the research community as user-friendly tools with good support for the users. Following this observation, we are currently working on creating an R

package implementation of the exon-based differential expression method, developed in Publication IV. Finally, as many analysis methods are typically available for the same purpose, the choice of the method can be a daunting task for an individual researcher. Some original method publications include a comparison to other popular methods, typically illustrating their superiority compared to others. Unfortunately, these comparisons are typically based on a single or very few data sets and thus their general reliability often remains relatively low. In the general absence of good benchmarking data sets with known ground truth, unbiased extensive comparison studies, also taking the practical usability of the methods into account, are extremely valuable, such as the one presented in Publication V.

Regarding the other methods and tools developed within this thesis, GeneFuncster (developed in Publication I) remains publicly available online and is routinely used by many researchers at the Turku Centre for Biotechnology at the University of Turku, and others. The tool has been expanded to include the Reactome pathway database [18] in order to compensate for the discontinued free content updates to the popular KEGG pathway database, previously used as the primary source of pathway information within GeneFuncster. Support for additional organisms has also been added. Based on the user feedback, in the future it would be very interesting to add features for enabling an easy comparison between several enrichment analysis runs. Biclustering (topic in Publication II) continues to be an interesting research topic within bioinformatics. Oghabian et al. [70] recently compared various biclustering methods and concluded that biclustering algorithms in general can discover more relevant genes compared to one-way clustering methods. Naulaerts et al. [66] also illustrated the applicability of biclustering and related frequent itemset mining techniques for different bioinformatics application domains. BiclusterMiner, developed in Publication II continues to be available for download online. In the future, it would be interesting to expand the method to work on the real numbers instead of the discretized input currently supported. BiForce Toolbox for epistasis analysis, developed in Publication III, has been downloaded 350 times since its publication and is still actively used. The software was originally developed for bioinformaticians who have the skill to install and operate software in computer cluster environment. As it is important to make the developed tools available for those lacking the skills and high performance computing equipment we are currently working with a web server based version of the toolbox. The current status of the epistasis analysis has been recently discussed in [107] where the authors concluded that epistasis detection studies so far have shown that large interaction terms between pairwise SNPs are unlikely to exist. However, they envision the next step in the epistasis research is to focus on meta-analyses integrating data over several previously carried out GWAS and to consider multilocus epistatic variance in addition to analyzing locus pairs. Such a multilocus approach has been recently introduced in [43], for example.

The current rapid accumulation of data representing various high-throughput data types in public repositories presents both a considerable challenge and a huge opportunity for the bioinformatics research in the future. Moving from

the analysis of individual data types towards integrative approaches holds the promise of building increasingly accurate models of the complex biological systems. This kind of integrative analysis will require novel sophisticated and efficient analysis methods. In order to fully take an advantage of the integration of these large data sets, the development of user-friendly tools with efficient visualization approaches and easily interpretable result reports will become increasingly important in the future.

## ACKNOWLEDGEMENTS

Although the seven publications included in this thesis have been published relatively recently, between 2011 and 2015, the work towards the thesis essentially started when I began my career as a bioinformatician at the Turku Centre for Biotechnology in 2004. The knowledge and experience gained from having worked for more than a decade with the analysis of large high-throughput data sets forms the basis on which the included publications build on. The work presented here has been carried out at the Department of Information Technology, University of Turku, in very close collaboration with Turku Centre for Biotechnology and its Sequencing and Microarray and Bioinformatics Units.

During the years I have been privileged to work with and learn from a large number of talented scientists and it would be an impossible task to name them all here. I would like to thank the past and present personnel of the Turku Centre for Biotechnology and especially the colleagues at the Microarray and Sequencing Centre and the Bioinformatics Unit with whom it has been a great pleasure to work over the years. I would like to thank Stephen Rudd, Bioinformatics Group leader at the time, who first saw the potential in me and gave me a position at the Centre. My greatest gratitude goes to Docent Attila Gyenesei whose encouragement and support has been crucial during this thesis project. Docent Laura Elo is warmly thanked for her role as the second supervisor of my thesis. I also thank Professor Riitta Lahesmaa and Doctors Juhani Soini and Riikka Lund who as the heads of Microarray and Sequencing Unit have given me their trust and the freedom to use my creativity in developing bioinformatics services and the related tools and pipelines. Professors Mauno Vihinen and Garry Wong are acknowledged for reviewing my thesis and giving valuable feedback and comments. I also wish to express my gratitude to all the coauthors, who have contributed to this thesis in the joint publications.

Finally, I warmly thank my family for their loving support over the years. I feel blessed and grateful for all the challenges and treasures that I find on my path during this amazing adventure called life.

Turku, December 2015
Asta Laiho

## REFERENCES


[1] File: Microarray exp horizontal (zh-cn).svg. http://commons.wikimedia.org [website], Nov 2014.

[2] http://www.genome.jp/kegg/pathway.html [website], Nov 2014.

[3] National center for biotechnology information, united states national library of medicine. ncbi dbsnp build 142 for human. [website] http://www.ncbi.nlm.nih.gov/projects/snp, Nov 2015.

[4] Affymetrix. *Affymetrix Genotyping Console 4.2 User Manual*, 2014.

[5] Agilent Technologies. *Agilent Genomic Workbench 7.0 Product Overview*, 2012.

[6] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[7] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–9, Jan 2015.

[8] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from rna-seq data. *Genome Res*, 22(10):2008–17, Oct 2012.

[9] Simon Andrews. Fastqc: a quality control tool for high throughput sequence data. [website] http://www.bioinformatics.babraham.ac.uk/projects/fastqc, Nov 2010.

[10] Riyue Bao, Lei Huang, Jorge Andrade, Wei Tan, Warren A Kibbe, Hongmei Jiang, and Gang Feng. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform*, 13(Suppl 2):67–82, 2014.

[11] Yoav Benjamini and Yosef Hochbergh. Controlling the false dicovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57:289–300, 1995.

[12] Michael Berthold, Christian Borgelt, Frank Hoppner, and Frank Klawonn. *Guide to Intelligent Data Analysis*. Springer, 2010.

[13] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–20, Aug 2014.

[14] Alan P Boyle, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, J Michael Cherry, and Michael Snyder. Annotation of functional variation in personal genomes using regulomedb. *Genome Res*, 22(9):1790–7, Sep 2012.

[15] Stuart M. Brown, editor. *Next-Generation DNA Sequencing Informatics*. Cold Spring Harbor Laboratory Press, 2013.

[16] Benilton S Carvalho, Thomas A Louis, and Rafael A Irizarry. Quantifying uncertainty in genotype calls. *Bioinformatics*, 26(2):242–9, Jan 2010.

[17] Heather J Cordell. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10(6):392–404, Jun 2009.

[18] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, Bijay Jassal, Steven Jupe, Lisa Matthews, Bruce May, Stanislav Palatnik, Karen Rothfels, Veronica Shamovsky, Heeyeon Song, Mark Williams, Ewan Birney, Henning Hermjakob, Lincoln Stein, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res*, 42(Database issue):D472–7, Jan 2014.

[19] Nicolas Delhomme, Ismaël Padioleau, Eileen E Furlong, and Lars M Steinmetz. easyrnaseq: a bioconductor package for processing rna-seq data. *Bioinformatics*, 28(19):2532–3, Oct 2012.

[20] Glynn Dennis, Jr, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003.

[21] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, Florence Jaffrézic, and French StatOmique Consortium. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Brief Bioinform*, 14(6):671–83, Nov 2013.

[22] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[23] Sorin Draghici. *Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition*. CRC Press, 2011.

[24] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10:48, 2009.

[25] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23(1):70–86, Jun 2002.

[26] Sara El-Metwally, Taher Hamza, Magdi Zakaria, and Mohamed Helmy. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol*, 9(12):e1003345, 2013.

[27] Laura L Elo, Leo Lahti, Heli Skottman, Minna Kyläniemi, Riitta Lahesmaa, and Tero Aittokallio. Integrating probe-level expression changes across generations of affymetrix arrays. *Nucleic Acids Res*, 33(22):e193, 2005.

[28] Daniel Fischer, Asta Laiho, Attila Gyenesei, and Anu Sironen. Identification of reproduction related gene polymorphisms using whole transcriptome sequencing in the large white pig population. *G3 (Bethesda)*, Apr 2015.

[29] Paul Flicek, Ikhlak Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Carlos García-Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Monika Komorowska, Eugene Kulesha, Ian Longden, Thomas Maurel, William M McLaren, Matthieu Muffato, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet Singh Riat, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sheppard, Daniel Sobral, Kieron Taylor, Anja Thormann, Stephen Trevanion, Simon White, Steven P Wilder, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Jennifer Harrow, Javier Herrero, Tim J P Hubbard, Nathan Johnson, Rhoda Kinsella, Anne Parker, Giulietta Spudich, Andy Yates, Amonida Zadissa, and Stephen M J Searle. Ensembl 2013. *Nucleic Acids Res*, 41(Database issue):D48–55, Jan 2013.

[30] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[31] Diane Gilbert-Diamond and Jason H Moore. Analysis of gene-gene interactions. *Curr Protoc Hum Genet*, Chapter 1:Unit1.14, Jul 2011.

[32] Claudia Gonzaga-Jauregui, James R Lupski, and Richard A Gibbs. Human genome sequencing in health and disease. *Annu Rev Med*, 63:35–61, 2012.

[33] Pauliina Halimaa, Ya-Fen Lin, Viivi H Ahonen, Daniel Blande, Stephan Clemens, Attila Gyenesei, Elina Häikiö, Sirpa O Kärenlampi, Asta Laiho, Mark G M Aarts, Juha-Pekka Pursiheimo, Henk Schat, Holger Schmidt, Marjo H Tuomainen, and Arja I Tervahauta. Gene expression differences between noccaea caerulescens ecotypes help to identify candidate genes for metal phytoremediation. *Environ Sci Technol*, 48(6):3344–53, Mar 2014.

[34] Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11:422, 2010.

[35] M A Harris, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, G M Rubin, J A Blake, C Bult, M Dolan, H Drabkin, J T Eppig, D P Hill, L Ni, M Ringwald, R Balakrishnan, J M Cherry, K R Christie, M C Costanzo, S S Dwight, S Engel, D G Fisk, J E Hirschman, E L Hong, R S Nash, A Sethuraman, C L Theesfeld, D Botstein, K Dolinski, B Feierbach, T Berardini, S Mundodi, S Y Rhee, R Apweiler, D Barrell, E Camon, E Dimmer, V Lee, R Chisholm, P Gaudet, W Kibbe, R Kishore, E M Schwarz, P Sternberg, M Gwinn, L Hannick, J Wortman, M Berriman, V Wood, N de la Cruz, P Tonellato, P Jaiswal, T Seigfried, R White, and Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–61, Jan 2004.

[36] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief Bioinform*, 13(3):281–91, May 2012.

[37] Heidi Hyytiäinen, Pekka Juntunen, Thomas Scott, Leena Kytömäki, Reija Venho, Asta Laiho, Sini Junttila, Attila Gyenesei, Joana Revez, and Marja-Liisa Hänninen. Effect of ciprofloxacin exposure on dna repair mechanisms in campylobacter jejuni. *Microbiology*, 159(Pt 12):2513–23, Dec 2013.

[38] Illumina. Genomestudio data analysis software. Data Sheet, 2013.

[39] Illumina Inc. Truseq stranded mrna lt sample prep kit - questions and answers [website] http://support.illumina.com, Feb 2015.

[40] Sini Junttila, Asta Laiho, Attila Gyenesei, and Stephen Rudd. Whole transcriptome characterization of the effects of dehydration and rehydration on cladonia rangiferina, the grey reindeer lichen. *BMC Genomics*, 14:870, 2013.

[41] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–80, Jan 2004.

[42] Maaria Kankare, Tiina Salminen, Asta Laiho, Laura Vesala, and Anneli Hoikkala. Changes in gene expression linked with adult reproductive di-

apause in a northern malt fly species: a candidate gene microarray study. *BMC Ecol*, 10:3, 2010.

[43] Hanni P Kärkkäinen, Zitong Li, and Mikko J Sillanpää. An efficient genome-wide multilocus epistasis search. *Genetics*, Sep 2015.

[44] Anna Koskinen, Asta Laiho, Heikki Lukkarinen, Pekka Kääpä, and Hanna Soukka. Maternal hyperglycemia modifies extracellular matrix signaling pathways in neonatal rat lung. *Neonatology*, 98(4):387–96, 2010.

[45] Vanessa M Kvam, Peng Liu, and Yaqing Si. A comparison of statistical methods for detecting differentially expressed genes from rna-seq data. *Am J Bot*, 99(2):248–56, Feb 2012.

[46] Alex C Lam, Joseph Powell, Wen-Hua Wei, Dirk-Jan de Koning, and Chris S Haley. A combined strategy for quantitative trait loci detection by genome-wide association. *BMC Proc*, 3 Suppl 1:S6, 2009.

[47] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultra-fast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[48] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart M G Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–43, Apr 2013.

[49] Benjamin Lewin, Jocelyn Krebs, Elliott Goldstein, and Stephen Kilpatrick. *Lewin's GENES XI*. Jones and Bartlett Learning, 2014.

[50] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, Jul 2009.

[51] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, Aug 2009.

[52] Jun Li and Robert Tibshirani. Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. *Stat Methods Med Res*, 22(5):519–36, Oct 2013.

[53] Yuwen Liu, Jie Zhou, and Kevin P White. Rna-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–4, Feb 2014.

[54] David J Lunn, John C Whittaker, and Nicky Best. A bayesian toolkit for genetic association studies. *Genet Epidemiol*, 30(3):231–47, Apr 2006.

[55] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. Gage: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10:161, 2009.

[56] Henryk Maciejewski. Gene set analysis methods: statistical models and methodological differences. *Brief Bioinform*, 15(4):504–18, Jul 2014.

[57] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform*, 1(1):24–45, 2004.

[58] Priyakshi Mahanta, Hasin Ahmed Afzal, and Juhal K. Kalita. Discretization in gene expression data analysis: a selected survey. In *CCSEIT '12 Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*. Association for Computing Machinery, 2012.

[59] Tobias Maier, Marc Guell, and Luis Serrano. Correlation of mrna and protein complex biological samples. *FEBS letters*, 583(24):3966–3973, Dec 2009.

[60] J. Marioni, C. Mason, S Mane, Stephens, and Y M., Gilad. *Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays*. Genome research, 2008.

[61] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20(9):1297–303, Sep 2010.

[62] Ivan Merelli, Andrea Calabria, Paolo Cozzi, Federica Viti, Ettore Mosca, and Luciano Milanesi. Snpranker 2.0: a gene-centric data mining tool for diseases associated snp prioritization in gwas. *BMC Bioinformatics*, 14 Suppl 1:S9, 2013.

[63] Alison M Meynert, Morad Ansari, David R FitzPatrick, and Martin S Taylor. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics*, 15:247, 2014.

[64] Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter, and Frank Stahl. Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol*, 24(1):22–30, Feb 2013.

[65] Elisa Närvä, Juha-Pekka Pursiheimo, Asta Laiho, Nelly Rahkonen, Maheswara Reddy Emani, Miro Viitala, Kirsti Laurila, Roosa Sahla, Riikka Lund, Harri Lähdesmäki, Panu Jaakkola, and Riitta Lahesmaa. Continuous hypoxic culturing of human embryonic stem cells enhances ssea-3 and myc levels. *PLoS One*, 8(11):e78847, 2013.

[66] Stefan Naulaerts, Pieter Meysman, Wout Bittremieux, Trung Nghia Vu, Wim Vanden Berghe, Bart Goethals, and Kris Laukens. A primer to frequent itemset mining for bioinformatics. *Brief Bioinform*, 16(2):216–31, Mar 2015.

[67] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and snp calling from next-generation sequencing data. *Nat Rev Genet*, 12(6):443–51, Jun 2011.

[68] Intawat Nookaew, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in saccharomyces cerevisiae. *Nucleic Acids Res*, 40(20):10084–97, Nov 2012.

[69] National Library of Medicine (NLM) and National Institutes of Health (NIH) the National Human Genome Institute (NHGRI). Geneed web [website] http://geneed.nlm.nih.gov/, 11 2014.

[70] Ali Oghabian, Sami Kilpinen, Sampsa Hautaniemi, and Elena Czeizler. Biclustering methods: biological relevance and application in gene expression analysis. *PLoS One*, 9(3):e90801, 2014.

[71] Jan Oosting. beadarraysnp: Normalization and reporting of illumina snp bead arrays. R package, 2014.

[72] G. Parmigiani, E.S. Garett, R.A. Irizarry, and S.L. Zeger, editors. *The Analysis of Gene Expression Data*. Springer, 2003.

[73] Walter Pavicic, Taina T Nieminen, Annette Gylling, Juha-Pekka Pursiheimo, Asta Laiho, Attila Gyenesei, Heikki J Järvinen, and Päivi Peltomäki. Promoter-specific alterations of apc are a rare cause for mutation-negative familial adenomatous polyposis. *Genes Chromosomes Cancer*, 53(10):857–64, Oct 2014.

[74] Amela Prelić, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–9, May 2006.

[75] Kim D Pruitt, Tatiana Tatusova, Garth R Brown, and Donna R Maglott. Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic Acids Res*, 40(Database issue):D130–5, Jan 2012.

[76] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, Sep 2007.

[77] Nusrat Rabbee. Rlmm: A genotype calling algorithm for affymetrix snp arrays. R package., 2005.

[78] Mark Reimers. Making informed choices about microarray data analysis. *PLoS Comput Biol*, 6(5):e1000786, May 2010.

[79] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47, Apr 2015.

[80] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, Jan 2010.

[81] Scott F Saccone, Raphael Bolze, Prasanth Thomas, Jiaxi Quan, Gaurang Mehta, Ewa Deelman, Jay A Tischfield, and John P Rice. Spot: a web-based tool for using biological databases to prioritize snps after a genome-wide association study. *Nucleic Acids Res*, 38(Web Server issue):W201–9, Jul 2010.

[82] Ramona Schmid, Patrick Baum, Carina Ittrich, Katrin Fundel-Clemens, Wolfgang Huber, Benedikt Brors, Roland Eils, Andreas Weith, Detlev Mennerich, and Karsten Quast. Comparison of normalization methods for illumina beadchip humanht-12 v3. *BMC Genomics*, 11:349, 2010.

[83] Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–4, Mar 2011.

[84] Holger Schwender. Multiple testing using sam and efron's empirical bayes approaches. R package., 2012.

[85] SEQC/MAQC-III Consortium. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat Biotechnol*, 32(9):903–14, Sep 2014.

[86] Barkur S. Shastry. Snps: Impact on gene function and phenotype. In Anton A. Komar, editor, *Single Nucleotide Polymorphisms*, volume 578 of *Methods in Molecular Biology™*, pages 3–22. Humana Press, 2009.

[87] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

[88] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14:91, 2013.

[89] Russell Steve, Meadows Lisa A., and R. Russell Roslin. *Microarray Technology in Practice*. Elsevier Inc., 2008.

[90] Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo, and Jill P Mesirov. Gsea-p: a desktop application for gene set enrichment analysis. *Bioinformatics*, 23(23):3251–3, Dec 2007.

[91] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in rna-seq: a matter of depth. *Genome Res*, 21(12):2213–23, Dec 2011.

[92] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[93] Kaisa J Teittinen, Toni Grönroos, Mataleena Parikka, Sini Junttila, Annemari Uusimäki, Asta Laiho, Hanna Korkeamäki, Kalle Kurppa, Hannu Turpeinen, Marko Pesu, Attila Gyenesei, Mika Rämet, and Olli Lohi. Sap30l (sin3a-associated protein 30-like) is involved in regulation of cardiac development and hematopoiesis in zebrafish embryos. *J Cell Biochem*, 113(12):3843–52, Dec 2012.

[94] Kaisa J Teittinen, Asta Laiho, Annemari Uusimäki, Juha-Pekka Pursiheimo, Attila Gyenesei, and Olli Lohi. Expression of small nucleolar rnas in leukemic cells. *Cell Oncol (Dordr)*, 36(1):55–63, Feb 2013.

[95] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nat Biotechnol*, 31(1):46–53, Jan 2013.

[96] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–11, May 2009.

[97] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–5, May 2010.

[98] Victor Trevino, Francesco Falciani, and Hugo A Barrera-Saldaña. Dna microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med*, 13(9-10):527–41, 2007.

[99] Virginia G Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21, Apr 2001.

[100] Erwin L van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends Genet*, 30(9):418–26, Sep 2014.

[101] Saran Vardhanabhuti, Steven J Blakemore, Steven M Clark, Sujoy Ghosh, Richard J Stephens, and Dilip Rajagopalan. A comparison of statistical

tests for detecting differential expression using affymetrix oligonucleotide microarrays. *OMICS*, 10(4):555–66, 2006.

[102] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory Biosci*, 131(4):281–5, Dec 2012.

[103] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164, Sep 2010.

[104] Lucas D Ward and Manolis Kellis. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*, 40(Database issue):D930–4, Jan 2012.

[105] W-H Wei, S Knott, C S Haley, and D-J de Koning. Controlling false positives in the mapping of epistatic qtl. *Heredity (Edinb)*, 104(4):401–9, Apr 2010.

[106] Wen-Hua Wei, Gib Hemani, Attila Gyenesei, Veronique Vitart, Pau Navarro, Caroline Hayward, Claudia P Cabrera, Jennifer E Huffman, Sara A Knott, Andrew A Hicks, Igor Rudan, Peter P Pramstaller, Sarah H Wild, James F Wilson, Harry Campbell, Nicholas D Hastie, Alan F Wright, and Chris S Haley. Genome-wide analysis of epistasis in body mass index using multiple human populations. *Eur J Hum Genet*, 20(8):857–62, Aug 2012.

[107] Wen-Hua Wei, Gibran Hemani, and Chris S Haley. Detecting epistasis in human complex traits. *Nat Rev Genet*, 15(11):722–33, Nov 2014.

[108] Alexander G Williams, Sean Thomas, Stacia K Wyman, and Alisha K Holloway. Rna-seq data: Challenges in and recommendations for experimental design and analysis. *Curr Protoc Hum Genet*, 83:11.13.1–11.13.20, 2014.

[109] Thomas D Wu and Serban Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–81, Apr 2010.

[110] Makiko Yoshida and Asako Koike. Snpinterforest: a new method for detecting epistatic interactions. *BMC Bioinformatics*, 12:469, 2011.

[111] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PLoS One*, 9(1):e78644, 2014.